

John von Neumann Institute for Computing (NIC)

Computational Nanoscience: Do It Yourself!

edited by

Johannes Grotendorst

Stefan Blügel

Dominik Marx

Winter School, 14 - 22 February 2006

Forschungszentrum Jülich, Germany

Lecture Notes

organized by

John von Neumann Institute for Computing

Forschungszentrum Jülich

Ruhr-Universität Bochum

NIC Series

Volume 31

ISBN 3-00-017350-1

Die Deutsche Bibliothek – CIP-Cataloguing-in-Publication-Data
A catalogue record for this publication is available from Die Deutsche
Bibliothek.

Publisher: NIC-Directors
Distributor: NIC-Secretariat
Research Centre Jülich
52425 Jülich
Germany
Internet: www.fz-juelich.de/nic

Printer: Graphische Betriebe, Forschungszentrum Jülich

© 2006 by John von Neumann Institute for Computing
Permission to make digital or hard copies of portions of this work
for personal or classroom use is granted provided that the copies
are not made or distributed for profit or commercial advantage and
that copies bear this notice and the full citation on the first page. To
copy otherwise requires prior specific permission by the publisher
mentioned above.

NIC Series Volume 31
ISBN 3-00-017350-1

Preface

Fostering education and training of young researchers in important fields of computational science is a major objective of the John von Neumann Institute for Computing (NIC). In the series of Winter Schools offered by NIC every second year in February/March this is the first School with hands-on training. This year the School is about Computational Nanoscience which has become a very active field of research in the NIC user community. The selected simulation codes play an outstanding role in many NIC research projects which use the supercomputing facilities provided by the Central Institute for Applied Mathematics (ZAM) of Forschungszentrum Jülich.

Computer simulation becomes more and more important in understanding materials and molecular systems. The nanometer scale in particular is governed by the laws of quantum mechanics, which calls for electronic structure theory in order to address questions related to stability of structures, chemical processes or spectral properties. This hands-on NIC Winter School focuses on the application of modern electronic structure calculations and dynamical simulation techniques covering aspects of solid state physics, surface and nanoscience, chemical reactions and dynamics, magnetism and electron transport as well as the structure and properties of large molecules and clusters. The School will provide a practical introduction to the theory behind and handling of pertinent software packages through practicals and tutorials in small groups using four codes. The full-potential linearized augmented plane wave code FLEUR and the Korringa-Kohn-Rostocker Green function code KKR-GF, the ab initio (Car-Parrinello) molecular dynamics simulation package CPMD, and the highly efficient quantum chemistry code TURBOMOLE. Although very different in concept and application focus, all these codes are well-known prototypical representatives and are used in various supercomputer centres around the world.

These lecture notes provide state-of-the-art information on methodological foundations, advanced concepts and computing properties by internationally leading experts. In two special articles an introduction to parallel computing and parallel linear algebra methods is given.

This School about "electronic structure at work" is tailored for PhD students with a solid background in basic quantum mechanics coming from physics, chemistry, material science and related disciplines. The School combines lectures in the morning, which elucidate the theoretical concepts and numerical procedures, with practical sessions in the afternoon. Within these practicals, bridges will be built from simple test examples up to realistic research problems. It is the aim of this NIC Winter School that participants gain a broad but nevertheless deep understanding of what electronic structure methods at the computational cutting edge can currently achieve.

The scientific programme was worked out by Johannes Grotendorst (Forschungszentrum Jülich), Stefan Blügel (Forschungszentrum Jülich) and Dominik Marx (Ruhr-Universität Bochum). The practical sessions are organized in three simulation labs with access to the

supercomputer JUMP (Jülich multiprocessor) at ZAM.

Because of the limited resources for the computer labs the number of participants is restricted to about 50 participants. We received within two weeks a wide-spread national and international resonance of 90 applications from 22 mainly European countries. Applicants for the School were selected on the basis of scientific background and excellence. Most of the participants have submitted an abstract for the poster session. This reflects the attractiveness of the programme and demonstrates the expectation of the participants to play an active role in this high-level scientific School. We are sure that the School is stimulating for the students, the lecturers and the tutors in the simulation labs.

Many individuals and institutions have significantly contributed to the success of the School. First of all we are very grateful to the lecturers for preparing extended lecture notes in due time, in spite of the heavy work load they all have to carry. Without their effort such an excellent reference book on Computational Nanoscience would not have been possible. We also thank the tutor teams of the three simulation labs FLEUR/KKR-GR (Institute for Solid State Research (IFF), Forschungszentrum Jülich), CPMD (Chair of Theoretical Chemistry, Ruhr-Universität Bochum) and TURBOMOLE (ZAM, Forschungszentrum Jülich) for preparing the codes and the illustrating examples on JUMP.

We would like to thank the Forschungszentrum Jülich, the main sponsor of the School, and the Ψ_k Network of the European Science Foundation (ESF) for their financial support. We are greatly indebted to several staff members of the Forschungszentrum Jülich for their most valuable help with the local arrangements, namely Rüdiger Esser (finance), Monika Marx (web pages and poster), Oliver Bückner (technical support) and last but not least the School's secretaries Bettina Scheid (local service) and Yasmin Abdel-Fattah (registration and accommodation). Particular thanks go to Monika Marx for the editing and composition of the lecture notes, the beauty of this book is entirely her merit.

Jülich and Bochum
February 2006

Johannes Grotendorst
Stefan Blügel
Dominik Marx

Contents

Methodological Foundations

Introduction to Hartree-Fock and CI Methods

<i>Volker Staemmler</i>	1
1 Introduction	1
2 The Molecular Hamiltonian	1
3 Born-Oppenheimer Approximation	3
4 Requirements for the Electronic Wave Function	4
5 Slater Determinants	6
6 Hartree-Fock Theory for Closed Shell States	8
7 Roothaan SCF	11
8 Configuration Interaction	13

Basis Sets, Accuracy, and Calibration in Quantum Chemistry

<i>Thomas Müller</i>	19
1 Introduction	19
2 Basic Approximations	20
3 Errors	28
4 Extrapolation Schemes	36
5 Calibration and Reference Data	39
6 Summary	41

Introduction to Density Functional Theory and Exchange-Correlation Energy Functionals

<i>Robert O. Jones</i>	45
1 Foreword	45
2 Introduction	45
3 Density Functional Formalism	48
4 Exchange-Correlation Energy Functionals	56
5 Concluding Remarks	68

The Pseudopotential Plane Wave Approach

<i>Bernd Meyer</i>	71
1 Introduction	71
2 Why Using Plane Waves?	72
3 Pseudopotentials	78

The Full-Potential Linearized Augmented Plane Wave Method	85
<i>Stefan Blügel and Gustav Bihlmayer</i>	
1 Introduction	85
2 Kohn-Sham Approach in a Nutshell	87
3 APW-Like Concepts to Solve the Kohn-Sham Equations	99
4 The FLAPW Method	105
5 The FLAPW Method for Specialized Geometries	118
6 Where has the CPU Time Gone?	122
The Korringa-Kohn-Rostoker (KKR) Green Function Method	
I. Electronic Structure of Periodic Systems	131
<i>Phivos Mavropoulos and Nikos Papanikolaou</i>	
1 Introduction and Historical Survey	131
2 Definition and General Properties of the Green Function	133
3 Single-Site Scattering	137
4 Multiple-Scattering Theory	140
5 KKR Representation of the Green Function	143
6 Description of the Full-Potential	145
7 Total Energy	147
8 Screened (Tight-Binding) KKR Method	149
9 Two-Dimensional Systems: Finite-Thickness Slabs and Half-Infinite Crystals	154
10 Self-Consistency Algorithm and Energy Contour Integration	155
Molecular Dynamics - Vision and Reality	
<i>Godehard Sutmann</i>	159
1 Introduction	159
2 Models and Methods for Particle Interactions	162
3 The Integrator	171
4 Conclusions	187
An Introduction to Ab Initio Molecular Dynamics Simulations	
<i>Dominik Marx</i>	195
1 Introduction	195
2 Theoretical Foundations	199
3 Summary and Outlook	235
 Advanced Concepts and Techniques	
Beyond Hartree-Fock: MP2 and Coupled-Cluster Methods for Large Systems	
<i>Christof Hättig</i>	245
1 Introduction	245
2 Second Order Methods for Ground and Excited States	246
3 The Resolution-of-the-Identity Approximation for ERIs	255

4	Parallel Implementation of RI-MP2 and RI-CC2 for Distributed Memory Architectures	261
5	RI-MP2 Calculations for the Fullerenes C ₆₀ and C ₂₄₀	266
6	Geometry Optimizations for Excited States with RI-CC2: The Intramolecular Charge Transfer States in Aminobenzonitrile Derivatives	270
7	Summary	274
The Korrington-Kohn-Rostoker (KKR) Green Function Method		
II. Impurities and Clusters in the Bulk and on Surfaces		
	<i>Peter H. Dederichs, Samir Lounis, and Rudolf Zeller</i>	279
1	Green Function Method	279
2	KKR Green Function Method for Impurities	281
3	Full-Potentials, Forces and Lattice Relaxations	283
4	Impurities in Metals and Semiconductors	287
5	Non-Collinear Configurations of 3d Impurities on Ferromagnetic Surfaces	290
6	Quantum Corrals on the Cu(111) Surface	294
7	Conclusions	297
The Optimized Effective Potential Method and LDA + U		
	<i>Stefan Kurth and Stefano Pittalis</i>	299
1	Introduction	299
2	Kohn-Sham Density Functional Theory	300
3	Orbital Functionals and the Optimized Effective Potential Method	302
4	Orbital Functionals for Exchange and Correlation	312
5	The LDA+U Method	320
6	Numerical Results	322
Many-Body Perturbation Theory: The GW Approximation		
	<i>Christoph Friedrich and Arno Schindlmayr</i>	335
1	Introduction	335
2	Theory	338
3	Implementation and Applications	343
4	Summary	349
Time-Dependent Density Functional Theory		
	<i>Nikos L. Doltsinis</i>	357
1	Introduction	357
2	Theory	358
3	Example Calculations	366
Free Energy and Rare Events in Molecular Dynamics		
	<i>Nikos L. Doltsinis</i>	375
1	Introduction	375
2	Constraint Techniques	376
3	Metadynamics	384

**Molecular Dynamics Beyond the Born-Oppenheimer Approximation:
Mixed Quantum–Classical Approaches**

<i>Nikos L. Doltsinis</i>	389
1 Introduction	389
2 Born-Oppenheimer Approximation	390
3 Mixed Quantum–Classical Approach	392
4 Approaches to Nonadiabatic Dynamics	394

Computing Properties

Surface Phase Diagrams from Ab Initio Thermodynamics

<i>Bernd Meyer</i>	411
1 Introduction	411
2 Gibbs Free Surface Energy	413
3 Gibbs Free Energies vs Total Energies	415
4 Limits for the Chemical Potentials	415
5 Restricted Thermodynamics	416
6 Example: Polar O–Terminated ZnO(000 $\bar{1}$) Surface	416

Spin-Polarized DFT Calculations and Magnetism

<i>Rudolf Zeller</i>	419
1 Introduction	419
2 Spin-Density-Functional Theory	420
3 Stoner Model for Ferromagnetism	422
4 Self-Consistency Iterations	429
5 Nanowires	432
6 Beyond the Stoner Model	435

Non-Collinear Magnetism: Exchange Parameter and T_C

<i>Gustav Bihlmayer</i>	447
1 Introduction	447
2 Vector-Spin Density Functional Theory	448
3 Magnetism and Exchange Interactions	454
4 Magnetic Anisotropy	462

Ab Initio Description of Electronic Transport

<i>Daniel Wortmann</i>	469
1 Introduction	469
2 Single-Particle View on Quantum Transport	471
3 DFT Calculations of Transport Setups	484
4 Summary	490

Parallel Computing and Numerical Methods

Introduction to Parallel Computing

Bernd Mohr

491

1	Introduction	491
2	Programming Models	494
3	MPI	495
4	OpenMP	498
5	Parallel Debugging	499
6	Parallel Performance Analysis	500
7	Summary	501

Parallel Linear Algebra Methods

Bernd Körfgen and Inge Gutheil

507

1	Introduction	507
2	The Poisson Problem	508
3	Vibration of a Membrane	516
4	Conclusion	519

Introduction to Hartree-Fock and CI Methods

Volker Staemmler

Chair of Theoretical Chemistry
Ruhr-Universität Bochum
Universitätsstr. 150, 44780 Bochum, Germany
E-mail: volker.staemmler@theochem.rub.de

This lecture presents a short introduction to the most elementary concepts of wave function based quantum chemistry. In the first part, the molecular Hamiltonian is specified and the Born-Oppenheimer approximation is made. Thereafter, the most important properties of electronic wave functions are discussed: normalization, antisymmetry and inclusion of the electronic spin. This leads to the concept of Slater determinants constructed from spin orbitals. How the molecular orbitals can be calculated by means of the Hartree-Fock approach is shown for the simplest case of a closed shell state. Finally, a presentation of the configuration interaction (CI) method which is used both for improving Hartree-Fock wave functions and for treating excited states concludes the lecture.

1 Introduction

In this lecture we will present a short introduction to the most elementary methods of wave function based quantum chemistry. We will primarily discuss the basic concepts and will not present long derivations or technical implementations in detail. The subject of this lecture is covered in all textbooks on electronic structure theory, a small selection of titles is contained in the list of references.¹⁻⁷ In particular, the presentation given in Ref. 7 is close to the one of the present lecture.

Wave function based quantum chemical methods can be characterized as being "ab initio" methods. This means that in these methods mathematical approximations to the full N-electron Schrödinger equation are constructed without the help of any adjustable parameters (except for the fundamental constants of physics) that can be fitted to reproduce experimental data. It is further characteristic for ab initio methods that they can be systematically improved towards the exact solution of the Schrödinger equation and that they possess an intrinsic criterion for the quality of the current approximation.

In the following sections we will use atomic units throughout

$$e = 1, m = 1, \hbar = 1, 4\pi\epsilon_0 = 1$$

Further, we will (nearly consistently) adopt the convention to use upper case letters for N-electron quantities and lower case letters for one-electron quantities. Only in some cases, when this convention too strongly contradicts the common use of symbols, we will deviate from it.

2 The Molecular Hamiltonian

Our aim is to solve the Schrödinger equation for the system under consideration (isolated atom, molecule, nanoparticle or extended system). In general, this system consists of a

certain number K of nuclei and a number N of electrons. The Hamiltonian, which determines the motion of the system, as well as its wave function will depend on the coordinates of all the nuclei and all the electrons. We will denote the coordinates of the nuclei by upper case letters $\underline{R}$ and those of the electrons by lower case letters $\underline{r}$. By underlining R and r we indicate that we consider the coordinates of all nuclei or electrons.

The Schrödinger equation for our system is given by

$$H(\underline{R}, \underline{r})\Psi(\underline{R}, \underline{r}) = E\Psi(\underline{R}, \underline{r}) \quad (1)$$

In the simplest case, i.e. if no external electrostatic or magnetic fields are present and if we restrict the interaction between the particles to the Coulomb interaction, the molecular Hamilton operator $H(\underline{R}, \underline{r})$ reads

$$H(\underline{R}, \underline{r}) = T_n + T_{el} + V_{n,el} + V_{el,el} + V_{n,n} \quad (2)$$

where

$$T_n = -\frac{1}{2} \sum_{\alpha}^K \frac{\Delta_{\alpha}}{M_{\alpha}} \quad (3)$$

(Δ_{α} being the Laplace operator for the coordinates of the α -th nucleus) is the operator of the kinetic energy of the nuclei,

$$T_{el} = -\frac{1}{2} \sum_i^N \Delta_i \quad (4)$$

is the operator of the kinetic energy of the electrons,

$$V_{n,el} = - \sum_i^N \sum_{\alpha}^K \frac{Z_{\alpha}}{r_{i\alpha}} \quad (5)$$

is the Coulomb attraction between the nuclei possessing the charges Z_{α} and the electrons,

$$V_{el,el} = \sum_i^N \sum_j^N \frac{1}{r_{ij}} \quad (6)$$

is the Coulomb repulsion between the electrons, and

$$V_{n,n} = \sum_{\alpha}^K \sum_{\beta}^L \frac{Z_{\alpha} Z_{\beta}}{r_{\alpha\beta}} \quad (7)$$

is the Coulomb repulsion between the nuclei. It must be noted that this Hamiltonian does not contain:

- a) External electric or magnetic fields. It is, however, quite straightforward to include the respective terms in the Hamiltonian. One has simply to take the classical expressions for the interaction of electric or magnetic moments with the corresponding fields and to replace the formulas with their quantum mechanical equivalents.
- b) Relativistic effects. The Hamiltonian given above is only valid for "slow" particles, i.e. for particles moving with a velocity much smaller than the velocity of light. This is generally sufficient for light elements, $Z < 40$, and also for the valence electrons of heavier elements. A rigorous theory of relativistic effects in quantum mechanical systems is rather complicated and beyond the scope of the present lecture.

- c) Spin dependent terms. Since the Hamiltonian (Eq. (2)) does not contain any terms which depend on the spins of the nuclei or the electrons it cannot be used to describe spin properties like spin-orbit or spin-spin couplings. However, as we will see in the next sections, some spin properties are introduced indirectly by the permutation symmetry of the wave function.

3 Born-Oppenheimer Approximation

For a system with K nuclei and N -electrons, the Schrödinger equation (Eq. (1)) is an eigenvalue equation in $3K + 3N$ independent variables, the cartesian coordinates of all the particles. Since it is in general impossible to solve such a problem exactly, either analytically or numerically, it is necessary to find approximate solutions, which are accurate enough to provide useful information.

The first of such approximations, which has been proposed 1927 by Born and Oppenheimer,^{8,9} consists in a separation of the motion of the light and therefore "fast" electrons from that of the "slow" nuclei. We write the total wave function $\Psi(\underline{R}, \underline{r})$ in the form of a product of an electronic wave function $\Phi_{el}(\underline{r}; \underline{R})$ and a wave function $\chi_n(\underline{R})$, which describes the motion of the nuclei:

$$\Psi(\underline{R}, \underline{r}) = \Phi_{el}(\underline{r}; \underline{R}) \cdot \chi_n(\underline{R}) \quad (8)$$

By means of this ansatz, the total Schrödinger equation (Eq. (1)) is decomposed into one equation for the electronic wave function $\Phi_{el}(\underline{r}; \underline{R})$

$$H_{el}\Phi_{el}(\underline{r}; \underline{R}) = E_{el}(\underline{R})\Phi_{el}(\underline{r}; \underline{R}) \quad (9)$$

with the electronic Hamiltonian

$$H_{el} = H - T_n = T_{el} + V_{n,el} + V_{el,el} + V_{n,n} \quad (10)$$

and one equation for the wave function $\chi_n(\underline{R})$ describing the motion of the nuclei

$$[T_n + E_{el}(\underline{R}) + C(\underline{R})]\chi_n(\underline{R}) = E\chi_n(\underline{R}) \quad (11)$$

The independent variables in the electronic wave function $\Phi_{el}(\underline{r}; \underline{R})$ are the coordinates $\underline{r}$ of the electrons; however, $\Phi_{el}(\underline{r}; \underline{R})$ depends also parametrically on the coordinates $\underline{R}$ of the nuclei, because the electronic Hamiltonian is a function of the positions $\underline{R}$ of the nuclei. That means that the electronic Schrödinger equation has to be solved for a given nuclear geometry ("clamped nuclei approximation"). The consequence is that the electronic energy, i.e. the eigenvalue $E_{el}(\underline{R})$ of the electronic Schrödinger equation (Eq. (9)), is not a constant, but depends also on the nuclear geometry. This geometry dependent electronic energy $E_{el}(\underline{R})$ plays the role of the potential energy in the Schrödinger equation (Eq. (11)) for the nuclear motion. It is therefore generally termed "potential energy surface" (PES). The energy E in Eq. (11), on the other hand, is a pure number, the energy eigenvalue of the total Schrödinger equation (Eq. (1)).

It has to be noted that concepts such as "potential energy curve" or "potential energy surface" $E(\underline{R})$ as well as geometrical structures etc. are only defined within the Born-Oppenheimer approximation!

Comments:

- a) One has also tried an ansatz similar to Eq. (8) in which the electronic wave function is independent of the nuclear geometry $\underline{R}$, the so-called "crude adiabatic approximation". But this scheme is so inaccurate that it is completely useless.
- b) The term $C(\underline{R})$ in Eq. (11) is called the "adiabatic diagonal correction". It is an operator acting on the nuclear wave function $\chi_n(\underline{R})$. Its explicit form is given by

$$C(\underline{R}) = \int \Phi_{el}^*(\underline{r}; \underline{R}) T_n \Phi_{el}(\underline{r}; \underline{R}) d\underline{r} - \sum_{\alpha} \frac{1}{M_{\alpha}} \int \Phi_{el}^*(\underline{r}; \underline{R}) \nabla_{\alpha} \Phi_{el}(\underline{r}; \underline{R}) d\underline{r} \cdot \nabla_{\alpha} \quad (12)$$

Because of the nuclear masses in the denominators in Eq. (12) this term is generally quite small and is mostly neglected altogether. Sometimes one distinguishes between the "Born-Oppenheimer" approximation, if $C(\underline{R})$ is neglected, and the "adiabatic approximation" if it is included.

- c) In most cases the Born-Oppenheimer approximation is a very good approximation. Generally, it is only necessary to go beyond it in cases of degeneracies or near degeneracies or at very high energies for the nuclear motion.
- d) Finally, if one has to go beyond the Born-Oppenheimer approximation, one can extend the ansatz (Eq. (8)) by using a linear combination of products of electronic and nuclear wave functions.

4 Requirements for the Electronic Wave Function

In the following, we will stay within the Born-Oppenheimer approximation and will only be concerned with the electronic Schrödinger equation (Eq. (9)). For simplicity, we will drop the subscript "el" and the nuclear coordinates $\underline{R}$ from $\Phi_{el}(\underline{r}; \underline{R})$.

A wave function which shall be used for describing the electronic structure of the system under consideration has to satisfy three requirements, in addition to being an approximate or exact solution of the electronic Schrödinger equation (Eq. (9)).

- a) Normalization. As for all quantum mechanical wave functions describing stationary states we will assume that $\Phi(\underline{r})$ is normalized to unity, i.e.

$$\int \Phi^*(\underline{r}) \Phi(\underline{r}) d\underline{r} = 1 \quad (13)$$

where the integration is to be performed over the coordinates of all N electrons. The condition (Eq. (13)) simply means that the probability for finding the system somewhere in space is unity.

- b) Antisymmetry with respect to the permutation of two electrons. Since electrons are fermions, the only solutions of the electronic Schrödinger equation, which can be used for describing electronic systems, have to be antisymmetric with respect to a permutation of any two electrons in the system. Mathematically speaking, only wave functions which belong to the totally antisymmetric representation of the permutation group of N electrons are allowed to describe N -electron systems.

If we denote a permutation of the electrons i and j by the operator P_{ij} , the antisymmetry requirement can be formulated as

$$P_{ij} \Phi(1, \dots, i, \dots, j, \dots, N) = \Phi(1, \dots, j, \dots, i, \dots, N) = -\Phi(1, \dots, i, \dots, j, \dots, N) \quad (14)$$

where we have simply indicated the order of the electrons by noting their serial numbers.

More generally, one can define an "antisymmetrizer" $\mathcal{A}$ which generates a fully antisymmetric wave function from any trial wave function $\tilde{\Phi}$

$$\mathcal{A} = \mathcal{N} \sum_P \epsilon_P P \quad (15)$$

where $\mathcal{N}$ is a normalization constant and the sum runs over all permutations P of the permutation group of N elements, with ϵ_P being the parity of the permutation P .

- c) Electronic spin. Since the electronic Hamiltonian (Eq. (10)) does not contain any spin operators, it does commute with the operators S_z and S^2 of the z-component and the square of the total electronic spin

$$[H, S_z] = 0, [H, S^2] = 0 \quad (16)$$

where S_z and S^2 are the N-electron spin operators

$$S_z = \sum_i^N s_{z_i}, S^2 = \sum_i^N s_i^2 \quad (17)$$

This means that electronic wave functions have to be eigenfunctions of S_z and S^2 with the eigenvalues M_S and $S(S+1)$.

A rigorous quantum mechanical treatment of one-electron with spin is only possible by means of the Dirac equation in which four-component or at least two-component one-electron wave functions (spinors) have to be used. Such a treatment is beyond the scope of this lecture. Here, we use a simplified scheme by representing the wave function of an electron with spin by a product

$$\psi(x, y, z, s) = \phi(x, y, z) \cdot \eta(s) \quad (18)$$

of a spatial part $\phi(x, y, z)$ which depends on the three spatial coordinates x, y, z and a spin function $\eta(s)$ which depends on the spin of the electron. There are only two possible spin functions $\eta(s)$ which are commonly denoted by " α " and " β " and which can be chosen to be orthogonal. In the quantum chemical literature the beta-spin is generally denoted by a bar above the orbital while the alpha-spin is not indicated, i.e. the one-electron wave functions with spin are

$$\phi(x, y, z) \cdot \alpha = \phi \quad (19)$$

$$\phi(x, y, z) \cdot \beta = \bar{\phi} \quad (20)$$

We will denote the one-electron wave functions with spin, $\psi(x, y, z, s)$ as given in Eq. (18), as "spin orbitals", while the spinfree one-electron wave functions $\phi(x, y, z)$ are simply called "orbitals".

In the following we will try to construct approximate N-electron wave functions. Whenever possible, they should satisfy the above mentioned requirements. There is no problem with the normalization. The antisymmetry can also be achieved, at least formally, by applying the antisymmetrizer (Eq. (15)). Finally, the electronic spin is taken care of at the one-electron level by using spin orbitals instead of pure spatial orbitals. However, it is generally difficult to construct N-electron spin eigenfunctions. We will see in the next section that in most cases it is quite easy to obtain eigenfunctions of S_z (simply because this is a sum of one-electron operators), but not for S^2 .

5 Slater Determinants

A closer inspection of the electronic Hamiltonian (Eq. (10)) shows that it contains one-electron terms, i.e. terms that depend only on the coordinates of one electron, two-electron terms, and the nuclear repulsion $V_{n,n}$ which is independent of the electron coordinates

$$H = \sum_i h(i) + \sum_{i<j} g(ij) + V_{n,n} \quad (21)$$

$$h(i) = -\frac{1}{2}\Delta_i - \sum_{\alpha}^K \frac{Z_{\alpha}}{r_{i\alpha}}, \quad g(ij) = \frac{1}{r_{ij}} \quad (22)$$

with $V_{n,n}$ given by Eq. (7).

Since $V_{n,n}$ is just a constant with respect to the electronic coordinates, it causes no problems. If the two-electron terms $g(ij)$ were absent, H would be a sum of terms each depending only on the coordinates of one electron. Then H would be separable: its eigenfunctions would be just products of eigenfunctions of the one-particle Hamiltonian h and its eigenvalues sums of eigenvalues of h . Unfortunately, H contains the two-electron repulsion terms $g(ij)$, and, also unfortunately, they are by no means small and cannot be simply neglected.

Nevertheless, we start by constructing trial wave functions as products of one-electron wave functions

$$\Phi(1, \dots, N) = \psi_a(1) \cdot \psi_b(2) \cdot \dots \cdot \psi_N(N) \quad (23)$$

Here, the serial number "1" again stands for the (space and spin) coordinates of electron number 1, etc. In order to account for the spins of the electrons, the one-electron wave functions $\psi_a(1)$ etc. in the product (Eq. (23)) are chosen to be the spin orbitals given in Eq. (18). Eq. (23) is interpreted by saying that the first electron is "occupying" the spin orbital ψ_a , the second electron the spin orbital ψ_b and so on.

Now, the product (Eq. (23)) does not possess the required antisymmetry property, therefore we have to apply the antisymmetrizer $\mathcal{A}$ (Eq. (15)) and obtain

$$\Phi(1, \dots, N) = \mathcal{A}[\psi_a(1) \cdot \psi_b(2) \cdot \dots \cdot \psi_N(N)] \quad (24)$$

Since the antisymmetrizer (Eq. (15)) is nothing else than an operator generating a determinant out of a simple product we can write the trial N-electron wave function in the form of a determinant

$$\Phi(1, \dots, N) = |\psi_a(1)\psi_b(2)\dots\psi_N(N)| \quad (25)$$

with $|\dots|$ being the usual notation for a determinant, but includes the normalization constant $\mathcal{N}$. Eq. (25) is indeed a determinant, the rows of which are numbered by the electrons and the columns by the occupied spin orbitals (or the other way round). Because of the antisymmetry of determinants, the Pauli principle is automatically satisfied; Eq. (25) vanishes if two of the occupied spin orbitals are identical. Such determinants, in which the elements are not numbers, but orbitals or spin orbitals, are called "Slater determinants".¹⁰ Though the form of our trial wave function (Eq. (25)) looks quite simple, one should not forget that it is a linear combination of product wave functions with as many as $N!$ terms.

In most cases it is assumed that the (spatial) orbitals ϕ_a form an orthonormal set, i.e. that the overlap integral between two orbitals is equal to the Kronecker δ

$$(\phi_a|\phi_b) = \int \phi_a^* \phi_b d\tau = \delta_{ab} \quad (26)$$

where the integration is over the three spatial coordinates of the electron. Because of the orthogonality of the spin functions α and β , the spin orbitals ψ_a form an orthonormal set as well

$$(\psi_a|\psi_b) = \int \psi_a^* \psi_b d\tau = \delta_{ab} \quad (27)$$

where the integration $d\tau$ also includes summation over the two spin orientations. In this case the normalization constant $\mathcal{N}$ in Eq. (25) is given by $1/\sqrt{(N!)}$. In principle, it is not necessary to choose orthonormal orbitals ϕ_a , however, all formulas become much simpler if Eq. (26) and Eq. (27) are satisfied.

N-electron wave functions in the form of Slater determinants possess several important properties:

- a) By construction, they satisfy the antisymmetry property and the Pauli principle.
- b) They offer an easy interpretation, since nothing seems more natural than the idea that a certain electron is attributed to a certain one-electron function or, stated differently, that an electron is occupying a certain orbital and possesses either the spin α or β . The molecular orbital theory of chemical bonding is entirely based on this interpretation. Of course, other forms of N-electron wave functions have also been proposed, e.g. in the so-called "valence bond theory", but such a simple and natural interpretation is lacking most of them.
- c) Whenever spin orbitals of the form (18)-(20) are used, Slater determinants are eigenfunctions of the N-electron spin operator S_z with an eigenvalue M_S equal to one-half of the difference between the number of α and β electrons:

$$S_z \Phi = M_S \Phi, M_S = \frac{1}{2}(N_\alpha - N_\beta) \quad (28)$$

However, except for a few simple cases, Slater determinants are not eigenfunctions of the N-electron spin operator S^2 . One of the exceptions is the situation of a closed shell system, in which there is an even number of electrons $N = 2n$ and exactly n spatial orbitals are doubly occupied. Fortunately, the stable ground states of most molecules belong to this case which will be treated in more detail in the next section. Whenever a single Slater determinant is not an eigenfunction of S^2 , one can construct eigenfunctions as linear combinations of Slater determinants, in which the same (spatial) orbitals are occupied, but with different spin orientations. Such linear combinations have coefficients which are fixed by spin or spatial symmetry and are called "configuration state functions" (CSFs).

- d) In many applications, e.g. in Hartree-Fock theory or in configuration interaction, one needs matrix elements of the N-electron Hamiltonian H or other N-electron operators between different Slater determinants. Though one Slater determinant consists of as many as $N!$ products of N spin orbitals, the construction of such matrix elements is straightforward and leads to rather simple formulas, in particular if the spatial orbitals ϕ_a are chosen orthonormal (Eq. (26)).

We will not report the general formulas for these matrix elements, but will present only two special cases, namely the normalization integral and the energy expectation value for the determinant

$$\Phi = |\phi_1 \bar{\phi}_1 \phi_2 \bar{\phi}_2 \dots \phi_n \bar{\phi}_n| \quad (29)$$

describing a closed shell system with an even number of electrons, $N = 2n$, and n (spatial) orbitals ϕ_i are doubly occupied. If one assumes that the orbitals ϕ_i are orthonormal, one can derive

$$(\Phi|\Phi) = 1 \quad (30)$$

$$(\Phi|H|\Phi) = \sum_{i=1}^n 2(i|h|i) + \sum_{i,j=1}^n [2(ii|jj) - (ij|ji)] \quad (31)$$

where the following one- and two-electron integrals occur: One-electron integral

$$(i|h|i) = \int \phi_i^*(1) h(1) \phi_i(1) dr_1 \quad (32)$$

two-electron Coulomb-integral

$$(ii|jj) = \int \phi_i^*(1) \phi_i(1) \frac{1}{r_{12}} \phi_j^*(2) \phi_j(2) dr_1 dr_2 \quad (33)$$

and two-electron exchange-integral

$$(ij|ji) = \int \phi_i^*(1) \phi_j(1) \frac{1}{r_{12}} \phi_j^*(2) \phi_i(2) dr_1 dr_2 \quad (34)$$

The N -electron expectation value (Eq. (31)) contains, as expected, the one-electron expectation values $(i|h|i)$ for all occupied orbitals and the Coulomb interaction $(ii|jj)$ between the occupied orbitals. In addition, there is a "new" term, the exchange interaction $(ij|ji)$. While the Coulomb-integral $(ii|jj)$ corresponds exactly to the classical Coulomb interaction between two charge distributions $\phi_i^* \phi_i$ and $\phi_j^* \phi_j$, the exchange interaction has no classical analogue and is a consequence of the antisymmetry requirement to the quantum mechanical wave function Φ .

Formulas for the matrix elements of the overlap matrix and of the N -electron Hamiltonian, between different Slater determinants Φ_I and Φ_J , can be generated as well.¹¹ Such formulas are needed in the method of configuration interaction as well as in all other post Hartree-Fock approaches.

6 Hartree-Fock Theory for Closed Shell States

Let us again consider an electronic system containing an even number of electrons, $N = 2n$. As mentioned above, this situation is realized for most ground states of stable molecules or periodic systems. In such cases, a one-determinant wave function of the form

$$\Phi_{SCF} = |\phi_1 \bar{\phi}_1 \phi_2 \bar{\phi}_2 \dots \phi_n \bar{\phi}_n| \quad (35)$$

in which n (spatial) orbitals ϕ_i are occupied by two electrons each, once with α and once with β spin, will be in general a good approximation to the true ground state wave function

of the N-electron Hamiltonian. It can be shown that such a determinant is an eigenfunction of both S_z and S^2 with the eigenvalues $M_S = 0$ and $S(S + 1) = 0$, i.e. it is a pure singlet state with a spin quantum number $S = 0$. The Hartree-Fock or self-consistent field (SCF) approach consists essentially in making the ansatz (Eq. (35)) for the N-electron wave function of the system under consideration.^{12,13}

The orbitals ϕ_i which are occupied in Φ_{SCF} are in general not known, but have yet to be determined. One choice could be the eigenfunctions of the one-particle Hamiltonian h (Eq. (22)). But this would mean to neglect the electron-electron repulsion entirely which yields too poor an approximation. In order to determine a reasonable set of orbitals one has to apply the variation principle to the wave function (Eq. (35)) and to vary the energy expectation value calculated with Φ_{SCF} as a functional of the occupied orbitals ϕ_i . The formula for this expectation value is given in Eq. (31), however, this simple formula is only valid if the occupied orbitals form an orthonormal set. That means that the energy expectation value has to be varied under the constraint that the orbitals remain orthonormal during the variation. This can be most easily achieved by using the method of the Lagrangean multipliers.

Therefore, we have to search for the minimum of the functional

$$\mathcal{J}[\phi] = \sum_{i=1}^n 2(i|h|i) + \sum_{i,j=1}^n [2(ii|jj) - (ij|ji)] - \sum_{i,j=1}^n 2\lambda_{ij}[(\phi_i|\phi_j) - \delta_{ij}] \quad (36)$$

by varying the orbitals ϕ_i . The coefficients λ_{ij} are the unknown Lagrangean multipliers. Performing this variation one arrives at the following one-particle eigenvalue equation

$$[h + \sum_j^{occ} (2J_j - K_j)]\phi_i = \sum_j^n \lambda_{ij} \phi_j \quad \text{for each } \phi_i \quad (37)$$

In this equation there are two operators describing the electron-electron interaction: The "Coulomb operator" J_j whose action on an arbitrary one-electron wave function ω is given by

$$J_j(1)\omega(1) = \int \phi_j^*(2)\phi_j(2) \frac{1}{r_{12}} dr_2 \cdot \omega(1) \quad (38)$$

and the "exchange operator" K_j

$$K_j(1)\omega(1) = \int \phi_j^*(2)\omega(2) \frac{1}{r_{12}} dr_2 \cdot \phi_j(1) \quad (39)$$

The Coulomb operator J_j describes the classical Coulomb repulsion that the electron number 2 occupying the orbital ϕ_j is exercising on the electron number 1 in the orbital ω , while the non-classical exchange operator is again the consequence of the antisymmetry of the N-electron wave function. It has to be noted that the Coulomb operator is a local operator, while the exchange operator is a non-local integral operator containing the wave function ω in the integral kernel. That means that for calculating the action of this operator onto a wave function ω at a certain point in space one has to know the function ω everywhere in space. Of course, this property renders the evaluation of exchange terms more complicated than that of Coulomb terms.

The sum over j on the l.h.s. of Eq. (37) runs over all n orbitals which are occupied in the SCF wave function Φ_{SCF} ; the sum on the r.h.s. of Eq. (37) involves the still unknown Lagrangean multipliers λ_{ij} .

Eq. (37) can be further simplified by noting that the SCF wave function Φ_{SCF} Eq. (35), being a determinant, is invariant with respect to unitary transformations among the occupied orbitals ϕ_i . This freedom can be used in order to bring the matrix λ_{ij} of the Lagrangean multipliers to a diagonal form. If we denote its diagonal elements by ϵ_i and introduce the "Fock operator" F by means of

$$F = h + \sum_j^{occ} (2J_j - K_j) \quad (40)$$

the one-electron equation determining the orbitals ϕ_i reads

$$F\phi_i = \epsilon_i\phi_i \quad (41)$$

This is the famous Hartree-Fock equation, which plays the role of a one-electron Schrödinger equation for the orbitals ϕ_i . Its main properties can be briefly characterized as follows:

- a) Eq. (37) or Eq. (41) is an integro-differential equation, since the kinetic energy operator contained in h involves differentiation and the exchange operators K_j are integral operators. Eq. (41) has the form of an eigenvalue equation, similar to the original N-electron Schrödinger equation (Eq. (9)).
- b) Since the Coulomb and exchange operators in the Fock operator F (Eq. (40)) have to be constructed from the orbitals ϕ_i which are yet to be determined by solving the Hartree-Fock equation (Eq. (41)), iterative schemes are necessary. Generally, one starts with some guess, $\phi_i^{(0)}$, constructs the operators $J_j^{(0)}$, $K_j^{(0)}$, and $F^{(0)}$ (Eqs. (38)-(40)) with them and solves Eq. (41). This will yield a new set of orbitals and orbital energies, $\phi_i^{(1)}$ and $\epsilon_i^{(1)}$. From them, a new Fock operator $F^{(1)}$ is constructed. Then again Eq. (41) can be solved, yielding the next set of orbitals and orbital energies, and so on. This procedure is repeated till "self-consistency" is reached, i.e. till the orbitals used for constructing the Fock operator are identical with the ones obtained by the solution of Eq. (41). Because of this self-consistency requirement, the Hartree-Fock method has the alternative name "self-consistent field" (SCF) method. Therefore we have already used the subscript "SCF" for the wave function (Eq. (35)).
- c) The Hartree-Fock equation (Eq. (41)) has infinitely many solutions, ϕ_i and ϵ_i , or rather a finite number of M solutions if one is working in a finite space of M one-electron basis functions (see next section). The ground state of the N-electron system is obtained if the energetically lowest Hartree-Fock orbitals are occupied in Φ_{SCF} (Aufbau principle). Generally one distinguishes between
 - occupied orbitals ϕ_i , $i = 1, \dots, n$, i.e. those orbitals occupied in Φ_{SCF} which are used to construct the Coulomb and exchange operators, and
 - virtual orbitals ϕ_i , $i > n$, which are not occupied in Φ_{SCF} . Strictly speaking, the virtual orbitals are not "optimized", since the energy expectation value (Eq. (31)) does not depend on them at all. They are, however, frequently used for electronically excited states as well as in CI approaches (see Section 8).
- d) The eigenvalues ϵ_i of the Hartree-Fock equation are called "orbital energies". In general, they are negative for occupied (bound) orbitals and positive for virtual (continuum) orbitals. The orbital energies of the occupied orbitals can be identified with the

ionization potentials of the system

$$\epsilon_i = -IP(i) \quad (42)$$

(Koopmans' theorem).¹⁴ Eq. (42) states that the orbital energy ϵ_i of the i -th occupied orbital is equal to the negative value of the energy necessary to remove one-electron from orbital ϕ_i . However, this is only an approximation because in deriving the theorem it has been assumed that the orbitals do not change (relax) after ionization. The accuracy of Koopmans' theorem is in the order of 0.5 - 2.0 eV for valence orbitals, for core orbitals the errors are much larger.

- e) After the Hartree-Fock equations have been solved and the Hartree-Fock orbitals have been obtained, one has to evaluate the Hartree-Fock energy E_{SCF} as the expectation value of the N -electron Hamiltonian, calculated with Φ_{SCF} (Eq. (35)). In contrast to the naive interpretation, E_{SCF} is not equal to the sum of the orbital energies of the occupied orbitals, since in this sum the electron-electron interaction is double-counted. Instead one finds

$$E_{SCF} = \sum_i^n [(i|h|i) + \epsilon_i] \quad (43)$$

The Hartree-Fock theory presented so far is only applicable to closed shell states which can be described by a wave function of the type of Eq. (35). It is, however, possible to extend this procedure to open shell systems, but the formalism is more complicated than for closed shell states. A detailed discussion is beyond the aim of this lecture; we will only give the acronyms for the most widely used variants of open shell Hartree-Fock theory: ROHF (restricted open shell Hartree-Fock) and UHF (unrestricted Hartree-Fock) for simpler cases where a one-determinant open shell wave function can be used, and MC-SCF (multi-configuration SCF) and CASSCF (complete active space SCF) for cases in which a multi-determinant wave function is necessary.

7 Roothaan SCF

Though the Hartree-Fock theory has been formulated in the early 1930ths, for a long time numerical calculations could only be performed for atoms. The reason is that Eq. (41) is an integro-differential equation in the 3D space which could not be solved by e.g. finite difference methods. Numerical solutions were only possible for atoms where one can separate the angular from the radial part and is left with 1D radial equations.

The breakthrough for molecular calculations came in 1951 when Hall¹⁵ and Roothaan¹⁶ independently proposed to expand the Hartree-Fock orbitals into a set of atom centered basis functions. At first, atomic orbitals were used for that purpose,¹⁷ but it turned out that the calculation of the necessary one- and two-electron integrals was much too difficult and time consuming. The proposal of Boys¹⁸ to use Gaussian functions instead opened the way to the field of electronic structure calculations for molecules.

The idea of Roothaan is to expand the unknown Hartree-Fock orbitals ϕ_i into a finite set of M known functions χ_ν , called "basis functions"

$$\phi_i = \sum_{\nu=1}^M c_{i\nu} \chi_\nu \quad (44)$$

Now, the coefficients $c_{i\nu}$ determine the orbitals ϕ_i and have to be calculated. One can either use the variation principle again and vary the energy expectation with respect to these coefficients or one can project the Hartree-Fock equation (Eq. (41)) onto the basis functions χ_ν . In both cases one obtains a matrix eigenvalue equation of the form

$$\sum_{\nu}^M F_{\mu\nu} c_{i\nu} = \epsilon_i \sum_{\nu}^M S_{\mu\nu} c_{i\nu}, \quad \text{for all } \mu = 1, \dots, M \quad (45)$$

or in a matrix notation

$$\underline{F} \cdot \underline{c}_i = \epsilon_i \underline{S} \cdot \underline{c}_i \quad (46)$$

where $\underline{F}$ and $\underline{S}$ are the Fock and overlap matrices and $\underline{c}_i$ the coefficient vectors. The matrix elements of the overlap matrix are defined as in Eq. (26)

$$S_{\mu\nu} = (\mu|\nu) = (\chi_\mu|\chi_\nu) = \int \chi_\mu^* \chi_\nu d\tau \quad (47)$$

In general, the basis functions cannot be chosen in such a way that S is the unit matrix. The Fock matrix is constructed from one- and two-electron integrals

$$F_{\mu\nu} = (\mu|h|\nu) + \sum_j \sum_{\rho,\sigma} c_{j\rho} c_{j\sigma} [2(\rho\sigma|\mu\nu) - (\rho\nu|\mu\sigma)] \quad (48)$$

with the matrix elements of the one-electron Hamiltonian (one-electron integrals) are given by

$$(\mu|h|\nu) = (\chi_\mu|h|\chi_\nu) = \int \chi_\mu^* h \chi_\nu d\tau \quad (49)$$

and the two-electron integrals by

$$(\rho\sigma|\mu\nu) = \int \chi_\rho^*(1) \chi_\sigma(1) \frac{1}{r_{12}} \chi_\mu^*(2) \chi_\nu(2) d\tau_1 d\tau_2 \quad (50)$$

The solution of the Hartree-Fock equation (Eq. (41)) is now achieved by the solution of the matrix eigenvalue equation (Eq. (45)), which yields the orbital energies ϵ_i and the coefficient vectors $c_{i\nu}$. Obviously, the "analytic" form of Eq. (41) and the "algebraic" form of Eq. (45) of the Hartree-Fock equation are only equivalent if $M \rightarrow \infty$, i.e. if a complete one-electron basis set is employed ("SCF limit").

Of course, the necessary numerical effort as well as the accuracy which can be achieved depend crucially on the quality and the size of the basis set χ_ν . The numerical effort is easy to estimate: There are M^4 two-electron integrals $(\mu\nu|\rho\sigma)$ for a basis with M basis functions χ_ν and the solution of the Hartree-Fock equation (45) requires the diagonalization of a M -dimensional matrix which is a M^3 step. On the other hand, the design of good basis sets requires a lot of experience and is by no means trivial. Fortunately, reasonable basis sets, spanning a wide range from rather poor and fast to very accurate and time consuming, have been determined during the last 50 years for all atoms and most applications. They are available in the commercial program packages.

Today, one can distinguish between two technical implementations of the Roothaan SCF procedure:

- a) Conventional SCF: One chooses a basis set, calculates all one- and two-electron integrals $(\mu|\nu)$, $(\mu|h|\nu)$, $(\mu\nu|\rho\sigma)$ and stores them on a peripheral storage device (disk). All integrals are read in and processed in every SCF iteration. Since I/O is slow compared to CPU in modern computers, this method is strongly I/O bound. The large number of two-electron integrals requires much peripheral space. The largest basis sets that can be handled this way can contain about 800-1200 basis functions.
- b) Direct SCF¹⁹: The two-electron integrals $(\mu\nu|\rho\sigma)$ are not calculated and stored prior to the SCF iterations, but are recalculated in each iteration. Of course, only those integrals are calculated which are really needed. For instance, Eq. ((48) shows that it is not the integral $(\mu\nu|\rho\sigma)$ itself that is needed, but always a combination like

$$c_{j\rho}c_{j\sigma}[2(\rho\sigma|\mu\nu) - (\rho\nu|\mu\sigma)] \quad (51)$$

i.e. a product of a density matrix element $c_{j\rho}c_{j\sigma}$ and a combination of integrals. That means, it is not necessary to calculate the integral if the density matrix element is too small or does not change from one iteration to the next ("prescreening" and "updating"). This way one can use basis sets of several thousand functions without needing much peripheral space. With efficient prescreening techniques the necessary computer times are not larger than for the conventional procedure.

8 Configuration Interaction

The Hartree-Fock method yields, even in favorable cases and if large one-particle basis sets are employed, only an approximation to the exact solution of the electronic Schrödinger equation (9). In many cases, in particular when accurate results are to be obtained, one has to go beyond the Hartree-Fock method.

In general, we denote all effects beyond the Hartree-Fock level as "correlation" effects. We will especially define the correlation energy of a system as the difference between the "exact" energy eigenvalue of the N-electron Hamiltonian and the SCF energy:

$$E_{corr} = E - E_{SCF} \quad (52)$$

The conceptually simplest method for accounting for correlation effects is the method of configuration interaction (CI). The main idea is based on the following lemma (expansion theorem): Let $\phi_i, i = 1, \dots, \infty$ be a complete basis of the one-particle Hilbert space. Then it can be proven that all N-electron Slater determinants

$$\Phi_I(1, \dots, N) = |\phi_{I1}(1)\phi_{I2}(2)\dots\phi_{IN}(N)| \quad (53)$$

which can be constructed by placing the N electrons into N of these orbitals in all possible ways, span the full antisymmetric N-particle Hilbert space. That means that any N-electron wave function can be expanded into the set of determinants $\Phi_I(1, \dots, N)$ (Eq. (53)). This holds, of course, also for the eigenfunctions of the N-electron Hamiltonian, i.e. we can make the following ansatz for them

$$\Psi_i(1, \dots, N) = \sum_I C_{iI} \Phi_I(1, \dots, N) \quad (54)$$

where we have denoted the eigenfunctions by Ψ_i in order to distinguish them from the Slater determinants Φ_I . It has to be stressed, that the expansion (Eq. (54)) holds for any

complete set of one-electron functions, i.e. for any basis of the one-particle Hilbert space, but the expansion coefficients C_{iI} will of course depend on the choice of the basis.

The CI method uses this lemma in the following manner: As usual, one selects a set of atomic basis functions, $\chi_\mu, \mu = 1, \dots, M$. By means of a Hartree-Fock calculation or simply by an orthogonalization procedure a set of orthonormal molecular orbitals $\phi_i, i = 1, \dots, M$ is constructed as linear combinations of the χ_μ . By combining the spatial orbitals with the spin functions α and β (Eqs. (19),(20)) one obtains $2M$ spin orbitals ψ_i . They are used to build up Slater determinants in the form indicated in Eq. (53) or configuration state functions (CSFs or simply "configurations") as fixed linear combinations of such determinants. And finally, the eigenfunctions of the N-particle Hamiltonian are expanded into these determinants as shown in Eq. (54). Of course, in real calculations one can only use a finite number M of atomic basis functions and therefore the Φ_I span only a finite subspace of the full N-particle Hilbert space.

For the determination of the expansion coefficients C_{iI} (CI coefficients) the variation principle is invoked again: The expectation value of the N-electron Hamiltonian, calculated with the wave function Ψ_i , has to be minimized as a function of the C_{iI} . Since Ψ_i depends only linearly on the parameters C_{iI} , the application of the variation principle is particularly simple ("linear" or "Ritz" variation principle) and leads to a set of linear equations

$$\sum_J (H_{IJ} - ES_{IJ})C_J = 0, \quad \text{for all } I \quad (55)$$

The quantities H_{IJ} and S_{IJ} in these equations are the matrix elements of the N-particle Hamiltonian and the overlap matrix calculated with the Slater determinants Φ_I and Φ_J

$$H_{IJ} = (\Phi_I | H | \Phi_J), \quad S_{IJ} = (\Phi_I | \Phi_J) \quad (56)$$

They can be easily evaluated (see Section 5) as long as the orbitals ϕ_i form an orthonormal set.

Eq. (55) is a set of homogeneous linear equations which has only non-trivial solutions if the determinant of the coefficient matrix vanishes. This leads to the secular equation

$$|H_{IJ} - ES_{IJ}| = 0 \quad (57)$$

This is an algebraic equation of the order of the dimension, P say, of the matrices H_{IJ} and S_{IJ} . Its solutions are the eigenvalues E_i and the eigenvectors C_{iI} of the electronic N-particle Hamiltonian, projected onto the finite, P -dimensional N-particle Hilbert space spanned by the Φ_I . Of course, one will obtain exactly P eigenvalues and eigenvectors.

Solving the secular equations (55) is equivalent to diagonalizing the matrix H_{IJ} with the metric S_{IJ} .

Though the concept as well as the numerical implementation of the CI approach are extremely simple, the method has one decisive disadvantage: The number of possible determinants Φ_I is extremely large, even for a small number N of electrons and a small size M of the atomic basis set. Though many methods have been designed and tested during the last 40 years, both for an efficient evaluation of the matrix elements H_{IJ} and S_{IJ} and for diagonalizing very large matrices, the CI problem is still not solved.

Similar to the Hartree-Fock-Roothaan approach we can distinguish between two technical implementations of the CI method: In the "conventional" scheme, the P^2 matrix elements H_{IJ} (or the matrix element formulas) are calculated once and stored on a peripheral storage device ("formula tape"), before they are processed in the iterations needed to

diagonalize H_{IJ} . For large CI dimensions P , the formula tape is getting extremely lengthy. Therefore, modern CI techniques prefer the "direct" approach, in which only those matrix elements H_{IJ} are calculated in each iteration which are needed to generate a correction vector

$$\sigma_I^{(n+1)} = \sum_J \frac{H_{IJ}}{H_{II} - E^{(n)}} C_J^{(n)} \quad (58)$$

to the CI vector $C_I^{(n)}$ with the energy $E^{(n)}$ of the preceding iteration. This way it is possible to treat CI problems with 10^5 to 10^9 determinants.

At present, CI type methods are used for several purposes:

1) Full CI

This seems to be the most straightforward scheme: As described above, one selects a finite number M of atomic basis functions, orthogonalizes them (this is only a technical step in order to be able to use simpler formulas for the matrix elements H_{IJ} and S_{IJ}), constructs all possible determinants Φ_I (maybe CSFs, adapted to the spin and spatial symmetry of the system under consideration instead of determinants), and diagonalizes H_{IJ} . As indicated above, this is only possible for small systems and small atomic basis sets. Such calculations serve primarily as benchmarks for testing the quality of approximate schemes (truncated CI, coupled-cluster methods and so on).

2) Truncated CI

Most CI treatments are aiming at calculating correlation effects for the electronic ground state of the system. For such a purpose one generally starts from the Hartree-Fock wave function which is a good zeroth order approximation and calculates corrections to it. Full CI calculations make no sense for such cases since the vast majority of the configurations (or determinants) do not contribute at all to the wave function or the correlation energy. Rather, one has to try to select the "important" configurations and discard the unimportant ones.

In general, this is done in the following way (for the case of a closed shell ground state): One starts from a "reference" determinant, mostly the SCF wave function as given in Eq. (35)

$$\Phi_{SCF} = |\phi_1 \bar{\phi}_1 \phi_2 \bar{\phi}_2 \dots \phi_n \bar{\phi}_n| \quad (59)$$

Then one defines singly, doubly, triply, ... excited determinants, in which one, two, three, ... of the spin orbitals occupied in Φ_{SCF} are replaced by virtual orbitals. If we denote the occupied orbitals by ϕ_i and $\bar{\phi}_j$ and the virtual orbitals by ϕ_a and $\bar{\phi}_b$ we have

Singles (singly excited determinants)

$$\Phi(i \rightarrow a) = |\phi_1 \bar{\phi}_1 \phi_2 \bar{\phi}_2 \dots \phi_a \bar{\phi}_i \dots \phi_n \bar{\phi}_n| \quad (60)$$

and similarly by replacing $\bar{\phi}_i$ with $\bar{\phi}_a$.

Doubles (doubly excited determinants)

$$\Phi(ij \rightarrow ab) = |\phi_1 \bar{\phi}_1 \phi_2 \bar{\phi}_2 \dots \phi_a \bar{\phi}_i \dots \phi_b \bar{\phi}_j \dots \phi_n \bar{\phi}_n| \quad (61)$$

Triples and higher excited determinants are defined in an analogous way. The ansatz for the CI wave function reads

$$\Psi_{CI} = \Phi_{SCF} + \sum_{i,a} C(i \rightarrow a) \Phi(i \rightarrow a) + \sum_{ij,ab} C(ij \rightarrow ab) \Phi(ij \rightarrow ab) + \dots \quad (62)$$

CI expansions truncated after the singles, doubles, triples, ... are called CIS, CISD, CISDT, ... expansions.

If one is interested in correlation effects for the electronic ground state, only the lowest root of the corresponding secular problem is calculated and analyzed. The "excited" determinants, or better "substituted" determinants, serve only for improving the description of the ground state and do not describe spectroscopically excited states.

If SCF orbitals are used in the CI, as in Eqs. (59)-(62), one can show

$$(\Phi_{SCF} | H | \Phi(i \rightarrow a)) = 0 \quad (63)$$

(Brillouin theorem). Similarly, all matrix elements of H between Φ_{SCF} and triples or higher excited determinants are zero, since H contains only one- and two-electron operators. That means that in first order only the doubles contribute to the correlation effects. In higher orders, also the singles, triples,... do contribute, but their contributions are rather small. Therefore, in many applications truncated CI-expansions are used, the most popular approach is CISD. (The singles could be discarded as well, but their number is so small that they are mostly included.)

The problem with the CISD approach is that it is not size consistent. This disadvantage is remedied by means of the "coupled-cluster" approaches which enable at present the most accurate electronic structure calculations, at least for small systems.

3) CIS

The most widely used method for calculating properties of electronically excited states is CIS (CI truncated to singles). Again, one starts from the Hartree-Fock wave function (Eq. (58)) of the corresponding electronic ground state and performs a CI calculation in which all single excitations from all occupied into all virtual orbitals are included:

$$\Psi_{CIS} = \Phi_{SCF} + \sum_{i,a} C(i \rightarrow a) \Phi(i \rightarrow a) \quad (64)$$

In contrast to the CISD approach designed for calculating correlation energies for ground states, now, the higher roots of the secular equations are used as approximations for singly excited electronic states. In this way, one can obtain a reasonably accurate description of the optical spectrum of the system.

4) Multi-reference CI (MR-CI)

In many cases, in particular in open shell systems, for excited states, or in cases of degeneracies or near degeneracies, a one-determinant Hartree-Fock wave function is not a good description, i.e. not a good zeroth order wave function, for the state under consideration. In such cases, one is forced to start from a multi-reference wave function (e.g. MC-SCF = multi configuration SCF or CASSCF = complete active space SCF) and continue with a CI on top of this reference.

Here is not enough place to go into more detail concerning the different CI versions or the other post-Hartree-Fock methods, like coupled-cluster approaches and perturbation theory. The reader is referred to the textbooks on electronic structure theory, in particular to the Ref. 5 and 7.

Exercises and Problems

- 1) Derive Eq. (12) using Eqs. (8) and (9). Hint: Multiply (8) from left with Φ_{el}^* and integrate over the electronic coordinates $\underline{r}$.
- 2) Write down explicitly the determinantal wave function (Eq. (25)) for a system with three electrons occupying three spin orbitals ψ_a , ψ_b , and ψ_c . How many spin orbital products does this wave function contain?
- 3) Prove Eq. (28) by using the eigenvalue equation of the one-electron spin operator s_z :

$$s_z \alpha = \frac{1}{2} \alpha, s_z \beta = -\frac{1}{2} \beta \quad (65)$$

- 4) Show that the three two-electron wave functions

$$|\phi_a(1)\phi_b(2)| \quad (66)$$

$$\frac{1}{\sqrt{2}}[|\phi_a(1)\bar{\phi}_b(2)| + |\bar{\phi}_a(1)\phi_b(2)|] \quad (67)$$

$$|\bar{\phi}_a(1)\bar{\phi}_b(2)| \quad (68)$$

are products of an antisymmetric space and a symmetric spin function, while

$$\frac{1}{\sqrt{2}}[|\phi_a(1)\bar{\phi}_b(2)| - |\bar{\phi}_a(1)\phi_b(2)|] \quad (69)$$

is the product of a symmetric space and an antisymmetric spin function. Determine the M_S values for these wave functions. Can you guess the spin quantum numbers S for them?

- 5) Derive formulas for the energy expectation values of the two-electron wave functions given in problem 4.
- 6) Derive Eq. (37).
- 7) Derive Koopmans' theorem (Eq. (42)). Assume that the orbitals for the ion are the same as for the ground state of the neutral molecule, write down the one-determinant wave function for the ion, derive the energy expectation value for the ion and calculate the difference

$$IP(i) = E(\text{ion}) - E_{SCF} \quad (70)$$

- 8) Derive the matrix form (Eq. (45)) of the Hartree-Fock equation (41) by projecting to the basis function χ_μ . Hint (as in problem 1)): Insert ansatz Eq. (44) into Eq. (41), multiply from left by χ_μ^* and integrate over the coordinates of the electron.
- 9) Derive the secular equations (55). Hint: Write down the expectation value of the N-electron Hamiltonian H with the wave function (Eq. (54)) and vary the coefficients C_I .
- 10) Set up all possible Slater determinants Φ_I for a three-electron system (e.g. the Li atom), using a set of four (spatial) orbitals $\phi_i, i = 1, 2, 3, 4$. Determine their M_S values and their excitation levels with respect to the "ground state" determinant

$$|\phi_1\bar{\phi}_1\phi_2| \quad (71)$$

- 11) How many Slater determinants can be constructed for H_2O in a basis set of 24 basis functions? How many of them are doubly excited with respect to the Hartree-Fock determinant?
- 12) In the region of the equilibrium distance R_e , the ground state wave function for the H_2 molecule can be fairly well approximated by a single Hartree-Fock determinant

$$\Psi_1 = |\phi_1 \bar{\phi}_1| \quad (72)$$

where

$$\phi_1 = \frac{1}{\sqrt{2(1+S)}}(\chi_a + \chi_b) \quad (73)$$

is the lowest molecular orbital, expressed as a linear combination of the two 1s atomic orbitals at the two hydrogen atoms. (S is the overlap integral between χ_a and χ_b). Show that this wave function does not dissociate correctly into two H atoms. Hint: Consider the limit

$$\phi_1 \rightarrow \frac{1}{\sqrt{2}}(\chi_a + \chi_b) \quad \text{for } R \rightarrow \infty \quad (74)$$

Show further that the two-configuration wave function

$$\Psi_2 = \frac{1}{\sqrt{2}}(|\phi_1 \bar{\phi}_1| - |\phi_2 \bar{\phi}_2|) \quad (75)$$

$$\phi_2 = \frac{1}{\sqrt{2(1-S)}}(\chi_a - \chi_b) \quad (76)$$

dissociates correctly.

References

1. R. McWeeny, *Methods of Molecular Quantum Mechanics*, Academic Press (1992).
2. F. Pilar, *Elementary Quantum Chemistry*, McGraw-Hill (1990).
3. I. N. Levine, *Quantum Chemistry*, Prentice-Hall (1991).
4. J. P. Lowe, *Quantum Chemistry*, Academic Press (1993).
5. A. Szabo and N. S. Ostlund, *Modern Quantum Chemistry*, Academic Press (1982).
6. W. J. Hehre, L. Radom, J. A. Pople, and P. v. R. Schleyer, *Ab initio Molecular Orbital Theory*, Wiley (1986).
7. F. Jensen, *Introduction to Computational Chemistry*, Wiley (1999).
8. M. Born and J. R. Oppenheimer, *Ann. Physik* **84**, 457 (1927).
9. W. Kolos and L. Wolniewicz, *J. Chem. Phys.* **41**, 3663 (1964).
10. J. C. Slater, *Phys. Rev.* **34**, 1293 (1929).
11. P. O. Löwdin, *Phys. Rev.* **97**, 1474, 1490, 1509 (1957).
12. D. R. Hartree, *Proc. Cambridge Phil. Soc.* **24**, 89, 111 (1928).
13. V. Fock, *Z. Physik* **61**, 126 (1930).
14. T. A. Koopmans, *Physica* **1**, 104 (1933).
15. G. G. Hall, *Proc. Roy. Soc. (London)*, A205, 541 (1951).
16. C. C. J. Roothaan, *Rev. Mod. Phys.* **23**, 69 (1951).
17. M. P. Barnett and C. A. Coulson, *Phil. Trans. Roy. Soc. London* A243, 221 (1951).
18. S. F. Boys, *Proc. Roy. Soc. (London)*, A200, 542 (1950).
19. J. Almlöf, K. Fägri Jr., and K. Korsell, *J. Comput. Chem.* **3**, 385 (1982).

Basis Sets, Accuracy, and Calibration in Quantum Chemistry

Thomas Müller

John von Neumann Institute for Computing
Central Institute for Applied Mathematics
Forschungszentrum Jülich
52425 Jülich, Germany
E-mail: th.mueller@fz-juelich.de

This lecture primarily deals with aspects of quantum chemical *ab initio* cluster codes without periodic boundary conditions. The most important parameters, which any user of quantum chemical program packages must select when setting up a problem is the choice of the model Hamiltonian and the one-electron and N-electron basis set. As there is no general fool-proof recipe available that covers the general case, we approach the problem by discussing the most relevant approximations and their implications. The aim is to provide the reader with a basic idea of the various approximations and to what extent results and conclusions maybe affected. Finally it is anticipated that most readers will not carry out quantum chemical calculations for their own sake but rather use them as a tool to extract data to be processed further for the investigation of the problem in question. Here, selecting an appropriate theoretical method and being aware of their potential shortcomings is even more important as not to jump to entirely false conclusions. Following a short review of the basic models and approximations, the major errors are discussed in some detail. Extrapolation and calibration methods are presented and illustrative case studies are discussed.

1 Introduction

Although the general aim is to strive for (highly) accurate solutions of the many-body Schrödinger equation this is possible only for the smallest systems because the resource requirements in terms of computer time, memory and disk space increase rather unfavourably with the size of the system. However, in many cases there is no need to carry out calculations at high accuracy since experimental problems associated with larger systems limit the resolution and correspondingly theoretical treatment of lower accuracy may be perfectly sufficient. Frequently, even semiquantitative answers are all that is needed and occasionally apparently simple problems are hard to solve accurate enough. Thus the notion of accuracy depends upon the specific problem in question.

A central aspect of quantum chemistry is the explicit or implicit determination of state-specific potential energy surfaces (PES), predicting geometries, spectra and chemical reactions. As these properties solely depend upon *relative* energies accurate calculations need to reproduce the shape of the exact PES. In fact most quantum chemical models are not generally applicable to arbitrary sections of the PES. Properties, which primarily sample regions of the electron density of little importance for the energy require some care. The requirements for their calculation may be drastically different from those for the energy. Hyperpolarizabilities, which determine various nonlinear optical effects, for example depend on the outer charge density distribution and are quite sensitive to dynamic electron correlation effects.

Being aware about the conceptual limitations of methods and to anticipate the applicability of a given approach is essential. It is advisable to judge the accuracy of quantum chemical calculations not simply by comparison to experiment. Experimental data frequently cannot be compared to quantum chemical results without further corrections for effects not included in the (quantum chemical) model Hamiltonian and the processing of the experimental raw data itself relies frequently on assumptions, which may or may not hold accurately enough. This might render an interpretation of the deviation impossible. It is quite tempting to adjust the various parameters of a quantum chemical calculation such that the results finally agree with the available experimental data one has set out to reproduce or to explain. This is - apart from possibly leading to entirely wrong conclusions - entirely useless, since it prevents us from analysing the result and thereby attaining a better understanding of the underlying problem beyond merely generating a collection of numbers.

The preferred approach is to assess the quality of a calculation within the theoretical approach. Comparison with experimental data follows as the second step. Constructing series of calculations in order to detect or to eliminate possible sources of error as well as using series of calculations of increasing accuracy for suitably chosen reference systems to calibrate the methods whenever necessary. Finally, extrapolation techniques may prove helpful in order to overcome practical limitations associated with the wave function expansion.

2 Basic Approximations

2.1 Hamiltonian

The standard molecular Hamiltonian used in electronic structure calculations in atomic units contains the electronic kinetic energy $\mathcal{T}_e$, electron-electron repulsion $\mathcal{V}_{ee}$, electron-nuclei attraction $\mathcal{V}_{ne}$ and the nuclear-nuclear repulsion energy V_{nn} and runs

$$\mathcal{H} = -\frac{1}{2} \sum_i \nabla_i^2 + \sum_{i>j} \frac{1}{|\mathbf{r}_i - \mathbf{r}_j|} - \sum_{iA} \frac{Z_A}{|\mathbf{r}_i - \mathbf{R}_A|} + \sum_{A>B} \frac{Z_A Z_B}{|\mathbf{R}_A - \mathbf{R}_B|} \quad (1)$$

$$= -\frac{1}{2} \sum_i \nabla_i^2 + \sum_{i>j} r_{ij}^{-1} - \sum_{iA} \frac{Z_A}{r_{iA}} + \sum_{A>B} \frac{Z_A Z_B}{R_{AB}} \quad (2)$$

$$= \mathcal{T}_e + \mathcal{V}_{ee} + \mathcal{V}_{ne} + V_{nn}. \quad (3)$$

Z denotes the nuclear charge, $\mathbf{R}$ and $\mathbf{r}$ are nuclear and electronic position vectors, respectively. Summations run over all electrons (index i, j) and nuclei (index A, B). This equation implies several simplifications: (i) the particles interact non-relativistically solely through Coulombic interaction (ii) the (clamped-nucleus) Born-Oppenheimer Approximation and (iii) the particles are described as point charges neglecting the finite size of the nuclei.

Relativistic contributions to the total energy, which scale as Z^4/c^2 with c the speed of light (≈ 137 a.u.), must be included for the heavier elements in at least some approximate manner. The impact of spin-orbit coupling may be of similar importance as electron correlation for the heavier elements and cannot be straightforwardly ignored (cf. Section 3.6).

The Born-Oppenheimer approximation^{1,2}, sometimes referred to as *adiabatic* approximation³, decouples electron motion from the motion of the nuclei. It works well wherever the gradient of the electronic wave function with respect to the nuclear coordinates is small. This assumption may be violated close to avoided crossings between different electronic states, where the characteristics of the participating states rapidly change. Proper treatment of these regions, which are of great importance for the understanding of the dynamics of photochemical reactions, invariably require the inclusion of *non-adiabatic* coupling terms⁴.

2.2 Wave Function Expansion

The eigenvalues and eigenfunctions of the time-independent Schrödinger equation

$$\mathcal{H}\Psi = E\Psi \quad (4)$$

are sought. As analytical solutions are unavailable except for the one-electron case an expansion in an N-particle basis set $\{\Phi\}$ offers a solution. For an initial trial wave function $\hat{\Psi}(\alpha)$, which depends on a set of parameters $\{\alpha\}$ the energy functional runs

$$\hat{E}(\alpha) = \frac{\langle \hat{\Psi}(\alpha) | \mathcal{H} | \hat{\Psi}(\alpha) \rangle}{\langle \hat{\Psi}(\alpha) | \hat{\Psi}(\alpha) \rangle}. \quad (5)$$

By virtue of the variation theorem the approximation $\hat{E}(\alpha)$ is an upper bound to the exact energy of Eq. (4) and the wave function $\hat{\Psi}(\alpha)$ will converge on average to the exact wave function Ψ . Using a linear expansion in α and making the energy stationary

$$\hat{\Psi}(\alpha) = \sum_i \alpha_i \Phi_i \quad (6)$$

with respect to α yields a generalized eigenvalue problem

$$\mathbf{H}\mathbf{c} = E\mathbf{S}\mathbf{c} \quad (7)$$

$$\mathbf{H}_{ij} = \langle \Phi_i | \mathcal{H} | \Phi_j \rangle \quad (8)$$

$$\mathbf{S}_{ij} = \langle \Phi_i | \Phi_j \rangle \quad (9)$$

$\mathbf{H}_{ij}$ and $\mathbf{S}_{ij}$ denote the hamiltonian and overlap matrix elements with respect to the N-particle basis functions Φ_i and Φ_j . The Dirac notation implies $\langle \Phi_i | \mathcal{O} | \Phi_j \rangle = \int \Phi_i^* \mathcal{O} \Phi_j d\tau$. As the basis may be chosen orthonormal, the overlap matrix is replaced by the unit matrix. A complete N-particle basis set yields the exact solution of Eq. (4). However, such a basis would be infinite in size and rather impractical to use.

Guidelines for the construction of the N-particle basis are the efficient representation of the N-particle space and the rapid evaluation of the respective matrix elements, i.e. the computation of a vast number of multi-dimensional N-electron integrals appearing in Eq. (8). Conventionally, the N-electron basis is constructed from antisymmetrized products of orthonormal spinorbitals $\phi_{ki}(\mathbf{r}_i)$, denoted Slater determinants Φ .

$$\Phi_i = \mathcal{A} \prod_k^N \phi_{ki}(\mathbf{r}_i) \quad (10)$$

The spinorbitals are a product of a spatial part $\tilde{\phi}_{ki}(\mathbf{r}_i)$ times a spin part $\sigma_i(\omega_i)$. The spatial part is expanded in linear combinations of atom-centered one-electron basis functions χ_μ . The favourable mathematical properties of Gaussian Type Orbitals (GTOs) for integral evaluation make them the preferred choice for $\chi_\mu(\mathbf{r})$.

$$\phi_i = \sigma_i \sum_{\mu} C_{\mu i} \chi_{\mu}(\mathbf{r}_i) = \sigma_i \omega \tilde{\phi}_i(\mathbf{r}) \quad (11)$$

$$\chi_{\mu}(\mathbf{r}) = x^{l_{\mu}} y^{m_{\mu}} z^{n_{\mu}} e^{-\alpha_{\mu} r^2} \quad (12)$$

The major challenge is the evaluation of the two-electron integrals over GTOs (Eq. (13)) that finally occur as the basic building blocks of the matrix elements $\langle \Phi_i | \mathcal{H} | \Phi_j \rangle$. ($\dots$) and $[\dots]$ indicate integrals over spatial orbitals and spinorbitals, respectively.

$$(\mu\nu|\kappa\lambda) = \int \chi_{\mu}(\mathbf{r}_1) \chi_{\nu}(\mathbf{r}_1) \frac{1}{|\mathbf{r}_1 - \mathbf{r}_2|} \chi_{\kappa}(\mathbf{r}_2) \chi_{\lambda}(\mathbf{r}_2) d\mathbf{r}_1 d\mathbf{r}_2 \quad (13)$$

$$(ij|kl) = \int \tilde{\phi}_i(\mathbf{r}_1) \tilde{\phi}_j(\mathbf{r}_1) \frac{1}{|\mathbf{r}_1 - \mathbf{r}_2|} \tilde{\phi}_k(\mathbf{r}_2) \tilde{\phi}_l(\mathbf{r}_2) d\mathbf{r}_1 d\mathbf{r}_2 \quad (14)$$

$$[ij|kl] = \int \phi_i(\mathbf{r}_1) \phi_j(\mathbf{r}_1) \frac{1}{|\mathbf{r}_1 - \mathbf{r}_2|} \phi_k(\mathbf{r}_2) \phi_l(\mathbf{r}_2) d\tau_1 d\tau_2 \quad (15)$$

$$[ij|kl] = \langle ik|jl \rangle \quad \langle ij||kl \rangle = \langle ij|kl \rangle - \langle ij|lk \rangle \quad (16)$$

$$(ij|kl) = (ji|kl) = (ij|lk) = (ji|lk) = (kl|ij) = (lk|ij) = (kl|ji) = (lk|ji) \quad (17)$$

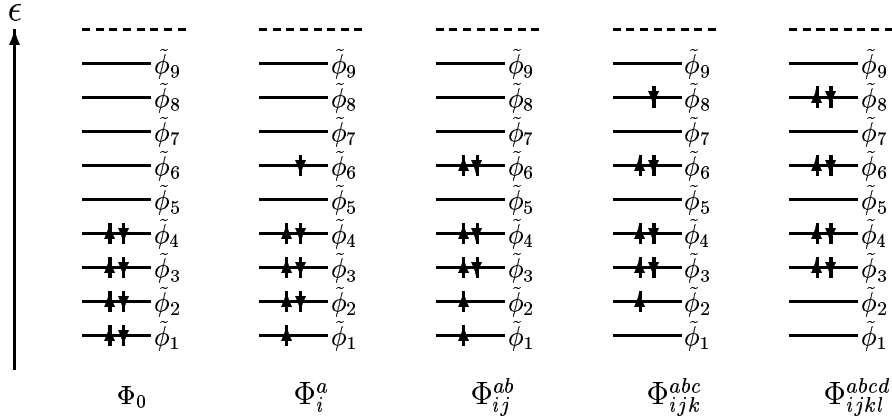


Figure 1. Schematic construction of closed-shell ground-state (Φ_0), singly (Φ_i^a), doubly (Φ_{ij}^{ab}), triply (Φ_{ijk}^{abc}) and quadruply (Φ_{ijkl}^{abcd}) excited determinants.

Figure 1 schematically depicts the construction of the N-particle basis set. The molecular orbital (MO) coefficients $C_{\mu i}$ are usually obtained from Hartree-Fock (HF) or multi-configurational self-consistent field (MCSCF) calculations. Whereas in the HF method $\hat{\Psi}(\alpha)$ is expanded in a single Slater determinant the MCSCF method expands $\hat{\Psi}(\alpha)$ into a small expansion of usually less than 50000 terms (cf. Eq. (6)). In either case *all* parameters

$\{\alpha\}$ and $C_{\mu i}$ are fully optimized. The MOs $\tilde{\phi}_i$ are ordered according to their eigenvalues (orbital energies ϵ_i). The ground (or reference) state is constructed by filling the energetically lowest-lying orbitals with two electrons each (closed shell case). Singly, doubly, triply and quadruply excited determinants are obtained by "exciting" electrons from some occupied to unoccupied orbitals. Continuing up to N-fold excitations generates the full N-particle basis. Slater determinants are in general no eigenfunctions to S^2 . However, spin-adapted configurations (CSFs) may be constructed by taking the appropriate linear combinations of Slater determinants.

The number of CSFs that can be constructed from n molecular orbitals and N electrons subject to a spin-multiplicity of $S = b/2$ is given by⁵:

$$N_{csf} = \frac{b+1}{n+1} \binom{n+1}{a} \binom{n+1}{n-a-b} \quad \text{where } N = 2a + b \quad S = b/2 \quad (18)$$

Table 1 collects some representative numbers for the size of a full N-particle basis for singlet states ($S = 0$). The case $n = N$ corresponds typically to basis sets used in HF calculations whereas $n = 2N$ is about the minimum size for basis sets suitable for electron correlation methods. Evidently, using the linear expansion of the wave function (Eq. (6)) in terms of all possible CSFs in a given basis set $\{\phi\}$ (denoted full configuration interaction (FCI)) is ruled out by the *factorial* growth of the N-particle space, except for the very smallest systems with at most 10 to 20 electrons in small basis sets. FCI calculations can be carried out for expansion lengths beyond 1 billion terms⁶. *The truncation of the one- and N-particle basis are the most important errors in quantum chemical calculations.*

2.3 Non-Dynamical Versus Dynamical Electron Correlation

Truncating the N-electron basis to a single determinant and variational optimization of the MO coefficients defines the HF method. Since the canonical HF orbitals form the basis for most advanced electron correlation methods, the electron correlation energy is defined as the difference between the exact non-relativistic energy at 0 K and its HF counterpart in a *complete* basis $\{\Phi\}$:

$$E_{corr} = E_{exact} - E_{HF,\infty} \quad (19)$$

Table 1. Total size of the N-particle basis for singlet states as function of the number of electrons N and molecular orbitals n : $N_{csf} = \frac{1}{n+1} \binom{n+1}{N/2}^2$.

N	n=N N_{csf}	n=2N N_{csf}
4	20	336
8	1 764	866 320
12	226 512	3 405 278 800
16	34 763 300	16 226 413 117 200
20	5 924 217 936	86 391 974 193 251 584
24	1 081 724 803 600	494 452 245 428 329 102 096

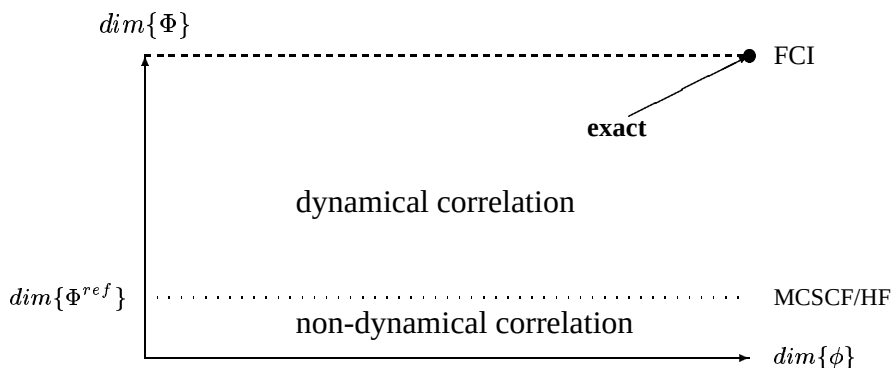


Figure 2. Schematic representation of basis set truncation effects. The dotted line indicates the reference configuration space dimension and the dashed line the dimension of the FCI space. Given a fixed basis set $\{\phi\}$ various methods are points along a vertical line.

E_{exact} is apart from one- and two-electron systems unknown and must be obtained from experimental data corrected for relativistic effects. In some cases Quantum Monte Carlo techniques may provide accurate reference data. Considering electron correlation as the inadequacy of expanding the N -electron wave function in a single Slater determinant (or a single CSF) two different effects can be identified.

The HF wave function by itself may be even qualitatively incorrect due to near-degeneracy effects: low-lying configurations strongly interact with each other resulting in large off-diagonal elements of $\mathbf{H}$ (Eq. (7)), so that the wave function is of *multi-configurational* nature. This effect is termed *non-dynamical* electron correlation. It is frequently of no importance for closed-shell molecules close to their equilibrium geometries but rapidly gains importance for open-shell systems, in excited states, transition metal compounds, for molecular systems at strained structures as well as a consequence of bond disruption or bond breaking. Computing chemical reaction paths accurately requires the balanced (or unbiased) description of all intermediate structures so that the minimum configuration space is the joined set of all near-degenerate configurations along the path. Treating several electronic states on the same footing requires similar measures to ensure unbiased treatment.

Dynamical electron correlation arises because - even though qualitatively correct - the HF mean-field model cannot catch the instantaneous electron-electron interaction correctly. The mathematical structure of the Hamiltonian enforces Kato's cusp condition⁷: there is a singularity in the Hamiltonian close to the coalescence point, where the interelectronic distance of a pair of electrons $r_{ij} \rightarrow 0$ vanishes. To cancel this singularity the wave function must contain linear terms r_{ij} so that there is a cusp at the coalescence point. To describe this cusp in terms of products one-electron functions requires high-angular momentum basis functions (angular correlation). Dynamical correlation effects may be accounted for by a variety of methods. Single-reference methods imply a qualitatively correct HF reference state whereas more general multi-reference methods can cope with any multi-configurational reference wave function.

Electron correlation methods aim at approximating the FCI result, i.e. expanding the wave function in the full N-electron basis at a given one-electron basis set. Different methods can be classified by the way the FCI space is truncated (cf. Figure 2).

2.4 One-Electron Basis Sets

For the ground state of Helium it has been empirically found that the energy contribution of an individual orbital with quantum numbers nml solely depends upon the principal quantum number. This finding suggests the construction of basis set hierarchies such that the next level of accuracy should include all orbitals of the corresponding principal quantum number. Thus, the minimum basis set e.g. for carbon amounts to the $n=1,2$ (2s1p). Incremented by one set of polarization functions $n=3$ (1s1p1d) gives the double zeta (DZ), adding another set $n=4$ (1s1p1d1f) yields the triple zeta (TZ) basis etc. These atomic basis functions are built from contracted GTOs (CGTOs) i.e. fixed linear combinations of primitive GTOs. So far only the number and angular momentum CGTOs is fixed. Contraction length, contraction coefficients and exponents of the individual primitive GTOs are determined by minimizing the averaged energy over one or several atomic states at HF (for minimum basis sets) and correlated level. Further optimization might also involve diatomics especially of the heavier elements in order to improve the applicability to molecular calculations. The exponents of the primitive GTOs are often chosen to (partially) form an even-tempered series, i.e. $\alpha_i = \alpha_{i-1}\beta$.

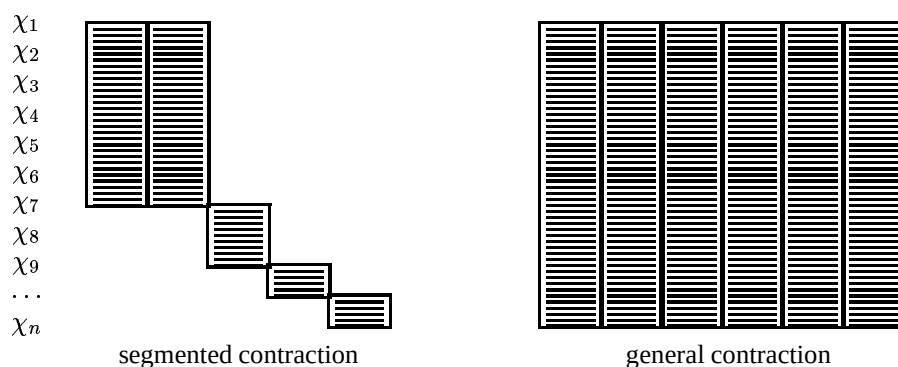


Figure 3. Segmented vs. generally contracted basis sets. All primitive GTOs χ_μ belong to the same angular momentum quantum number $l = m_\mu + l_\mu + n_u$.

There are two types of basis sets denoted generally contracted and segmented. Whereas segmented basis sets aim at minimizing the contraction length and thus the number of primitive integrals to evaluate, the general contractions exploit reusing intermediate quantities for an entire contraction set at the expense of large contraction lengths. Most integral codes favour one or the other basis set type.

Initially, the concept to add all basis functions belonging to a given principle number simultaneously was taken advantage of by the atomic-natural orbital (ANO) basis sets⁸ They form generally contracted basis sets and the contraction coefficients are obtained as the eigenvectors of the one-electron density matrix with the largest eigenvalues (occupation

numbers). Thus given the number of primitives and their exponents the density matrix averaged over several low-lying atomic states based on some electron correlation method in the primitive GTO basis is computed. The NOs with the highest occupation numbers are used to define the contraction coefficients of the basis set. The ANO basis sets are now available for almost the entire periodic table^{10,9,11}.

The segmented *correlation consistent basis sets* (cc-pVXZ, cc-pCVXZ, aug-cc-pVXZ, X=D,T,Q,5,6) by Dunning and co-workers^{12,13}. They have been shown to form a balanced sequence of increasing accuracy and have been often used for extrapolation of properties to the complete basis set (CBS) limit. Finally, there are segmented basis sets introduced by Ahlrichs and co-workers^{14,15}, whose main emphasis was to provide compact basis sets well-suited for correlated molecular calculations. Many basis sets are available from <http://www.emsl.pnl.gov/forms/basisform.html>.

Obviously, the choice of the one-electron basis is crucial for the quality of a calculation. Turning primarily to "simple properties" such as geometries and vibrational frequencies, their calculation at the uncorrelated (HF) level of theory can be carried out reasonably accurate (vibrational frequencies with empirical scaling only) with DZ quality basis sets. Using electron correlation methods (including DFT!) requires usually at least TZ quality basis sets. Turning to properties, which sample the electron density distribution in regions spatially different from those relevant for the energy gives rise to more stringent basis set requirements. Static electric multipole moments, sample $\langle r^n \rangle$ so that not only very diffuse basis functions but also those of *higher angular momentum* are required. In addition the result is sensitive to the level of dynamic electron correlation treatment. To describe Rydberg states also calls for very diffuse basis functions, even to obtain qualitatively correct results. Similarly, there is no point in correlating the non-valence electrons of a molecule while using a standard polarized valence basis set, as they are insufficiently flexible to incorporate core and core-valence correlation at all. However, the core-electrons contribute substantially to the correlation energy as they are spatially close to each other, thereby forcing energetically favourable but (with inadequate basis sets) artificial valence electron density distributions. In summary, correlating core electrons without basis sets augmented by tight basis functions including those of high angular momentum, at best is a waste of computer time and at worst deteriorates the quality of the calculation.

2.5 Electron Correlation Methods: N-Electron Basis Truncation

Electron correlation methods aim at approximating the exact solution within a given one-electron basis set by truncating the full N-electron basis in some systematic way. Basically three different methods with many individual variants are in use: configuration interaction, perturbation theory and coupled-cluster methods.

The multi-reference configuration interaction (MRCI) method expands the wave function linearly in the N-electron basis. Truncation is achieved by defining a reference space and restricting the configuration space to single and double excitations out of each configuration contained in the reference space, which is chosen such as to deal with the non-dynamical electron correlation.

$$\Psi_{MRCI} = \sum_i^{ref} c_i^{ref} \Psi_i^{ref} + \sum_{ia} c_i^a \Psi_i^a + \sum_{ij}^{ab} \Psi_{ij}^{ab} \quad (20)$$

There are several variants of this method. Internally contracted MRCI¹⁶ restricts the reference space to a single multi-configurational wave function thereby making the size of the configuration space independent of the size of the reference configuration space. Other variants treat only a part of the configuration space variationally while the remaining configurations enter through perturbation theory¹⁷.

Perturbation theory splits the Hamilton operator into a zeroth order approximation and a fluctuation potential. Non-degenerate (or single-reference) Møller-Plesset (MP) perturbation theory uses the Fock operator as zeroth order approximation, thus all eigenvalues and eigenvectors are available straight from the initial HF calculation. Working in the canonical HF orbital basis, energies up to third order are available straight away. The notation $\langle 0|\mathcal{O}|n\rangle$ denotes the matrix element between the ground state determinant and any excited determinant (solutions of $\mathcal{F}$, eigenvalues E_i , orbital energies ϵ_i).

$$\mathcal{H} = \mathcal{H}_0 + \mathcal{H}' = \mathcal{F} + \mathcal{H}' \quad (21)$$

$$E = E_0 + \sum_{i=1}^{\infty} \tilde{E}_i \quad (22)$$

$$E_0 = \langle 0|\mathcal{H}_0|0\rangle = \sum_i 2\epsilon_i \quad (23)$$

$$\tilde{E}_1 = \langle i|\mathcal{H}'|i\rangle = E^{HF} - E_0 \quad (24)$$

$$\tilde{E}_2 = \sum_{n \neq 0} \frac{|\langle 0|\mathcal{H}'|n\rangle|^2}{E_0 - E_n} = \sum_{\substack{i < j \\ a < b}} \frac{|\langle \Phi_{ij}^{ab}|\mathcal{H}'|\Phi_0\rangle|^2}{\langle \Phi_{ij}^{ab}|\mathcal{H}_0 - E_0|\Phi_{ij}^{ab}\rangle} = \sum_{\substack{i < j \\ a < b}} \frac{|\langle ij||ab\rangle|^2}{\epsilon_i + \epsilon_j - \epsilon_a - \epsilon_b} \quad (25)$$

$$\tilde{E}_3 = \sum_{nm \neq 0} \frac{\langle 0|\mathcal{H}'|n\rangle \langle n|\mathcal{H}'|m\rangle \langle m|\mathcal{H}'|0\rangle}{(E_0 - E_n)(E_0 - E_m)} - E_{HF} \sum_{n \neq 0} \frac{|\langle 0|\mathcal{H}'|n\rangle|^2}{(E_0 - E_n)^2} \quad (26)$$

Important variants thereof include CASPT2¹⁸ using a single-reference multi-configurational zeroth order wave function, which has been successfully applied to a wide variety of ground and excited states¹⁹.

coupled-cluster (CC) methods invoke a non-linear expansion of the wave function.

$$\Psi_{CC} = e^{\mathcal{T}} \Psi_0 = e^{\mathcal{T}} \Phi_0 \quad (27)$$

$$e^{\mathcal{T}} = 1 + \mathcal{T} + \frac{1}{2}\mathcal{T}^2 + \dots = \sum_{k=0}^{\infty} \frac{1}{k!} \mathcal{T}^k \quad (28)$$

$$\mathcal{T} = \sum_i \mathcal{T}_i \quad (29)$$

$$\mathcal{T}_1 \Phi_0 = \sum_{ia} t_i^a \tau_i^a \Phi_0 = \sum_{ia} t_i^a \Phi_i^a \quad (30)$$

$$\mathcal{T}_2 \Phi_0 = \sum_{ijab} t_{ij}^{ab} \tau_{ij}^{ab} \Phi_0 = \sum_{ijab} t_{ij}^{ab} \Phi_{ij}^{ab} \quad (31)$$

The cluster operators $\mathcal{T}_n$ are linear combinations of excitation operators τ weighted with cluster amplitudes t . Applied to a reference state (here a closed shell Slater determinant)

they generate all n th excited Slater determinants with respect to the reference state Ψ_0 . Eq. (27) expands thus to

$$\Psi_{CC} = (1 + \mathcal{T}_1 + \mathcal{T}_2 + \mathcal{T}_1^2 + \mathcal{T}_1\mathcal{T}_2 + \mathcal{T}_2^2 + \mathcal{T}_3 + \dots)\Psi_0 \quad (32)$$

All these wave function based expansions - in principle - can be improved systematically in order to approach the FCI limit: MRCI may either increase the excitation level or enlarge the reference space. Perturbation theory can be continued up to high order and in the coupled-cluster method the cluster operator (Eq. (29)) maybe truncated at N^{th} excitation level. As the MPn series may diverge, the coupled-cluster method is more expensive but more reliable. A serious drawback of MPn and CC approaches is the limitation to single reference cases.

Density functional theory is somewhat special as electron correlation and exchange are treated through a more or less well-founded correlation-exchange functional of the one-electron density. There is no direct way connecting wave function expansions with DFT. Owing to the limited knowledge about the properties of the *exact* exchange-correlation functional, there can be done little about *systematic* improving DFT towards the FCI limit. In terms of computational efficiency DFT is by far superior to Hartree-Fock and electron correlation methods (maybe with exception of *local* MP2²⁰). Compared to electron correlation methods, DFT results display a considerable spread in accuracy depending upon the specific case as well as the utilized functional whereas they are almost always superior to HF.

3 Errors

3.1 Basis Set Truncation Errors

To this end we can define the apparent error as the difference between the true solution of the electronic Schrödinger equation (corresponding to the experimental value corrected for non-BO effects and possibly relativistic corrections). This can further be subdivided into the basis set error associated with the limited size of the 1-electron particle basis and the n -electron error associated with the incompleteness of the n -electron basis. In the limit of a complete basis set the basis set error vanishes and solely the n -electron error remains denoted the *intrinsic error* of the corresponding model.

One-electron basis set truncation errors may be very large, especially if basis sets entirely inadequate to the given problem or the desired accuracy are chosen. Under the assumption, that non-dynamical electron correlation plays no role, good theoretical estimates

method	pCVDZ [10s2p2d]	pCVTZ	pCVQZ	pCV5Z	pCV6Z	56-expln.
MP2	-382.7	-477.8	-510.7	-523.1	-528.7	-536.4
CCSD	-387.8	-478.2	-507.1	-516.7	-520.6	-525.9
CCSD(T)	-400.4	-498.0	-528.5	-538.7	-542.7	-548.3

Table 2. Recovered electron correlation in mE_h for N_2 at equilibrium distance $R_{N-N} = 109.77 pm$. Estimated basis set limits: -537.2 (MP-R12), -526.2 (CCSD-R12), -548.2 (CCSD(T)-R12). Taken from Ref. 23. All electrons correlated. 56-expln=(216 E(pCV6Z)-125 E(pCV5Z))/91.

method	pVDZ	pVTZ	pVQZ	pV5Z	45 expln.	est. intr. error
MP2	213.36	229.23	235.78	238.36	241.1	+12.7
MP3	191.31	205.76	212.14	214.31	216.6	-11.8
MP4	203.58	221.13	227.98	230.41	232.9	+4.5
CCSD	193.67	207.98	214.17	216.28	218.5	-9.9
CCSD(T)	200.69	216.70	223.17	225.42	227.8	-0.6
IMRCI	201.96	217.90	224.19	226.31	228.5	+0.1
IMRCI+Q	200.74	216.62	223.14	225.37	227.7	-0.7

Table 3. Computed Dissociation energy in kcal/mol for N_2 .
Experiment: 228.4 kcal/mol. 45-extrapolation= $\frac{125D_e(pV5Z)-64D_e(pVQZ)}{61}$. Except extrapolation taken from Ref. 24. Estimated intrinsic error taken as the difference between experimental value and 45-extrapolation.

of the CBS limit can be obtained from the so-called R12-methods, which expand the wave functions explicitly including terms linear in R12. This allows for a more accurate description of the electron-electron cusp (Katos cusp condition). In particular the relatively cheap MP-R12 method^{21,22} has been frequently used for benchmarking the CBS limit. Table 2 demonstrates that only by employing very large basis sets up to including i functions for the nitrogen molecule it is possible to recover more than 98% of the correlation energy. Hence, only extrapolation to the basis set limit (cf. Section 4.1) is a practical way to (almost) quantitatively take care of electron correlation. The differences in the estimated basis set limits indicate that the impact of N-electron basis set truncation is of the order of $10 mE_h$.

Systematic expansion of the N-electron basis towards the FCI limit is in principle possible by the MPn perturbation series or by coupled-cluster methods truncating the cluster operator at higher and higher excitations. From the practical point of view either approach is limited to rather small basis sets and small systems. Fortunately, energy *differences* converge more rapidly. In fact, IMRCI+Q/IMRCI and CCSD(T) are almost equivalent whereas CCSD is clearly insufficient and MPn displays an oscillating behaviour going from low to higher order. Note, that even at the cc-pV5Z level the one-electron basis set error is about 2-3 kcal/mol.

3.2 Basis Set Superposition Error

The basis set superposition error (BSSE) is a spurious contribution to the interaction energy arising from the improved description of each fragment in the total basis as compared to the fragment basis alone. It is ultimately a consequence of the one-electron basis set incompleteness. Evidently, the BSSE error will vanish asymptotically as the complete one-electron basis is approached. BSSE appears whenever the molecular geometry is changed and is particularly important for the computation of interaction energies. Without correction an artificial increased binding energy is obtained. In particular when using modest basis sets, BSSE can result in substantial distortions of the PES.

A conceptually simple method to account for BSSE is the counterpoise correction method²⁵ in which the energies of the fragments are computed in the full basis of the entire complex and subtracted from the energy of the entire system. Although there is some debate about its accuracy it is the procedure of choice to account for BSSE (cf. a

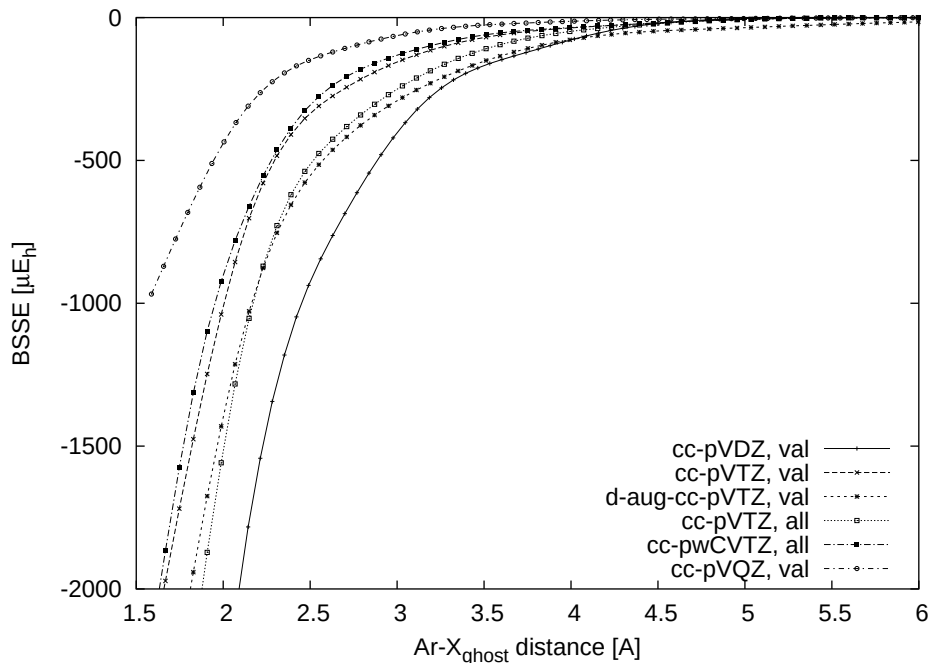


Figure 4. Basis set superposition error for Ar augmented by an Ar ghost basis centered at distance d computed at the CISD level of theory. Van der Waals radius of Ar is 1.88 Å, empirical well depth for Ar_2 is estimated to 99.7 cm^{-1} ($453.6\mu E_h$) at $d=3.757\text{\AA}$.

review on counterpoise theory²⁶).

Investigations on N_2 indicate a BSSE error at experimental equilibrium distance of about 0.8 kcal/mol for TZ and about 0.4 kcal/mol for QZ basis sets. Accurate extrapolation to the complete basis set limit therefore must correct for BSSE prior to extrapolation. BSSE, however, also can be viewed as a basis set incompleteness indicator and thus used to correct the binding energy by adding an *empirical* fraction of the BSSE to correct for basis set incompleteness.

Figure 4 illustrates the BSSE error for Ar in presence of an Ar ghost basis using a variety of Dunning’s correlation consistent basis sets at the CISD level of theory. BSSE is moderately method dependent but the trends discussed here hold in general. Increasing the cardinal number of a given basis reduces the BSSE error substantially. Valence electron correlation basis sets and core valence basis sets show a similar BSSE error. Adding diffuse functions strongly increases the BSSE and it vanishes much more slowly with increasing distance to the ghost basis. As a note of caution, correlating the core electrons in a valence basis almost doubles the BSSE error (without improving the results). Although the BSSE error is somewhat method-dependent the conclusions are generally applicable.

The argon dimer forms a van-der-Waals complex kept stable solely by dispersion, which is solely an dynamic electron correlation effect. To properly describe such kind of complexes several sets of diffuse functions are required. At the experimental equilibrium distance of Ar_2 the BSSE correction (twice the Ar-X interaction energy) amounts to

about $250\mu E_h$, i.e. about 50% overbinding for Ar_2 due to BSSE. As increasing the level of electron correlation treatment increases the well depth while increasing the basis set size on the one hand side improves the description of electron correlation effects and on the other hand side decreases the well depth due to reduced BSSE error, one can easily be trapped by error compensation effects. An MP2 calculation with the aug-cc-pVTZ basis accidentally agrees quite well with the experimental well depth.

3.3 Core-Core and Core-Valence Electron Correlation

Core-core (CC) correlation are denoted electron correlation effects solely restricted to the core electrons while core-valence (CV) correlation describes the relaxation of the core electrons upon valence electron correlation and vice versa. Although core and core-valence electron correlation in absolute terms contributes largely to the total correlation energy (N_2 : $E_{corr}^{val} = 408mE_h$, $E_{core}^{all} = 548mE_h$ ²³), its differential effects on the electron distribution of the valence shells are modest. For compounds of first and second row elements the effects are fairly small. The dissociation energy of N_2 e.g. changes by about 1.5 kcal/mol²⁷, and the depth of the van-der-Waals minimum for Ar_2 is affected by about $4\mu E_h$ ²⁸ corresponding to about 1% of the valence electron correlation contribution. Hence, for the light atoms CC and CV correlation effects need to be considered for high-accuracy calculations, only.

Note, that accounting for CC and CV correlation requires large basis sets and the treatment of many electrons rendering calculations demanding. However, even though solely correlating the valence electrons, the basis set must contain all AOs representing the inner-shell electrons. This sparked the idea of representing the inner-shell electrons by a suitably parametrized potential as to reduce the basis set size denoted effective core potential (ECPs) in contrast to all-electron methods. ECPs are applicable if the relaxation of the core electrons upon chemical binding or valence electron correlation can be neglected. Of course ECPs of neighbouring atoms must not overlap.

There are two different classes of ECPs: the pseudo orbital type^{29,30}, which replaces the valence orbitals by nodeless pseudo-potentials resembling the HF valence orbitals as much as possible (in the sense of shape and orbital energies) and the core electrons by a potential (containing angular momentum projection operators $\mathcal{P}_l = |l\rangle\langle l|$) of the form

$$U(r) = U_L(r) + \sum_{l=0}^{L-1} U_l(r)\mathcal{P}_l \quad (33)$$

where $L - 1$ is the maximum angular momentum of the core electrons. The potentials $U_l(r)$ and pseudo orbitals are fitted to an analytic form involving Gaussians. The ab initio

model potentials (AIMP) method^{31,32} uses valence orbitals of the correct nodal structure.

$$\mathcal{H} = \sum_i -\frac{1}{2}\nabla_i^2 + \sum_A -\frac{Z_{eff}^A}{r_{Ai}} + V_{coul}^A(i) + V_{exch}^A(i) + P^A + \sum_{i>j} r_{ij}^{-1} + V_{nn} \quad (34)$$

$$V_{coul}^A(i) = -\frac{Z_{core}^A}{r_{Ai}} + 2 \sum_c^{core(A)} J_c(i) \approx \sum_k A_k r_i^{n_k} \exp(-\alpha_k r_i^2)/r_i \quad (35)$$

$$V_{exch}^A(i) = - \sum_c^{core(A)} K_c(i) = \text{non-local representation} \quad (36)$$

$$P^A = \sum_c^{core(A)} 2\epsilon_c |\Phi_c\rangle \langle \Phi_c| \quad (37)$$

Partitioning the orbital space into core and valence orbitals the effective core potential is derived by a local representation of the Coulomb potential and a spectral representation of the exchange potential, i.e. the core electrons are replaced by the HF potential. In order to ensure the orthogonality of core and valence orbitals and to avoid collapsing the valence orbitals into the core space, a level shifter is added shifting the core orbitals to positive energies, so that they give a positive contribution to the valence orbital energy unless the valence orbitals are orthogonal to the core orbitals. The specification of AIMP involves apart from the local and non-local potential expansions a basis set expansion of the core orbitals, thereby allowing to combine AIMP with arbitrary basis sets. The quality of AIMP depends upon the quality of the reference calculation. However, to achieve the correct nodal structure of the valence orbitals the valence basis sets are larger than those required for the pseudo-orbital method and include tight GTOs.

3.4 Size-Extensivity Errors

Although the general approaches towards the inclusion of dynamical electron correlation discussed so far yield identical results going to infinite order and infinite excitation, respectively, truncated methods display shortcomings.

An important issue is the correct scaling of the correlation energy with the number of particles. The correlation energy of a system consisting of n identical non-interacting molecules should scale linearly with number of independent subsystems n . CC and MPn in fact give the correct answer. This is not generally true for truncated CI methods. A related property is size-consistency, i.e. the energy of two noninteracting systems must be the sum of the individual components.

$$E_{AB}(R_{AB} \rightarrow \infty) = E_A + E_B \quad (38)$$

A method may be size-extensive but not size-consistent. Hence, it is desirable to apply a size-extensive electron correlation method to a size-consistent reference wave function. Since RHF is not size-consistent, correlation treatments based upon RHF reference wave function may not necessarily be size-consistent at the correlated level either.

The lack of size-extensivity of truncated CI in fact renders it useless for systems containing many electrons, correct relative energies require an unbiased treatment of electron correlation throughout electronic states and molecular configurations. Owing to the lack of

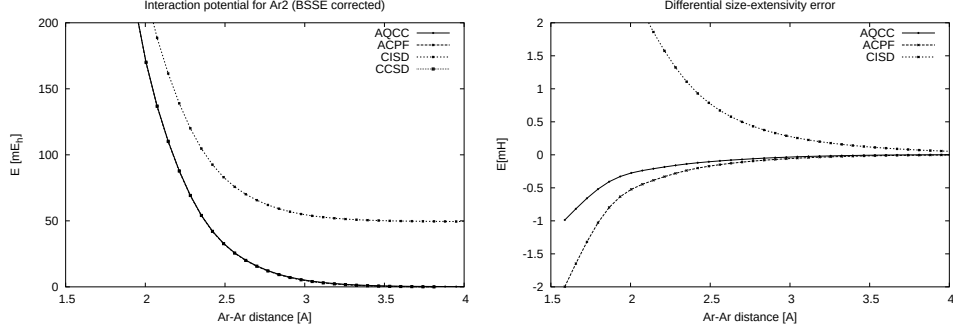


Figure 5. Ar-Ar interaction potential corrected for BSSE computed at CCSD, CISD, ACPF and AQCC level of theory using the cc-CVTZ basis set, all electrons correlated. As the basis lacks diffuse functions, the interaction potential is completely repulsive and corresponds to the closed-shell single reference case at all distances. The differential size-extensivity error defined as the difference to the CCSD energy shifted such that $\Delta E(r \rightarrow \infty) = 0$. Van der Waals radius of Ar is 1.88 Å, empirical well depth for Ar_2 is estimated to 99.7 cm^{-1} ($453.6 \mu E_h$) at $d=3.757 \text{ Å}$.

generally applicable methods for multi-configurational reference wave functions, approximately size-extensive MRCI methods have been developed. The widely used a posteriori Davidson correction scales the computed correlation energy with the weight of the reference configuration in the final wave function

$$\Delta E_{Davidson} = (E - E_0)(1 - c_0^2) \quad (39)$$

where E_0 denotes the energy of the reference wave function and c_0^2 is the weight of the relaxed reference wave function in the final wave function. Alternatively, the size-extensivity of the MRCI method may be approximately restored by optimizing the renormalized functional, where Ψ^{ex} is the component of the wave function not included in the reference space (intermediate normalization).

$$E_{corr} = \frac{\langle \Psi | \mathcal{H} - E_0 | \Psi \rangle}{1 + g \langle \Psi^{ex} | \Psi^{ex} \rangle} \quad (40)$$

The scaling factor g depends upon the number of correlated electrons n_e and defines various approximations such as MRAQCC³³ ($g = 1 - \frac{(n_e-3)(n_e-2)}{n_e(n_e-1)}$) and MRACPF³⁴ ($g = \frac{2}{n_e}$). $g = 1$ corresponds to MRCI.

Whereas a MRCI treatment is variational and does not diverge even though the reference space might be inadequate, this does not apply to the size-extensivity corrected variants. They depend upon the proper choice of the reference space and similar to MPn intruder states and near-degeneracies with configurations not included in the reference space produces artifacts. Hence, for scanning large portions of the potential energy surface MRCI though suffering from the lack of size-extensivity is more robust than MRAQCC or MRACPF.

Whereas CCSD is size-consistent, i.e. the interaction potential goes for large distances to zero, this does neither apply to CISD nor to its size-extensivity corrected counterparts (Figure 5, left) although it may be small in some cases. However, using the supermolecule

approach amounts to shifting the zero point of the potential to the Ar-Ar potential at large distances. Since the Ar_2 case is an ideal single reference case the degree of size-extensivity can be approximately judged by the difference between the CCSD and the AQCC/ACPF/CISD curves (Figure 5, right).

3.5 Orbital Resolution

The separation of MO coefficient optimization (HF or MCSCF) from the dynamical electron correlation treatment introduces a dependence upon the MO basis. The problem of orbital resolution is connected to the partitioning of the MO space. Whereas the HF energy is invariant to unitary transformation among the occupied or unoccupied orbitals this does not apply to electron correlation methods: Møller-Plesset perturbation theory requires the canonical HF orbitals, CASSCF calculations are invariant only with respect to rotations among frozen, inactive, active and virtual orbitals etc. Whereas some of the orbital dependencies (such as for MP2) are rather related to the implementation, others are inherent due to the choice of the method-dependent configuration space selection. Whenever the MO basis is partially invariant to orbital rotations in the MO optimization step they need to be "resolved" by solving an eigenvalue problem of some operator with non-degenerate eigenvalues within the invariant subspace such as a Fock matrix or the one-electron density matrix. However, accidental degeneracies may occur and if the degenerate orbitals in the subsequent electron correlation treatment are split into different orbital groups spurious results (such as spikes on a PES) may occur. The remedy is to use a different orbital resolution operator.

Regarding the computation of excitation energies by choosing the MO basis from a state-specific orbital optimization introduces a bias in favour of this state in the subsequent dynamical electron correlation treatment. This can be observed as a systematic error in the excitation energies. For this reason commonly state-averaged orbitals are employed, that is the energy functional that is minimized during orbital optimization is a weighted average over all states of interest.

Since the ground and excited state PES are quite different at different levels of theory, at certain molecular configurations we are close to avoided crossings at one level of theory while far away from it at a different level. If this occurs for the orbital optimization step in contrast to the inclusion of dynamical electron correlation, the MO basis may be even qualitatively completely wrong thereby reducing the accuracy of the subsequent electron correlation treatment drastically. To overcome this bias by inclusion of a sufficiently large configuration space requires substantial effort if at all successful.

3.6 Relativistic Effects

Relativistic effects can be defined as anything arising from the finite speed of light ($c \approx 137\text{au}$) as compared to $c = \infty$ for the non-relativistic case. The most fundamental starting point for molecular many electron relativistic treatments is the Dirac-Breit Hamiltonian $\mathcal{H}^{DB}$ with possibly additional quantum electrodynamical corrections.

$$\mathcal{H}^{DB} = \mathcal{H}^{DC} + \mathcal{H}^{Breit} \quad (41)$$

$$= \sum_i h_i^{Dirac} + \sum_{i>j} 1/r_{ij} + \mathcal{H}^{Gaunt} + \mathcal{H}^{retard} \quad (42)$$

$\mathcal{H}^{DB}$ can be viewed as the Dirac-Coulomb Hamiltonian $\mathcal{H}^{DC}$ plus perturbative treatment of two-electron relativistic corrections (Gaunt term and retarded Hamiltonian or derived corrections from QED). $\mathcal{H}^{DC}$ treats one-electron relativistic effects exactly in terms of a linear combination of N one-electron Dirac Hamiltonians h^{Dirac} while retaining the non-relativistic electron-electron repulsion term. The relativistic four-component single-particle wave functions can be decomposed into two two-components termed "large" and "small". These terms are related by the kinetic energy balance condition. In contrast to the non-relativistic case the Dirac-Breit Hamiltonian cannot be rigorously derived and is not fully Lorentz invariant so that purists may consider a relativist theory of many-electron systems not yet available.

The spectrum of the Dirac-Breit Hamiltonian is qualitatively characterized by having the bound states bracketed by continuum states at $-2c^2 \geq E > -\infty$ and $\infty > E \geq 0$. The presence of a negative continuum leads to the variational collapse (Brown-Ravenhall disease). Either by imposing appropriate boundary conditions or by modifying the Hamiltonian such as to project out the undesired negative continuum prevents the variational collapse (no-pair Hamiltonian³⁵). Once the appropriate relativistic Hamiltonian is chosen in principle the standard non-relativistic methods can be applied analogously except for working with a four-component wave function. Expanding the large and small components in atom-centered basis sets requires very large basis sets as to fulfill the kinetic energy balance condition. This puts serious constraints upon the range of applicability so that the fully relativistic approach is often simplified by quasirelativistic approximations or the use of pseudopotentials especially since we have to consider electron correlation and relativistic effects simultaneously.

In the non-relativistic limit the small component vanishes and only the large component is retained. Separating large and small components by unitary transformations of the relativistic Hamiltonian and truncating the resulting Hamiltonian at order c^{-2} yields the Pauli and Breit-Pauli two-component form starting from $\mathcal{H}^{DC}$ and $\mathcal{H}^{DB}$, respectively. Eliminating magnetic interactions (spin-orbit coupling) only the mass-velocity correction and the Darwin term are retained.

$$\mathcal{H}_{massvel} = -\frac{1}{8c^2} \sum_i p_i^4 \quad (43)$$

$$\mathcal{H}_{Darwin} = \frac{\pi}{2c^2} \sum_{iA} \delta \mathbf{r}_{iA} \quad (44)$$

$$(45)$$

Choosing instead the four-component no-pair Hamiltonian³⁵ as starting point and applying the Douglas-Kroll transformation³⁶ finally yields the spin-free relativistic one-component Douglas-Kroll Hamiltonian. The latter is variationally stable and can be employed in all-electron treatments. All-electron one-component treatments of systems containing many heavy atoms is computationally demanding due to the large size of relativistic basis sets although much cheaper than their fully relativistic counterparts. Relativistic ECPs whether of pseudopotential or AIMP type are usually derived from relativistic Cowan-Griffin³⁷ or Douglas-Kroll all-electron calculations. The parametrization of RECPs may also accommodate error compensation such as deficiencies in the reference data and inappropriate basis sets. Due to this empirical element they should be used only in combination with the valence basis set used for parametrization. AIMP on the other hand may be com-

bined with any basis set sufficiently flexible to represent the nodal structure of the valence orbitals.

The main consequences of relativistic effects are (i) radial contraction of all s and $p_{1/2}$ atomic orbitals as well as the inner-shell $p_{3/2}$ AOs (ii) the spin-orbit splitting for all AOs with non-vanishing orbital angular momentum and (iii) the radial expansion and energetic destabilization of most d and all f type AOs³⁸. Relativistic effects influence the chemistry of 5d block elements, lanthanoids and in particular the actinoids owing to relativistic changes to the valence shell structure. This makes them potentially offering a very rich not yet explored chemistry. Particular large effects are found for gold and gold chemistry (aurophilic attraction)³⁹. Similar applies to strong closed-shell interactions, which are primarily a consequence of relativistic effects⁴⁰.

4 Extrapolation Schemes

4.1 One-Electron Basis Set Extrapolation

Of particular importance is the possibility to extrapolate to the complete basis set limit. This amounts to extrapolating horizontally to the right in Figure 2. To allow for extrapolation schemes it is mandatory to have a hierarchical sequence of basis sets that converge uniformly to the CBS limit. The cc-pVXZ, cc-pCVXZ basis sets form such a sequence. Since the HF energy and the correlation energy have different convergence characteristics separate extrapolation of each term is important.

For the correlation energy there is some theoretical motivation on the functional form establishing a relationship between maximum angular momentum of the basis set and the correlation energy ΔE_{corr} of He at the CI level and for many-electron atoms at the MP2 level. The asymptotic behaviour of the contribution by all orbitals of angular momentum l by a basis saturated individually for each l is given by

$$\Delta E_{corr,l} = a(l + \frac{1}{2})^{-4} \quad (46)$$

Given a (saturated) basis set with maximum quantum number l , the correlation energy is of the form

$$\Delta E_{corr} = \Delta E_{corr,\infty} + b(L + 1)^{-3} \quad (47)$$

Identifying $L + 1$ with the cardinal number X ($D=2, T=3, \dots$) of the cc basis sets and using a two-point extrapolation with two successive basis sets we obtain

$$\Delta E_{corr,\infty} = \frac{\Delta E_{corr,X+1}(X + 1)^3 - \Delta E_{corr,X}X^3}{(X + 1)^3 - X^3} \quad (48)$$

It has been empirically found, that the most reliable results are obtained by two-point fits using the two largest practical basis sets²³. Comparison with correlated calculations using R12 methods indicate the correct asymptotic basis set limit for the correlation energy⁴². The CBS limit of the correlation energy should be independent of the chosen pair of basis sets. In practice these authors have found that extrapolations involving the $X=D$ basis sets are inferior. As the number of basis functions N in the basis sets grow as X^3 , the correlation energy scales as $O(1/N)$ and the computational effort for integral evaluation

scales as $O(N^4)$ so that the relation between correlation energy and CPU time consumption is $\Delta E_{corr} \approx ct_{cpu}^{-4}$.

The extrapolation of the HF energy to the CBS limit has no theoretical motivation. Most popular are the exponential and power forms, e.g.

$$E_{HF} = E_{HF,\infty} + Be^{-\alpha X} \quad (49)$$

$$E_{HF} = E_{HF,\infty} + BX^{-\alpha} \quad (50)$$

$$(51)$$

Apart from the general finding that the HF energy converges much more rapidly than the correlation energy with cardinal number X , conclusions are ambiguous about the extrapolation error⁴³.

Application of this extrapolation formula to reaction enthalpies, weak interactions⁴⁴, spectroscopic constants and dipole moments⁴⁵ have been encouraging and a statistical analysis of the extrapolated data to experimental as well as reference data based on the MP-R12^{21,22} and the CC-R12 method⁴¹ indicate that *chemical accuracy*, i.e. $\approx 4 \text{ kJ/mol}$ can be achieved already at the 34-extrapolation level. 23-extrapolations, though already largely improving the quality of the results compared to the raw data, are far from chemical accuracy. Similar applies to the raw data itself even for large basis sets. The exponential extrapolation advocated by Dunning and coworkers tends to overestimate the rate of convergence.

$$\Delta E_X^{corr} = E_\infty^{corr} + ae^{bX} \quad (52)$$

It is somewhat problematic that a good estimate of the CBS limit requires large basis sets thereby excluding application to large molecules. Truhlar et. al. extrapolated HF and correlation energies by the form⁴⁶

$$E_{HF} = E_{HF,\infty} + A_{HF}X^{-\alpha} \quad (53)$$

$$E_{corr} = E_{corr,\infty} + A_{corr}X^{-\beta} \quad (54)$$

$$(55)$$

The constants α and β are considered universal and have been obtained by fitting the $X=D,T$ energies to the CBS limits estimated by larger efforts. This method allows to extrapolate to the CBS limit using relatively small basis sets⁴⁷. However, since double zeta type basis sets are known to be insufficient for electron correlation, the parametrization must take care of it, which adds an element of statistical uncertainty as error compensation is implicitly exploited, which may or may not work on systems not included in the test suite used for parametrization.

In order to derive a basis set convergence for molecular properties, which may be largely different from the electronic energy, one starts from an expression involving the energy derivative of the Hamiltonian

$$\mathcal{H}(\lambda) = \mathcal{H}(\lambda = 0) + \lambda \mathcal{V} \quad (56)$$

so that a given property is associated with some perturbation of the Hamiltonian. Upon expanding the correlation energy as

$$E_{\lambda,X}^{corr} = \sum_{n=0}^{\infty} \sum_{k=0}^{\infty} \frac{C_n^{(k)} \lambda^k}{k!} X^{-n} \quad (57)$$

the coefficients $C_n^{(k)}$ govern the convergence of the k th order property. It is a priori not possible to predict the relative importance of the terms scaling as X^{-n} , so that asymptotic scaling may not be achieved with basis sets of cardinal numbers, for which it is possible to perform calculations and each property requires careful investigations.

A note of caution must be supplied as well: the extrapolations solely models the errors associated with the expansion of the electron-electron cusp. Other errors such as basis set superposition error, size-extensivity error or the use of an ill-chosen basis (e.g. lack of diffuse basis functions for multipole moments) must be corrected for separately. Secondly, it is necessary to point out that these extrapolation schemes have been used mostly in conjunction with single-reference systems at the MP or CC level of theory. Since the above scheme considers dynamic electron correlation only, it is tempting to extend it to multi-reference cases by defining the electron correlation energy with respect to energy of the reference wave function. However, the rather foggy distinction between dynamic and non-dynamic electron correlation and thereby its dependence upon reference configuration space definition, suggests a cautious attitude.

4.2 N-Electron Basis Set Extrapolation

Retaining a given one-electron basis somewhat similar extrapolations to the FCI limit, i.e. in vertical direction in Figure 2 would be desirable. Similar to the hierarchical one-electron basis sets, a systematic, uniform improvement of the N-electron basis is mandatory. Møller-Plesset perturbation series can be ruled out due to the tendency to diverge at high order without having an a priori error indicator at hand. Truncated approximately size-extensive MRCI cannot be easily extended in practical terms. Thus, coupled-cluster methods seem to provide the only possible route to this task. Truncating the cluster operator $\mathcal{T}$ heavily such as e.g. in CCD, CC2 may yield extrapolations largely benefitting from error compensation (such as e.g. MP2). CCSD(T) and CCSDT, however, already come quite close to the FCI limit and continuing to higher order cluster expansions may allow accurate extrapolations. Due to the introduction of new techniques CC codes capable of going to high order truncations of the cluster operator may become available in near future.

4.3 Gaussian-n (Gn) Models

In actual problems, however, the ultimate aim is to extrapolate to the FCI result at CBS limit (i.e. the exact result) including corrections for zero-point vibration and relativistic effects as to model the experimental situation, i.e. extrapolation to the upper right corner of Figure 2. Although effects due to one- and N-electron basis set truncation are in general non-additive, extrapolation models can be devised, which assume the additivity of certain contributions. As the model is devised such as to reproduce the data (energies, ionization potentials, electron affinities, proton affinities) of a test suite of molecules, by virtue of the choice of basis sets and level of correlation treatment, error compensation is implicitly made use of. The G3 model⁵⁰ uses a HF/6-31G(d) optimized equilibrium structure and scaled harmonic frequencies (scale factor 0.8929) to compute the zero-point energy. The geometry is refined at the all-electron MP2/6-31G(d) level. A series of single-point calculations is carried out at this molecular geometry to estimate the correlation energy starting with an MP4/6-31G(d) calculation as reference point. Corrections are computed for diffuse

functions, higher polarization functions, correlation contributions beyond MP4 through quadratic CI (a CC variant) and corrections for basis set effects and non-additivity of diffuse vs. polarization function extensions computed at all-electron MP2 level. Spin-orbit corrections for atomic species taken from experiment or accurate theoretical calculations are added along with the zero point energy. Finally, so-called higher level corrections - a parametrized expression depending upon the number of valence electron pairs optimized to reproduce the test suite results - are added. Average deviations for energies, ionization potentials, electron affinities, proton affinities are between 4 and 8 kJ/mol. The Gn models ($n=1,2,3$)⁴⁸⁻⁵⁰ are available for first and second row elements, only. Moreover, due to assuming additivity for various contributions and parametrized higher level correction this model is likely to benefit largely from error compensation, so that it would be difficult to extend it to the entire periodic table. There are also variants G2(MP2)⁵¹ and G3(MP2)⁵², which estimate basis set effects at MP2 level. The associated computational effort goes as $G3(MP2) < G2(MP2) < G3 < G2$ ⁵². Note in particular, that the Gn methods are tied to certain basis sets and targets, for which these basis sets are inadequate will invariably lead to unreliable results.

5 Calibration and Reference Data

A remarkable side effect of extrapolation schemes parametrized to fit experimental data (thereby relying on error compensation) is the availability of extensive test suites of molecules, for which reliable experimental data or accurate data from other sources exist. This includes e.g. the G(n) test sets⁴⁹. There are many such reference data concerning a variety of properties frequently used for parametrization of density functionals. Although the agreement with the reference data may be stunning regarding the comparatively small effort, they constitute unbounded extrapolation for systems not included in the test suite. In addition error estimates are unavailable as the implicit error compensation may add up effects of opposite sign, which in turn depend upon the system in question. Large discrepancies between such extrapolations and the experimental data may indicate experimental errors as well as a failure of the extrapolation procedure as well.

More reliable calibration procedures avoid tying themselves to experimental data but rather invoke a sequence of calculations of increasing accuracy in terms of one- and N-electron basis sets, chosen to allow separating the various errors and judging to what extent additivity is a safe assumption.

A considerable problem associated with quantum chemical calculation comprises the decision whether a given method is applicable to a given problem. In many cases there is no simple a priori yes/no answer to the problem but instead a series of test calculations is required. Solely comparing the computed results with some experimental or otherwise derived reference data is frequently of little help, since we cannot distinguish between error compensation and correct results. In other words even despite of good agreement with reference data, we cannot connect it to the parameters determining the outcome of the quantum chemical calculations. In addition, we face the problem of having either no reference data at hand or the derivation of the reference data themselves relied on theoretical models that may not be appropriate. Thus multiple sources of error generally hamper any reasonable interpretation of the comparison.

Therefore, a generally profitable approach is to perform *series* of quantum chemical

calculations, which are set up such as to deduce the effect of a single source of error as accurately as possible. Since it is not save to assume straightforwardly the additivity of different error sources calibration or benchmarking certain sources of error against a judicial chosen reference is of great value for the understanding on the properties of quantum chemical methods. In particular systematic investigations on the impact of certain parameters, provide a considerable insight into the methods and are to some extent transferable.

Of crucial importance is to disentangle the one-electron basis set truncation error from the intrinsic (i.e. systematic) error of a given method. Since it is usually not possible to repeat extensive calibration calculations for large molecules, it is sensible to define a set of representative reference molecules. Setting up a series of calculations we aim at elucidating the intrinsic (or systematic) error of quantum chemical methods with respect to the desired property. Provided, the reference set was chosen appropriately, the results will be mostly transferable to the larger systems.

Of particular value are benchmarks probing the full N-electron basis set limit within a given (incomplete) one-electron basis. Here FCI calculations are to be mentioned in the first place - as they are exact within the given one-electron basis set, which should be of double zeta if possible triple zeta quality for meaningful comparisons. Second to it are coupled-cluster techniques including quadruple or higher excitations in particular the R12 variants, which converge rapidly to the one-electron basis set limit. At least for true single-reference cases they practically reproduce the FCI data. For the general case and small molecules size-extensivity corrected MR-SDCI such as MR-AQCC and MR-ACPF may provide reference data. Often it is necessary to extrapolate to the CBS limit in the course of the analysis.

5.1 Error Indicators

Hence, scrutinizing solely the final results of a series of calculations will indicate the non-applicability of a given method by divergence or convergence to different method-dependent limits. To understand the failure of a method requires to verify whether the basic assumptions of the method hold for the specific case.

Provided the MOs are obtained from a qualitatively correct SCF calculation it is still the question for diagnostics on the applicability of the method. In case of CI one can simply examine the weight of the reference wave function: if configurations with large weights appear from outside the reference space, the reference space must be appropriately adjusted. For Coupled-Cluster methods commonly the $\mathcal{T}_1$ diagnostic is used, which amounts to the scaled norm of the single excitation cluster amplitudes. The reasoning for the choice of this indicator is that the single excitations reflect orbital bias best. CCSD calculations with $\mathcal{T}_1 > 0.02$ call for caution; inclusion of higher excitations will increase the robustness of the CC treatment. Møller-Plesset perturbation theory is less effective. Looking at the norm of correction vector and energy of MP2 solely may erroneously indicate applicability. Including MP3 and monitoring the change in correlation energy increment and norm of the correction vector is more sensitive and more representative of the convergence of the MPn series. The approximately size-extensively corrected MRCI functionals such as MRACPF and MRAQCC are also fairly sensitive to intruder states, which in serious cases may force them to diverge (this depends upon the wave function optimization scheme).

Finally, Figure 6 displays the characteristic behaviour of various standard techniques on the dissociation of the N_2 molecule. Note, that methods relying on the closed shell

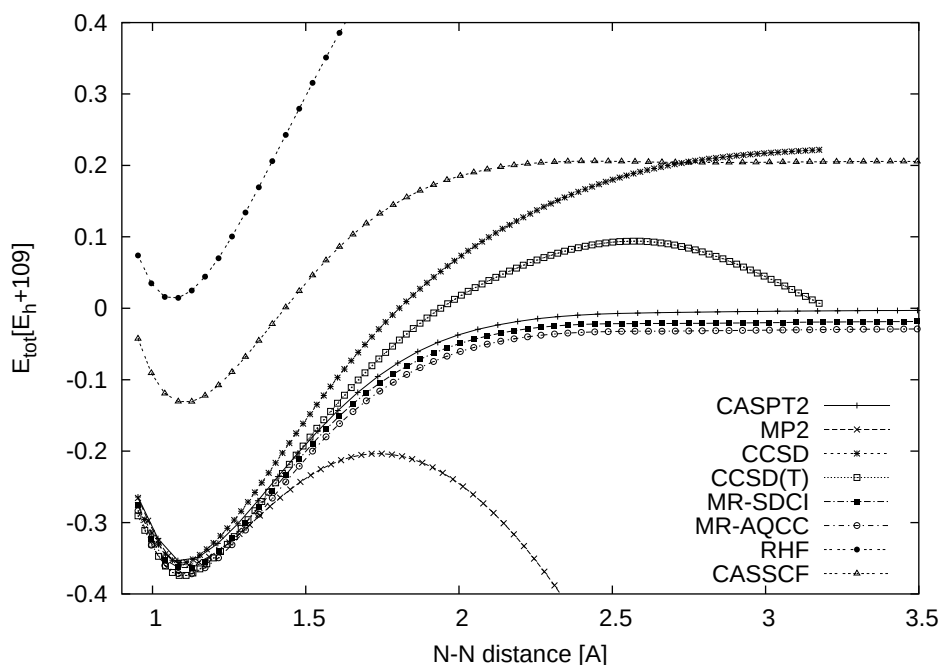


Figure 6. Total energies computed at various levels of theory: RHF, CASSCF (valence CAS, 176 CSFs), MP2, CCSD, CCSD(T), CASPT2 (valence CAS), MR-CISD (valence CAS reference space), MR-AQCC (valence CAS reference space); cc-pVTZ basis set, 1s electrons kept frozen; experimental dissociation energy $0.364 E_h$ and bond length $1.074 \text{ \AA}$.

HF reference (MP2, CCSD, CCSD(T)) fail badly as the RHF reference performs poorly. Methods relying on an appropriate zeroth order MCSCF reference wave functions (MR-CISD, MRAQCC, CASPT2) describe the dissociation qualitatively correct.

6 Summary

This brief survey has described the most relevant ingredients and practical error sources in quantum chemical (cluster) calculations using standard quantum chemical methods such as HF, MCSCF, perturbation theory, coupled-cluster methods and configuration interaction. It has been pointed out that the most important error sources are, apart from using a single-reference method in a multi-reference case, the choice of one- and N-electron basis set. The concept of non-dynamical versus dynamical electron correlation has been explained. The molecular orbital basis may give rise to a bias in favour of one state or be even qualitatively completely inadequate for a given purpose. Improving the MO basis is much more important than high-level electron correlation treatment. Orbital rotations mixing MOs that are treated on a different footing in the dynamical correlation treatment lead to artifacts. Relativistic effects are important for the chemistry of the heavy elements. However, even for the second and third-row elements relativistic effects have to be considered for highly accurate calculations. Highly accurate data can be obtained only from extrapolations to

the CBS limit, high-level electron correlation treatment and elimination of size-extensivity and BSSE errors.

Cheap extrapolation models aiming at the applicability to a wide range of systems almost invariably contain an element of statistics, which ties the model explicitly and/or implicitly to a set of reference data thereby benefitting from error compensation.

A critical attitude towards the quality of a calculation is very beneficial. Decomposition of the total error into its various sources and calibration of some suitably chosen reference systems should always be taken into account whenever there is some fundamental doubt about the results given the required level of accuracy.

Acknowledgments

Some of data have been computed with the Quantum Chemistry Packages COLUMBUS⁵³ and MOLCAS⁵⁴.

References

1. M. Born, R. Oppenheimer, *Ann. Phys.* **84**, 457 (1927).
2. M. Born, K. Huang, *The Dynamical Theory of Crystal Lattices*, Oxford University Press (1954).
3. C.J. Ballhausen, A.E. Jansen, *Ann. Rev. Phys. Chem.* **23**, 15 (1972).
4. D.R. Yarkony in *Conical Intersections*, eds. W. Domcke, D.R. Yarkony, H. Köppel, World Scientific (2004).
5. I. Shavitt, *Int. J. Quantum Chem.* **S11**, 131 (1977).
6. J. Olsen, P. Jørgensen, J. Simon, *Chem. Phys. Lett.* **169**, 463 (1990).
7. T. Kato, *Pure Appl. Math.* **10**, 151 (1957).
8. J. Almlöf, P.R. Taylor, *J. Chem. Phys.* **86** 4070 (1987).
9. P.O. Widmark, P.-Å. Malmqvist, B.O. Roos, *Theoret. Chim. Acta* **77**, 291 (1990), P.O. Widmark, B.J. Persson, B.O. Roos, *Theoret. Chim. Acta* **79**, 419 (1991), R. Pou-Ameriga, M. Merchán, P.O. Widmark, B.O. Roos, *Theoret. Chim. Acta* **92**, 149 (1995).
10. K. Pierloot, B. Dumez, P.O. Widmark, B.O. Roos, *Theoret. Chim. Acta* **90**, 87 (1991).
11. B.O. Roos, R. Lindh, P.A. Malmqvist, V. Veryazov, P.O. Widmark, *J. Phys. Chem. A* **108** 2851 (2005).
12. T.H. Dunning, Jr., *J. Chem. Phys.* **90** 1007 (1989).
13. T.H. Dunning, Jr., K.A. Peterson, A.K. Wilson, *J. Chem. Phys.* **114** 9244 (2001).
14. A. Schäfer, C. Huber, R. Ahlrichs, *J. Chem. Phys.* **100** 5029 (1994).
15. F. Weigend, R. Ahlrichs, *Phys. Chem. Chem. Phys.* **7** 3297 (2005).
16. H.-J. Werner, P.J. Knowles, *J. Chem. Phys.* **89**, 5803 (1988).
17. R.J. Buenker, S.D. Peyerimhoff, W. Butscher, *Mol. Phys.* **35**, 771 (1978).
18. K. Andersson, P.-Å. Malmqvist, B.O. Roos, *J. Chem. Phys.* **96**, 1218 (1992).
19. B.O. Roos, K. Andersson, M.P. Fülscher, P.-Å. Malmqvist, K. Serrano-Andres, K. Pierloot, M. Merchán, *Advances in Quantum Chemistry* **93**, 219 (1996).
20. M. Schütz, G. Hetzer, H.-J. Werner, *J. Chem. Phys.* **111**, 5691 (1999).
21. W. Klopper, W. Kutzelnigg, *Chem. Phys. Lett.* **134**, 17 (1986), *ibid. J. Chem. Phys.* **94**, 1985 (1991).

22. W. Klopper, Chem. Phys. Lett. **186**, 583 (1991).
23. A. Halkier, T. Helgaker, P. Jørgensen, W. Klopper, H. Koch, J. Olsen, A.K. Wilson, Chem. Phys. Lett. **286** 243 (1998).
24. T.H. Dunning, Jr, K.A. Peterson, J. Phys. Chem. **99** 3898 (1995).
25. S.F. Boys, F. Bernadi, Mol. Phys. **19**, 553 (1970).
26. F.B. van Duijneveldt, J.G.C.M. van Duijneveldt-van de Rijdt, J.H. van Lenthe, Chem. Rev. **94** 1873 (1994).
27. H.J. Werner, P.J. Knowles, J. Chem. Phys. **94** 1264 (1991).
28. T. van Mourik, A.K. Wilson, T.H. Dunning, Jr., Mol. Phys. **96** 529 (1999).
29. P. J. Hay, W.R. Wadt, J. Chem. Phys. **82**, 270 (1985).
30. L.R. Kahn, P. Baybutt, D.J. Truhlar, J. Chem. Phys. **65** 3826 (1976).
31. V. Bonifacic, S. Huzinaga, J. Chem. Phys. **60** 2779 (1974).
32. S. Huzinaga, L. Seijo, Z. Barandiaran, M. Klobukowski, J. Chem. Phys. **86** 2132 (1987).
33. P.G. Szalay, R.J. Bartlett, Chem. Phys. Lett. **214** 481 (1993).
34. R.J. Gdanitz, R. Ahlrichs, Chem. Phys. Lett. **143** 413 (1988).
35. B.A. Heß, Phys. Rev. A **33** 3742 (1986).
36. M. Douglas, N.M. Kroll, Ann. Phys. (N.Y.) **82**, 89 (1974).
37. R.D. Cowan, D.C. Griffin, J. Opt. Soc. Am. **66**, 1010 (1976).
38. P. Pyykkö, Chem. Rev. **88** 563 (1988).
39. P. Pyykkö, Angew. Chem. Int. Ed. **43** 4412 (2004).
40. P. Pyykkö, Chem. Rev. **97** 597 (1997).
41. J. Noga, W. Kutzelnigg, J. Chem. Phys. **101**, 7738 (1994).
42. T. Helgaker, W. Klopper, H. Koch, J. Noga, J. Chem. Phys. **106** 9639 (1997).
43. A. Halkier, T. Helgaker, P. Jørgensen, W. Klopper, J. Olsen, Chem. Phys. Lett. **302**, 437 (1999).
44. A. Halkier, W. Klopper, T. Helgaker, P. Jørgensen, P.R. Taylor, J. Chem. Phys. **111** 9157 (1999).
45. A. Halkier, W. Klopper, T. Helgaker, P. Jørgensen, J. Chem. Phys. **111** 4424 (1999).
46. D.G. Truhlar, Chem. Phys. Lett. **294** 45 (1998).
47. P.L. Fast, M.L. Sanchez, D.G. Truhlar, J. Chem. Phys. **111** 2921 (1999).
48. L.A. Curtiss, C. Jones, G.W. Trucks, K. Raghavachari, J.A. Pople, J. Chem. Phys. **93**, 2537 (1990).
49. L.A. Curtiss, K. Raghavachari, G.W. Trucks, J.A. Pople, J. Chem. Phys. **94**, 7221 (1991).
50. L.A. Curtiss, K. Raghavachari, P.C. Redfern, V. Rassolov, J.A. Pople, J. Chem. Phys. **109**, 7764 (1998).
51. L.A. Curtiss, K. Raghavachari, J.A. Pople, J. Chem. Phys. **98**, 1293 (1993).
52. L.A. Curtiss, K. Raghavachari, P.C. Redfern, V. Rassolov, J.A. Pople, J. Chem. Phys. **110**, 4703 (1999).
53. H. Lischka, R. Shepard, R.M. Pitzer, I. Shavitt, M. Dallos, Th. Müller, P.G. Szalay, M. Setz, G.S. Kedziora, S. Yabushita, Z. Zhang, Phys. Chem. Chem. Phys. **3** 664 (2001).
54. MOLCAS version 6.5, G. Karlström, R. Lindh, P.A. Malmqvist, B.O. Roos, U. Ryde, V. Veryazov, P.O. Widmark, M. Cossi, B. Schimmelpfennig, P. Neogrady, L. Seijo, Comp. Mat. Sci. **28** 222 (2003).

Introduction to Density Functional Theory and Exchange-Correlation Energy Functionals

R. O. Jones

Institute for Solid State Research
Forschungszentrum Jülich
52425 Jülich, Germany
E-mail: r.jones@fz-juelich.de

Density functional calculations of cohesive and structural properties of molecules and solids can be performed with less computational effort than by using other methods of comparable accuracy. We outline the theory and discuss the local density approximations basic to most applications, and we discuss ways to modify these approximations to yield improved results.

1 Foreword

The NIC Winter School 2006 focuses on Computational Nanoscience, particularly on understanding the properties of materials so that we can “tailor” new ones with novel properties. A glance at the program shows the dominant role played by the density functional (DF) formalism in this area. I have been asked to present an introduction to this theory and, in a second lecture, to discuss approximations to the exchange and correlation energy. These lecture notes cover these topics, but the first lecture will follow a more historical (and more lively) approach to DF theory. The theory now plays a ubiquitous role in condensed matter calculations *and* in chemistry, but its general acceptance was slow in coming, and there has been much rewriting of history since its arrival. The following notes are derived in part from earlier articles¹ and provide a background for other talks in this School.

2 Introduction

The description of many-particle systems covers much of physics and chemistry. In computational nanoscience we focus on atoms, molecules, and solids, on their structural and cohesive properties, and on how they interact with each other. We usually emphasize the properties of specific materials rather than universal properties of all, and our interest on structural and binding means that the total energy E of a system and its components is very important. The density functional (DF) formalism is our main theme here, but we shall note how its perspective differs from that of other methods used to discuss many-particle systems.

If we wish to discuss the properties of an interacting system, such as the electrons in a molecule or solid, it is natural to consider the many-electron wave function, $\Psi(\vec{r}_1, \vec{r}_2, \dots, \vec{r}_N)$, where the $\vec{r}_i$ denote the particle coordinates and spins. This has been discussed in this School by Prof. V. Staemmler. Hartree² made one of the earliest approximations for Ψ in writing the many-electron wave function as a product of single-particle functions, i.e.

$$\Psi(\vec{r}_1, \vec{r}_2, \dots) = \psi_1(\vec{r}_1) \dots \psi_N(\vec{r}_N) \quad (1)$$

Each of the functions $\psi_i(\vec{r}_i)$ satisfies a one-electron Schrödinger equation with a potential term arising from the average field of the other electrons, i.e.

$$\left(-\frac{\hbar^2}{2m}\nabla^2 + V_{\text{ext}} + \Phi_i\right) \psi_i(\vec{r}) = \epsilon_i \psi_i(\vec{r}), \quad (2)$$

where the Coulomb potential Φ_i is given by Poisson's equation,

$$\nabla^2 \Phi_i = 4\pi e^2 \sum_{\substack{j=1 \\ i \neq j}}^N |\psi_j|^2, \quad (3)$$

and V_{ext} is the potential due to the nuclei. Fermi statistics can be incorporated by replacing the product wave function by a single determinantal function.^{3,4} This ‘‘Hartree-Fock’’ (HF) approximation leads to an additional, nonlocal exchange term in the Schrödinger equation, but does not change the single-particle picture, with the wave function described in terms of orbitals with particular spins and occupation numbers. The inclusion of ‘‘exchange’’ effects improves the total energy, and the HF approximation remains an indispensable benchmark in molecular physics.

In 1960 Coulson,⁵ however, noted that ‘‘it is now perfectly clear that a single configuration (Slater determinant) wave function must inevitably lead to a poor energy’’. A linear combination of such determinants would be better, and the inclusion of ‘‘correlation effects’’ by this approach (‘‘configuration interaction’’, CI) improves the many-particle wave function, from which most properties of interest can be calculated. Some such approaches will be discussed in this School by Dr. C. Hättig. The explosive increase in the number of configurations with increasing electron number means, however, that only systems with relatively few electrons can be calculated with high accuracy. The complexity of the resulting solutions can also hinder a simple interpretation of the results.

Thomas⁶ and Fermi⁷ had proposed much earlier a scheme based on the density of electrons in the system, $n(\vec{r})$. The Thomas-Fermi (TF) scheme assumes that the motions of the electrons are uncorrelated, and that the corresponding kinetic energy can be described by a local approximation based on the results for free electrons, $[n(\vec{r})]^{5/3}$. The potential of the system can be determined from solving Poisson's equation, and the requirement of constant chemical potential leads to the TF equation for $n(\vec{r})$. This approach gives a poor description of real systems, but in its focus on the *density* it is a forerunner of later DF theories, and its mathematical properties have received much attention.⁸ In general, the integrated quantity $n(\vec{r}) = N \int d\vec{r}_2 \dots \int d\vec{r}_N \Psi^*(\vec{r}, r_2 \dots r_N) \Psi(\vec{r}, r_2 \dots r_N)$ should be easier to describe than the precise details of the wave function, Ψ .

Shortly after the original work, Dirac⁹ proposed that exchange effects be included by incorporating a term derived from the exchange energy density in a homogeneous system,

$$\epsilon_x^{\text{Dirac}} = -\frac{3}{4} \left(\frac{3}{\pi}\right)^{\frac{1}{3}} n^{\frac{1}{3}} = \frac{0.458}{r_s} \text{ a.u.}, \quad (4)$$

where $r_s = (3/(4\pi n))^{1/3}$ is a measure of the average interelectronic distance. The use of an approximate exchange potential in addition to the Hartree term in Eq. (2) was suggested by Slater,¹⁰ who showed that the exchange potential in a system of variable density could be approximated by a term with the same density dependence. However, it is *crucial* to appreciate that the argument of Slater does not depend on the exchange energy in a

homogeneous system. The exchange interaction means that an electron of a given spin will be surrounded by a region where the density of electrons with the same spin (say, spin up) is reduced. The effect of the “exchange hole” can be estimated by replacing it by a uniform density inside a sphere of radius r_0 , and zero elsewhere. Since the hole contains a single electron,¹⁰ $r_0 = (3/(4\pi n_\uparrow))^{1/3}$, and we obtain an exchange energy with the same density dependence as that of Dirac.

Gáspár¹¹ adopted the Dirac approximation to the exchange energy, and derived a slightly different exchange potential by using a variational approach. This method reproduced the HF eigenfunctions and eigenvalues for the Cu^+ ion very well, showing that an approximation based on results for a homogeneous system could give a satisfactory description even if the density is very far from uniform.

The above work has been essential to the development of modern DF theory. As an approximation to the (nonlocal) exchange potential, the simple local density (LD) approximation provided the basis for much of the work on the electronic structure of solids in the 1950’s and early 1960’s. The calculation of electronic band structures and related quantities is, of course, one of the success stories of modern condensed matter physics. The experience obtained in solving Schrödinger-like equations with local effective potentials and the improvement in numerical methods and computers during the past decades have been decisive in carrying out density functional calculations in recent years.

Hohenberg and Kohn¹² showed that ground state properties, particularly the total energy E , of a system of interacting particles could be related to the density distribution. For a nonmagnetic, nondegenerate ground state of a system of electrons in the presence of a scalar field V_{ext} , they showed that the density uniquely determines the potential, up to an arbitrary additive constant. Their proof has been simplified and extended by Levy,¹³ and the TF equation may be derived from this formalism as an approximation. Its application to the total energy of the interacting system results in a minimum principle, and the determination of the energy of the ground state can be found by the solution of single-particle equations.¹⁴ In practice, total energy calculations require approximations to be made for the exchange-correlation energy, E_{xc} , and Kohn and Sham¹⁴ showed that the local density (LD) approximation,

$$E_{\text{xc}}^{\text{LD}} = \int d\vec{r} n(\vec{r}) \varepsilon_{\text{xc}}[n(\vec{r})], \quad (5)$$

could be applied to the limiting cases of a slowly varying density and very high densities. Here $\varepsilon_{\text{xc}}[n]$ is the exchange and correlation energy per particle of a homogeneous electron gas with density n . The exchange contribution, ε_{x} , has the same form as that of Gáspár, and a variety of approximations exist for the correlation term, ε_{c} .

The limiting cases noted above are not realized in atoms, molecules or solids, and Kohn and Sham commented that “we do not expect an accurate description of chemical bonding” with the LD approximation.¹⁴ A decade passed before the first attempts were made to test its ability to describe the bonds in molecules, and it is remarkable that these showed that the LD approximation could generally reproduce ground state geometries, vibration and phonon frequencies, and moments of the density very well. The DF formalism remains the basis of most of the parameter-free calculations in extended systems, and has found widespread application to molecules and clusters. Most calculations have used the local density approximation [Eq. (5)] or modifications of it.

The practical necessity of approximating E_{xc} leads to an essential difference in perspective between the density functional and CI approaches. The goal of the latter is an exact numerical solution of the Schrödinger equation, which would yield exact answers for most quantities of interest, whereas even an exact solution of the density functional equations will only reflect the accuracy of the approximation for E_{xc} . We shall discuss some of the many approximations that have been developed. In order to appreciate some qualitative features of the density in a real system, we show in Figure 1 the spherically averaged density in the ground state of the carbon atom. The density falls monotonically from the nucleus, with the radial oscillations often expected. The charge density in small molecules also tends to be relatively featureless, with maxima at the nuclei, saddle points along the bonds and a generally monotonic decay away from both. Furthermore, the electron density in a molecule or solid shows relatively small departures from the overlapped densities of its atomic constituents. Energy differences result from subtle changes in relatively featureless density distributions, and it is not surprising that it is difficult to find simple prescriptions for determining them.

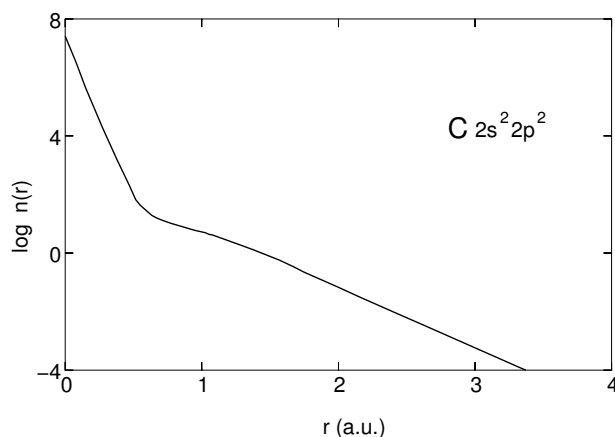


Figure 1. Spherical average of density in ground state of C atom

3 Density Functional Formalism

3.1 The Density as Basic Variable

Ground state (GS) properties of a system, including the energy E , can be expressed as functionals of the GS electron density, i.e. they are determined by a knowledge of the density alone. We show that $E[n]$ satisfies a variational principle. The Thomas-Fermi (TF) is a special case of this formalism.

The basic theorems of the density functional formalism were derived by Hohenberg and Kohn,¹² and Levy¹³ provided a simpler and more general derivation. We consider N

electrons moving in an external potential, $V_{\text{ext}}(\vec{r})$, i.e. the Hamiltonian is

$$H = T + V_{\text{ee}} + \sum_{i=1}^N V_{\text{ext}}(\vec{r}_i), \quad (6)$$

where T and V_{ee} are the kinetic and electron-electron interaction operators, respectively. For all “ N -representable” densities, $n(\vec{r})$,^a i.e. those obtainable from some antisymmetric wave function, $\psi(\vec{r}_1, \vec{r}_2, \dots, \vec{r}_N)$, Levy defined the functional,

$$F[n] = \min_{\psi \rightarrow n} \langle \psi | T + V_{\text{ee}} | \psi \rangle, \quad (7)$$

where the minimum is taken over all ψ that give the density n . $F[n]$ is universal in the sense that it refers neither to a specific system nor to the external potential $V_{\text{ext}}(\vec{r})$. If we denote E_{GS} , ψ_{GS} , and $n_{\text{GS}}(\vec{r})$ to be the ground state energy, wave function and density, respectively, then the two basic theorems of DF theory are:

$$E[n] \equiv \int d\vec{r} V_{\text{ext}}(\vec{r})n(\vec{r}) + F[n] \geq E_{\text{GS}} \quad (8)$$

for all N -representable $n(\vec{r})$, and

$$\int d\vec{r} v(\vec{r})n_{\text{GS}}(\vec{r}) + F[n_{\text{GS}}] = E_{\text{GS}} \quad (9)$$

To prove the variational principle [Eq. (8)] we introduce the notation $\psi_{\text{min}}^n(\vec{r})$ for a wave function that minimizes $F[n]$ in Eq. (7), so that

$$F[n] = \langle \psi_{\text{min}}^n | T + V_{\text{ee}} | \psi_{\text{min}}^n \rangle \quad (10)$$

Writing $V = \sum_i V_{\text{ext}}(\vec{r}_i)$, we have

$$\int d\vec{r} V_{\text{ext}}(\vec{r})n(\vec{r}) + F[n] = \langle \psi_{\text{min}}^n | V + T + V_{\text{ee}} | \psi_{\text{min}}^n \rangle \geq E_{\text{GS}}, \quad (11)$$

according to the minimum property of the GS. This proves the inequality (8). Using the minimum property once more we find

$$E_{\text{GS}} = \langle \psi_{\text{GS}} | V + T + V_{\text{ee}} | \psi_{\text{GS}} \rangle \leq \langle \psi_{\text{min}}^{n_{\text{GS}}} | V + T + V_{\text{ee}} | \psi_{\text{min}}^{n_{\text{GS}}} \rangle. \quad (12)$$

We subtract the interaction with the external potential and obtain

$$\langle \psi_{\text{GS}} | T + V_{\text{ee}} | \psi_{\text{GS}} \rangle \leq \langle \psi_{\text{min}}^{n_{\text{GS}}} | T + V_{\text{ee}} | \psi_{\text{min}}^{n_{\text{GS}}} \rangle. \quad (13)$$

On the other hand, the definition of $\psi_{\text{min}}^{n_{\text{GS}}}$ yields the reverse relation between the two sides of Eq. (13). This is possible only if

$$\langle \psi_{\text{GS}} | T + V_{\text{ee}} | \psi_{\text{GS}} \rangle = \langle \psi_{\text{min}}^{n_{\text{GS}}} | T + V_{\text{ee}} | \psi_{\text{min}}^{n_{\text{GS}}} \rangle. \quad (14)$$

^aHohenberg and Kohn¹² worked in the space of V -representable densities, i.e. those that can be realized for some external potential, $V_{\text{ext}}(\vec{r})$. This space is a subspace of N -representable densities.

Then we have

$$\begin{aligned}
E_{\text{GS}} &= \int d\vec{r} V_{\text{ext}}(\vec{r}) n_{\text{GS}}(\vec{r}) + \langle \psi_{\text{GS}} | T + V_{\text{ee}} | \psi_{\text{GS}} \rangle \\
&= \int d\vec{r} V_{\text{ext}}(\vec{r}) n_{\text{GS}}(\vec{r}) + \langle \psi_{\text{min}}^{n_{\text{GS}}} | T + V_{\text{ee}} | \psi_{\text{min}}^{n_{\text{GS}}} \rangle \\
&= \int d\vec{r} V_{\text{ext}}(\vec{r}) n_{\text{GS}}(\vec{r}) + F[n_{\text{GS}}]
\end{aligned} \tag{15}$$

These are the basic theorems. An important result also follows from Eq. (14). If the ground state GS is nondegenerate, $\psi_{\text{min}}^{n_{\text{GS}}} = \psi_{\text{GS}}$. If the ground state is degenerate, $\psi_{\text{min}}^{n_{\text{GS}}}$ is equal to one of the GS wave functions, and the others can also be obtained. The GS charge density then determines the GS wave function(s), from which all GS properties can be calculated. These properties are therefore functionals of the density.

It has been said that Wilson¹⁵ gave a shorter, if less general, proof of the definitive role of the electron density shortly after the DF formalism was first published. It is based on the “cusp condition”¹⁶ satisfied by the density at the nucleus

$$Z_{\alpha} = \frac{-1}{2\langle n(0) \rangle} \left[\frac{\partial \langle n(r_{\alpha}) \rangle}{\partial r_{\alpha}} \right]_{r_{\alpha}=0}, \tag{16}$$

where the brackets $\langle \dots \rangle$ denote a spherical average. The density n then determines uniquely the charges Z_{α} as well as the positions of the nuclei and hence the Hamiltonian.

These theorems provide a general method for calculating GS properties. If an approximation can be found for $F[n]$, we need to minimize $E[n]$ in Eq. (13) for the potential $v(\vec{r})$ of interest. This leads to corresponding approximations for E_{GS} and $n_{\text{GS}}(\vec{r})$. The same procedure applies to a ground state property X if we have an approximation for the functional $X[n]$.

3.2 Thomas-Fermi (TF) Approximation

Here the electrons are treated as independent particles, and the electron-electron interaction energy arises solely from the electrostatic energy,

$$E_{\text{es}}[n] = \frac{e^2}{2} \int d\vec{r} \int d\vec{r}' \frac{n(\vec{r})n(\vec{r}')}{|\vec{r} - \vec{r}'|}. \tag{17}$$

Furthermore, we assume that the kinetic energy is given by

$$T[n] = \int d\vec{r} t[n(\vec{r})], \tag{18}$$

where $t[n]$ is the kinetic energy density for a system of noninteracting electrons with density n . We have

$$t[n] = 2 \frac{1}{(2\pi)^3} \int_{|k| \leq k_F} d\vec{k} \frac{\hbar^2 k^2}{2m}, \tag{19}$$

where $2(4\pi/3)k_F^3/(2\pi)^3 = n$. This gives

$$T_0[n] = C_k \int d\vec{r} \{n(\vec{r})\}^{\frac{5}{3}}, \tag{20}$$

where $C_k = 3\hbar^2(3\pi^2)^{\frac{2}{3}}/10m$.

We now minimize $E(n)$ under the condition that the number of electrons is kept constant,

$$\int d\vec{r} n(\vec{r}) = N. \quad (21)$$

Using the method of Lagrange multipliers, we require that

$$E[n] + \lambda N = T[n] + E_{\text{es}}[n] + \int d\vec{r} n(\vec{r}) \{V_{\text{ext}}(\vec{r}) + \lambda\} \quad (22)$$

has a minimum. The corresponding Euler equation is the TF equation,

$$\frac{5}{3}C_k n(\vec{r})^{\frac{2}{3}} + e^2 \int d\vec{r}' \frac{n(\vec{r}')}{|\vec{r} - \vec{r}'|} + V_{\text{ext}}(\vec{r}) + \lambda = 0. \quad (23)$$

Decades of study have shown that the TF model provides a rough description of, for example, the charge density and the electrostatic potential, but it has severe deficiencies. The charge density is infinite at the nucleus, and it does not decay exponentially far from the nucleus of an atom, but as r^{-6} . It has also been shown that TF theory does not result in atoms binding to form molecules or solids. The absence of shell structure in the TF atom means that the observed periodic variation of many properties with changing atomic number cannot be reproduced. In fact, the atoms *shrink* with increasing atomic number Z (as $Z^{-1/3}$),¹⁷ and the model never predicts ferromagnetism.¹⁸

3.3 Single-particle Description of a Many-particle System

Many of the drawbacks of the Thomas-Fermi approach can be traced to the approximate treatment of the kinetic energy. The task of finding good approximations to the energy functional is greatly simplified by using a different separation introduced by Kohn and Sham,¹⁴

$$E[n] = T_0[n] + \int d\vec{r} n(\vec{r}) \left(V_{\text{ext}}(\vec{r}) + \frac{1}{2}\Phi(\vec{r}) \right) + E_{\text{xc}}[n]. \quad (24)$$

T_0 is the kinetic energy of a system with density n in the absence of electron-electron interactions, Φ is the classical Coulomb potential for electrons, and the remainder E_{xc} defines the exchange-correlation energy. T_0 differs from the true kinetic energy T , but it is of comparable magnitude and is treated *exactly* in this approach. This removes many of the deficiencies of the Thomas-Fermi approximation, such as the lack of a shell structure of atoms or the absence of chemical bonding in molecules and solids. All terms in Eq. (24) but the exchange-correlation energy E_{xc} can be evaluated exactly, so that the (unavoidable) approximations for E_{xc} play a central role in the following discussion.

The variational principle applied to Eq. (24) yields

$$\frac{\delta E[n]}{\delta n(\vec{r})} = \frac{\delta T_0}{\delta n(\vec{r})} + V_{\text{ext}}(\vec{r}) + \Phi(\vec{r}) + \frac{\delta E_{\text{xc}}[n]}{\delta n(\vec{r})} = \mu, \quad (25)$$

where μ is the Lagrange multiplier associated with the requirement of constant particle number.^b If we compare this with the corresponding equation for a system with an effective potential $V(\vec{r})$ but *without* electron-electron interactions,

$$\frac{\delta E[n]}{\delta n(\vec{r})} = \frac{\delta T_0}{\delta n(\vec{r})} + V(\vec{r}) = \mu, \quad (26)$$

we see that the mathematical problems are identical, provided that

$$V(\vec{r}) = V_{\text{ext}} + \Phi(\vec{r}) + \frac{\delta E_{\text{xc}}[n]}{\delta n(\vec{r})}, \quad (27)$$

Where the last term is referred to as the exchange-correlation potential V_{xc} . The solution of Eqs. (26), (27) can be found by solving the Schrödinger equation for noninteracting particles,

$$\left(-\frac{1}{2}\nabla^2 + V(\vec{r})\right) \psi_i(\vec{r}) = \epsilon_i \psi_i(\vec{r}), \quad (28)$$

yielding

$$n(\vec{r}) = \sum_{i=1}^N |\psi_i(\vec{r})|^2. \quad (29)$$

The condition (27) can be satisfied in a self-consistent procedure.

The solution of this system of equations leads then to the energy and density of the lowest state, and all quantities derivable from them. The formalism has also been generalized to the lowest state with a given set of quantum numbers.¹⁹ In this case E_{xc} depends on the values of the quantum numbers, and the density variations must remain within the space corresponding to the given quantum numbers. Instead of seeking these quantities by determining the wave function of the system of interacting electrons, the density functional method reduces the problem to the solution of a single-particle equation of Hartree form. In contrast to the HF potential,

$$V_{\text{HF}} \psi(\vec{r}) = \int d\vec{r}' V_{\text{HF}}(\vec{r}, \vec{r}') \psi(\vec{r}'), \quad (30)$$

the effective potential, $V(\vec{r})$ is *local*, and the equations are no more complicated to solve than Hartree's. The kinetic energy, the electrostatic interaction between core and valence electrons, and between valence electrons are treated exactly. Only the exchange energy, E_{x} , and the even smaller correlation contribution require approximation. This is in marked contrast to the Thomas-Fermi and related methods, where the large kinetic energy term is approximated.

We note here the problem of the “self-interaction correction” (SIC), where the Coulomb energy of interaction of an electron with itself must be cancelled by a contribution to E_{xc} .

^bThe functional derivative $\delta F[n]/\delta n(\vec{r})$ is defined so that the first order change in the functional due to a change in its argument $n(\vec{r})$ is:

$$\delta F = \int d\vec{r} \frac{\delta F}{\delta n(\vec{r})} \delta n(\vec{r})$$

The importance of SIC is obvious immediately if we consider a single-electron system such as the hydrogen atom. We also note the scaling condition on the exchange energy²⁰

$$E_x [\lambda^3 n(\lambda \vec{r})] = \lambda E_x [n(\vec{r})], \quad (31)$$

which has proved to be useful in constructing functional approximations.

3.4 Spin Systems

The generalization of the single-particle equations to spin systems or an external magnetic field requires the consideration of the spin indices, α , of the one-electron operators $\psi_\alpha(\vec{r})$. In general, this requires the replacement of V_{ext} by a spin-dependent potential, $V_{\text{ext}}^{\alpha\beta}(\vec{r})$, and the charge density $n(\vec{r})$ by the density matrix

$$n_{\alpha\beta}(\vec{r}) = \langle \Psi | \psi_\beta^\dagger(\vec{r}) \psi_\alpha(\vec{r}) | \Psi \rangle \quad (32)$$

All ground state properties are functionals of the density matrix $n_{\alpha\beta}$, and E is stationary with respect to variations in $n_{\alpha\beta}$, provided²¹

$$\sum_\alpha \int d\vec{r} n_{\alpha\alpha} = N. \quad (33)$$

The application of the variational principle requires that $n_{\alpha\beta}$ is N -representable, which is guaranteed for all non-negative, differentiable $n_{\alpha\beta}$ that satisfy Eq. (33).²² The effective spin-dependent potential in the single-particle equations is

$$V_{\text{xc}}^{\alpha\beta} = \frac{\delta E_{\text{xc}}}{\delta n_{\alpha\beta}(\vec{r})}. \quad (34)$$

The original derivation of the Hohenberg-Kohn theorem leads to analytical problems in the case of spin systems. These authors showed that the ground state density $n(\vec{r})$ determined the ground state wave function $\Psi(\vec{r}_1, \vec{r}_2, \dots)$ uniquely, which determines the potential $V(\vec{r})$ up to an additive constant. For spin DF theory, however, von Barth and Hedin²¹ showed that the uniqueness of the potential is not guaranteed, and Capelle and Vignale²³ found that this is true in general. A similar result was found by Eschrig and Pickett.²⁴ Kohn et al.²⁵ estimate that the practical consequences for spin DF calculations are not significant. Since the potential and magnetic field are unambiguous, and the kinetic, Coulomb, and exchange-correlation energies are defined in terms of Ψ , but independently of $n(\vec{r})$ and the spin density. A result that is useful in constructing functional approximations is “spin scaling”,²⁶ which applies to the exchange energy in spin DF theory,

$$E_x^{\text{SDF}}[n_\uparrow, n_\downarrow] = \frac{1}{2} (E_x^{\text{DF}}[2n_\uparrow] + E_x^{\text{DF}}[2n_\downarrow]). \quad (35)$$

The approximation used most widely in total energy calculations of spin systems is the local spin density (LSD) approximation,

$$E_{\text{xc}}^{\text{LSD}} = \int d\vec{r} n(\vec{r}) \varepsilon_{\text{xc}}[n_\uparrow(\vec{r}), n_\downarrow(\vec{r})], \quad (36)$$

where $\varepsilon_{\text{xc}}[n_\uparrow, n_\downarrow]$ is the exchange and correlation energy per particle of a homogeneous, spin-polarized electron gas with spin-up and spin-down densities $n_\uparrow$ and $n_\downarrow$, respectively.

Numerous electron gas calculations have been performed over the years, and some parameterizations of $\varepsilon_{\text{xc}}[n_{\uparrow}, n_{\downarrow}]$ will be discussed below. Different calculations lead to somewhat different results, but all are free of adjustable parameters. The $X\alpha$ -approximation is obtained if ε_{xc} in Eq. (5) is replaced by the exchange energy per electron and multiplied by $3\alpha/2$, i.e.

$$E_{\text{x}}^{\text{X}\alpha} = -\frac{3}{2}\alpha C \int d\vec{r} \left\{ (n_{\uparrow}(\vec{r}))^{4/3} + (n_{\downarrow}(\vec{r}))^{4/3} \right\}, \quad (37)$$

where $C = 3(3/4\pi)^{1/3}$. The α -dependence of energy differences for a given atom or molecule is weak for values near $2/3$, the exchange energy value (LSDX). There are systematic differences between the two sets of calculations, with the latter approximation leading to overestimates of the relative stability of states with larger spin densities.

3.5 Exchange-Correlation Energy, E_{xc}

The numerical advantages of the approach described are obvious. Efficient methods exist for solving single-particle Schrödinger equations with a local effective potential, and there is no restriction to small systems. We have noted, however, that E_{xc} , is defined as the difference between the exact energy and other contributions that may be evaluated numerically exactly. In practice, it is necessary to make approximations for this term, and we now examine it in detail.

The crucial simplification in the density functional scheme is the relationship between the interacting system, whose energy and density we seek, and the fictitious, noninteracting system for which we solve Eqs. (28), (29). This can be studied by considering the interaction $\lambda/|\vec{r} - \vec{r}'|$ and varying λ from 0 (noninteracting system) to 1 (physical system). This is done in the presence of an external potential, V_{λ} ,²⁷ such that the ground state of the Hamiltonian

$$H_{\lambda} = -\frac{1}{2}\nabla^2 + V_{\text{ext}}(\vec{r}) + V_{\lambda} + \lambda V_{\text{ee}} \quad (38)$$

has density $n(\vec{r})$ for all λ . The exchange-correlation energy of the interacting system can then be expressed in terms of an integral over the coupling constant λ .^{19,28}

$$E_{\text{xc}} = \frac{1}{2} \int d\vec{r} n(\vec{r}) \int d\vec{r}' \frac{1}{|\vec{r} - \vec{r}'|} n_{\text{xc}}(\vec{r}, \vec{r}' - \vec{r}), \quad (39)$$

with

$$n_{\text{xc}}(\vec{r}, \vec{r}' - \vec{r}) \equiv n(\vec{r}') \int_0^1 d\lambda \left(g(\vec{r}, \vec{r}', \lambda) - 1 \right). \quad (40)$$

The function $g(\vec{r}, \vec{r}', \lambda)$ is the pair-correlation function of the system with density $n(\vec{r})$ and Coulomb interaction λV_{ee} . The exchange-correlation hole, n_{xc} , describes the effect of the interelectronic repulsions, i.e. the fact that an electron present at the point $\vec{r}$ reduces the probability of finding one at $\vec{r}'$. The xc-energy may then be viewed as the energy resulting from the interaction between an electron and its exchange-correlation hole. This picture is, of course, the same as that used by Slater.¹⁰

We note three points: Firstly, since $g(\vec{r}, \vec{r}')$ tends to unity as $|\vec{r} - \vec{r}'| \rightarrow \infty$, the above separation into electrostatic and exchange-correlation energies can be viewed as an approximate separation of the consequences of long- and short-range effects, respectively, of the Coulomb interaction. We may then expect that the total interaction energy will be less sensitive to changes in the density, since the long-range part can be calculated exactly.

The second observation¹⁹ arises from the isotropic nature of the Coulomb interaction, V_{ee} , and has important consequences. A variable substitution $\vec{R} \equiv \vec{r}' - \vec{r}$ in Eq. (39) yields

$$E_{xc} = \frac{1}{2} \int d\vec{r} n(\vec{r}) \int_0^\infty dR R^2 \frac{1}{R} \int d\Omega n_{xc}(\vec{r}, R). \quad (41)$$

Eq. (41) shows that the xc-energy depends only on the spherical average of $n_{xc}(\vec{r}, R)$, so that approximations for E_{xc} can give an *exact* value, even if the description of the nonspherical parts of n_{xc} is quite inaccurate. Thirdly, from the definition of the pair-correlation function, there is a sum-rule that requires that the xc-hole contains one electron, i.e. for all $\vec{r}$,

$$\int d\vec{r}' n_{xc}(\vec{r}, \vec{r}' - \vec{r}) = -1. \quad (42)$$

This means that we can consider $-n_{xc}(\vec{r}, \vec{r}' - \vec{r})$ as a normalized weight factor, and define locally the radius of the xc-hole,

$$\left\langle \frac{1}{\vec{R}} \right\rangle_{\vec{r}} = - \int d\vec{r}' \frac{n_{xc}(\vec{r}, R)}{|\vec{R}|}. \quad (43)$$

This leads to

$$E_{xc} = -\frac{1}{2} \int d\vec{r} n(\vec{r}) \left\langle \frac{1}{\vec{R}} \right\rangle_{\vec{r}}. \quad (44)$$

showing that, provided the sum-rule [Eq. (42)] is satisfied, the exchange-correlation energy depends only weakly on the details of n_{xc} .¹⁹ In fact, we can say that it is determined by the first moment of a function whose second moment we know exactly. This argument simply makes more precise the ideas of Slater concerning the exchange hole and its consequences.

The spin analogy of the exact expression for E_{xc} [Eqs. (39), (40)] is

$$E_{xc} = \frac{1}{2} \sum_{\alpha\beta} \int d\vec{r} n_\alpha(\vec{r}) \int d\vec{r}' \frac{n_\beta(\vec{r}')}{|\vec{r} - \vec{r}'|} \int_0^1 d\lambda (g_{\alpha\beta}(\vec{r}, \vec{r}'; \lambda) - 1). \quad (45)$$

The atomic density distribution shown in Figure 1 makes it clear that arguments based on small departures from homogeneity cannot be applied. However, we have seen that a partial cancellation of errors can be expected provided that the sum-rule [Eq. (42)] is satisfied. It is instructive to compare the form of the exchange hole found in LSD calculations with that determined exactly, i.e. in a Hartree-Fock calculation. This is shown in Figure 2 for two representative values of $\vec{r}$ in the nitrogen atom. The approximate and exact holes are qualitatively different; the approximate hole is spherically symmetric and centred on $\vec{r}$, while the exact hole is very asymmetric around its centre $\vec{r}'$. The spherical averages are, however, remarkably similar, and the values of the exchange energy differ by only a few percent. The large differences in the exchange holes arise almost completely from the nonspherical components, which contribute nothing to the exchange energy.

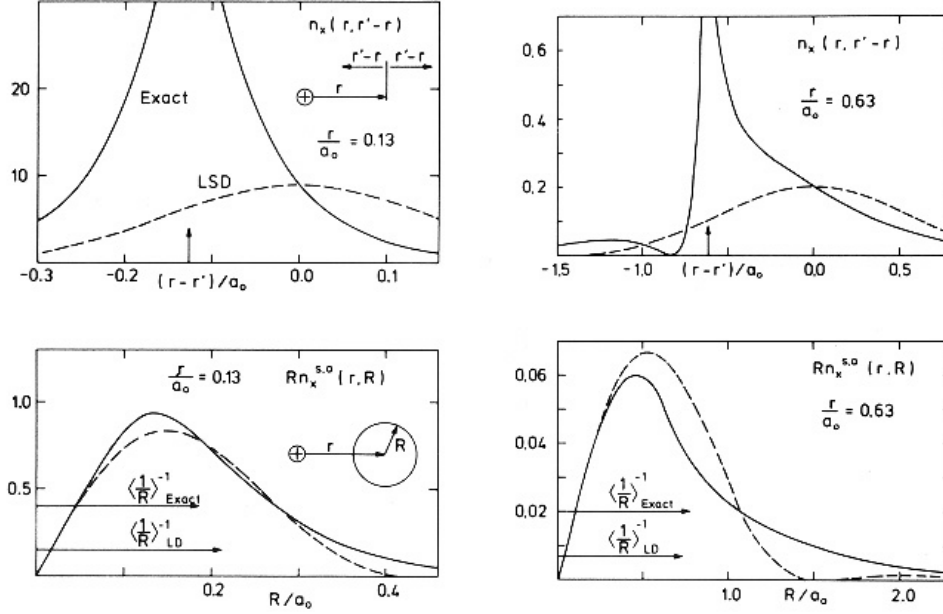


Figure 2. Exact (solid line) and approximate (broken line) exchange hole $n_x(\vec{r}, \vec{r}' - \vec{r})$ [Eq. (40)] for a spin up electron in a nitrogen atom for $r = 0.13$ and 0.63 a.u. The top figure shows the hole along a line through the nucleus and the electron. The arrow indicates the nuclear position and $\vec{r} - \vec{r}' = 0$ gives the electron position. The exact hole has a large weight at the nucleus, while the approximate hole is centered on the electron. The lower figure shows the spherical average of the hole around the electron. The area under the curve is proportional to the exchange energy. The figure also shows the value of $\langle 1/\bar{R} \rangle_{\vec{r}}$ defined in Eq. (43).

4 Exchange-Correlation Energy Functionals

The local density (LD) and local spin density (LSD) approximations have played a continuing role in DF calculations, and most functional approximations in current use are based on them. In this section we discuss these approximations as used in actual calculations, the reasons for the errors that result, and some of the modifications introduced to correct for them.

4.1 LD and LSD Approximations

The LD approximation [Eq. (5)] has an exchange component ε_x , given by the Dirac formula [Eq. (4)], and a much more complicated correlation component ε_c . The latter is commonly based on the results of quantum Monte Carlo (MC) calculations for a homogeneous electron gas of different densities.^{29,30} The LSD approximation [Eq. (36)] also requires results for a spin-polarized system and a means of interpolating for one with partial polarization.

Perdew and Zunger³¹ proposed the formula

$$\varepsilon_c^{\text{PZ}} = \begin{cases} A \ln r_s + B + C r_s \ln r_s + D r_s & r_s \leq 1, \\ \gamma / (1 + \beta_1 \sqrt{r_s} + \beta_2 r_s) & r_s > 1. \end{cases} \quad (46)$$

For dense electron systems ($r_s \leq 1$) the random phase approximation provides the parameters for the LD approximation, and the parameters for a fully polarized gas can be obtained by scaling arguments. Other parameters were obtained by fitting to the results of Ceperley and Alder.²⁹

Another form that has found widespread use is due to Vosko et al.³²

$$\begin{aligned} \frac{\varepsilon_c^{\text{VWN}}}{A} = & \ln \left(\frac{r_s}{F(\sqrt{r_s})} \right) + \frac{2b}{\sqrt{4c - b^2}} \arctan \left(\frac{\sqrt{4c - b^2}}{2\sqrt{r_s} + b} \right) \\ & - \frac{bx_0}{F(x_0)} \left[\ln \left(\frac{\sqrt{r_s} - x_0}{F(\sqrt{r_s})} \right) + \frac{2(b - 2x_0)}{\sqrt{4c - b^2}} \arctan \left(\frac{\sqrt{4c - b^2}}{\sqrt{r_s} + b} \right) \right]. \end{aligned} \quad (47)$$

Here $F(x) = x^2 + bx + c$ and the fitting parameters, which differ for the polarized and unpolarized cases, are obtained from the data of Ceperley and Alder.²⁹

Interpolation between unpolarized (U) and fully polarized (P) results is usually based on a formula involving the spin polarization ζ ,

$$\varepsilon_{\text{xc}}(n, \zeta) = f(\zeta) \varepsilon_{\text{xc}}^{\text{U}} + (1 - f(\zeta)) \varepsilon_{\text{xc}}^{\text{P}}(n) \quad (48)$$

where

$$\zeta = \frac{n_{\uparrow}(\vec{r}) - n_{\downarrow}(\vec{r})}{n_{\uparrow}(\vec{r}) + n_{\downarrow}(\vec{r})}. \quad (49)$$

Functional forms of $f(\zeta)$ based on the Hartree-Fock and random phase approximations have been suggested by von Barth and Hedin²¹ and Vosko et al.,³² respectively.

4.2 Results of LD and LSD Calculations – General Trends

We have noted above that the electron density in molecules and solids is generally close to a superposition of atomic densities. These are far from uniform and even have a cusp at the nucleus. The density is so far from uniform that the use of the LD and LSD approximations was treated initially by many with great reserve, particularly in the world of chemistry. We would not be discussing them today, however, if they did not have real value. In fact, they almost always lead to the correct *picture* of binding trends across and down the periodic table. Structures, bond lengths, and vibration frequencies in many systems are reproduced satisfactorily, and the deviations from measurements of these and other quantities are often quite systematic. We now discuss some of the deviations.

Binding energies of molecules and cohesive energies of solids are usually overestimated. In weakly bound systems, e.g. where closed-shell molecules bind, these overestimates are often large, and the bond lengths are too short. In finite systems (atoms, surfaces, ...) V_{xc} decays exponentially rather than with the inverse power of distance (the “image potential”). This is one of the consequences of the incomplete self-interaction correction. Negative ions (such as H^-) are often unbound, although it may be possible to estimate binding energies in some cases. In semiconductors the energy gap between valence and

conduction bands is less than the measured values. This topic will be discussed in this school by Dr. A. Schindlmayr. The so-called strongly correlated systems, such as the insulating 3d-transition metal oxides, are not described well by LSD calculations, which predicts them to be metallic. Partially filled *d*- and *f*-bands, in general, require modifications of the LD and LSD approximations. Such problems will be discussed in this school by Dr. S. Kurth.³³ Some calculated energy differences are in spectacular disagreement with experiment and provide us with a way to understand the defects of the LD and LSD approximations. We now focus on promotion energies in atoms, i.e. different states of lowest symmetry of the same atom.

4.3 LSD Calculations – Sources of Error

The above arguments show that the LSD approximation can give a reasonable description of systems where the density is far from homogeneous. As experience with molecular calculations developed, however, it became clear that LSD calculations gave rise to some persistent, occasionally systematic errors for a variety of systems.¹ The binding energies of *sp*-bonded molecules, for example, is often overestimated by ~ 1 eV per bond, and discrepancies of this size are simply unacceptable.

When atoms bind to form a molecule or solid, there is generally a change in the relative *s*, *p*, *d* ... contributions to the wave function. Such energy transfers can be studied directly in atoms by calculating the energy required to change an *s*-electron into a *p*-electron, for example the $^4S (2s^2 2p^3)$ ground state of N into the $^4P (2s^1 2p^4)$ excited state. Since these are the lowest lying states of these symmetries, density functional calculations are justified.¹⁹

In Figure 3 we compare experimental *sp*-transfer energies for first row atoms and ions with the results of LSD calculations.³⁴ For atoms we define

$$\Delta_{sp} = E(1s^2 2s 2p^{n-1}) - E(1s^2 2s^2 2p^{n-2}) \quad (50)$$

with a similar definition for the ions. While the overall trends are given well, the absolute deviations in ΔE_{sp} are of variable quality, with particularly large deviations in the cases of *O* and *N*. Also shown are the HF^{35,36} and LSDX ($X\alpha$) results. The trends are given fairly well by these calculations, but there are significant errors, and the comparison with experiment shows differences between the first and second halves of the row. To obtain insight into these problems we study some simple model systems, focusing on the exchange energy, where the presence of explicit formulas simplifies the discussion greatly.

The HF exchange energy can be expressed in terms of exchange integrals,³⁷

$$I_{ij} = e^2 \int d\vec{r} \int d\vec{r}' \frac{\Phi_i^*(\vec{r}) \Phi_j(\vec{r}) \Phi_i(\vec{r}') \Phi_j^*(\vec{r}')}{|\vec{r} - \vec{r}'|}, \quad (51)$$

where Φ_i and Φ_j are HF orbitals. Such an integral depends strongly on the nodal structure of Φ_i and Φ_j . If Φ_i and Φ_j have different *l*- and *m*-quantum numbers, for example, the integrand oscillates and I_{ij} is reduced. Since the LSD approximations for exchange (referred to as LSDX) and correlation are expressed only in terms of the charge density, we cannot expect them to include effects of the nodal structure in a precise way.

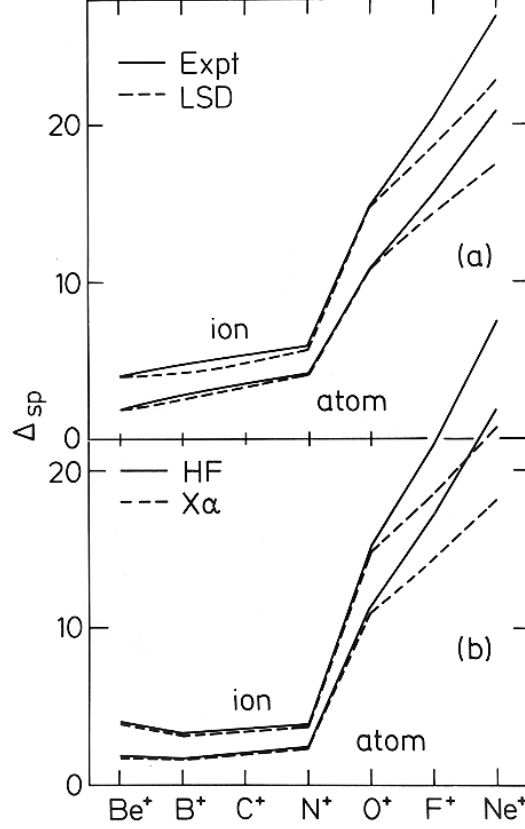


Figure 3. The sp transfer energies Δ_{sp} for the first row atoms and ions. (a) experimental and LSD results, (b) Hartree-Fock and LSDX ($X\alpha$) results. The energies are in eV.

As an example, we consider the sp -transfer in an F atom, where an $s_{\downarrow}$ electron is transferred to a $p_{\downarrow}$ orbital,³⁴

$$1s^2 2s(\uparrow\downarrow) 2p(\uparrow\uparrow\downarrow\downarrow) \rightarrow 1s^2 2s(\uparrow) 2p(\uparrow\uparrow\downarrow\downarrow\downarrow). \quad (52)$$

The change in the exchange energy due to this process is³⁷

$$\Delta E_x = -\frac{9}{25}G^2(2p, 2p) + \frac{2}{3}G^1(2p, 2s), \quad (53)$$

where the Slater integrals G^k are defined by

$$G^k(i, j) = e^2 \int_0^\infty dr r^2 \int_0^\infty dr' (r')^2 \frac{r_{<}^k}{r_{>}^{k+1}} \phi_i(r) \phi_j(r) \phi_i(r') \phi_j(r'). \quad (54)$$

Here $r_{<}$ ($r_{>}$) is the smaller (larger) of r and r' , and $\phi_i(r)$ is the radial part of $\Phi_i(\vec{r})$. The first term in Eq. (53) is the exchange interaction between a p -electron with $m = -1$ and two p -electrons with $m = 1$ and 0 . The second term is the interaction between an s -electron and two p -electrons. The integrand of Eq. (51) corresponding to the p - p interaction

has two nodal planes as a function of $\vec{r}$, while there is only a single nodal plane for the s - p interaction. The latter is larger as a result. Using realistic values for the Slater integrals, we find that $\Delta E_x \sim 6$ eV. The LSDX calculation, however, gives similar radial extents for the s - and p -orbitals. If we assume that the radial parts are *identical* and neglect the small nonspherical corrections, the LSDX approximation predicts that the exchange energy is *unchanged* by the sp -transfer, and the LSDX value for Δ_{sp} differs from the HF result by 6 eV. The LSDX and LSD (both exchange and correlation) approximations show similar trends, although the deviation between the LSD result and experiment is reduced to 2.6 eV, since the large change in the exchange energy is compensated in part by a change in ΔE_c of opposite sign.

A simple model problem. In order to gain insight into these problems, Gunnarsson and Jones³⁴ performed extensive LSDX and LSD calculations for atoms and compared the results with HF calculations and experiment, respectively. The differences between the LSDX and HF results are reproduced remarkably well by model calculations where we isolate the effects of the l -dependence of the orbitals by assuming that the s , p and d orbitals have the same *radial* dependence. With this assumption G^k only depends on k . For the $3s$ orbital in calcium, the numerical relations

$$G^0 = \frac{G^1}{0.680} = \frac{G^2}{0.516} = \frac{G^3}{0.414} = \frac{G^4}{0.344} \quad (55)$$

and

$$\int d\vec{r} \varepsilon_x[n(\vec{r}), 0] = 0.451G^0 \quad (56)$$

were found, where $n(\vec{r})$ is the charge density due to a Ca $3s$ electron. Similar coefficients were obtained for the orbitals of other elements of the iron series. Eq. (56) shows that the unphysical self-interaction, $-G^0/2$, is cancelled to within about 10%. In Figure 4(a) we compare the LSDX and HF exchange energies as a function of the number N of spin up electrons. In the HF case, the shells are filled in the order s , p , d and within each shell in the order $m, m-1, \dots, -m$. In the LSDX case the small nonspherical contributions are neglected. The LSDX and HF results agree remarkably well.

Figure 4(b) shows results for the interelectronic exchange energy, for which the self-interaction has been subtracted. With the filling order s , p , d , the LSDX and HF results are in rather good agreement, with the magnitude of the LSDX results being somewhat larger. Figure 4(b) also shows results for occupations where a subshell is left empty. This does not influence the LSDX results in the present model, but it has a pronounced effect on the HF results. For instance, the curve “ s -shell empty” is lower than the curve corresponding to the s , p , d filling, since the p - p exchange interaction is smaller than the s - p interaction. It follows from the orthogonality of the HF orbitals that the exchange hole contains one electron, and a similar sum rule is also satisfied by the LSD and LSDX approximations.¹⁹ Aspects of orthogonality and node formation are then included in all these schemes, although Figure 4(b) shows that the sum rule does not *guarantee* a good description.

This model calculation illustrates two important conclusions: (i) If we occupy the orbitals with the minimum number of nodal planes consistent with the Pauli principle, the trends in the interelectronic exchange energies are reproduced well by the LSDX approximation. The absolute value is overestimated in all systems considered by Gunnarsson and Jones.³⁴ (ii) The energy for the transfer from such a state to a state with one additional

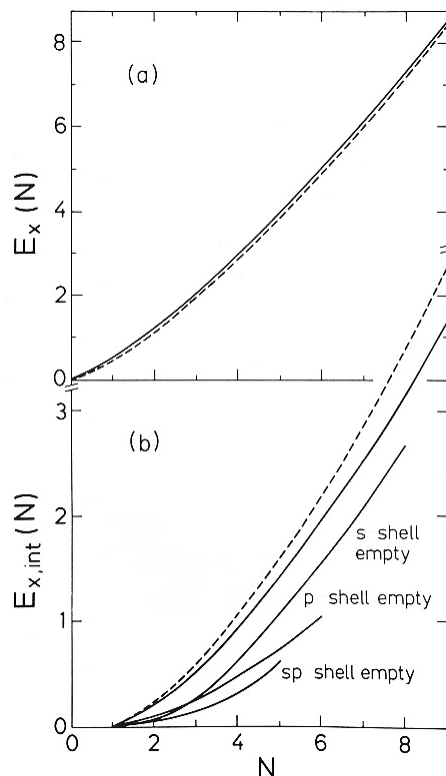


Figure 4. Exchange energies in model described in text. (a) E_x^{LSD} (dashed curve) and E_x^{HF} (solid curve) as a function of the number of electrons N (in units of $-G^0$); (b) the interelectronic exchange $E_{x,\text{int}}$ for these two approximations. In the HF approximation, we show the dependence for different schemes of occupying the orbitals.

node is often underestimated substantially in the LSDX approximation. These results were derived from atomic calculations, but the arguments can be extended to small molecules.³⁴

This shows that the nodal structure of the wave functions can greatly affect the accuracy of the LSD approximation, and we have identified classes of problems where the LSD results must be treated with caution. For states with the minimum number of nodal planes consistent with the orthogonality of the orbitals, the LSD approximation usually leads to a moderate overestimate of the exchange-correlation energy. For states with additional nodal planes the exchange-correlation energy is often greatly overestimated. In atoms, the depopulation of s -orbitals can lead to large errors, and similar effects may be expected in bonding situations where sp or sd hybridization reduces the s occupancy.

4.4 A Simplification of LD Calculations: Harris Functional

Before discussing some of the suggestions for overcoming problem with the LD and LSD approximations, we shall mention one way of *simplifying* LD calculations. In 1972, Kim

and Gordon^{38,39} suggested the following extension of the Thomas-Fermi approach: The density of a system comprising closed-shell atoms or molecules is approximated by overlapping the HF densities of the free atoms (or molecules), and the energy is then calculated using the TF approximations for the electrostatic [Eq. (17)] and kinetic energy [Eq. (18)] terms and Eq. (5) for the exchange-correlation energy. The energies of the constituents are estimated in the same way, and the binding energies of rare gas molecules so obtained agree remarkably well with experiment.

The same approach has been studied in the DF context by Harris.⁴⁰ For systems where the density is close to the overlapped density of the constituents, one can derive an expression for the binding energy for a given geometry,

$$E_R \sim \sum_n a_n \tilde{\epsilon}_n - \int d\vec{r} n_f(\vec{r}) \left[\frac{1}{2} \phi_f(\vec{r}) + V_{xc}^{n_f}(\vec{r}) \right] + E_{xc}[n_f] + E_N^R, \quad (57)$$

where n_f is the sum of the densities of the fragments, ϕ_f and $V_{xc}^{n_f}$ are the corresponding Coulomb and exchange-correlation potentials, and E_N is the internuclear repulsion. The $\tilde{\epsilon}_n$ are eigenvalues of the potential

$$\tilde{V}(\vec{r}) = \phi_f(\vec{r}) + V_{xc}^{n_f}(\vec{r}) + V_{\text{ext}}(\vec{r}), \quad (58)$$

where V_{ext} is the external potential. This expression has the same form as that of Kim and Gordon, with the important difference that the change in kinetic energy (for bonding between atoms A and B with potentials V_A and V_B) is

$$\begin{aligned} \delta T_\epsilon = 2 \sum_n \delta \tilde{\epsilon}_n - \int d\vec{r} n(\vec{r}_1) [\tilde{V}(\vec{r}) - V_A(\vec{r}_1)] \\ - \int d\vec{r} n(\vec{r}_2) [\tilde{V}(\vec{r}) - V_B(\vec{r}_2)], \end{aligned} \quad (59)$$

instead of the corresponding Thomas-Fermi expression.

The results of this scheme agree satisfactorily with self-consistent DF calculations for He₂, Be₂, C₂, N₂, F₂ and Cu₂.⁴⁰ It provides a useful alternative in cases where more detailed calculations would be prohibitive, and it has been adopted in some molecular dynamics calculations, where more detailed calculations are performed at regular intervals to ensure that the trajectories are reliable. It is, however, not self-consistent, and spin plays no role.

4.5 Modifications of the LSD Approximation: Pair-correlation Functions

We have seen that the total energy of a system is insensitive to many details of the pair-correlation function $g(\vec{r}, \vec{r}')$, which occurs in Eq. (40). It is nevertheless tempting to modify the pair-correlation function to remove some of its more unpleasant features. The expression for the exchange-correlation energy [Eq. (41)] shows that the differences between the contributions of the exact and approximate exchange holes are largely due to the non-spherical components of the hole. Since these do not contribute to E_{xc} , total energies and total energy differences can be remarkably good, even in systems where the density distribution is far from uniform. In the LD approximation, we assume that the xc-hole $n_{xc}(\vec{r}, \vec{r} - \vec{r}')$ depends only on the charge density at the electron. It would be more

appropriate to assume⁴¹ that n_{xc} depends on a suitable average $\overline{n(\vec{r})}$,

$$n_{xc}(\vec{r}, \vec{r}^{\dagger} - \vec{r}) = \overline{n(\vec{r})} \int d\lambda [g_h(\vec{r} - \vec{r}^{\dagger}, \lambda, \overline{n(\vec{r})}) - 1]. \quad (60)$$

It is possible to choose the weight function which determines $\overline{n(\vec{r})}$ so that the functional reduces to the exact result in the limit of almost constant density. The approximation [Eq. (60)] satisfies the sum rule [Eq. (42)]. Other prescriptions for the weight function have been proposed.

An alternative approximation is obtained if we keep the proper prefactor $n(\vec{r}^{\dagger})$ in Eq. (40), leading to the so-called “weighted density” (WD) approximation,

$$n_{xc}(\vec{r}, \vec{r}^{\dagger} - \vec{r}) = n(\vec{r}^{\dagger}) G(|\vec{r} - \vec{r}^{\dagger}|, \tilde{n}(\vec{r})), \quad (61)$$

where $\tilde{n}(\vec{r})$ is chosen to satisfy the sum rule [Eq. (42)]. The WD models provide a link between ε_{xc} , the density-density correlation function, and the response function of a homogeneous system. As an example, we now discuss an analytic form⁴² that is computationally simpler and can be chosen to give exact results in certain limits. We assume that

$$G(\vec{r}, n) = C(n) (1 - \exp[-\lambda(n)/|r|^5]), \quad (62)$$

where C and λ are parameters to be determined. The functional G behaves as $|\vec{r}|^{-5}$ for large distances, which is needed to obtain an image potential. For a homogeneous system with density n , we require that the model functional should both fulfil the sum rule for $\tilde{n}(\vec{r}) = n$ and give the exact xc-energy. This leads to two equations

$$\int d\vec{r} G(|\vec{r}|, n) = -1 \quad (63)$$

$$\frac{e^2}{2} \int d\vec{r} \frac{1}{|\vec{r}|} G(|\vec{r}|, n) = \varepsilon_{xc}(n), \quad (64)$$

which are sufficient to determine the two parameters $C(n)$ and $\lambda(n)$.

This functional is exact in several limiting cases: (i) for a homogeneous system, (ii) for one-electron systems such as the hydrogen atom it gives an exact cancellation of the electron self-interaction, (iii) for an atom it gives the correct behaviour of the xc-energy density far from the nucleus. $\varepsilon_{xc}(\vec{r}) = -e^2/2r$, (iv) far outside the surface it gives the image potential $\varepsilon_{xc}(z) = -e^2/(4z)$. The LSD approximation gives qualitatively incorrect answers for (iii) and (iv), and the cancellation in (ii) is only approximate. Since (ii) is satisfied, this approximation provides a “self-interaction correction” in the sense we discuss below. The errors in the WD approximation for the total energy of atoms are about an order of magnitude smaller than those in the LSD approximation, but the *sp*- and *sd*-transfer energies are little improved over LSD results.¹

4.6 Self-interaction Corrected (SIC) Approximations

In the DF formalism each electron interacts with itself via the Coulomb electrostatic energy, and this unphysical interaction would be exactly cancelled by a contribution from the xc-energy. In the LSD approximation this cancellation is imperfect, but numerically rather good. The incorrect treatment of the self-interaction in approximate functionals has

led some workers to consider self-interaction corrected (SIC) functionals. Within the LSD approximation the SIC functional takes the form

$$E_{\text{SIC}} = E_{\text{LSD}}[n_{\uparrow}(\vec{r}), n_{\downarrow}(\vec{r})] - \sum_{i\sigma} \delta_{i\sigma}, \quad (65)$$

where E_{LSD} is the LSD energy functional and

$$\delta_{i\sigma} = \frac{e^2}{2} \int d\vec{r} \int d\vec{r}' \frac{n_{i\sigma}(\vec{r})n_{i\sigma}(\vec{r}')}{|\vec{r} - \vec{r}'|} + E_{\text{xc}}^{\text{LSD}}(n_{i\sigma}, 0) \quad (66)$$

is the self-interaction correction (SIC) for the orbital i with spin σ and charge density $n_{i\sigma}(\vec{r})$. The first term in Eq. (66) is the self-interaction energy, and the second is the LSD approximation to the xc-energy of a fully spin-polarized system with density $n_{i\sigma}(\vec{r})$. This functional is exact for a one-electron system. It leads to the equation

$$\left[-\frac{\hbar^2}{2m}\nabla^2 + V(\vec{r}) + V_{i\sigma}^{\text{SIC}}(\vec{r})\right]\phi_{i\sigma} = \sum_j \lambda_{ij}^{\sigma} \phi_{j\sigma}, \quad (67)$$

where $V(\vec{r})$ is the effective potential (27) entering in a normal LSD calculation, and $V_{i\sigma}^{\text{SIC}}(\vec{r})$ is additional potential resulting from the term Eq. (66). This potential is orbital-dependent and the Lagrange parameters λ_{ij}^{σ} are introduced to ensure that the solutions $\phi_i(\vec{r})$ (referred to as “local orbitals”) are orthogonal. The SIC approximation is not invariant under a unitary transformation of the orbitals, and different basis sets lead to different total energies.

Unfortunately, ionization and transfer energies of atoms are not generally better in the LSD-SIC approximation, at least if non-spherical corrections are neglected. The improved total energies in SIC calculations can, in fact, be traced to a much better treatment of the innermost core electrons,⁴³ which play a relatively minor role in most chemical and physical processes of interest.

4.7 Modifications of the LSD Approximation: Gradient Expansions

Hohenberg and Kohn¹² introduced the local density approximation in the context of DF theory, but they also pointed out the need for modifications in systems where the density is not homogeneous. One such modification was the approximation

$$E_{\text{xc}} = E_{\text{xc}}^{\text{LSD}} - \frac{1}{4} \int d\vec{r} \int d\vec{r}' K_{\text{xc}} \left(\vec{r} - \vec{r}', n\left(\frac{\vec{r} + \vec{r}'}{2}\right) \right) \left\{ n(\vec{r}) - n(\vec{r}') \right\}^2, \quad (68)$$

where the kernel K_{xc} is related to the dielectric function of a homogeneous medium. This approximation is exact in the limit of weak density variations

$$n(\vec{r}) = n_0 + \Delta n(\vec{r}), \quad (69)$$

where $|\Delta n(\vec{r})| \ll n_0$, but the results for real systems were not encouraging. For free atoms the energy is infinite,⁴¹ indicating that the sum-rule [Eq. (42)] is not satisfied. In spite of this, gradient expansions have played an extremely important role in DF theory and its applications. We shall now discuss some of the forms used.

4.7.1 Generalized Gradient Approximations (GGA)

The second-order generalized gradient approximation to the exchange-correlation energy is written

$$E_{\text{xc}}^{\text{GGA}}(n) = \int d\vec{r} \varepsilon_{\text{xc}}(n, |\nabla n|, \nabla^2 n). \quad (70)$$

Many forms of such functionals have been suggested, and some of the most important (and all their parameters) are given by Filippi et al.⁴⁴ These authors also compared these forms with the exact results for an exactly soluble system of two electrons in an external harmonic potential. We shall discuss some functionals that are widely used in actual calculations. We use some standard abbreviations:

$$k_F = (3\pi n)^{1/3}; \quad k_s = \left(\frac{4}{\pi} k_F\right)^{1/2}; \quad s = \frac{|\nabla n|}{2k_F n}; \quad t = \frac{|\nabla n|}{2k_s n}; \quad r_s = \left(\frac{3}{4\pi n}\right)^{1/3} \quad (71)$$

For the exchange contribution, Becke⁴⁵ suggested the following form:

$$\varepsilon_x = \varepsilon_x^{\text{LD}} \left[1 - \frac{\beta}{2^{1/3} A_x} \frac{x^2}{1 + 6\beta x \operatorname{arcsinh}(x)} \right], \quad (72)$$

where $x = 2(6\pi^2)^{1/3} s = 2^{1/3} |\nabla n|/n^{4/3}$, $A_x = \frac{3}{4}(3/\pi)^{1/3}$, and $\beta = 0.0042$. The parameter β was optimized to give exchange energies of noble gas atoms using Hartree-Fock orbitals). This functional has been coupled with the LYP functional for correlation:⁴⁶

$$\varepsilon_c = -a \frac{1}{1 + dn^{1/3}} \left\{ n + bn^{-2/3} \left[C_F n^{5/3} - 2t_W + \frac{1}{9} \left(t_W + \frac{1}{2} \nabla^2 n \right) \right] \exp(-cn^{-1/3}) \right\}, \quad (73)$$

where

$$t_W = \frac{1}{8} \left(\frac{|n|^2}{n} - \nabla^2 n \right) \quad (74)$$

and $C_F = 3/10(3\pi^2)^{2/3}$, $a = 0.04918$, $b = 0.132$, $c = 0.2533$, $d = 0.349$. It surprised many that the combination of Becke exchange with either the VWN or LYP correlation functionals outperformed correlated *ab initio* methods (MP2 and QCSID) in calculations of atomization energies for 32 molecules.⁴⁷ The importance of using a non-local exchange has been underscored by many subsequent calculations.⁴⁸

The question of ways to improve exchange-correlation functionals has led to a clear dichotomy of views. The first implies that the search for a functional is so difficult that we should develop a reasonable form and fit its adjustable parameters to experimental data. The BLYP and B3LYP functionals are of this type. From this standpoint, DF theory is then “of semi-empirical nature”.⁴⁹ Experimental data from up to 407 atomic and molecular systems have been used,⁴⁹ in this case with a functional with 15 adjustable parameters. The second approach seeks to build on the positive features of the LSD approximation by incorporating exact constraints and hoping that the added features will automatically improve the description of reality. We shall focus on functionals in the second category.

Perdew, Burke, and Ernzerhof⁵⁰ have described a functional form (PBE) that has several attractive features:

$$E_{\text{xc}} = \int d\vec{r} n(\vec{r}) \varepsilon_x^{\text{LD}}(n(\vec{r})) F_{\text{xc}}(n, \zeta, s), \quad (75)$$

where

$$F_x(s) = 1 + \kappa - \frac{\kappa}{1 + \mu^2/\kappa}, \quad (76)$$

$\mu = \beta(\pi^2/3) = 0.21951$, and $\beta = 0.066725$. The correlation energy has the form

$$E_c = \int d\vec{r} n(\vec{r}) [\varepsilon_c^{\text{LSD}}(n, \zeta) + H(n, \zeta, t)] \quad (77)$$

with

$$H(n, \zeta, t) = \left(\frac{e^2}{a_0} \right) \gamma \phi^3 \ln \left\{ 1 + \frac{\beta \gamma^2}{t} \left[\frac{1 + At^2}{1 + At^2 + A^2 t^4} \right] \right\}. \quad (78)$$

The spin-scaling factor $\phi(\zeta) = [(1 + \zeta)^{2/3} + (1 - \zeta)^{2/3}] / 2$, $\gamma = (1 - \ln 2) / \pi^2 = 0.031091$, and

$$A = \frac{\beta}{\gamma} [\exp(-\varepsilon_c^{\text{LD}}[n] / (\gamma \phi^3 e^2 / a_0)) - 1]^{-1}. \quad (79)$$

The PBE form has several advantages, including: (1) In the slowly varying limit ($t \rightarrow 0$), H is given by the second-order gradient expansion, (2) In the rapidly varying limit ($t \rightarrow \infty$), $H \rightarrow -\varepsilon_c^{\text{LD}}$, and correlation vanishes, (3) It cancels the logarithmic singularity of $\varepsilon_c^{\text{LD}}$ in the high density limit. The accuracy of the PBE functional for atoms and molecules has been compared with results of LSD, BLYP, and B3LYP functions has been made by Ernzerhof and Scuseria.⁵¹ The PBE functional performed as well as B3LYP for the properties considered by these authors.

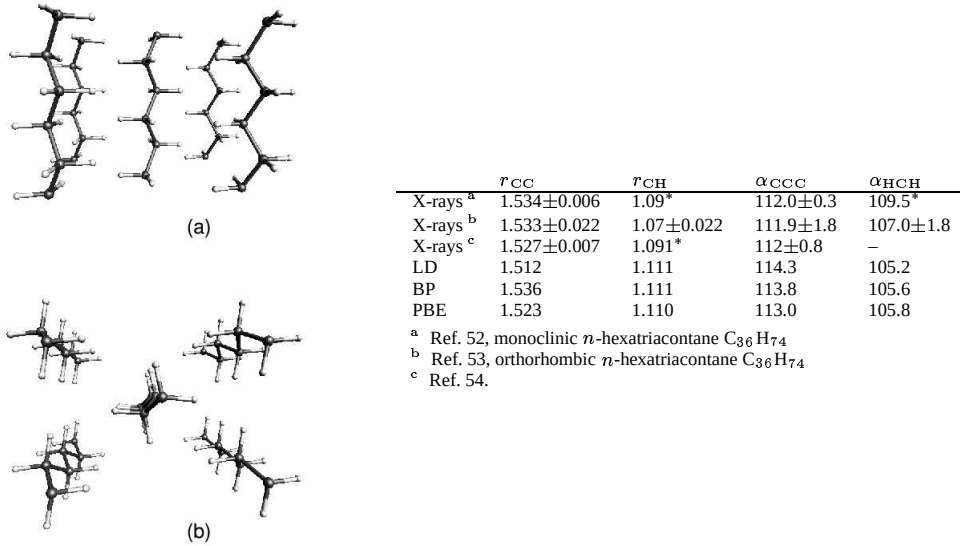


Figure 5. Left: Structure of crystalline polyethylene: (a) normal to chain axis, (b) along chain axis. Right: Structural parameters of crystalline PE from x-ray diffraction, and calculated structures for a single chain. Bond lengths are in Å, angles in degrees. Values marked with an asterisk (*) were assumed, and not measured directly.

Generalized gradient approximations generally lead to improved bond angles, lengths, and energies. In particular, the strengths of hydrogen bonds and other weak bonds between closed shell systems are significantly better than local density results. However, the self-interaction problem remains, and some asymptotic requirements for isolated atoms are not satisfied (V_{xc}^{GGA} vanishes exponentially far from the nucleus, $V_{xc}^{GGA}(\vec{r} \rightarrow 0) \rightarrow -\infty$).

An example of the differences between LSD and PBE calculations is shown in Figure 5.⁵⁵ Polyethylene crystallizes in a structure with parallel chains, and the table shows that the structure of the individual chains is described rather well, with the C-C bonds being slightly longer in PBE and BP⁵⁶ calculations than in those using LD. There are, however, striking differences in the *interchain* separations. The LD calculations lead to a pronounced minimum with a shortest distance between C atoms in different chains (3.67 Å) less than the measured value (4.59 Å),⁵⁴ while PBE leads to a modest overestimate (4.66 Å).

4.7.2 Meta-GGA

The next step in the development of gradient approximations is to incorporate the kinetic energy density (or the Laplacian of the density). A version based on the PBE form was described by Perdew et al.,⁵⁷ who added the variables $\tau_{\uparrow}$ and $\tau_{\downarrow}$, where

$$\tau_{\sigma}(\vec{r}) = \sum_i^{\text{occ.}} \frac{1}{2} |\nabla \psi_{i\sigma}(\vec{r})|^2 \quad (80)$$

is the kinetic energy density for the occupied Kohn-Sham orbitals. However, this and other forms initially developed included parameter(s) found by fitting to experimental data. This last feature was avoided in the recent work of TPSS,⁵⁸ whose form satisfied the requirement that the exchange potential be finite at the nucleus for ground state one- and two-electron densities. This is a constraint satisfied by LSD, but lost in GGA. Extensive numerical tests for atoms, molecules, solids, and jellium surfaces showed generally very good results.

4.7.3 Hybrid Schemes: Combination of HF and DF Schemes

In section (4.3) we saw just how poor the exchange energy differences could be between states whose wave functions have different nodal structures. It has also been noted for many years (see, for example, Ref. 27) that errors in the LD descriptions of exchange and correlation tend to balance. This suggests that a combination of Hartree-Fock exchange and DF calculations could be useful.

$$E_{xc}^{\text{hybrid}} = \alpha E_x^{\text{HF}} + E_c, \quad (81)$$

where α can be chosen to satisfy particular criteria. A formal justification for such hybrid schemes was given by Görling and Levy.⁵⁹ The B3LYP functional,⁴⁸ which is widely used in the chemical community, is an example. Here three adjustable parameters are used to fit calculated values to a molecular data base.

4.7.4 Optimized Effective Potential (OEP)

In the Kohn-Sham implementation of DF theory we describe the kinetic energy functional for the system of non-interacting electrons in terms of the orbitals $\psi_{i\sigma}(\vec{r})$. The best *local*

effective potential (a requirement of DF theory) can be determined variationally for each σ such that, when substituted into a stationary single-particle equation it leads to a set of eigenfunctions that minimize $E[\psi_{i\sigma}(\vec{r})]$. This was first suggested by Sharp and Horton⁶⁰ following the local density description of exchange,¹⁰ and it is referred to as the optimized effective potential (OEP).⁶¹ Its determination requires the solution of an integral equation for each $V_{xc,\sigma}(\vec{r})$, which is a demanding exercise. How this can be done, as well as possible simplifications, are discussed in this School by Dr. S. Kurth.

5 Concluding Remarks

The density functional theory is the basis of most of the calculations in computational nanoscience that are free of adjustable parameters. It is clear from the present discussion that “DF calculations” cover a colourful variety of functional approximations, and not all are “free of adjustable parameters”. The development of approximations to the exchange-correlation functionals over the past 20 years has improved the performance of DF calculations, and John Perdew is no doubt not alone in thinking that progress up the “Jacob’s ladder”⁶² (Figure 6) will continue until energy differences can be determined to within ~ 1 kcal/mol (“chemical accuracy”). The numerical cost increases as one climbs, and this may not necessarily bring more information. Most problems in “computational nanoscience” are performed part way up the ladder, and this situation will probably remain true for some time.

DF calculations should always be performed with all critical facilities intact. The familiar program “packages” will almost always lead to an answer, but it is easy to obtain

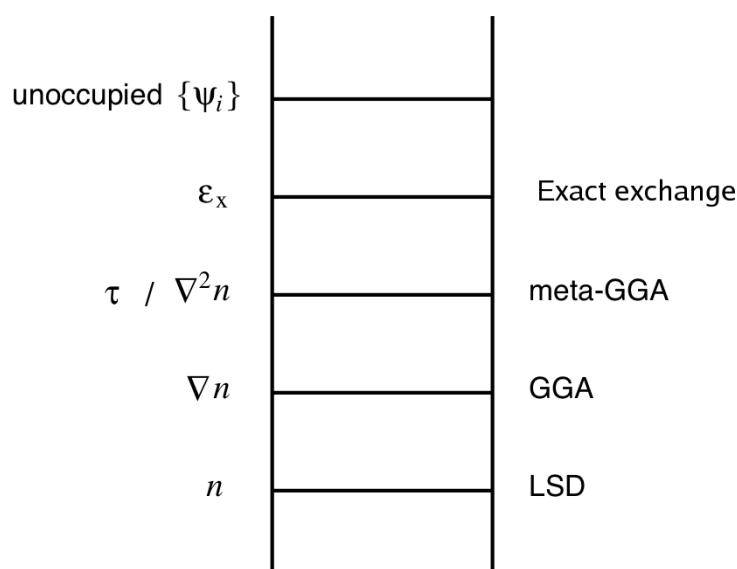


Figure 6. “Jacob’s ladder” of DF schemes according to Perdew and collaborators.

answers void of physical content. It is odd to have to conclude this manuscript on such a note, but years of refereeing articles in the field suggest that it is necessary.

References

1. R. O. Jones and O. Gunnarsson, *Rev. Mod. Phys.* **61**, 689 (1989).
2. D. R. Hartree, *Proc. Cambridge Philos. Soc.* **24**, 89 (1928).
3. J. C. Slater, *Phys. Rev.* **35**, 210 (1930).
4. V. Fock, *Z. Physik*, **61**, 126 (1930).
5. C. Coulson, *Rev. Mod. Phys.* **32**, 170 (1960).
6. L. H. Thomas, *Proc. Cambridge Philos. Soc.* **23**, 542 (1927).
7. E. Fermi, *Z. Physik*, **48**, 73 (1928).
8. E. H. Lieb, *Rev. Mod. Phys.* **48**, 553 (1976).
9. P. A. M. Dirac, *Proc. Cambridge Philos. Soc.* **26**, 376 (1930).
10. J. C. Slater, *Phys. Rev.* **81**, 385 (1951), *ibid.* **82**, 538 (1951).
11. R. Gáspár, *Acta. Phys. Hung.* **3**, 263 (1954).
12. P. Hohenberg and W. Kohn, *Phys. Rev.* **136**, B864 (1964).
13. M. Levy, *Proc. Nat. Acad. Sci. USA* **76**, 6062 (1979).
14. W. Kohn and L. J. Sham, *Phys. Rev.* **140**, A1133 (1965).
15. E. B. Wilson (Harvard University).
16. T. Kato, *Comm. Pure. Appl. Math.* **10**, 151 (1957). The cusp condition applies to systems with Coulomb interactions.
17. E. H. Lieb and B. Simon, *Phys. Rev. Lett.* **31**, 681 (1973).
18. O. Gunnarsson and H. Hjelmberg, *Phys. Scr.* **11**, 97 (1975).
19. O. Gunnarsson and B. I. Lundqvist, *Phys. Rev. B* **13**, 4274 (1976).
20. M. Levy and J. P. Perdew, *Phys. Rev. A* **32**, 2010 (1985).
21. U. von Barth and L. Hedin, *J. Phys. C: Solid State Phys.* **5**, 1629 (1972).
22. E. H. Lieb, *Int. J. Quantum Chem.* **24**, 243 (1983).
23. K. Capelle and G. Vignale, *Phys. Rev. Lett.* **86**, 5546 (2000).
24. H. Eschrig and W. E. Pickett, *Solid State Commun.* **118**, 123 (2001).
25. W. Kohn, A. Savin, and C. A. Ullrich, *Int. J. Quantum Chem.* **100**, 20 (2004).
26. G. L. Oliver and J. P. Perdew, *Phys. Rev. A* **20**, 397 (1979).
27. J. Harris and R. O. Jones, *J. Phys. F* **4**, 1170 (1974).
28. D. C. Langreth and J. P. Perdew, *Solid State Commun.* **17**, 1425 (1975).
29. D. M. Ceperley and B. J. Alder, *Phys. Rev. Lett.* **45**, 566 (1980).
30. G. Ortiz and P. Ballone, *Phys. Rev. B* **50**, 1391 (1994).
31. J. P. Perdew and A. Zunger, *Phys. Rev. B* **23**, 5048 (1981).
32. S. Vosko, L. Wilk, and M. Nusair, *Can. J. Phys.* **58**, 1200 (1980).
33. See, for example, V. I. Anisimov, F. Aryasetiawan, and A. I. Lichtenstein, *J. Phys.: Condens. Matter* **9**, 767 (1997).
34. O. Gunnarsson and R. O. Jones, *Phys. Rev. B* **31**, 7588 (1985).
35. S. Fraga, K. M. S. Saxena, and B. W. N. Lo, *At. Data*, **3**, 323 (1971).
36. G. Verhaegen and C. M. Moser, *J. Phys. B* **3**, 478 (1970).
37. J. C. Slater, *Quantum Theory of Atomic Structure*, Vol. II, McGraw-Hill, New York (1960), Appendix 21.
38. R. G. Gordon and Y. S. Kim, *J. Chem. Phys.* **56**, 3122 (1972).

39. Y. S. Kim and R. G. Gordon, J. Chem Phys. **60**, 1842 (1974).
40. J. Harris, Phys. Rev. A, **31**, 1770 (1985).
41. O. Gunnarsson, M. Jonson, and B. I. Lundqvist, Phys. Rev. B **20**, 3136 (1979).
42. O. Gunnarsson and R. O. Jones, Phys. Scr. **21**, 394 (1980).
43. O. Gunnarsson and R. O. Jones, Solid State Commun. **37**, 249 (1981).
44. C. Filippi, C. J. Umrigar, and M. Taut, J. Chem. Phys. **100**, 1290 (1994).
45. A. D. Becke, Phys. Rev. A **33**, 3098 (1988).
46. C. Lee, W. Yang, and R. G. Parr, Phys. Rev. B **37**, 785 (1988).
47. B. G. Johnson, P. M. W. Gill, and J. A. Pople, J. Chem. Phys. **97**, 7846 (1992), *ibid.* **98**, 5612 (1993).
48. See, for example, A. D. Becke, J. Chem. Phys. **98**, 5648 (1998).
49. A. D. Boese and N. C. Handy, J. Chem. Phys. **114**, 5497 (2001).
50. J. P. Perdew, K. Burke, and M. Ernzerhof, Phys. Rev. Lett. **77**, 3865 (1996); *ibid.* **78**, 1386 (1997) (E).
51. M. Ernzerhof and G. E. Scuseria, J. Chem. Phys. **110**, 5029 (1999).
52. H. M. M. Shearer and V. Vand, Acta Cryst. **9**, 379 (1956).
53. P. W. Teare, Acta Cryst. **12**, 294 (1959).
54. S. Kavesh and J. M. Schultz, J. Polym. Sci. Part A-2, **8**, 243 (1970).
55. B. Montanari, P. Ballone, and R. O. Jones, J. Chem. Phys. **108**, 6947 (1998).
56. This functional uses the Becke approximation for exchange (Ref. 45) and the correlation energy functional of J. P. Perdew, Phys. Rev. B **33**, 8822 (1986).
57. J. P. Perdew, S. Kurth, A. Zupan, and P. Blaha, Phys. Rev. Lett. **82**, 2544 (1999).
58. J. Tao, J. P. Perdew, V. N. Staroverov, and G. E. Scuseria, Phys. Rev. Lett. **91**, 146401 (2003). See also J. P. Perdew, J. Tao, V. N. Staroverov, and G. E. Scuseria, J. Chem. Phys. **120**, 6898 (2004).
59. A. Görling and M. Levy, J. Chem. Phys. **106**, 2675 (1997).
60. R. T. Sharp and G. K. Horton, Phys. Rev. **90**, 317 (1953).
61. J. D. Talman and W. F. Shadwick, Phys. Rev. A **14**, 36 (1976).
62. "And he dreamed, and behold a ladder set up on the earth, and the top of it reached to heaven: and behold the angels of God ascending and descending on it." Genesis 28:12 (King James Version).

The Pseudopotential Plane Wave Approach

Bernd Meyer

Chair of Theoretical Chemistry
Ruhr-Universität Bochum
44780 Bochum, Germany
E-mail: bernd.meyer@theochem.rub.de

The pseudopotential plane wave approach has become one of the most widely used methods for calculating ground state properties of extended systems within the framework of density functional theory. The simplicity of plane waves leads to very efficient numerical schemes for solving the Kohn–Sham equations, and the employment of pseudopotentials guarantees that the wave functions can be expanded in a relatively small set of plane waves. In this article the basic expressions for plane-wave-based total energy calculations and a short overview of the different types of ab initio pseudopotentials will be given.

1 Introduction

In the previous chapters the complicated many-body problem of strongly interacting electrons and nuclei has been mapped within the framework of the Born–Oppenheimer approximation and density functional theory onto a problem of single-particles moving in an effective external potential for a set of fixed nuclei. Our aim now is to develop a practical numerical scheme to solve the resulting single-particle Kohn–Sham equations for extended systems like crystalline solids or liquids.

The most common approach to tackle this problem is to expand the single-particle eigenstates of the Kohn–Sham equations into a set of basis functions. The Schrödinger equation then transforms into an algebraic equation for the expansion coefficient which may be solved by various well-established numerical methods. In this chapter we will show that plane waves are a particular well suited set of basis functions for extended systems. Plane waves are the exact eigenfunctions of the homogeneous electron gas. Therefore, plane waves are the natural choice for a basis expansion of the electron wave functions for simple metals where the ionic cores can be viewed as rather small perturbations to the homogeneous electron gas (“nearly free electron” metals, see for example Ref. 1). Plane waves are orthonormal and energy-independent. Hence, upon a basis set expansion the Schrödinger equation transforms into a simple matrix eigenvalue problem for the expansion coefficients. A further advantage of plane waves is that they are not biased to any particular atom. Any region in space is treated on an equal footing so that calculations do not have to be corrected for a basis set superposition error. Since plane waves do not depend on the positions of the atoms, the Hellmann–Feynman theorem can be applied directly to calculate atomic forces. Even for a non-complete basis set the Pulay terms are identical zero.

In practical calculations only plane waves up to a certain cutoff wave vector are included in the basis set. The convergence of the calculations with respect to the basis set size is therefore controlled by a single parameter and can be checked simply by increasing the length of the cutoff wave vector. However, due to the nodal structure of the valence wave

functions in the core region of the atoms a prohibitively large number of plane waves would be needed for a good representation of these fast oscillations. For plane wave approaches to be of practical use we have to replace the Coulomb potential of the electron–nucleus interaction by pseudopotentials. By introducing pseudopotentials we are able to achieve two goals: First, we can remove the core electrons from our calculations. The contribution of the core electrons to the chemical bonding is negligible but they contribute most to the total energy of the system (typically a thousand times more than the valence electrons). Hence, the removal of the core electrons from the calculation means that total energy differences between ionic configurations can be taken between much smaller numbers so that the required accuracy for the total energy calculations will be much less demanding than in an all–electron calculation. Second, by introducing pseudopotentials we replace the true valence wave functions by so-called pseudo wave functions which match exactly the true valence wave functions outside the ionic core region but are nodeless inside. These pseudo wave functions can be expanded using a much smaller number of plane wave basis states. A further advantage of pseudopotentials is that relativistic effects can be incorporated easily into the potential while further treating the valence electrons non–relativistically.

In spite of introducing pseudopotentials, the number of basis functions N_{pw} needed for an accurate calculation is still an order of magnitude larger than for approaches using localized orbitals. This disadvantage, however, is more than compensated by the possibility to evaluate many expressions with the help of the Fast Fourier Transform (FFT) algorithm. The most time consuming step in solving the single–particle Schrödinger equations is to apply the Hamilton operator to the valence wave functions. In a traditional matrix representation of the Hamilton operator this step scales quadratically with the number of basis functions. With plane waves and the FFT algorithm this operation reduces to a $N_{\text{pw}} \ln(N_{\text{pw}})$ scaling. Hence, for large systems the use of plane wave basis functions will become much more efficient than localized basis sets. Furthermore, the total charge density and the Hartree potential are easily calculated in a plane wave representation.

In Section 2 we will introduce the basic notation for describing infinitely extended periodic systems. This can only be a very brief summary. More details can be found in any textbook on solid state physics, for example in the books of Ashcroft and Mermin¹ or Kittel². We will then show why it is useful to expand the valence wave functions using a basis set of plane waves. In Section 3 the basic ideas underlying present ab initio pseudopotentials will be given. More details on the plane wave pseudopotential method can be found in the extremely useful review articles of Denteneer and van Haeringen³, Pickett⁴, Payne et al.⁵ and Marx and Hutter⁶.

2 Why Using Plane Waves?

2.1 Supercells

Although we have simplified the complicated many–body problem of interacting electrons in the Coulomb potentials of fixed nuclei to a set of single–particle equations, the calculation of the one–electron wave functions for an extended (or even infinite) system is still a formidable task. To make the problem tractable we assume that our system of interest can be represented by a box of atoms which is repeated periodically in all three spacial directions. The box shall be described by three vectors $\mathbf{a}_1$, $\mathbf{a}_2$, and $\mathbf{a}_3$. The volume of the

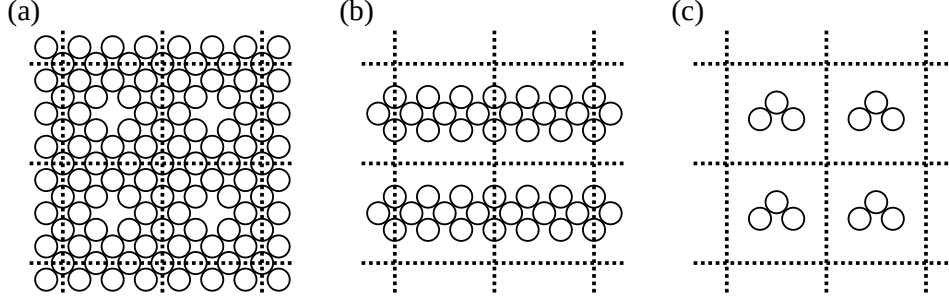


Figure 1. Schematic illustration of a supercell geometry (a) for a vacancy in a bulk crystalline solid, (b) for a surface, and (c) for an isolated molecule. The boundaries of the supercells are shown by dashed lines.

box is given by

$$\Omega_c = \mathbf{a}_1 \cdot (\mathbf{a}_2 \times \mathbf{a}_3) \quad . \quad (1)$$

The three vectors define a lattice in real space. General lattice vectors $\mathbf{T}$ are multiples of the primitive lattice vectors:

$$\mathbf{T} = N_1 \mathbf{a}_1 + N_2 \mathbf{a}_2 + N_3 \mathbf{a}_3 \quad , \quad (2)$$

where N_1, N_2, N_3 can be any integer number. The box can be, for example, either the primitive unit cell of a crystal or a large supercell containing a sufficient number of independent atoms to mimic locally an amorphous solid or a liquid phase. By using supercells also atomic point defects, surfaces or isolated molecules can be modeled as illustrated schematically in Figure 1. It is essential to make the supercells large enough to prevent the defects, surfaces or molecules in neighboring cells from interacting appreciably with each other. The independence of the configurations can be checked systematically by increasing the volume of the supercell until the computed quantity of interest has converged.

2.2 Fourier Representations

The translational symmetry of the atomic arrangements can now be exploited to reduce the computational cost for solving the Kohn–Sham equations. The effective potential (as well as the electron density) is a periodic function with the periodicity of the lattice, i.e.

$$V_{\text{eff}}(\mathbf{r} + \mathbf{T}) = V_{\text{eff}}(\mathbf{r}) \quad (3)$$

for any lattice vector $\mathbf{T}$ of the form of Eq. (2). Therefore V_{eff} can be expanded into a Fourier series

$$V_{\text{eff}}(\mathbf{r}) = \sum_{\mathbf{G}} V_{\text{eff}}(\mathbf{G}) e^{i\mathbf{G}\mathbf{r}} \quad , \quad V_{\text{eff}}(\mathbf{G}) = \frac{1}{\Omega_c} \int_{\Omega_c} V_{\text{eff}}(\mathbf{r}) e^{-i\mathbf{G}\mathbf{r}} d^3\mathbf{r} \quad . \quad (4)$$

The sum runs over all vectors $\mathbf{G}$ which fulfill the condition $\mathbf{G} \cdot \mathbf{T} = 2\pi M$ for all lattice vectors $\mathbf{T}$ with M being an integer number. The vectors $\mathbf{G}$ form a lattice, the so-called reciprocal lattice, which is generated by the three primitive vectors $\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3$ defined by¹

$$\mathbf{a}_i \cdot \mathbf{b}_j = 2\pi \delta_{ij} \quad , \quad i, j = 1, 2, 3 \quad . \quad (5)$$

The volume of the unit cell of the reciprocal lattice is given by

$$\mathbf{b}_1 \cdot (\mathbf{b}_2 \times \mathbf{b}_3) = \frac{(2\pi)^3}{\Omega_c} \quad . \quad (6)$$

2.3 Bloch's Theorem

The solutions of a single-particle Schrödinger equation with a periodic potential are by no means themselves necessarily periodic. However, the eigenstates can be chosen in such a way that associated with each wave function ψ is a wave vector $\mathbf{k}$ to hold

$$\psi(\mathbf{r} + \mathbf{T}) = e^{i\mathbf{k}\mathbf{T}} \psi(\mathbf{r}) \quad (7)$$

for every lattice vectors $\mathbf{T}$ (Bloch's theorem¹). From now on all eigenstates of the single-particle Schrödinger equation will be labeled with its corresponding vector $\mathbf{k}$. From the form of the exponential factor in Eq. (7) it is obvious that the values of $\mathbf{k}$ can be restricted to within one unit cell of the reciprocal lattice. By convention this unit cell is usually taken to be the first Brillouin zone (BZ)¹. Different solutions to the same vector $\mathbf{k}$ will be labeled with the band index j .

Bloch's theorem is often stated in an alternative form. The property in Eq. (7) is equivalent to the statement that all eigenfunctions $\psi_{\mathbf{k}j}$ of a single-particle Schrödinger equation with a periodic potential can be written as a periodic function $u_{\mathbf{k}j}$ modulated by a plane wave with wave vector $\mathbf{k}$ ¹:

$$\psi_{\mathbf{k}j}(\mathbf{r}) = e^{i\mathbf{k}\mathbf{r}} u_{\mathbf{k}j}(\mathbf{r}) \quad . \quad (8)$$

This allows us to restrict the calculation of the eigenfunctions to within one unit cell. The form of the eigenfunctions in all other unit cells is determined by applying Eq. (7). From now on we will assume that the eigenfunctions are normalized with respect to a single unit cell:

$$\int_{\Omega_c} |\psi_{\mathbf{k}j}(\mathbf{r})|^2 d^3\mathbf{r} = 1 \quad . \quad (9)$$

Since the functions $u_{\mathbf{k}j}$ are periodic they can be expanded in a set of plane waves. Together with the exponential prefactor we get:

$$\psi_{\mathbf{k}j}(\mathbf{r}) = \sum_{\mathbf{G}} c_{\mathbf{G}}^{\mathbf{k}j} e^{i(\mathbf{k}+\mathbf{G})\mathbf{r}} \quad . \quad (10)$$

Before we make use of the plane wave expansion of the wave functions we write the Kohn-Sham equations of density functional theory in the notation of Bloch-states:

$$\left(-\frac{\hbar^2}{2m} \Delta + V_{\text{eff}}(\mathbf{r}) \right) \psi_{\mathbf{k}j}(\mathbf{r}) = \varepsilon_{\mathbf{k}j} \psi_{\mathbf{k}j}(\mathbf{r}) \quad (11)$$

with

$$V_{\text{eff}}(\mathbf{r}) = V_{\text{ext}}(\mathbf{r}) + V_{\text{H}}[n(\mathbf{r})] + V_{\text{xc}}[n(\mathbf{r})] \quad (12)$$

and

$$n(\mathbf{r}) = 2 \frac{\Omega_c}{(2\pi)^3} \sum_j \int_{\text{BZ}} |\psi_{\mathbf{k}j}(\mathbf{r})|^2 \Theta(E_{\text{F}} - \varepsilon_{\mathbf{k}j}) d^3\mathbf{k} \quad . \quad (13)$$

V_{ext} , V_{H} and V_{xc} are the external potential of the nuclei, the Hartree and the exchange–correlation potential, respectively. By the factor of 2 in Eq. (13) we take the electron spin into account. Θ is a step function which is one for positive and zero for negative arguments. E_{F} is the Fermi energy up to which single–particle states have to be occupied. The Fermi energy is defined by the number of electrons N_e in the unit cell:

$$\int_{\Omega_c} n(\mathbf{r}) d^3\mathbf{r} = N_e \quad . \quad (14)$$

For an insulator the Fermi energy lies in a band gap. Hence, at each $\mathbf{k}$ –point exactly $N_e/2$ bands will be occupied. For metals one or more bands cross the Fermi energy so that the number of occupied states will change between $\mathbf{k}$ –points.

2.4 $\mathbf{k}$ –Point Sampling

By making use of Bloch’s theorem we have transformed the problem of calculating an infinite number of electronic states extended infinitely in space to one of calculating a finite number of eigenstates at an infinite number of $\mathbf{k}$ –points which are extended over a single unit cell. At first glance this seems to be only a minor improvement since still an infinite number of calculations are needed for the different $\mathbf{k}$ –points. However, the electronic wave functions at $\mathbf{k}$ –points which are close together will be very similar. Hence it is possible to represent the wave functions of a region of $\mathbf{k}$ –space by the wave function at a single $\mathbf{k}$ –point. We thus define a regular mesh of N_{kpt} $\mathbf{k}$ –points and replace the integral over the Brillouin zone by a discrete sum over the chosen $\mathbf{k}$ –point mesh:

$$\frac{\Omega_c}{(2\pi)^3} \int_{\text{BZ}} \dots \Theta(E_{\text{F}} - \varepsilon_{\mathbf{k}j}) d^3\mathbf{k} \quad \longrightarrow \quad \frac{1}{N_{\text{kpt}}} \sum_{\mathbf{k}} f_{\mathbf{k}j} \dots \quad (15)$$

The $f_{\mathbf{k}j}$ are occupation numbers which are either one or zero. Several schemes to construct such $\mathbf{k}$ –point meshes have been proposed in the literature^{7–9}. Within this approximation the electronic states at only a finite number of $\mathbf{k}$ –points are needed to calculate the charge density and hence the total energy of the solid. The error induced by this approximation can be reduced systematically by increasing the density of the $\mathbf{k}$ –point mesh. For insulators it turns out that usually only a small number of $\mathbf{k}$ –points is required to get good converged results. For increasing size of the supercell the volume of the Brillouin zone becomes smaller and smaller (see Eq. (6)). Therefore, with increasing supercell size less and less $\mathbf{k}$ –points are needed. From a certain supercell size on it is often justified to use just a single $\mathbf{k}$ –point, which is usually taken to be $\mathbf{k}=0$ (Γ –point approximation). For metallic systems, on the other hand, much denser $\mathbf{k}$ –point meshes are required in order to get a precise sampling of the Fermi surface. In these cases the convergence with respect to the $\mathbf{k}$ –point density can often be accelerated by introducing fractional occupation numbers^{10–13}.

2.5 Jellium Model

Why is it reasonable to expand the electronic wave functions using a basis set of plane waves? Let us first consider the very simple example of the jellium model (also called the Sommerfeld theory of metals¹). In this model only the valence electrons are considered and the charges of the remaining ionic cores are assumed to be spread out into a uniform

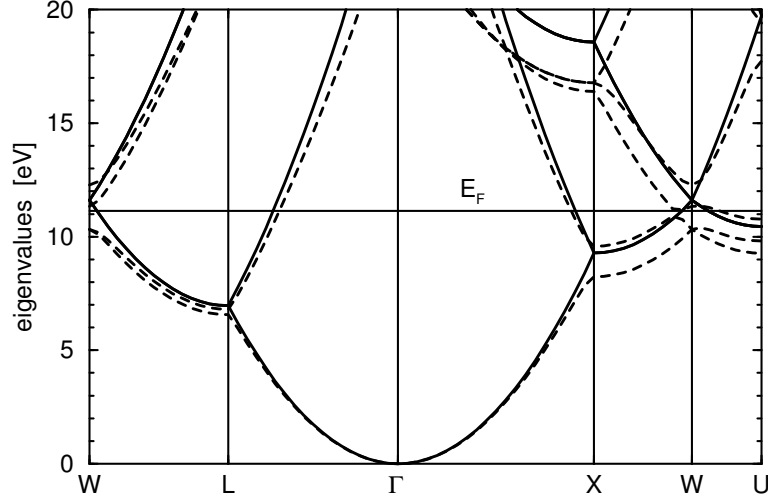


Figure 2. Band structure for a jellium model assuming the valence electron density and the cubic face-centered lattice of the “nearly-free electron” metal Al (solid line). The corresponding band structure of a full DFT calculation is shown by dashed lines. The energy levels are plotted along some high symmetry lines in the first Brillouin zone. The horizontal line indicates the Fermi energy.

positive background distribution. The external effective potential becomes constant and is set to zero. For the jellium model the Kohn–Sham equations can be solved easily. The eigenstates of the single-particle Schrödinger equation

$$-\frac{\hbar^2}{2m}\Delta\psi_{\mathbf{k}j}^{\text{jel}}(\mathbf{r}) = \varepsilon_{\mathbf{k}j}^{\text{jel}}\psi_{\mathbf{k}j}^{\text{jel}}(\mathbf{r}) \quad (16)$$

are plane waves to a single reciprocal lattice vector $\mathbf{G}_j$ with the eigenvalues describing a simple quadratic dispersion relation:

$$\psi_{\mathbf{k}j}^{\text{jel}}(\mathbf{r}) = \frac{1}{\sqrt{\Omega_c}} e^{i(\mathbf{k}+\mathbf{G}_j)\mathbf{r}} \quad , \quad \varepsilon_{\mathbf{k}j}^{\text{jel}} = \frac{\hbar^2}{2m} \|\mathbf{k} + \mathbf{G}_j\|^2 \quad . \quad (17)$$

The labeling of the reciprocal lattice vectors $\mathbf{G}_j$ is done in such a way that the eigenvalues are in increasing order with increasing band index j . With these solutions the electron density becomes constant, and the Fermi energy is given by:

$$n(\mathbf{r}) = \frac{N_e}{\Omega_c} \quad , \quad E_F = \frac{\hbar^2}{2m} \left(3\pi^2 \frac{N_e}{\Omega_c} \right)^{2/3} \quad . \quad (18)$$

In Figure 2 the band structure of the jellium model is shown for a cubic face-centered lattice (fcc) assuming the unit cell volume and the number of valence electrons of Al. The folding of the free-electron parabola into the Brillouin zone of the fcc lattice already creates a surprisingly complex band structure. Note that some of the energy bands are highly degenerate since sometimes several different reciprocal lattice vectors lead to the same eigenvalue.

2.6 Nearly Free Electrons

If we assume that the true external potential of the ionic cores of the atoms represents only a weak perturbation to the jellium model, then, taking the arguments from perturbation theory, it would be quite natural to describe the corresponding wave functions by a linear combination of a few plane waves, and the band structure should deviate not too much from the jellium result. Thinking of the Coulomb-nature of the ionic core potentials this seems to be not a very realistic assumption. However, comparing the true band structure of Al from a full DFT calculation with the free electron bands in Figure 2 reveals a surprising agreement. Some of the degeneracies of the free electron bands are lifted, but overall there is a rather close match between the two calculations.

How can this be understood? Though the eigenvalues are very similar, the corresponding eigenfunctions are quite different. In the jellium calculation they are nodeless single plane waves. In the full DFT calculation, on the other hand, the valence wave functions are composed of the $3s$ and $3p$ Al valence orbitals with nodes and fast oscillations in the ionic core region. These nodes and oscillations are due to the requirement that the valence wave functions are orthogonal to the core state. Effectively this leads to a repulsion of the $3s$ and $3p$ electrons from the core region (which can be viewed as a Pauli repulsion) so that overall the ionic core of the Al atom indeed behaves more like a weak perturbation for the valence electrons. However, the nodes and oscillations of the valence wave functions are restricted to the small ionic core region. Outside they match rather well the plane waves of the jellium calculation. Hence it should be possible to replace the ionic core potential by a much weaker potential that reproduces exactly the true wave functions outside the core region to the same eigenvalues, but with nodeless wave functions inside the ionic core. This is the original idea of introducing pseudopotentials. In the first applications pseudopotentials were constructed by explicitly projecting out the core electron contributions (so-called Phillips–Kleinman pseudopotentials¹⁴) or were purely empiric¹⁵. Today pseudopotentials are constructed from ab initio calculations for isolated atoms. We will come back to this point in the next section where we introduce pseudopotentials from a slightly different point of view. For the moment we will just assume that we can always introduce an appropriate pseudopotential so that the corresponding pseudo wave functions can be represented by a small set of plane waves.

2.7 Fourier Representation of the Kohn–Sham Equations

In a plane wave representation of the wave functions the Kohn–Sham equations assume a particular simple form. If we insert Eq. (10) into Eq. (11), multiply from left with $\exp(-i(\mathbf{k} + \mathbf{G}')\mathbf{r})$ and integrate over $\mathbf{r}$ we get the matrix eigenvalue equation

$$\sum_{\mathbf{G}} \left(\frac{\hbar^2}{2m} \|\mathbf{k} + \mathbf{G}\|^2 \delta_{\mathbf{G}'\mathbf{G}} + V_{\text{eff}}(\mathbf{G}' - \mathbf{G}) \right) c_{\mathbf{G}}^{\mathbf{k}j} = \varepsilon_{\mathbf{k}j} c_{\mathbf{G}'}^{\mathbf{k}j} \quad . \quad (19)$$

In practical calculations the Fourier expansion (Eq. (10)) of the wave functions is truncated by keeping only those plane wave vectors $(\mathbf{k} + \mathbf{G})$ with a kinetic energy lower than a given cutoff value E_{pw} :

$$\frac{\hbar^2}{2m} \|\mathbf{k} + \mathbf{G}\|^2 \leq E_{\text{pw}} \quad . \quad (20)$$

The convergence of all calculations with respect to the basis set size can be tested simply by increasing step by step the plane wave cutoff energy.

The electron density in Fourier representation is given by

$$n(\mathbf{G}) = \frac{2}{N_{\text{kpt}}} \sum_{\mathbf{k}j} f_{\mathbf{k}j} \sum_{\mathbf{G}'} \left(c_{\mathbf{G}'-\mathbf{G}}^{\mathbf{k}j} \right)^* c_{\mathbf{G}'}^{\mathbf{k}j} . \quad (21)$$

Since we have truncated the wave functions at a maximum wave vector it is obvious from Eq. (21) that the electron density has only non-vanishing Fourier components up to twice the length of this cutoff wave vector. In Fourier space the calculation of the Hartree potential is particularly simple. It is given by

$$V_{\text{H}}(\mathbf{G}) = 4\pi e^2 \frac{n(\mathbf{G})}{\|\mathbf{G}\|^2} . \quad (22)$$

As the electron density, the Hartree potential has a finite Fourier expansion. To calculate the exchange–correlation potential we have to Fourier transform the electron density to real–space, evaluate the given functional and Fourier transform back the result.

2.8 Fast Fourier Transformation (FFT)

The main advantage of working with plane waves is that the evaluation of various expressions can be speeded up significantly by using FFTs. In particular, since the wave functions and the electron density have a finite Fourier representation this can be done without any loss in accuracy, as long as we use in our real–space Fourier grid twice as many grid points in each spacial direction than the number of points in the Fourier space grid¹⁶. For example, the calculation of the electron density according to Eq. (21) scales quadratically with the number N_{pw} of plane waves. However, if we Fourier transform the wave functions to real–space (which scales with $N_{\text{pw}} \ln(N_{\text{pw}})$), calculate $|\psi_{\mathbf{k}j}(\mathbf{r})|^2$ on the real–space Fourier grid (N_{pw} scaling) and then Fourier transform back the result we significantly reduce the computational cost. Along the same arguments we can also reduce the number of calculations for the evaluation of the term $\sum_{\mathbf{G}} V_{\text{eff}}(\mathbf{G}' - \mathbf{G}) c_{\mathbf{G}}^{\mathbf{k}j}$ in Eq. (19) from a N_{pw}^2 to a $N_{\text{pw}} \ln(N_{\text{pw}})$ scaling.

3 Pseudopotentials

3.1 Frozen–Core Approximation

Most physical and chemical properties of crystals depend to a very good approximation only on the distribution of the valence electrons. The core electrons do not participate in the chemical bond. They are strongly localized around the nucleus, and their wave functions overlap only very little with the core electron wave functions from neighboring atoms. Therefore, the distribution of the core electrons basically does not change when the atoms are placed in a different chemical environment. It is thus justified to assume the core electrons to be “frozen” and to keep the core electron distribution of the isolated atom in the crystal environment.

The first advantage of the frozen–core approximation is that now less electrons have to be treated and less eigenstates of the Kohn–Sham equations have to be calculated. The second advantage is that the total energy scale is largely reduced when the core electrons are

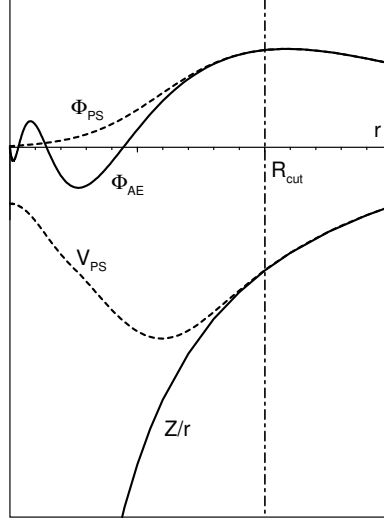


Figure 3. Schematic illustration of an atomic all-electron wave function (solid line) and the corresponding atomic pseudo wave function (dashed line) together with the respective external Coulomb potential and pseudopotential.

removed from the calculation which makes the calculation of energy differences between atomic configurations numerically much more stable.

In principle one might just take the distribution of the core electrons and combine their Hartree potential with the Coulomb potential of the nucleus to an ionic core potential. However, this is not very useful since the valence wave functions still have to maintain their nodal structure in order to be orthogonal to the core states. Much more practical is to replace immediately the ionic core potential by a pseudopotential which will lead to nodeless valence wave functions, as we will show in the following.

3.2 Normconserving Pseudopotentials

Present day pseudopotentials are constructed from ab initio calculations for isolated atoms. Let us assume we have solved the Kohn–Sham equations for a single atom of the chemical species for which we would like to generate a pseudopotential. This can be done easily since due to the spherical symmetry of atoms the wave functions can be written as a product of a radial function and a spherical harmonic. The Schrödinger equation then reduces to one-dimensional differential equations for the radial functions which can be integrated numerically. A typical result for a radial function from such an “all-electron” atom calculation together with the corresponding external Coulomb potential is shown in Figure 3. Our aim is now to replace the effective all-electron potential within a given sphere with radius R_{cut} by a much weaker new potential with a nodeless ground state wave function to the same energy eigenvalue as the original all-electron state which matches exactly the all-electron wave function outside R_{cut} (depicted with dashed lines in Figure 3).

Why should this be possible at all? This can be understood by the following line of arguments. The radial Schrödinger equation for a fixed potential and fixed energy ε (not

necessarily an eigenvalue) to the angular momentum l is a one-dimensional ordinary linear second order differential equation which has two linear independent solutions. However, only one of the two solutions, $\Phi_l(r)$, is regular for $r \rightarrow 0$. The logarithmic derivative

$$L_l(\varepsilon) = \left. \frac{d}{dr} \ln \Phi_l(r; \varepsilon) \right|_{R_{\text{cut}}} = \frac{\Phi'_l(R_{\text{cut}}; \varepsilon)}{\Phi_l(R_{\text{cut}}; \varepsilon)} \quad (23)$$

is therefore a well defined function of the energy ε . On the other hand, for a given energy ε and logarithmic derivative L_l at R_{cut} the solution of the radial Schrödinger equation inside and outside the sphere is uniquely defined (save a constant factor). This follows directly from the properties of one-dimensional second order differential equations. However, the solution is only regular for $r \rightarrow 0$ if the energy ε and the logarithmic derivative L_l fulfill the relation Eq. (23). From this observation we can conclude that if we modify the potential inside the atomic sphere in such a way that the relation $L_l(\varepsilon)$ is not changed, the wave functions outside the sphere remain unchanged.

For the energy of the eigenstate of our all-electron calculation $\varepsilon_l^{\text{AE}}$ this can be done in the following way: we replace the all-electron wave function Φ_l^{AE} inside the sphere by an arbitrary smooth nodeless function Φ_l^{PS} with the same logarithmic derivative at R_{cut} as the original all-electron function. Since the Φ_l^{PS} is nodeless we can simply invert the radial Schrödinger equation with this new function and with the eigenvalue $\varepsilon_l^{\text{AE}}$ of the all-electron calculation to get the potential that has exactly the required property! In fact, we have a quite large extend of freedom to setup the new pseudo wave function Φ_l^{PS} , and over the last decades many different recipes have been published how it could be done¹⁷⁻²⁴. One further important requirement is the so-called normconserving condition. The all-electron and the pseudo wave function inside the atomic sphere must have the same norm to guarantee that both wave functions generate identical electron densities in the outside region. Next to this condition, the additional degrees of freedom in generating a suitable pseudopotential can be employed to make the pseudo wave functions as smooth as possible²⁵.

Up to now we have reproduced the logarithmic derivative of the effective all-electron potential only for the reference energy $\varepsilon_l^{\text{AE}}$. However, if we change the chemical environment of our atom, the eigenstates will be at a slightly different energy. Therefore, for a pseudopotential to be useful it has to be able to reproduce the logarithmic derivative of the all-electron potential over a whole energy range. The wider this energy range the more “transferable” to other chemical environments is the pseudopotential. As it has been shown, in particular the normconserving condition guarantees such a transferability. Furthermore, the pseudopotential should be as “soft” as possible. By this we mean that the number of plane waves required to expand the pseudo wave functions should be as small as possible. Both properties, transferability and softness, are closely related to the cutoff radius R_{cut} and compete with each other. Low cutoffs give pseudopotentials with a very good transferability. However, increasing R_{cut} makes the pseudopotentials softer. Usually one has to find a compromise between the two requirements. An upper limit for R_{cut} is given by half the distance to the next nearest atom in the configuration for which we want to apply the pseudopotential. If we exceed this value there won’t be any region between the neighboring atoms left where we recover the true all-electron wave functions. Hence, we can not expect anymore to get an accurate description of the chemical bond between the two atoms.

3.3 Fully Nonlocal Pseudopotentials

Since the logarithmic derivative in Eq. (23) depends on the angular momentum l we have to construct a separate pseudopotential $V_l^{\text{PS}}(r)$ for each value of l . The full pseudopotential for our atom therefore has to be a nonlocal operator. This is done in the following way:

$$\hat{V}^{\text{PS}} = V_{\text{loc}}^{\text{PS}}(r) + \sum_l V_{\text{nl},l}^{\text{PS}}(r) \hat{P}_l \quad , \quad V_{\text{nl},l}^{\text{PS}}(r) = V_l^{\text{PS}}(r) - V_{\text{loc}}^{\text{PS}}(r) \quad . \quad (24)$$

The pseudopotential $V_l^{\text{PS}}(r)$ to one specific angular momentum (usually the highest value of l for which a pseudopotential has been generated) is taken to be the so-called local part of the pseudopotential $V_{\text{loc}}^{\text{PS}}(r)$. The nonlocal components $V_{\text{nl},l}^{\text{PS}}(r)$ are defined as the differences between the original l -dependent $V_l^{\text{PS}}(r)$ and this local part of the pseudopotential. Since all $V_l^{\text{PS}}(r)$ are identical outside of R_{cut} the nonlocal components of the pseudopotential are strictly confined within R_{cut} . $\hat{P}_l$ is a projection operator which picks out the l -th angular momentum component from the subsequent wave function. By this construction it is guaranteed that when the full pseudopotential operator $\hat{V}^{\text{PS}}$ is applied to a general wave function each angular momentum component of the wave function experiences only its corresponding part $V_l^{\text{PS}}(r)$ of the potential.

Since the projection operators $\hat{P}_l$ act only on the angular variables of the position vector $\mathbf{r}$ the pseudopotential $\hat{V}^{\text{PS}}$ is still a local operator with respect to the radius r . The form (Eq. (24)) is therefore called a semilocal pseudopotential. For numerical efficiency, however, it would be desirable to have the pseudopotential in a fully nonlocal form:

$$\hat{V}^{\text{PS}} = V_{\text{loc}}^{\text{PS}}(r) + \sum_{ij} |\beta_i > B_{ij} < \beta_j| \quad . \quad (25)$$

The $\beta_i(\mathbf{r})$ are suitably chosen projection function which are strictly localized within R_{cut} . Kleinman and Bylander²⁶ have given a prescription how a semilocal potential of the form (Eq. (24)) can be transformed into a fully nonlocal representation. As has been shown by Vanderbilt²⁷ it is also possible to construct directly from an atomic all-electron calculation a fully nonlocal potential. Basically all present plane-wave-based total energy codes employ pseudopotentials in the form of Eq. (25).

3.4 Vanderbilt Ultrasoft Pseudopotentials

Very difficult to treat within a pseudopotential scheme are all elements with nodeless valence states (in particular those with $2p$ and $3d$ valence electrons). For those elements the pseudo and the all-electron wave functions are almost identical. Since these valence electrons are strongly localized in the ionic core region, many plane waves are required for a good representation of their wave function which often makes calculations for such elements prohibitively expensive.

To circumvent this problem Vanderbilt has introduced a new type of pseudopotentials, the so-called ultrasoft pseudopotentials, in which the normconserving requirement has been relaxed^{27,28}. Instead of representing the full valence wave function by plane waves, only a small portion of the wave function is calculated within the Vanderbilt ultrasoft pseudopotential scheme (see dashed line in Figure 4). This allows to reduce substantially the plane wave cutoff energy in the calculations. The price to pay, however, is that the Fourier

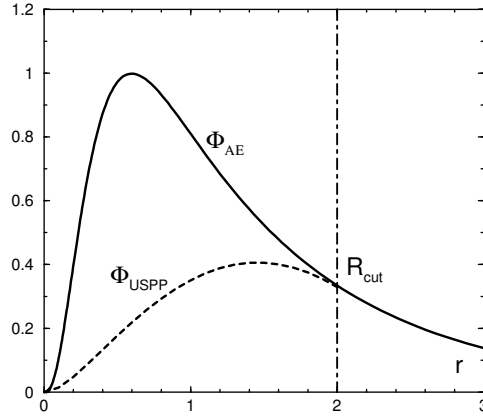


Figure 4. Illustration of a strongly localized valence wave function inside the atomic core region and the modified wave function in the Vanderbilt ultrasoft pseudopotential scheme.

representation of the Kohn–Sham equation becomes more complicated. First, when the electron density is calculated we have to add back the part of the electron distribution which is represented by the difference between the solid and the dashed line in Figure 4 (the so-called augmentation charges). Second, due to the relaxation of the normconserving condition, the Bloch eigenstates $\psi_{\mathbf{k}j}$ will be not orthonormal anymore. An overlap matrix has to be introduced and the eigenvalue problem (Eq. (19)) will transform into a generalized eigenvalue equation. Third, the nonlocal part of the pseudopotential becomes density-dependent. Fourth, due to these modification additional terms in the force calculation have to be evaluated. However, the gain in computational cost by lowering the plane wave cutoff energy outweigh in many cases the additional computational effort which is required by these modifications.

References

1. N.W Ashcroft and N.D. Mermin, *Solid State Physics*, (Saunders College Publishing, Philadelphia, 1976).
2. C. Kittel, *Introduction to Solid State Physics*, (Wiley & Sons, New York, 1976).
3. P.J.H. Denteneer and W. van Haeringen, *J. Phys. C: Solid State Phys.* **18**, 4127 (1985).
4. W. Pickett, *Comput. Phys. Rep.* **9**, 115 (1989).
5. M.C. Payne, M.P. Teter, D.C. Allan, T.A. Arias, and J.D. Joannopoulos, *Rev. Mod. Phys.* **64**, 1045 (1992).
6. D. Marx and J. Hutter, in *Modern Methods and Algorithms of Quantum Chemistry*, ed. J. Grotendorst, (NIC, FZ Jülich, 2000), see <http://www.theochem.rub.de/go/cprev.html>.
7. D.J. Chadi and M.L. Cohen, *Phys. Rev. B* **8**, 5747 (1973).
8. H.J. Monkhorst and J.D. Pack, *Phys. Rev. B* **13**, 5188 (1976).
9. J. Moreno and J.M. Soler, *Phys. Rev. B* **45**, 13891 (1992).
10. C.-L. Fu and K.-M. Ho, *Phys. Rev. B* **28**, 5480 (1983).

11. M.J. Gillan, J. Phys.: Condens. Matter **1**, 689 (1989).
12. M. Methfessel and A.T. Paxton, Phys. Rev. B **40**, 3616 (1989).
13. P.E. Blöchl, O. Jepsen, and O.K. Andersen, Phys. Rev. B **49**, 16223 (1994).
14. J.C. Phillips and L. Kleinman, Phys. Rev. **116**, 287 (1959).
15. M.L. Cohen and V. Heine, *Solid State Physics, Vol. 24*, (Academic, New York, 1970).
16. W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery, *Numerical Recipes*, (Cambridge University Press, 1992).
17. D.R. Hamann, M. Schlüter, and C. Chiang, Phys. Rev. Lett. **43**, 1494 (1979).
18. G.B. Bachelet, D.R. Hamann, and M. Schlüter, Phys. Rev. B **26**, 4199 (1982).
19. G. Kerker, J. Phys. C **13**, L189 (1980).
20. D. Vanderbilt, Phys. Rev. B **32**, 8412 (1985).
21. N. Troullier and J.L. Martins, Solid State Comm. **74**, 613 (1990).
22. N. Troullier and J.L. Martins, Phys. Rev. B **43**, 1993 (1991).
23. S. Goedecker, M. Teter, and J. Hutter, Phys. Rev. B **54**, 1703 (1996).
24. C. Hartwigsen, S. Goedecker, and J. Hutter, Phys. Rev. B **58**, 3641 (1998).
25. A.M. Rappe, K.M. Rabe, E. Kaxiras, and J.D. Joannopoulos, Phys. Rev. B **41**, 1227 (1990).
26. L. Kleinman and D.M. Bylander, Phys. Rev. Lett. **48**, 1425 (1982).
27. D. Vanderbilt, Phys. Rev. B **41**, 7892 (1990).
28. K. Laasonen, A. Pasquarello, R. Car, C. Lee, and D. Vanderbilt, Phys. Rev. B **47**, 10142 (1993).

Full-Potential Linearized Augmented Planewave Method

Stefan Blügel and Gustav Bihlmayer

Institute for Solid State Research
Forschungszentrum Jülich
52425 Jülich, Germany
E-mail: {s.bluegel, g.bihlmayer}@fz-juelich.de

The full-potential linearized augmented planewave (FLAPW) method has emerged as a widely used very robust and precise state-of-the-art *ab initio* electronic structure technique with reasonable computational efficiency to simulate the electronic properties of materials on the basis of density functional theory (DFT). Due to the high precision it is widely accepted that it provides the density functional answer to the problem. The shape of the charge density, the one-electron potential and the wavefunction is taken into account with high accuracy. The FLAPW method is an all-electron algorithm which is universally applicable to all atoms of the periodic table, in particular to transition metals and rare-earths and to multi-atomic systems with compact as well as open structures. Due to the all-electron nature of the method, magnetism is included rigorously and nuclear quantities e.g. isomer shift, hyperfine field, electric field gradient (EFG), and core level shift are calculated routinely. Also open structures such as surfaces, clusters, organic and inorganic molecules as well as wires can be treated without problems. The capability of calculating atomic forces exerted on the atoms opens the path to structure optimization. In this chapter, a short introduction to first-principles methodology is given, the FLAPW method is reviewed, modern extensions of the LAPW basis set are discussed, extensions of the method to geometries suitable to applications in nanoscience such as the film and wire geometries are presented. Details for the practical users e.g. important parameters controlling the accuracy of the results and an analysis of the CPU time is given for the FLEUR-code, a particular implementation of the FLAPW method.

1 Introduction

Societal requests for environment prediction and protection, the durability of chemicals, the vision of new applications in information technology such as autonomous robots, bio-diagnostic systems, or faster information processing, as well as demands on the sustainable and efficient use of resources and energies translate in a huge demand on understanding, predicting, modeling and simulating the properties, chemical reactions, synthesis and growth processes of emergent quantum materials. Modern solid state materials have a multiplicity of novel properties exhibiting for example a rapid phase response to external stimuli such as light, pressure, magnetic field or electrical conductivity so that manifold uses are possible even today or can be expected for the future. Materials of this sort are often multicomponent systems such as magnetic tunneljunctions (e.g. NiMnSb|MgO|Co₂MnSn), high-temperature superconductors (e.g. HgBa₂Ca₂Cu₂O₈), or perovskite-type materials with complex magnetic structures. A particularly rich arsenal of assets for material design and tailoring of material properties is provided when the surface of materials is provided as templates for fabrication. Nanostructures down to the atomic scale made of single atoms or of small molecules can be manufactured to form chains and clusters or structures with specific electronic properties by employing the tip of a scanning tunneling microscope (STM) or relying on the instruments of self-assembly. Nanostructured thin film systems are decisive functional units in electronic devices, sensors and in biological systems. The

existence of particular surface and interface alloys and the complex interplay between morphological, structural, magnetic and electronic features in nanostructured systems stand as examples for a wide field of phenomena which are largely not understood, while offering exceptional technological opportunities at the same time.

During the past ten years, first-principles calculations based on the density functional theory (DFT)¹ in the local (spin-) density approximation (LDA) or in the generalized gradient approximation (GGA) (for a review see Refs. 2–4) emerged as the most powerful framework to respond to the demands mentioned above on a microscopic level. By first-principles or *ab initio*, respectively, is meant, that the parameters of the theory are fixed by the basic assumptions and equations of quantum mechanics and, for our discussion, density functional theory. The overwhelming success of the density functional theory for the description of the ground-state properties of large material classes including insulators, semiconductors, semimetals, half-metals, simple metals, transition-metals and rare-earths in bulk, at surfaces and as nanostructures makes it the unchallenged foundation of any modern electronic structure theory. The wide applicability combined with the predictive power of the approach turned it to the “standard model” in material science. In principle, the only input needed for the theory are the atomic numbers of the constituent atoms of a system, all other properties follow as a direct consequence of the density functional equations. In practice, the definition has to be modified since one is always limited to some set of model systems. These limitations might include system size, crystal structure, neglect of disorder, low or zero temperature, or any number of other restrictions on the “phase space” to probe. While some of these restrictions and limitations are burdensome, the goal of calculations is not merely to obtain numbers, but rather insight. By focusing on well-defined, but restricted models, by working on chemical trends rather than on isolated case studies, by investigating systems in hypothetical non-equilibrium structures or follow simulations in idealized environments, which may not be realized in experiments, one is able to develop different levels of understanding of the system in question and may hopefully learn which aspects of the problem are important. In the density functional theory we work in an effective one-particle picture: the wavefunctions are solutions of the Kohn-Sham equations and the interaction of the particles is taken into account by a self-consistent field, which depends on the density of the particles. The adequacy and limitations of this approach have been extensively discussed in the literature and pioneered by^{5,6}. Successful extensions of the theoretical framework treat quasiparticle excitations of weakly correlated electron systems by many-body perturbation theory based on Green function techniques in the so-called *GW* approximation for the electronic self-energy (for details see chapter “Many-Body Perturbation Theory: The *GW* Approximation” by C. Friedrich and A. Schindlmayr) or by time-dependent density functional theory (TDDFT)^{7,8}. The treatment of strongly correlated electron systems is currently being explored investigating the LDA+U approximation, the optimized effective potential method (OEP)⁹ and by combining the dynamical mean field theory (DMFT)^{10,11} with the LDA.

In this chapter we review the full-potential linearized augmented planewave (FLAPW) method^{12,13} to solve the density functional equations for a crystalline solid, ultrathin film and one-dimensional system (a review is given by D. J. Singh¹⁴). The method originates from the APW method proposed by Slater^{15–17}. Great progress of the APW methodology was achieved as the concept of linear methods^{18–22}, was introduced by Andersen, and first applied by Koelling and Arbman using a model potential within the muffin-tin approxi-

mation. The linearized APW (LAPW) method reconciled the linear-algebra formulation of the variational problem with the convergence properties of the original formulation and allowed a straight forward extension of the method to the treatment of crystal potentials of general shape. The treatment of the potential and charge density without shape approximation^{23,24} and the implementation of the total energy¹³ led to the development of FLAPW bulk^{12,24-30} film codes^{12,30-32}. It was during this time that the power and accuracy of the method were demonstrated to the community, largely through a series of calculations of surface and adsorbate electronic structures (for a review see Wimmer *et al.*³³). These and other demonstrations established the FLAPW method as the method of choice for accurate electronic structure calculations for a broad spectrum of applications.

Constant conceptual and technical developments and refinements such as the proposal and implementation of the scalar-relativistic approximation (SRA)³⁴, the spin-orbit interaction by second variation³⁶, and the possibility to calculate forces^{37,38} acting on the ions to carry out structure optimizations, Quasi-Newton methods³⁹ to accelerate the self-consistency iterations, the iterative diagonalization techniques⁴⁰⁻⁴², the proposal of new efficient basis sets, the LAPW+LO⁴³ and APW+lo⁴⁴ basis, in which the APW basis is amended by local orbitals (lo), the extension of the method to non-collinear magnetism⁴⁵, to the wire geometry⁴⁶, to calculations of the quasiparticle self-energy in the *GW* approximation⁴⁷, and the recent formulation and application of the scattering problem in semi-infinite crystals⁴⁸⁻⁵⁰ has made APW-like methods, and for our discussion the FLAPW method, a robust, versatile and flexible method, at reasonable computational expense. It is an all-electron method, that means, one works with a true crystal potential, which diverges as $1/r$ at the nucleus, as opposed to the pseudo-potential (for a review see Refs. 51, 52), in which the singularity is removed. Due to the all-electron nature of the method, magnetism is included rigorously and nuclear quantities⁵³ e.g. isomer shift, hyperfine field, electric field gradient (EFG), and core level shift are calculated routinely. The method and the breadth of applications has benefited from the large growth of available computing power and parallelization strategies.

This chapter starts with a quick overview to the Kohn-Sham ansatz outlining the general aspects of the first-principles methodology followed by an introduction of the APW-like concepts to solve the Kohn-Sham equation for a periodic solid. Then, the FLAPW method is described in detail for bulk solids followed by a short description of adaptations of the FLAPW methods to systems of reduced dimensions, the film and wire geometry. Finally an analysis of the CPU time distribution across the different steps of an *ab initio* calculation with an FLAPW program is presented. Strategies to speed-up the calculations in the light of symmetry and parallelization concepts are described.

2 Kohn-Sham Approach in a Nutshell

2.1 Total Energy and Force

In the density functional theory, the total energy $E[\{\mathbf{R}\}, \{\psi_i\}]$ of a system of interacting atoms and electrons is a functional of the atomic positions $\{\mathbf{R}\}$ and the electron density $n(\mathbf{r})$. The electron density can be expressed in terms of M occupied single-particle orbitals

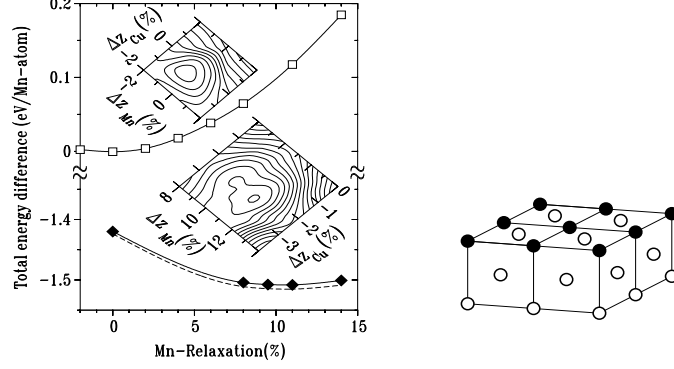


Figure 1. Example: structural optimization of Mn and Cu surface atoms in a Cu(100)c(2 × 2)Mn surface alloy. Right figure: Schematic representation of the substitutional surface alloy film of one monolayer thickness (● indicates the Mn atoms) grown as overlayer on a fcc (001) substrate (○). Left figure: Total energy per Mn atom vs. the buckling relaxation Δz_{Mn} of Mn in relative units with respect to the theoretical interlayer spacing of Cu, $d_{Cu} = 1.76$ Å. The open squares represent the nonmagnetic and the solid diamonds the ferromagnetic results. The solid lines (for Cu atoms fixed at the ideally terminated positions $\Delta z_{Cu} = 0$) and dashed line (the top Cu atom is always at its optimally relaxed position) are the fitting polynomials. The upper (lower) inset shows the contour plot of the nonmagnetic (ferromagnetic) total energy with respect to the buckling of Mn and Cu. The minimum, which determines the optimal structure is found in the inner circle. The contour interval is 1 meV. The energy of the nonmagnetic solution at 0% relaxation was chosen as the origin of the total energy scale (taken from Ref. 54).

$\psi_i(\mathbf{r})$:

$$n(\mathbf{r}) = \sum_{i(occ)}^M |\psi_i(\mathbf{r})|^2, \quad (1)$$

where i labels the states. If the total energy functional $E[\{\mathbf{R}\}, \{\psi_i\}]$ is minimized with respect to the electronic degrees of freedom $\{\psi_i\}$, we recover the Born-Oppenheimer surface $\Phi[\{\mathbf{R}\}]$:

$$\Phi[\{\mathbf{R}\}] = \min_{\{\psi_i\}} E[\{\mathbf{R}\}, \{\psi_i\}], \quad (2)$$

on which the atoms move. The derivative of $\Phi[\{\mathbf{R}\}]$ with respect to the atomic position $\mathbf{R}^\mu$ gives the force $\mathbf{F}^\mu$,

$$\mathbf{F}^\mu = -\nabla_{\mathbf{R}^\mu} \Phi[\{\mathbf{R}\}] \quad (3)$$

exerted on the atom μ , which ties electronic structure to structural optimization and molecular dynamics calculations. The energy functional is divided into several terms:

$$E[\{\mathbf{R}\}, \{\psi_i\}] = E_{kin}[\{\psi_i\}] + E_H[\{\psi_i\}] + E_{xc}[\{\psi_i\}] + E_{ext}[\{\mathbf{R}\}, \{\psi_i\}] + E_{ion}[\{\mathbf{R}\}], \quad (4)$$

where E_{kin} is the kinetic energy of non-interacting electrons, E_H is the Hartree energy, i.e. the classical Coulomb energy of the electrons, and E_{xc} is the exchange-correlation energy which contains terms coming from the Pauli principle (exchange hole), from correlations due to the repulsive Coulombic electron-electron interaction and from the contribution to the kinetic energy of interacting electrons³. E.g. In the local density approximation $E_{xc}[n]$

is written in the form $E_{\text{xc}}[n] = \int d\mathbf{r} n(\mathbf{r}) \varepsilon_{\text{xc}}(n(\mathbf{r}))$. Then, E_{ext} is the interaction energy of the electrons with the ions, e.g. described by the $1/r$ potential as in all-electron methods or by pseudo-potentials, and E_{ion} is the classical Coulomb energy of the ions.

2.2 The Kohn-Sham Equations

The single-particle wavefunctions $\psi_i(\mathbf{r})$ are obtained by minimization of the total energy with respect to the wavefunctions subject to the normalization constraint

$$\int d\mathbf{r} |\psi_i(\mathbf{r})|^2 = 1. \quad (5)$$

This leads to the Kohn-Sham equations⁵⁵, an eigenvalue problem for the eigenfunctions $\psi_i(\mathbf{r})$ and the eigenvalues ε_i :

$$\hat{H}[n] \psi_i[n] = \varepsilon_i[n] \psi_i[n], \quad (6)$$

where all quantities depend on the electron density n . According to the form of the total energy Eq. (4), the Hamiltonian $\hat{H}$ is a sum of corresponding terms and the eigenvalue problem is written in the form:

$$(\hat{T}_0 + \hat{V}_{\text{ext}} + \hat{V}_H + \hat{V}_{\text{xc}}) \psi_i(\mathbf{r}) = \varepsilon_i \psi_i(\mathbf{r}) \quad (7)$$

In the real space representation the individual terms are the following:

$$\text{kinetic energy :} \quad \hat{T}_0 = -\frac{\hbar^2}{2m} \Delta_{\mathbf{r}} \quad (8)$$

$$\text{external-potential :} \quad V_{\text{ext}}(\{\mathbf{R}\}, \mathbf{r}) = \sum_{\mu} \frac{e^2 Z^{\mu}}{|\mathbf{r} - \mathbf{R}^{\mu}|} \quad (9)$$

$$\text{Hartree potential :} \quad \Delta_{\mathbf{r}} V_H(\mathbf{r}) = 4\pi e^2 n(\mathbf{r}) \quad (10)$$

$$\text{xc-potential (LDA) :} \quad V_{\text{xc}}(\mathbf{r}) = \frac{\delta}{\delta n(\mathbf{r})} \int d\mathbf{r} n(\mathbf{r}) \varepsilon_{\text{xc}}(n(\mathbf{r})) \quad (11)$$

In a pseudo-potential approach $\hat{V}_{\text{ext}}$ is replaced for each atom μ by a pseudo-potential $\hat{V}_{\text{ps}}$. The terms $\hat{V}_H[n]$ and $\hat{V}_{\text{xc}}[n]$ are local potentials and explicitly density dependent. Thus, the Hamiltonian $\hat{H}[n]$ and the wavefunctions $\psi_i([n], \mathbf{r})$ are also dependent on the electron density $n(\mathbf{r})$. Together with the expression Eq. (1) a self-consistency problem to obtain the charge density $n(\mathbf{r})$ is established, which is solved iteratively until the input density (used to define the potential terms in the Hamiltonian) is equal to the output density within the required accuracy. The number of self-consistency iterations N_{iter} is considerably reduced applying Quasi-Newton methods⁵⁶.

The external potential $\hat{V}_{\text{ext}}[\{\mathbf{R}\}]$ depends explicitly on the positions $\{\mathbf{R}\}$ of all atoms, which change at certain steps to optimize the atomic structure or every time-step of a molecular dynamics algorithm. Thus, the Hamiltonian $\hat{H}[\{\mathbf{R}\}]$ and the wavefunctions $\psi_i(\{\mathbf{R}\}, r)$ are also dependent on the atomic positions $\{\mathbf{R}\}$. After the self-consistency condition for the electron density has been fulfilled, the atom positions are moved by a molecular static or molecular dynamics time-step, $\{\mathbf{R}(t)\} \rightarrow \{\mathbf{R}(t + \Delta t)\}$. Thus, for N_{MD} molecular time steps the eigenvalue problem has to be solved $N_{\text{MD}} N_{\text{iter}}$ times. These arguments suggest a particular loop structure of a typical first-principles method

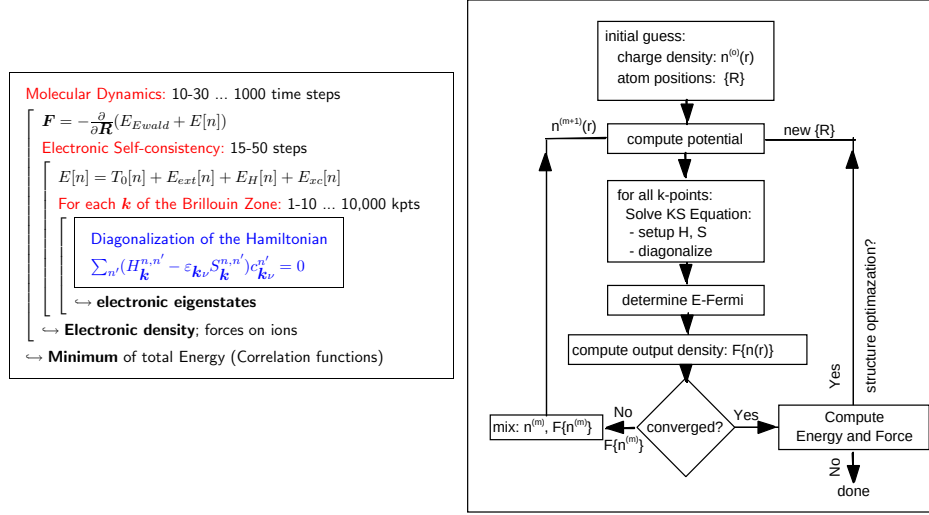


Figure 2. Right: Typical loop structure of a first-principles code based on density functional theory as applied to solid state materials. Left: Schematic flow-chart for self-consistent density functional calculations e.g. as realized by a FLAPW calculation.

and a particular sequence how the different elements are calculated. This is summarized in Figure 2.

Typical codes use LDA exchange-correlation potentials and energies of Hedin and Lundqvist⁵⁷ or Vosko, Wilk, and Nusair⁵⁸, or GGA functionals of Perdew *et al.*^{59,60} are given as analytical expressions of the density and their derivatives in case of the GGA.

2.3 Magnetism

If magnetism occurs, the ground state has a broken symmetry and the ground-state energy is described by functionals which depend on the vector-magnetization density $\mathbf{m}(\mathbf{r})$ as an additional field to the ordinary charge density $n(\mathbf{r})$, discussed so far. An additional term $\mu_B \underline{\sigma} \cdot \mathbf{B}_{xc}(\mathbf{r})$ appears in the Kohn-Sham equations Eq. (7), where $\mu_B = \frac{e\hbar}{2mc}$ is the Bohr magneton, $\mathbf{B}_{xc}$ is the magnetic xc-field an electron experiences, and $\underline{\sigma}$ are the Pauli spinors. Thus, calculating magnetic systems, one works in a two-dimensional spin-space and the basis functions $\psi_{i\sigma}$ carry an additional spin label $\sigma = \pm 1$. The Hamiltonian is a 2×2 matrix in spin-space and is now hermitian and not symmetric. Complex magnetic structures lower frequently the symmetry of the problem and more states have to be calculated or a much larger fraction of the BZ (see Section 2.6) has to be sampled, respectively, pushing the computational effort to the limits of modern supercomputers. In case of collinear magnetism, e.g. ferro-, ferri-, or antiferromagnetism, $\underline{\sigma} \cdot \mathbf{B}_{xc}$ reduces to $\underline{\sigma}_z \cdot B_{xc}$, the Hamiltonian is diagonal in spin space, the magnetization density m_z is then given by spin-up and -down densities, $m_z(\mathbf{r}) = n_{\uparrow}(\mathbf{r}) - n_{\downarrow}(\mathbf{r})$, and the effort of a magnetic calculation is just twice that of a nonmagnetic one. In general, the magnetic moment $\mathbf{M} = \int d\mathbf{r} \mathbf{m}(\mathbf{r})$ is a vector quantity, and the search of the magnetic structure can be done dynamically bearing similarities to the dynamical structure optimization combining

molecular dynamics and simulated annealing. Therefore, everything said in this chapter on structural optimization applies to both, the atomic and the magnetic structure. Throughout the paper, the spin label is dropped for convenience. More information on the treatment of magnetism can be found in the chapter “Non-collinear magnetism: exchange parameter and T_C ” by G. Bihlmayer.

2.4 The Eigenvalue Problem

In all-electron methods, eigenvalue problem Eq. (7) is solved for all occupied states i but typically subject to different boundary conditions. As shown schematically in Figure 3 we distinguish core electrons from valence electrons. The former have eigenenergies which are at least a couple of Rydbergs below the Fermi energy, the potential they experience is to an excellent approximation spherically symmetric and the wavefunctions have no overlap to neighboring atoms. The eigenvalue problem of these states is solved applying the boundary conditions of isolated atoms, which is numerically tackled by a shooting method. Valence electrons in a crystalline solid form electron bands and the eigenvalue problem is solved subject to the Bloch boundary conditions. The eigenstate is classified by the band index ν and a three-dimensional Bloch vector $\mathbf{k}$ within the first Brillouin zone, ($i \in \{\mathbf{k}\nu\}$). Some materials contain chemical elements with states (e.g. $5p$ states of $4f$ elements or W , p states of early transition metals) intermediate between band and core states and those are coined semicore states. These are high-lying and extended core states and particular care has to be taken on their treatment since their treatment as core states can cause significant errors in total energy, force and phonon calculations. According to the different treatment of the electrons, we decompose the charge density in the valence, semicore and core densities

$$n(\mathbf{r}) = n_{\text{val}}(\mathbf{r}) + n_{\text{sc}}(\mathbf{r}) + n_{\text{core}}(r), \quad (12)$$

the latter being spherically symmetric. The charge densities are calculated according to Eq. (1). Wavefunctions and energies of core states give access to hyperfine quantities such as isomer shifts, hyperfine fields and electric field gradient as well as chemical shifts of core levels.

There are many possible ways to solve the Kohn-Sham equations for valence electrons. Frequently, a variational method is chosen by which a wavefunction $\psi_{\mathbf{k}\nu}(\mathbf{r})$ of Bloch vector $\mathbf{k}$ and band index ν is sought as a linear combination of basis functions $\varphi_n(\mathbf{k}, \mathbf{r})$

$$\psi_{\mathbf{k}\nu}(\mathbf{r}) = \sum_{n=1}^N c_{\mathbf{k}\nu}^n \varphi_n(\mathbf{k}, \mathbf{r}) \quad (13)$$

satisfying the Bloch boundary conditions. $c_{\mathbf{k}\nu}^n$ are the expansion coefficients of the wavefunction (coefficient vector), and N is the number of basis functions taken into account. By this expansion, the eigenvalue problem

$$\hat{H}\psi_{\mathbf{k}\nu}(\mathbf{r}) = \varepsilon_{\mathbf{k}\nu}\psi_{\mathbf{k}\nu}(\mathbf{r}) \quad (14)$$

is translated into an algebraic eigenvalue problem of dimension N

$$(\mathbf{H}(\mathbf{k}) - \varepsilon_{\mathbf{k}\nu}\mathbf{S}(\mathbf{k}))\mathbf{c}_{\mathbf{k}\nu} = 0 \quad \forall \mathbf{k} \in \text{BZ} \quad (15)$$

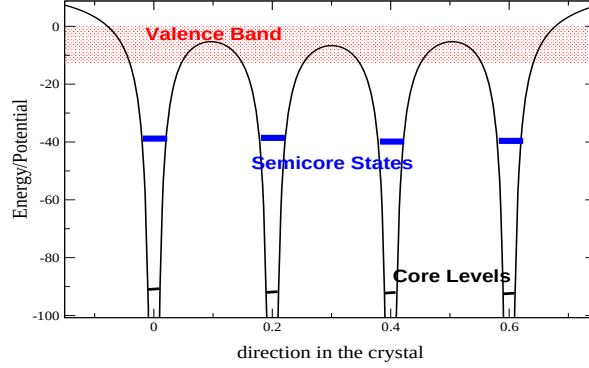


Figure 3. Schematic representation of the energy position of valence, semicore and core electrons in periodic potential.

for the coefficient vector $c_{\mathbf{k}\nu}^n$ corresponding to the eigenvalues $\varepsilon_{\mathbf{k}\nu}$. The Hamilton $H^{n,n'}(\mathbf{k})$ and overlap matrices $S^{n,n'}(\mathbf{k})$ are hermitian or real symmetric, depending on the point symmetry of the atomic structure. If the basis functions are orthonormal, i.e. $\langle \varphi_n | \varphi_{n'} \rangle = \delta^{n,n'}$, as for example in case of simple planewaves, then the overlap matrix S , defined as

$$S^{n,n'}(\mathbf{k}) = \int_{\Omega} \varphi_n^*(\mathbf{k}, \mathbf{r}) \varphi_{n'}(\mathbf{k}, \mathbf{r}) d^3r \quad (16)$$

becomes diagonal, $S^{n,n'}(\mathbf{k}) = \delta^{n,n'}$, and the generalized eigenvalue problem Eq. (15) becomes of standard type. Ω is the volume of the unit cell.

In general, the general eigenvalue problem is reduced to a standard one using the Cholesky decomposition. It can be shown (e.g. Stoer⁶¹), that any hermitian and positive definite matrix can be decomposed into a matrix product of a lower triangular matrix with only positive diagonal elements and its transposed. Clearly, the overlap matrix satisfies these conditions and can be written $S = LL^{tr}$. Therefore, Eq. (15) becomes

$$Hc_i = \varepsilon_i LL^{tr} c_i, \quad (17)$$

multiplying from the left with L^{-1} and introducing a unit matrix we finally find

$$P \mathbf{x}_i = \varepsilon_i \mathbf{x}_i, \quad (18)$$

after we have P defined as $P = L^{-1}H(L^{-1})^{tr}$ and $\mathbf{x}_i = L^{tr} c_i$. Thus, the generalized eigenvalue problem has been reduced to a simple one. The eigenvectors c_i can be obtained by the back-transformation, $c_i = (L^{tr})^{-1} \mathbf{x}_i$.

The choice of the most efficient numerical algorithm to solve Eq. (15) depends on the number of basis functions N and the number M of states ν taken into account. If $M/N > \sim 0.1$, direct numerical diagonalization schemes are employed, for example parallelized eigenvalue solvers taken from the ScaLAPACK library package. If $M/N < \sim 0.1$ or if N is too large to fit the eigenvalue problem into the memory of a computer the eigenvalue problem is solved iteratively. Any iterative solution of an eigenvalue problem can be divided into two parts: (i) the determination of the iterative improvement of the state vector

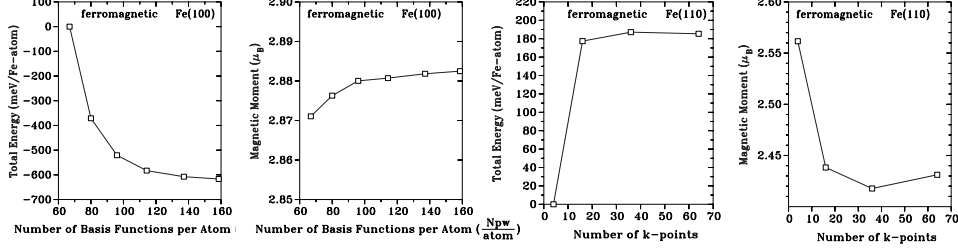


Figure 4. Test of convergence carried out by the FLAPW method of (absolute) total energy and magnetic moment as function (i) of the number of the LAPW basis functions (see two left figures) for a 7 layer Fe(100) film and (ii) number of special $\mathbf{k}$ -points in the IBZ (see two right figures) for an 11 layer Fe(110) film. The calculations of (i) were carried out for the $rk m$ -parameters $rk m = 7.5, 8.0, 8.5, 9.0, 9.5, 10.0$ corresponding to $N = 67, 80, 96, 114, 137, 158$ basis functions.

$c_{\mathbf{k}\nu}^{n,[m]}$ at iteration step m by multiplying the Hamiltonian with the state vector to obtain the update $c_{\mathbf{k}\nu}^{n,[m+1]}$:

$$c_{\mathbf{k}\nu}^{n,[m+1]} = \sum_{n'} H^{n,n'}(\mathbf{k}) c_{\mathbf{k}\nu}^{n',[m]}, \quad (19)$$

and (ii) the orthonormalization of the wavefunctions

$$\sum_n c_{\mathbf{k}\nu}^{n,[m+1]} c_{\mathbf{k}\nu'}^{n,[m+1]} = \delta_{\nu,\nu'}. \quad (20)$$

(iii) Frequently, each iteration step is accompanied by a direct sub-space diagonalization of a dimension proportional to M , on which Hamiltonian $\hat{H}$ is projected. If the multiplication of $\mathbf{H} \cdot \mathbf{c}$ can be made fast by expressing the Hamiltonian in terms of dyadic products or convolutions as in norm-conserving or ultra-soft pseudo-potentials minimizing thereby the number of multiplications, iterative methods become particular beneficial.

2.5 The CPU Time Requirement

The number of basis functions N is determined by the required precision P of a calculation and by the volume Ω of the unit cell or the number of atoms in the unit cell, N_A , respectively. The precision P is controlled by the finest real-space resolution the basis functions can resolve. For three-dimensional unit cells N scales as $N \propto P^3$. In general, the triple $(N_{\mathbf{k}}, M, N)$, the number of $\mathbf{k}$ -vectors in the BZ used, the number M of states ν considered, and the number of basis functions N are determined by the required precision of the calculation and by the volume of the unit cell. These parameters determine the CPU time and memory requirements of the calculations. Keeping the loop-structure in mind exhibited in Figure 2, typically the calculational CPU time scales as

$$\text{CPU} \propto N_{\text{MD}} \cdot N_{\text{iter}} \cdot N_{\mathbf{k}} \begin{cases} N^3 & \text{direct diagonalization} \\ M_{\text{iter}}(MN \ln N + NM^2) & \text{iterative diagonalization} \end{cases} \quad (21)$$

where M_{iter} gives the number of eigenvalue iterations. This gives just a gross estimate as for iterative methods based on the Car-Parrinello idea where self-consistency iterations

and eigenvalue iterations can be combined to directly minimize total energy functional $N_{\text{MD}} \cdot N_{\text{iter}} \cdot M_{\text{iter}}$ depends on many details. The scaling relation for precision scaling is:

$$\text{the number of } \mathbf{k}\text{-points: } N_{\mathbf{k}} \propto P_{\mathbf{k}}^3 \quad (22)$$

$$\text{the number of basis functions } N: N \propto P^3, \quad (23)$$

where $P_{\mathbf{k}}$ is the precision controlling the $\mathbf{k}$ -point summation, e.g. of the force, the total energy or the electron density. Assuming that the volume Ω of the unit cell is proportional to the number of atoms N_A , the scaling relation for volume scaling is:

$$\text{the number of } \mathbf{k}\text{-points: } N_{\mathbf{k}} \propto 1/N_A, \quad (24)$$

$$\text{the number of states } \nu: M \propto N_A, \quad (25)$$

$$\text{the number of basis functions } N: N \propto N_A. \quad (26)$$

From these considerations it is argued to develop electronic structure methods (see Section 2.8) with efficient basis sets to reduce their number N , to develop algorithms to accelerate the convergence (see Section 2.7) and to employ an efficient $\mathbf{k}$ -point integration scheme (see Section 2.6).

2.6 Brillouin Zone Integration and Fermi Energy

The calculation of the electron density, total energy, force or stress tensor for infinite periodic solids require the integration of functions over the Brillouin zone that depend on the Bloch vector and the energy band. These integrations stretch only over the occupied part of the band, i.e. over the region of the Brillouin zone where the band energy $\epsilon_{\nu}(\mathbf{k})$ (ν is the band index) is lower than the Fermi energy. Hence, the integrals are of the form

$$\frac{1}{V_{\text{BZ}}} \int_{\text{BZ}} \sum_{\nu, \epsilon_{\nu}(\mathbf{k}) < E_F} f_{\nu}(\mathbf{k}) d^3k, \quad (27)$$

where f is the function to be integrated, e.g. $f = 1$ for the total number of electrons, $f = \epsilon$ for the eigenvalue sum and so on. Numerically, these integrations are performed on a discrete mesh in the Brillouin zone. In fact the effort of the BZ integration is in practice significantly reduced by employing the point group symmetry, where the integration is reduced to the irreducible wedge of the BZ (IBZ). There are different methods that can be used to perform the integration, e.g. the special points method^{62,63} and the tetrahedron method⁶⁴⁻⁶⁶. The special points method is a method to integrate smoothly varying periodic functions of $\mathbf{k}$. The function to be integrated has to be calculated at a set of special points in the IBZ, each of which is assigned a weight. Thus, the BZ integration is transformed into a sum over a set of $\mathbf{k}$ -points. At each $\mathbf{k}$ -point a sharp energy cutoff is introduced to include only those states in the summation whose energy is below the Fermi energy. Thus, the integrals become:

$$\frac{1}{V_{\text{BZ}}} \int_{\text{BZ}} \sum_{\nu, \epsilon_{\nu}(\mathbf{k}) < E_F} f_{\nu}(\mathbf{k}) d^3k \longrightarrow \sum_{\mathbf{k} \in \text{IBZ}} \sum_{\nu, \epsilon_{\nu}(\mathbf{k}) < E_F} f_{\nu}(\mathbf{k}) w(\mathbf{k}). \quad (28)$$

Alternatively, this integration can be viewed as an integration over the whole Brillouin zone, where the function to be integrated is given by a product of the function f with a step function that cuts out the region of the Brillouin zone, where the band energy is

above the Fermi energy. Clearly, the resulting function does not satisfy the condition of being smoothly varying. Therefore, the special k-points method does not converge very quickly, and rather many k-points are needed to obtain accurate results. On the other hand this method is simple to implement, because the weights depend only on $\mathbf{k}$ and the band energy (via the step function) at each k-point. Another problem arises from this “sharp” differentiation between occupied and empty bands (parts of bands). Let’s consider a band that is very close to the Fermi energy at a certain k-point. During the iterations the energy of this band might rise above or drop below the Fermi energy. This leads to sudden changes in the charge density, which can slow down or even prevent the convergence of the density. These sudden changes are clearly a result of the discretization in momentum space. To avoid this problem, the sharp edges of the step function are smoothened, e.g. by introducing a so-called temperature broadening in the context of a the Fermi function $(e^{(\epsilon - E_F)/k_B T} + 1)^{-1}$ rather than the step function. The temperature T or energy $T k_B$ are additional external parameters adjusted to obtain the best convergence.

2.7 Achieving Self-Consistency

According to Section 2.2 the Kohn-Sham equation Eq. (7) are Schrödinger-like independent-particle equations which must be solved subject to the condition that the effective potential field $V_{\text{eff}}(\mathbf{r}) = V_{\text{ext}}(\mathbf{r}) + V_H(\mathbf{r}) + V_{\text{xc}}(\mathbf{r})$ and the density field $n(\mathbf{r})$ are consistent. The electron density $n_0(\mathbf{r})$ that minimizes the energy functional is a fix-point of the mapping

$$n'(\mathbf{r}) = F\{n(\mathbf{r})\}, \quad (29)$$

i.e. it solves

$$\mathcal{F}\{n_0(\mathbf{r})\} = 0, \quad \text{with } \mathcal{F}\{n(\mathbf{r})\} = F\{n(\mathbf{r})\} - n(\mathbf{r}). \quad (30)$$

(The same can be formulated for the potential.) Typically, the density is expanded into a large set of basis functions. Thus, in actual calculations, the charge density is a coefficient vector of dimension $N_Q \sim 8 * N$ (N defined as in Eq. (13) and Eq. (30) constitutes a system of N_Q nonlinear equations, which can be solved by iteration:

$$n^{m+1}(\mathbf{r}) = F\{n^m(\mathbf{r})\}. \quad (31)$$

A starting density $n^{(0)}(\mathbf{r})$ can be constructed by a superposition of atomic densities. A straight mapping as is suggested in Eq. (31) is in general divergent. Convergence can be achieved if the output density $F\{n^m(\mathbf{r})\}$ is mixed with the input density $n^m(\mathbf{r})$.

The simplest and slowest of such mixing schemes is the so-called “simple mixing”, which converges only linearly. The density for the next iteration is constructed as a linear combination of $n^{(m)}$ and $F\{n^m\}$ according to:

$$n^{(m+1)} = (1 - \alpha)n^{(m)} + \alpha F\{n^m\} = n^{(m)} + \alpha \mathcal{F}\{n^{(m)}\}, \quad (32)$$

where α is the so-called mixing parameter. If it is chosen small enough, the iteration converges and is very stable. However, for the type of systems one is interested in, α is very small, requiring many hundreds of iterations. In spin-polarized calculations different mixing parameters can be used for the charge and the magnetization density. Usually, the spin mixing parameter can be chosen far larger than the parameter for the charge density.

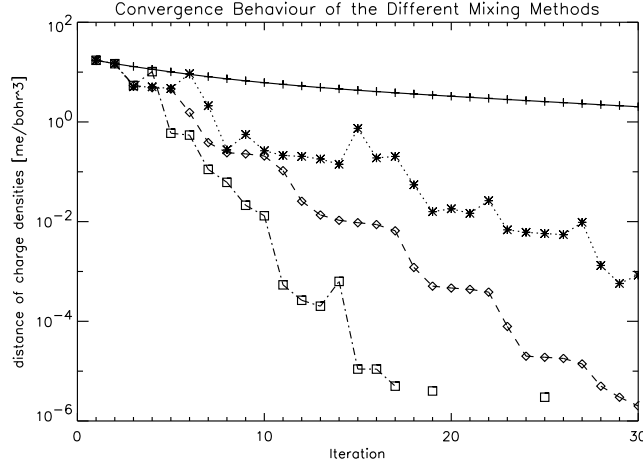


Figure 5. Comparison of the convergence of charge density calculated by different methods for a non-magnetic bcc Fe crystal using the FLAPW method. Calculations are carried out for mixing parameter $\alpha = 0.04$. + corresponds to simple mixing, and *, $\diamond$, and $\square$ to different Quasi-Newton methods: * Broyden's 1st method, $\diamond$ Broyden's 2nd method, $\square$ generalized Anderson method. The distance of the residual vector vs. number of iterations is plotted semi-logarithmically⁶⁷.

In the Newton-Raphson method, the functional $\mathcal{F}\{n\}$ is linearized around the approximate solution $n^{(m)}$.

$$\mathcal{F}\{n\} \approx \mathcal{F}\{n^{(m)}\} + \mathcal{J}\{n^{(m)}\}(n - n^{(m)}), \quad \mathcal{J}\{n^{(m)}(\mathbf{r})\} = \left. \frac{\partial \mathcal{F}\{n(\mathbf{r})\}}{\partial n(\mathbf{r}')} \right|_{n^{(m)}(\mathbf{r})}. \quad (33)$$

In actual calculations the Jacobian $\mathcal{J}$ is a $N_Q \times N_Q$ matrix. Similar to the well-known Newton method to find zeros of one-dimensional functions, the next approximation to n_0 , $n^{(m+1)}$, is determined from the requirement, that the linearized functional in Eq. (33) vanishes at $n^{(m+1)}$. Thus, $n^{(m+1)}$ is given by:

$$n^{(m+1)} = n^{(m)} - \left[\mathcal{J}\{n^{(m)}\} \right]^{-1} \mathcal{F}\{n^{(m)}\}. \quad (34)$$

In opposite to the simple mixing, the Newton-Raphson method converges quadratically. The major drawback of this method is the difficulty to evaluate the Jacobian. Even if the functional $\mathcal{F}\{n\}$ were known, the evaluation would be cumbersome due to the enormous size of $\mathcal{J}\{n\}$. In addition, the Jacobian has to be inverted where the amount of calculation scales with cube of the dimension. A further problem is that the convergence radius is rather small so that the method can only be used if $n^{(m)}$ is already very close to n_0 .

The development of the Quasi-Newton methods made it possible to exploit the advantages of the Newton-Raphson method, i.e. to make use of the information that is contained in the Jacobian, for problems where the Jacobian cannot be calculated or its determination is too demanding. Rather than computing the Jacobian each iteration, an approximate Jacobian is set up and improved iteration by iteration. From the linearization of $\mathcal{F}\{n\}$ in Eq. (33) we find the following condition for the Jacobian, which is usually called Quasi-

Newton condition:

$$\Delta n^{(m)} = \left[\mathcal{J}^{(m)} \right]^{-1} \Delta \mathcal{F}^{(m)} \quad (35)$$

$$\Delta n^{(m)} = n^{(m)} - n^{(m-1)}, \quad \Delta \mathcal{F}^{(m)} = \mathcal{F}\{n^{(m)}\} - \mathcal{F}\{n^{(m-1)}\}$$

Quasi-Newton methods converge super-linearly and have a larger convergence radius than the Newton-Raphson method. Since the Jacobian is built up iteration by iteration, the “history” of the previous iterations is memorized in $\mathcal{J}$, whereas the Jacobian of the Newton-Raphson method depends only on the previous iteration. In this sense the Newton-Raphson method is self-corrective³⁹, it “forgets” inadequately chosen corrections. The Quasi-Newton methods sometimes need to be restarted, if the iteration converges only slowly. This can happen if the starting density is very far from n_0 or when physical or numerical parameters that affect the calculations are changed during the iteration. Eq. (35) does not determine the Jacobian uniquely, instead Eq. (35) constitutes a system of N_Q equations for N_Q^2 unknowns. The various Quasi-Newton schemes differ by the ansatz how the new information is used to build the inverse Jacobian. The quality of the convergence is measured by the distance of the residual vector:

$$d_{n^{(m)}} = \|\mathcal{F}\{n^{(m)}\}\| = \|\mathcal{F}\{n^{(m)}\} - n^{(m)}\|. \quad (36)$$

2.8 The Electronic Structure Methods

The quest to solve the Kohn-Sham equation (7) efficiently for periodic solids, solids with surfaces and interfaces, clusters and molecules has lead to a wide spectrum of very successful and efficient electronic structure methods. Treating isolated clusters or molecules, methods based on localized orbitals are frequently selected going hand in hand with the chemical intuition of the system in question. Considering methods applicable to periodic solids, frequently algorithms are chosen where the Bloch boundary condition can be included in the basis set. Guiding principles to develop electronic structure methods are obtained by having a closer look at the mathematical nature of the Schrödinger-like Kohn-Sham equation Eq. (7) with the kinetic energy operator Δ and the $1/r$ singularity at the nucleus with the simultaneous necessity to calculate the xc-potential $V_{xc}[n](\mathbf{r})$ and the Hartree potential $V_H[n](\mathbf{r})$.

The planewave basis is obviously a very good choice, as the planewave is diagonal to the Laplace operator Δ appearing in both the kinetic energy operator and the Poisson equation to calculate the Hartree potential (see Eq. (8)), and for a function expanded in planewaves, its power is also completely expressible by a planewave expansion. This property is needed for calculating the charge density from the wave function. Thus, using a planewave basis set, the calculation of the kinetic energy, charge density, and the Hartree potential are obtained by simple algebraic expressions. The calculation of the $V_{xc}(\mathbf{r})$ can be best performed if the charge density is expressed in real-space. The discrete fast Fourier transformation (FFT) provides a fast algorithm to communicate between both spaces. However, planewave basis sets do not converge at the presence of the $1/r$ singularity. Thus, planewave basis sets can only be used in the context of a pseudopotential approximation to the true potential where the $1/r$ potential has been replaced by an appropriate smooth potential.

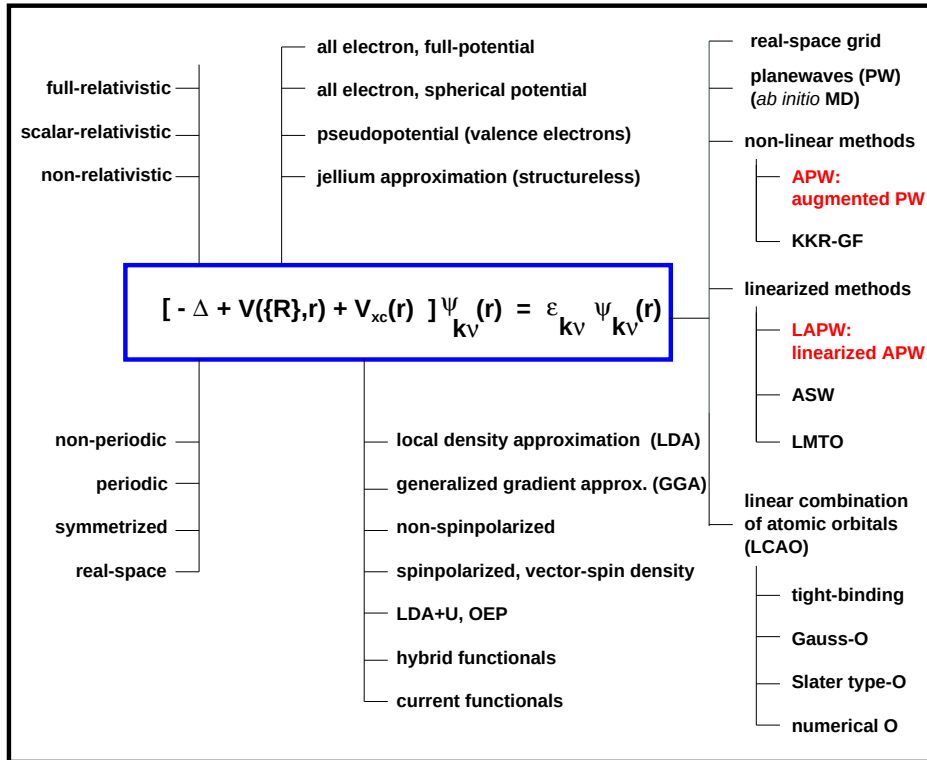


Figure 6. Very rough and schematic overview of electronic structure methods indicating a rich spectrum of methods developed for different purposes, applications, geometries and symmetries, chemical elements and materials requiring different approximations.

All-electron methods have to cope with the $1/r$ singularity. Since this singularity cannot be dealt with variationally, one typically works here with basis functions, which are the numerical solution of $(-\Delta + V_{\text{eff}} - E_l)\varphi = 0$ of the effective (spherical) potential containing the $1/r$ singularity, computed in a sphere around the atom at a given energy parameter E_l . These basis functions treat the singularity exactly. The matching of this wavefunction in such a sphere to the rest of the crystal outside the sphere divides the all-electron methods with regard to the eigenvalue dependence of the basis set into two groups: The nonlinear methods as for example the Korringa-Kohn-Rostocker (KKR) method (see lecture note “The Korringa-Kohn-Rostoker (KKR) Green Function Method” by Ph. Mavropoulos on page 131) and the APW method, and the linear methods, of which the most commonly used are the linear muffin-tin orbital method (LMTO)¹⁹, the augmented spherical⁶⁸ and the APW-based schemes, e.g. FLAPW method.

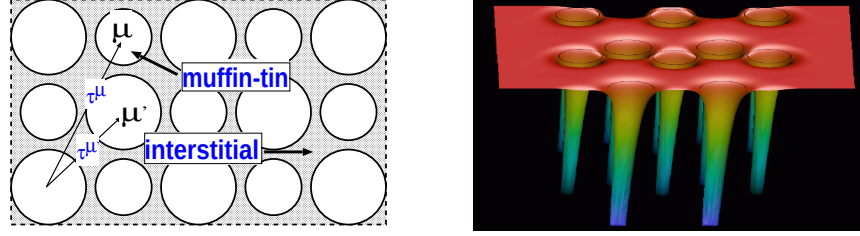


Figure 7. Left figure: Volume of unit cell partitioned into muffin-tin spheres of two different types of atoms and the interstitial region. Right figure: Actual self-consistent effective potential as obtained from an FLAPW calculation.

3 APW-Like Concepts to Solve the Kohn-Sham Equations

3.1 The APW Concept

There are many possible ways to solve the Kohn-Sham equations. Frequently, a variational method is chosen by which a wavefunction $\psi_{\mathbf{k},\nu}(\mathbf{r})$ of Bloch vector $\mathbf{k}$ and band index ν is sought as a linear combination of basis functions $\varphi(\mathbf{r})$ satisfying the Bloch boundary conditions. The most straightforward choice would be to expand the wavefunction into planewaves or Fourier series, respectively,

$$\psi(\mathbf{k}, \nu) = \sum_{|\mathbf{k}+\mathbf{G}| \leq K_{\max}} c_{\mathbf{k},\nu}^{\mathbf{G}} \exp[i(\mathbf{k} + \mathbf{G})\mathbf{r}]. \quad (37)$$

Here $\mathbf{G}$ are all reciprocal lattice vectors up to the largest value of $K_{\max}$ and $c_{\mathbf{k},\nu}^{\mathbf{G}}$ are variational coefficients. The planewave basis set has some important advantages: Planewaves are orthogonal, they are diagonal in momentum space and the implementation of planewave based methods is rather straightforward due to their simplicity. The credit goes to Slater¹⁵ having realized that owing to the singularity of the crystal potential at the nucleus, electron wavefunctions are varying very quickly near it, the planewave expansion would converge very slowly, large wavevectors ($K_{\max}$) would be needed to represent the wavefunctions accurately, which makes the set-up and diagonalization of the Hamiltonian matrix in terms of planewaves impracticable if not impossible. Even with the modern computer hardware, the planewaves are used only in the context of pseudopotential which allow an accurate description of the wavefunctions between the atoms, but avoid the fast oscillations near the core. Thus, less basis functions are needed.

In the APW method the space is partitioned into spheres centered at each atom site, the so-called muffin-tins (MTs), and into the remaining interstitial region (cf. Figure 7). The MT spheres do not overlap and they are typically chosen such that they nearly (to allow for structural relaxations) fill the maximal possible space. Inside the muffin-tins, the potential is approximated to be spherically symmetric, and in many implementations the interstitial potential is set constant. The restrictions to the potential are commonly called shape-approximations. Noting that planewaves solve the Schrödinger equation in a constant potential, Slater suggested to replace the Bessel functions $j_l(Kr)$ in the Rayleigh decomposition of the planewave inside the sphere by radial functions $u_l(K, r)$, which match the Bessel functions in value at the sphere radius R_{MT} and whose product with

the spherical harmonics $Y_L(\hat{\mathbf{r}})$ are the solutions in a spherical potential. It is this procedure what is understood by the term augmentation. Thus, the single wavefunctions $\psi_{\mathbf{k},\nu}(\mathbf{r})$ are expressed as trial functions

$$\psi_{\mathbf{k},\nu}(\mathbf{r}) = \sum_{|\mathbf{G}+\mathbf{k}| \leq K_{\max}} c_{\mathbf{k},\nu}^{\mathbf{G}} \varphi_{\mathbf{G}}(\mathbf{k}, \mathbf{r}) \quad (38)$$

in terms of the APW basis functions:

$$\varphi_{\mathbf{G}}(\mathbf{k}, \mathbf{r}) = \begin{cases} e^{i(\mathbf{k}+\mathbf{G})\mathbf{r}} & \text{interstitial region} \\ \sum_{lm} a_L^{\mu\mathbf{G}}(\mathbf{k}) u_l(r^\mu | E) Y_L(\hat{\mathbf{r}}^\mu) & \text{muffin-tin } \mu \end{cases} \quad (39)$$

The position $\mathbf{r}$ inside the spheres μ located at $\boldsymbol{\tau}^\mu$ (see Figure 7) is given with respect to the center of each sphere. L abbreviates the quantum numbers l and m and u_l is the regular solution of the radial Schrödinger equation

$$\left\{ -\frac{\hbar^2}{2m} \frac{\partial^2}{\partial r^2} + \frac{\hbar^2}{2m} \frac{l(l+1)}{r^2} + V(r) - E \right\} r u_l(r) = 0 \quad (40)$$

to the energy parameter E_l . Here, $V(r)$ is the spherical component of the potential $V(\mathbf{r})$. The coefficients

$$a_L^{\mu\mathbf{G}}(\mathbf{k}) = a_L^{\mu}(\mathbf{k} + \mathbf{G}) = 4\pi \exp(i\mathbf{k}\boldsymbol{\tau}^\mu) i^l Y_L^*(\hat{\mathbf{K}}) \frac{j_l(KR^\mu)}{u_l(R^\mu)}, \quad \mathbf{K} = \mathbf{k} + \mathbf{G} \quad (41)$$

are determined from the requirement, that the wavefunctions are continuous at the boundary of the muffin-tin spheres in order for the kinetic energy to be well-defined. The variational coefficients $c^{\mathbf{G}}$ uniquely determine the wavefunction in the interstitial region.

If E were kept fixed, used only as a parameter during the construction of the basis, the Hamiltonian could be set up in terms of this basis. This would lead to a standard secular equation for the band energies where for a given $\mathbf{k}$ -point in the Brillouin zone (BZ) a set of band energies E_ν are determined. Unfortunately, it turns out, that the APW basis does not offer enough variational freedom if E is kept fixed. An accurate description can only be achieved if the energies are set to the corresponding band energies $E_{\mathbf{k},\nu}$. In this case the Hamiltonian matrix H depends not only on $\mathbf{k}$, $H(\mathbf{k})$, but also on $E_{\mathbf{k},\nu}$, $H(E_{\mathbf{k},\nu})$, and the latter can no longer be determined by a simple diagonalization. Since the u_l 's depend then on the band energies, the solution of the secular equation becomes a nonlinear problem, which is computationally much more demanding than a secular problem. One way of solving this problem is to fix the energy E and scan over $\mathbf{k}$ to find a solution, i.e. find one band at the time, instead of diagonalizing a matrix to find all the bands at a given $\mathbf{k}$. Thus, in Slater's formulation of the method E enters as an additional non-linear variational parameter varying the shape of the functions u_l till the optimal shape is found for the band energies $E_{\mathbf{k},\nu}$ one has looked for. There are several other limitations connected to the APW method. One is rather obvious, when $u_l(R)$ in Eq. (41) becomes zero at the MT boundary, the radial function and the plane wave become decoupled, known as the asymptote problem. Others are beyond the scope of the chapter. Further information about the APW method can be found in the book by Loucks¹⁷, which also reprints several early papers including Slater's original publication¹⁵.

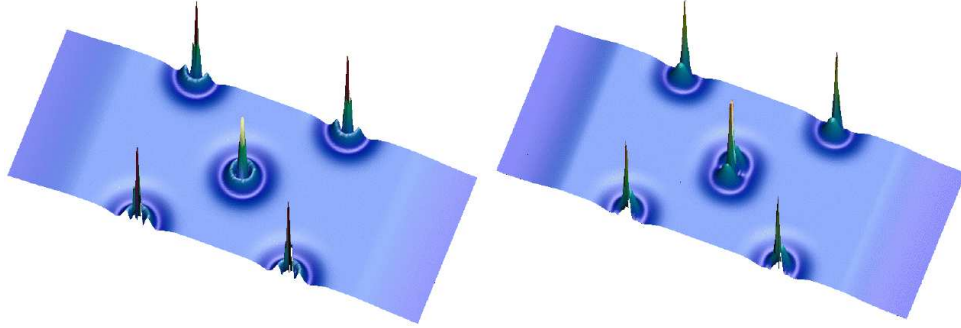


Figure 8. Square of the LAPW basis function generated for $\mathbf{G} = 0$ and $\mathbf{k}$ at the origin ($\bar{\Gamma}$ -point) (left) and boundary ($\bar{M}$ -point) (right) of the Brillouin zone of a 3-layer thin film of Cu(100). The cuts are taken in the $\{110\}$ plane. The basis functions are optimally suited to represent $4s$ states of Cu (left) and $4p$ states (right).

There is one remaining point. Please notice that the APW method produces per construction principle wavefunctions with a discontinuity in the slope at the muffin-tin boundary. Due to these discontinuous first derivatives the secular equation in the APW basis

$$\sum_{\mathbf{G}'} (\langle \varphi_{\mathbf{G}} | \mathbf{H} - \varepsilon_{\mathbf{k}\nu} | \varphi_{\mathbf{G}'} \rangle + \langle \varphi_{\mathbf{G}} | \mathbf{T}_S | \varphi_{\mathbf{G}'} \rangle) c_{\mathbf{k}\nu}^{\mathbf{G}'} = 0 \quad (42)$$

contains a second term due to the matrix elements $\langle \psi | -\nabla^2 | \psi \rangle$ of the kinetic energy operator $\mathbf{T}$ commonly defined as $\mathbf{T} = -\nabla^2$, which is replaced by $\langle \nabla \psi | \nabla \psi \rangle$, leading then via Green's theorem to the appearance of additional surface integrals $T_S \propto \int \psi^* \left[\left(\frac{\partial \psi}{\partial n} \right)_- - \left(\frac{\partial \psi}{\partial n} \right)_+ \right] dS$, where $+$ ($-$) indicates just outside and inside the muffin-tin sphere. The matrix elements of $\mathbf{T}_S$ are proportional to the difference of the logarithmic derivatives from the function u_l , $D(u_l|E) = \frac{u_l'(R)}{u_l(R)}$, and that of an empty sphere $D(j_l|E) = \frac{j_l'(R)}{j_l(R)}$, taken at the sphere boundary. The logarithmic derivatives are related to the phase shifts in scattering events. Thus, the second term in Eq. (42) can be interpreted describing the scattering of a planewave coming from the crystal at the sphere of the atoms. It is well-known that the logarithmic derivatives and the phase shifts are energy dependent quantities, which explains the explicit energy dependence of the APW Hamiltonian in particular, and all nonlinear electronic structure methods in general.

3.2 The LAPW Basis Functions

To avoid the problems connected with the APW method resulting from the energy dependence of the Hamiltonian, in the middle of the seventies linearized methods were invented by Andersen¹⁹ and Koelling and Arbmman²⁰. Based on an idea proposed by Marcus²², the basis functions u_l in the muffin-tins were supplemented by their energy derivatives $\dot{u}_l$, but both, u_l and $\dot{u}_l$, are now evaluated at a fixed energy E_l . The original energy dependence of the radial basis function is thereby replaced by the Taylor series:

$$u_l(E) = u_l(E_l) + (E - E_l)\dot{u}_l(E_l) + \dots \quad (43)$$

terminated after the linear term. In this way, the wavefunctions are affected by an error which is quadratic in the deviation of the eigenvalue E from the energy parameter E_l , the error in the eigenvalues enter only to fourth order²⁰. With this extension, the explicit form of the basis functions is now:

$$\psi_{\mathbf{G}}(\mathbf{k}) = \begin{cases} \exp(i(\mathbf{k} + \mathbf{G})\mathbf{r}) & \text{interstitial} \\ \sum_{l,m} (a_{lm}^{\mu,\mathbf{G}}(\mathbf{k})u_l^\mu(r) + b_{lm}^{\mu,\mathbf{G}}(\mathbf{k})\dot{u}_l^\mu(r))Y_{lm}(\hat{\mathbf{r}}^\mu) & \text{muffin-tin } \mu. \end{cases} \quad (44)$$

Examples of LAPW basis functions are shown in Figure 8. The values of the coefficients $a_{lm}^{\mu,\mathbf{G}}(\mathbf{k})$ and $b_{lm}^{\mu,\mathbf{G}}(\mathbf{k})$ are determined to ensure continuity in value and derivative of the basis functions across the muffin-tin boundary. (A detailed description of these coefficients will follow in Section 4.2.3.) Thereby, also the surface integrals $\int \psi^* \left(\frac{\partial \psi}{\partial n} \right) dS$ which were encountered in the APW method disappear. In this way, the energy dependence of the Hamiltonian is removed, simplifying the eigenvalue problem, Eq. (15), to a standard problem of linear algebra. Instead of working with u_l and $\dot{u}_l$ several LAPW implementations follow the ASW idea, working only with u_l but for two different energy parameters E_l and E'_l . As we see below working with u_l and $\dot{u}_l$ is rather elegant.

If $\hat{H}_{sp}^\mu$ denotes the spherical Hamiltonian in Eq. (40), $\dot{u}$ can be determined from the energy derivative of this equation at E_l :

$$\hat{H}_{sp}^\mu \dot{u}_l^\mu = E_l \dot{u}_l^\mu + u_l^\mu. \quad (45)$$

The normalization of the radial functions is usually chosen like: ^a

$$\int_0^{R^\mu} r^2 u_l^{\mu 2} dr = 1 \quad (46)$$

and the energy derivatives, $\dot{u}_l^\mu$, are orthogonal to the radial functions, i.e.

$$\int_0^{R^\mu} r^2 u_l^\mu \dot{u}_l^\mu dr = 0. \quad (47)$$

This is a relation, which will simplify the calculation of the elements of the Hamilton matrix.

Stimulated by the idea of the LAPW basis set, one may ask to improve the basis set by matching only the 1st derivative continuously, but also higher derivatives working with higher energy derivatives of u_l . This approach has actually been followed by Takeda and Kübler⁶⁹ using n energy parameters to match the wavefunction continuously till the $(n - 1)$ st derivative. However, it turned out that such wavefunctions are variationally very stiff and the convergence of the results with respect of the number of basis functions is rather slow. This can be understood by following this procedure up to the extreme where the wavefunction matches to all derivatives. Then we know, the u_l must be the Bessel function j_l or the planewave, respectively. We have already argued before that this requires an infinite number of planewaves to describe the wavefunction at the $1/r$ singularity. Thus, it is a great merit of the LAPW basis set, that the basis set is linear, but nearly as efficient as

^a In many LAPW codes, the electrons in the muffin-tin are treated in the scalar-relativistic approximation³⁵. This means that a two-component wavefunction is used and the normalization conditions are modified accordingly. For the continuity conditions, only the “large component” of the radial function is taken into account. To keep the formalism as simple as possible, in the following we will discuss only the non-relativistic case.

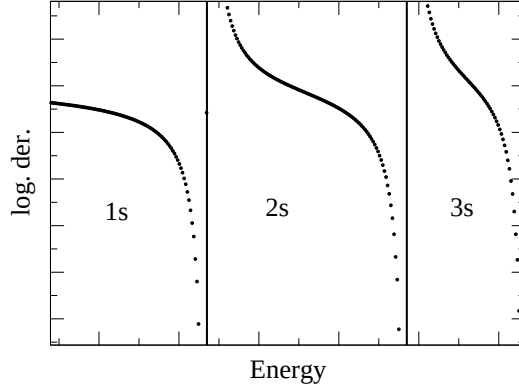


Figure 9. Schematic drawing of the logarithmic derivative, $\frac{u'_l(R)}{u_l(R)}$, for $l = 0$ as a function of the energy. The asymptotes indicate where the nodes of the wavefunction pass through the muffin-tin radius. They separate the branches labeled $1s$, $2s$ and $3s$.

the APW method. The speed of convergence with respect to the number of basis functions can even be improved by the introduction of local orbitals (see Section 3.3).

The energies E_l are chosen to minimize the linearization errors, i.e. in the center of gravity of the l -like bands. It should be noticed here, that the choice of the energy parameter in a certain sense also determines the nodal structure of the wavefunction. A basis function, where the $l = 1$ energy parameter is chosen to describe a $2s$ -like wavefunction in a certain muffin-tin, will not be suitable to describe a $3s$ or a $1s$ state. The energy parameter is then said to be within the $2s$ branch (cf. Figure 9). The flexibility of the basis function of course also depends on the size of the muffin-tin radius, R , so that with the choice of a smaller R in some cases two branches can be forced to “collapse” to a single branch⁷⁰. On the other hand, a smaller flexibility allows to separate core- from valence states in a calculation. Thus, in a typical calculation only high-lying valence states are calculated (e.g. $3s$, $3p$, $3d$), while very localized states (e.g. $1s$, $2s$, $2p$) are excluded from the calculation. These states are then treated in a separate, atomic like, calculation using the $l = 0$ part of the muffin-tin potential.

As a final point, we will address the question how large l should be in a realistic calculation. Since the a and b coefficients in Eq. (44) should ensure continuity across the muffin-tin boundary, the plane-wave cutoff, $G_{\max}$ and the l cutoff, $l_{\max}$, are normally chosen to match: A planewave with wavevector $G_{\max}$ (given in inverse atomic units) has $G_{\max}/\pi$ nodes per atomic unit. A spherical harmonic with $l = l_{\max}$ has $2l_{\max}$ nodes along a great circle on the muffin-tin sphere, i.e. there are $l_{\max}/(\pi R)$ nodes per atomic unit. Therefore, a reasonable choice of the cutoffs is $l_{\max} = RG_{\max}$, typically $l_{\max} = 8$ is chosen.

3.3 Local Orbitals: LAPW+LO and APW+lo

In certain materials high-lying core states, the so-called semicore states, pose a problem to LAPW calculations: they are too delocalized to be described as core electrons (contained entirely in the muffin-tin), but the energy parameter E_l , which would be needed for their

description, is already used to describe higher lying valence states. E.g. the $5p$ levels of La are too high in energy to be neglected in total energy calculations⁴³. Several possible strategies have been proposed to overcome this problem: The above mentioned reduction of the size of the muffin-tin radius⁷⁰, so-called two-window calculations⁷¹, and the use of local orbitals⁴³.

The local orbitals are an extension to the FLAPW basis, that can be used to improve the variational freedom for a specific purpose, e.g. to improve the representation of the semicore states. The extra basis functions are completely localized inside the muffin-tin spheres, i.e. their value and derivative falls to zero at the muffin-tin radius. Thus, no additional boundary conditions have to be satisfied. This can be achieved via a linear combination including three radial functions, the standard FLAPW functions u_l^μ and $\dot{u}_l^\mu$ plus a further radial function u_{lo}^μ . This new radial function is constructed in the same way as u_l^μ , but with a different energy parameter E_{lo}^μ . If the local orbitals are used to treat semicore states, this energy parameter is set to the energy of these states. The local orbitals can be used very specifically, e.g. if they are applied to the $5p$ semicore states of tungsten only local orbitals with p -character are added to the basis.

The three functions u_l^μ , $\dot{u}_l^\mu$ and u_{lo}^μ have to be combined, so that the value and the derivative of the local orbital go to zero at the muffin-tin radius. Additionally, the resulting radial functions can be required to be normalized. Hence, to determine the coefficients of the radial functions $\tilde{a}_{lo}^\mu$, $\tilde{b}_{lo}^\mu$ and $\tilde{c}_{lo}^\mu$ we make use of the following three conditions:

$$\tilde{a}_{lo}^\mu u_l^\mu(R^\mu) + \tilde{b}_{lo}^\mu \dot{u}_l^\mu(R^\mu) + \tilde{c}_{lo}^\mu u_{lo}^\mu(R^\mu) = 0, \quad (48)$$

$$\tilde{a}_{lo}^\mu \frac{\partial u_l^\mu}{\partial r}(R^\mu) + \tilde{b}_{lo}^\mu \frac{\partial \dot{u}_l^\mu}{\partial r}(R^\mu) + \tilde{c}_{lo}^\mu \frac{\partial u_{lo}^\mu}{\partial r}(R^\mu) = 0, \quad (49)$$

$$\int_0^{R^\mu} (\tilde{a}_{lo}^\mu u_l^\mu(R^\mu) + \tilde{b}_{lo}^\mu \dot{u}_l^\mu(R^\mu) + \tilde{c}_{lo}^\mu u_{lo}^\mu(R^\mu))^2 r^2 dr = 1, \quad (50)$$

where lo is the index of the local orbital, which is necessary because more than one local orbital can be added for each atom. The local orbitals are finally coupled to “fictitious” planewaves, $\mathbf{G}_{lo}$, in the same way as the FLAPW basis functions:

$$\varphi_{\mathbf{G}_{lo}}^{\mu, lo}(\mathbf{k}, \mathbf{r}) = \sum_m \left(a_{Lo}^{\mu, \mathbf{G}_{lo}}(\mathbf{k}) u_l^\mu + b_{Lo}^{\mu, \mathbf{G}_{lo}}(\mathbf{k}) \dot{u}_l^\mu + c_{Lo}^{\mu, \mathbf{G}_{lo}}(\mathbf{k}) u_{lo}^\mu \right) Y_L(\hat{\mathbf{r}}) \quad (51)$$

with

$$a_{Lo}^{\mu, \mathbf{G}_{lo}}(\mathbf{k}) = e^{i(\mathbf{k} + \mathbf{G}_{lo}) \cdot \boldsymbol{\tau}^\mu} \tilde{a}_{lo}^\mu 4\pi \frac{1}{W} i^l Y_L^*(\mathbf{k} + \mathbf{G}_{lo}) \quad (52)$$

and similarly for the b and c coefficients (cf. Eq. (79)).

It should be noted here that, although originally developed to treat semicore states, the LAPW+LO scheme allows also the treatment of higher lying states, that are far above the energy parameters E_l of conventional LAPW and which are impossible to describe with decent accuracy^b. Similar concepts have been explored by Krasovskii and Schattke^{72,73} in the extended LAPW method.

^bOf course accuracy refers here always to the linearization error in comparison to the APW method. Whether an accurate description of individual states, especially higher lying ones, is possible in DFT at all is not our concern here.

Another drawback of the linearized methods (as compared to the APW method) is the slower convergence of the results (e.g. eigenvalues) with respect to the number of basis functions that are used in the calculation. The condition of continuous first derivative at the muffin-tin boundary made the LAPW basis functions “stiffer” as compared to the APW’s. Although the LAPW’s are flexible in describing wavefunctions far from the linearization energy E_l , they provide a poorer basis close to E_l . This leads to an increasing number of LAPW’s, that have to be used to describe the wavefunction. An alternative way of linearizing the APW method has been proposed by Sjöstedt *et al.*⁴⁴: The APW basis functions (Eq. (39) but now evaluated at a fixed energy) are complemented within the muffin-tin by another type of local orbitals:

$$\varphi_{\mathbf{G}_{l_0}}^{\mu, l_0}(\mathbf{k}, \mathbf{r}) = \sum_m \left(a_{L_0}^{\mu, \mathbf{G}_{l_0}}(\mathbf{k}) u_l^\mu + b_{L_0}^{\mu, \mathbf{G}_{l_0}}(\mathbf{k}) \dot{u}_l^\mu \right) Y_L(\hat{\mathbf{r}}), \quad (53)$$

which are evaluated at the same energy as the APW’s. The coefficients a and b are chosen such, that this local orbital vanishes at the muffin-tin boundary. This introduces again a discontinuous first derivative at the muffin-tin radius, and the surface integral, that appeared in the APW method, has to be calculated. But in many cases this (small) additional numerical effort is compensated by a faster convergence with respect to the number of basis functions. Especially for structures with atoms with small muffin-tin radius (like O or N) this APW+lo method can lead to significant savings in computing time^{44, 74}. The question, up to which l the APW’s should be supplemented with these lo’s without introducing too large computational costs and from which l on a normal (L)APW basis set can be used without causing numerical errors has been investigated by Madsen *et al.*⁷⁴.

4 The FLAPW Method

4.1 The Concept of FLAPW

The full-potential LAPW method (FLAPW)^{24, 12} combines the choice of the LAPW basis set with the treatment of the full-potential and charge density without any shape-approximations in the interstitial region and inside the muffin-tins. This generalization is achieved by relaxing the constant interstitial potential V_I^0 and the spherical muffin-tin approximation $V_{MT}^0(r)$ due to the inclusion of a warped interstitial $\sum V_I^{\mathbf{G}} e^{i\mathbf{G}\mathbf{r}}$ and the non-spherical terms inside the muffin-tin spheres:

$$V(\mathbf{r}) = \begin{cases} \sum_{\mathbf{G}} V_I^{\mathbf{G}} e^{i\mathbf{G}\mathbf{r}} & \text{interstitial region} \\ \sum_L V_{MT}^L(r) Y_L(\hat{\mathbf{r}}) & \text{muffin-tin .} \end{cases} \quad (54)$$

This method became possible with the development of a technique for obtaining the Coulomb potential for a general periodic charge density without shape-approximations and with the inclusion of the Hamiltonian matrix elements due to the warped interstitial and non-spherical terms of the potential. The charge density n is represented analogously to Eq. (54), just exchanging V by n . Details of the solution of the Poisson equation for an arbitrarily shaped periodic potential are described in Section 4.6.

4.2 Construction of the Hamiltonian Matrix

The FLAPW Hamiltonian and overlap matrices consist of two contributions from the two regions into which space is divided.

$$H = H_I + H_{MT} \quad \text{and} \quad S = S_I + S_{MT} \quad (55)$$

Both contributions have to be computed separately.

4.2.1 Contribution of the Muffin-Tins

Writing the product of the radial functions u with the spherical harmonics as $\phi_L = u_l Y_L$, the contribution of the muffin-tin to the Hamiltonian matrix and the overlap matrix is given by:

$$H_{MT}^{\mathbf{G}'\mathbf{G}}(\mathbf{k}) = \sum_{\mu} \int_{MT\mu} \left(\sum_{L'} a_{L'}^{\mu\mathbf{G}'}(\mathbf{k}) \phi_{L'}^{\alpha}(\mathbf{r}) + b_{L'}^{\mu\mathbf{G}'}(\mathbf{k}) \dot{\phi}_{L'}^{\alpha}(\mathbf{r}) \right)^* \hat{H}_{MT\alpha} \left(\sum_L a_L^{\mu\mathbf{G}}(\mathbf{k}) \phi_L^{\alpha}(\mathbf{r}) + b_L^{\mu\mathbf{G}}(\mathbf{k}) \dot{\phi}_L^{\alpha}(\mathbf{r}) \right) d^3r. \quad (56)$$

(The overlap matrix $S_{MT}^{\mathbf{G}'\mathbf{G}}(\mathbf{k})$ is obtained by replacing $\hat{H}_{MT\alpha}$ by 1.) It is distinguished between the atom index μ and the atom type index $\alpha(\mu)$. In most applications they are symmetry equivalent atoms in the unit cell, i.e. some atoms can be mapped onto each other by space group operations. Clearly, these atoms must possess the same physical properties, e.g. the potential has to be equal. As a consequence, the Hamiltonian and the basis functions $\phi_L^{\alpha}(\mathbf{r})$ do not differ among the atoms of the same type. This fact is exploited in that the muffin-tin potential of an atom type is only stored once for the representative atom, and the matrices in Eq. (58) are also calculated for the representative only. $\hat{H}_{MT\alpha}$ is the scalar relativistic Hamiltonian operator. It can be split up into two parts, the spherical Hamiltonian $\hat{H}_{sp}$ (cf. Eq. (40)) and the nonspherical contributions to the potential V_{ns} .

$$\hat{H}_{MT\alpha} = \hat{H}_{sp}^{\alpha} + V_{ns}^{\alpha} \quad (57)$$

The above integrations contain the following type of matrix elements:

$$t_{L'L}^{\alpha\phi\phi} = \int_{MT\alpha} \phi_{L'}^{\alpha}(\mathbf{r}) \hat{H}_{MT\alpha} \phi_L^{\alpha}(\mathbf{r}) d^3r. \quad (58)$$

These matrix elements do not depend on the $A_L^{\mu\mathbf{G}}(\mathbf{k})$ and $B_L^{\mu\mathbf{G}}(\mathbf{k})$ coefficients. Thus, they are independent of the Bloch vector and need to be calculated only once per iteration. The functions ϕ_L^{α} and $\dot{\phi}_L^{\alpha}$ have been constructed to diagonalize the spherical part $\hat{H}_{sp}^{\alpha}$ of the muffin-tin Hamiltonian $\hat{H}_{MT\alpha}$:

$$\hat{H}_{sp}^{\alpha} \phi_L^{\alpha} = E_l \phi_L^{\alpha} \quad \text{and} \quad \hat{H}_{sp}^{\alpha} \dot{\phi}_L^{\alpha} = E_l \dot{\phi}_L^{\alpha} + \phi_L^{\alpha}. \quad (59)$$

Multiplying these equations with $\phi_{L'}^{\alpha}(\mathbf{r})$ and $\dot{\phi}_{L'}^{\alpha}(\mathbf{r})$ respectively and integrating over the muffin-tins gives

$$\begin{aligned} \langle \phi_{L'}^{\alpha} | \hat{H}_{sp}^{\alpha} \phi_L^{\alpha} \rangle_{MT\alpha} &= \delta_{ll'} \delta_{mm'} E_l & \langle \phi_{L'}^{\alpha} | \hat{H}_{sp}^{\alpha} \dot{\phi}_L^{\alpha} \rangle_{MT\alpha} &= \delta_{ll'} \delta_{mm'} \\ \langle \dot{\phi}_{L'}^{\alpha} | \hat{H}_{sp}^{\alpha} \phi_L^{\alpha} \rangle_{MT\alpha} &= 0 & \langle \dot{\phi}_{L'}^{\alpha} | \hat{H}_{sp}^{\alpha} \dot{\phi}_L^{\alpha} \rangle_{MT\alpha} &= \delta_{ll'} \delta_{mm'} E_l \langle \dot{u}_l^{\alpha} | \dot{u}_l^{\alpha} \rangle_{MT\alpha}, \end{aligned} \quad (60)$$

where the normalization condition for u_l^α has been used. So, only the expectation values of the nonspherical part of the potential are left to be determined. Since the potential is also expanded into a product of radial functions and spherical harmonics,

$$V^\alpha(\mathbf{r}) = \sum_{L''} V_{L''}^\alpha(r) Y_{L''}(\hat{\mathbf{r}}), \quad (61)$$

the corresponding integrals consist of a product of radial integrals and angular integrals over three spherical harmonics, the so-called Gaunt coefficients:

$$t_{L'L}^{\alpha\phi\phi} = \sum_{l''} I_{l'l''}^{\alpha uu} G_{l'l''}^{m'mm''} + \delta_{ll'} \delta_{mm'} E_l \quad (62)$$

with

$$G_{l'l''}^{mm'm''} = \int Y_{lm}^* Y_{l'm'} Y_{l''m''} d\Omega \quad \text{and} \quad I_{l'l''}^{\alpha uu} = \int u_{l'}^\alpha(r) u_{l''}^\alpha(r) V_{l''}^\alpha(r) r^2 dr \quad (63)$$

as well as similar expressions for $I_{l'l''}^{\alpha u\dot{u}}$ and others. The I matrices contain the radial integrals. Finally, the Hamiltonian and overlap matrix elements become

$$\begin{aligned} H_{MT}^{\mathbf{G}'\mathbf{G}}(\mathbf{k}) = & \sum_{\mu} \sum_{L'L} (a_{L'}^{\mu\mathbf{G}'}(\mathbf{k}))^* t_{L'L}^{\alpha\phi\phi} a_L^{\mu\mathbf{G}}(\mathbf{k}) + (b_{L'}^{\mu\mathbf{G}'}(\mathbf{k}))^* t_{L'L}^{\alpha\phi\phi} b_L^{\mu\mathbf{G}}(\mathbf{k}) \\ & + (a_{L'}^{\mu\mathbf{G}'}(\mathbf{k}))^* t_{L'L}^{\alpha\phi\dot{\phi}} b_L^{\mu\mathbf{G}}(\mathbf{k}) + (b_{L'}^{\mu\mathbf{G}'}(\mathbf{k}))^* t_{L'L}^{\alpha\phi\dot{\phi}} a_L^{\mu\mathbf{G}}(\mathbf{k}), \end{aligned} \quad (64)$$

$$S_{MT}^{\mathbf{G}'\mathbf{G}}(\mathbf{k}) = \sum_{\mu} \sum_L (a_L^{\mu\mathbf{G}'}(\mathbf{k}))^* a_L^{\mu\mathbf{G}}(\mathbf{k}) + (b_L^{\mu\mathbf{G}'}(\mathbf{k}))^* b_L^{\mu\mathbf{G}}(\mathbf{k}) \langle \dot{u}_l^\alpha | \dot{u}_l^\alpha \rangle_{MT\mu}. \quad (65)$$

4.2.2 The Interstitial Contribution

The interstitial contributions to the Hamiltonian and overlap matrix have the following form:

$$H_I^{\mathbf{G}\mathbf{G}'}(\mathbf{k}) = \frac{1}{\Omega} \int_I e^{-i(\mathbf{G}+\mathbf{k})\mathbf{r}} \left(-\frac{\hbar^2}{2m} \Delta + V(\mathbf{r}) \right) e^{i(\mathbf{G}'+\mathbf{k})\mathbf{r}} d^3r, \quad (66)$$

$$S_I^{\mathbf{G}\mathbf{G}'} = \frac{1}{\Omega} \int_I e^{-i(\mathbf{G}+\mathbf{k})\mathbf{r}} e^{i(\mathbf{G}'+\mathbf{k})\mathbf{r}} d^3r. \quad (67)$$

The potential is also expanded into planewaves in the interstitial region:

$$V(\mathbf{r}) = \sum_{\mathbf{G}'} V_{\mathbf{G}'} e^{-i\mathbf{G}'\mathbf{r}}. \quad (68)$$

Without the existence of the muffin-tin spheres the integration would stretch over the entire unit cell and the integration would become rather simple. The kinetic energy is diagonal in momentum space and the potential is local, diagonal in real space and of convolution form in momentum space.

$$H_I^{\mathbf{G}\mathbf{G}'}(\mathbf{k}) = \frac{\hbar^2}{2m} |\mathbf{G} + \mathbf{k}|^2 \delta_{\mathbf{G}\mathbf{G}'} + V_{(\mathbf{G}-\mathbf{G}')},$$

$$S_I^{\mathbf{G}\mathbf{G}'} = \delta_{\mathbf{G}\mathbf{G}'}.$$

However, these matrix elements are not as straightforward to calculate as they appear at first glance, because of the complicated structure of the interstitial region. The integrations have to be performed only in between the muffin-tins. Therefore, a step function $\Theta(\mathbf{r})$ has to be introduced, that cuts out the muffin-tins:

$$\Theta(\mathbf{r}) = \begin{cases} 1 & \text{interstitial region} \\ 0 & \text{muffin-tins} . \end{cases} \quad (69)$$

Using the step function the matrix elements can be written as:

$$\begin{aligned} H_I^{\mathbf{G}\mathbf{G}'}(\mathbf{k}) &= \frac{1}{\Omega} \int_{cell} e^{-i(\mathbf{G}-\mathbf{G}')\mathbf{r}} V(\mathbf{r}) \Theta(\mathbf{r}) d^3r \\ &+ \frac{1}{2}(\mathbf{G}' + \mathbf{k})^2 \frac{1}{\Omega} \int_{cell} e^{-i(\mathbf{G}-\mathbf{G}')\mathbf{r}} \Theta(\mathbf{r}) d^3r , \end{aligned} \quad (70)$$

$$S_I^{\mathbf{G}\mathbf{G}'} = \frac{1}{\Omega} \int_{cell} e^{-i(\mathbf{G}-\mathbf{G}')\mathbf{r}} \Theta(\mathbf{r}) d^3r . \quad (71)$$

In momentum space Eq. (70) becomes:

$$H_I^{\mathbf{G}\mathbf{G}'}(\mathbf{k}) = (V\Theta)_{(\mathbf{G}-\mathbf{G}')} + \frac{\hbar^2}{2m}(\mathbf{G}' + \mathbf{k})^2 \Theta_{(\mathbf{G}-\mathbf{G}')} , \quad (72)$$

$$S_I^{\mathbf{G}\mathbf{G}'} = \Theta_{(\mathbf{G}-\mathbf{G}')} , \quad (73)$$

where $\Theta_{\mathbf{G}}$ and $(V\Theta)_{\mathbf{G}}$ are the Fourier coefficients of $\Theta(\mathbf{r})$ and $V(\mathbf{r})\Theta(\mathbf{r})$ respectively. Apparently these coefficients are needed up to a cutoff of $2G_{\max}$. The step function can be Fourier transformed analytically:

$$\Theta_{\mathbf{G}} = \delta_{\mathbf{G},0} - \sum_{\mu} e^{-i\mathbf{G}\boldsymbol{\tau}^{\mu}} \frac{4\pi(R_{MT}^{\alpha})^3}{\Omega} \frac{j_1(GR_{MT}^{\alpha})}{GR_{MT}^{\alpha}} ,$$

where $\boldsymbol{\tau}^{\mu}$ indicates the position of atom μ . The Fourier transform of the product of $V(\mathbf{r})$ and $\Theta(\mathbf{r})$ is given by a convolution in momentum space:

$$(V\Theta)_{\mathbf{G}} = \sum_{\mathbf{G}'} V_{\mathbf{G}'} \Theta_{(\mathbf{G}-\mathbf{G}')} .$$

This convolution depends on both, $\mathbf{G}$ and $\mathbf{G}'$, therefore the numerical effort increases like $(G_{\max})^6$. However, $(V\Theta)_{\mathbf{G}}$ can be determined more efficiently, using Fast Fourier Transform (FFT). In Figure 10 it is shown schematically how $(V\Theta)_{\mathbf{G}}$ can be obtained using FFT. Using this scheme the numerical effort increases like $(G_{\max})^3 \ln(G_{\max})^3$ with $G_{\max}$.

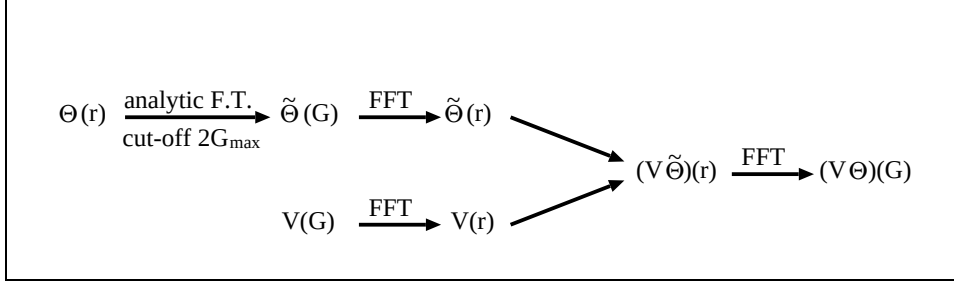


Figure 10. Schematic representation of the calculation of $(V\Theta)_G$. First $\Theta(\mathbf{r})$ is Fourier transformed analytically with a cutoff of $2G_{\max}$ yielding $\tilde{\Theta}_G$. Then $\tilde{\Theta}_G$ and V_G are fast Fourier transformed and multiplied on a real space mesh. Finally, the result $(V\tilde{\Theta})(\mathbf{r})$ is back-transformed to momentum space.

4.2.3 The Muffin-Tin a- and b-Coefficients

Within FLAPW the electron wavefunctions are expanded differently in the interstitial region and the muffin-tins. Each basis function consists of a planewave in the interstitial, which is matched to the radial functions and spherical harmonics in the muffin-tins. The coefficients of the function inside the spheres are determined from the requirement, that the basis functions and their derivatives are continuous at the sphere boundaries. These coefficients play an important role. In this section we will therefore discuss how the matching conditions can be solved and what properties they induce.

In many systems that the FLAPW method can be applied to some atoms are symmetry equivalent, i.e. these atoms can be mapped onto each other by a space group operation $\{\mathbf{R}|\mathbf{t}\}$. Such a group of atoms is called an atom type, represented by one of the atoms. Let $\{\mathbf{R}^\mu|\mathbf{t}^\mu\}$ be the operation that maps the atom μ onto its representative. This atom can now be assigned a local coordinate frame S^μ (cf. Figure 11), where the origin of S^μ is at the atoms position $\boldsymbol{\tau}^\mu$. The local frame is chosen such that the unit vectors of the local frame S^μ are mapped onto those of the global frame by $\mathbf{R}^\mu$ ($\mathbf{R}^\mu S^\mu = S^g$). The local frame of the representative atom S^α is only translated with respect to the global frame, i.e. the same rotation $\mathbf{R}^\mu$ maps S^μ onto S^α . The potential (and other quantities) inside the muffin-tins can now be written in terms of the local coordinate system. Due to the symmetry we find $V_{MT^\alpha}(\mathbf{r}^\alpha) = V_{MT^\mu}(\mathbf{r}^\mu)$, where $\mathbf{r}^\alpha$ and $\mathbf{r}^\mu$ are expanded in terms of the local frames S^α and S^μ respectively. As a consequence the radial functions $u_l(r)$ and the t-matrices are the same for all atoms of the same type. This way symmetry is exploited to save memory and computer time (during the calculation of the t-matrices).

Any planewave can be expanded into spherical harmonics via the Rayleigh expansion:

$$e^{i\mathbf{K}\mathbf{r}} = 4\pi \sum_L i^l j_l(rK) Y_L^*(\hat{\mathbf{K}}) Y_L(\hat{\mathbf{r}}), \quad (74)$$

where $r = |\mathbf{r}|$, $K = |\mathbf{K}|$ and $\mathbf{K}$ abbreviates $(\mathbf{G} + \mathbf{k})$. Looked at from the local frame $\mathbf{K}$ and $\boldsymbol{\tau}^\mu$ appear rotated, besides the origin of the local frame is shifted. Therefore, the planewave has the following form in the local frame:

$$e^{i(\mathbf{R}^\mu \mathbf{K})(\mathbf{r} + \mathbf{R}^\mu \boldsymbol{\tau}^\mu)}. \quad (75)$$

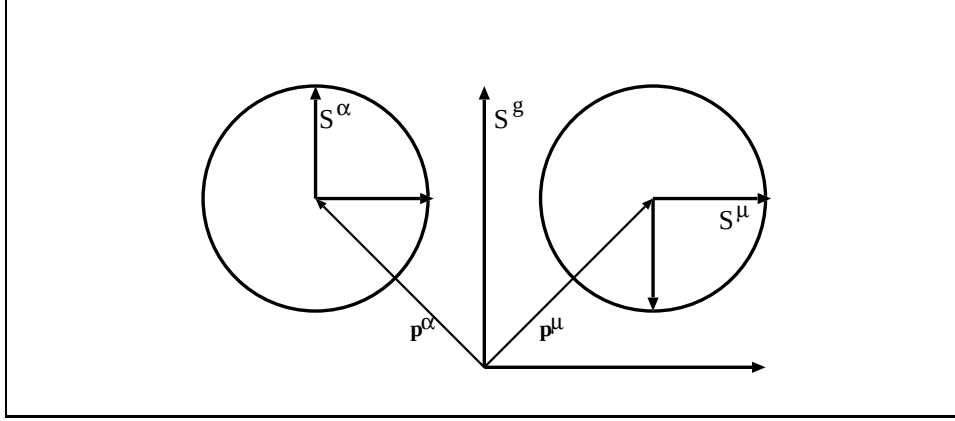


Figure 11. Local coordinate frames inside each muffin-tin.

Thus, the Rayleigh expansion of the planewave in the local frame is given by:

$$e^{i\mathbf{K}\cdot\mathbf{r}} = 4\pi \sum_L i^l j_l(rK) Y_L^*(\mathbf{R}^\mu \hat{\mathbf{K}}) Y_L(\hat{\mathbf{r}}). \quad (76)$$

The requirement of continuity of the wavefunctions at the sphere boundary leads to the equation:

$$\begin{aligned} & \sum_L a_L^\mu(\mathbf{k}) u_l(R_{MT^\alpha}) Y_L(\hat{\mathbf{r}}) + b_L^\mu(\mathbf{k}) \dot{u}_l(R_{MT^\alpha}) Y_L(\hat{\mathbf{r}}) \\ &= e^{i\mathbf{K}\cdot\mathbf{r}} 4\pi \sum_L i^l j_l(rK) Y_L^*(\mathbf{R}^\mu \hat{\mathbf{K}}) Y_L(\hat{\mathbf{r}}), \end{aligned} \quad (77)$$

where R_{MT^α} is the muffin-tin radius of the atom type α . The second requirement is, that the derivative with respect to r , denoted by $\partial/\partial r = ' ,$ is also continuous:

$$\begin{aligned} & \sum_L a_L^\mu(\mathbf{k}) u'_l(R_{MT^\alpha}) Y_L(\hat{\mathbf{r}}) + b_L^\mu(\mathbf{k}) \dot{u}'_l(R_{MT^\alpha}) Y_L(\hat{\mathbf{r}}) \\ &= e^{i\mathbf{K}\cdot\mathbf{r}} 4\pi \sum_L i^l K j'_l(rK) Y_L^*(\mathbf{R}^\mu \hat{\mathbf{K}}) Y_L(\hat{\mathbf{r}}). \end{aligned} \quad (78)$$

These conditions can only be satisfied, if the coefficients of each spherical harmonic $Y_L(\hat{\mathbf{r}})$ are equal. Solving the resulting equations for $A_L^\mu(\mathbf{k})$ and $B_L^\mu(\mathbf{k})$ yields:

$$\begin{aligned} a_L^\mu(\mathbf{k}) &= e^{i\mathbf{K}\cdot\mathbf{r}} 4\pi \frac{1}{W} i^l Y_L^*(\mathbf{R}^\mu \hat{\mathbf{K}}) \\ & \quad [\dot{u}_l(R_{MT^\alpha}) K j'_l(R_{MT^\alpha} K) - \dot{u}'_l(R_{MT^\alpha}) j_l(R_{MT^\alpha} K)] \\ b_L^\mu(\mathbf{k}) &= e^{i\mathbf{K}\cdot\mathbf{r}} 4\pi \frac{1}{W} i^l Y_L^*(\mathbf{R}^\mu \hat{\mathbf{K}}) \\ & \quad [u'_l(R_{MT^\alpha}) j_l(R_{MT^\alpha} K) - u_l(R_{MT^\alpha}) K j'_l(R_{MT^\alpha} K)]. \end{aligned} \quad (79)$$

The Wronskian W is given by:

$$W = [\dot{u}_l(R_{MT^\alpha})u'_l(R_{MT^\alpha}) - u_l(R_{MT^\alpha})\dot{u}'_l(R_{MT^\alpha})] . \quad (80)$$

4.3 Brillouin Zone Integration and Fermi Energy

In the current implementation of the FLAPW method the Fermi energy is determined in two steps. First the bands are occupied (at all $\mathbf{k}$ -points simultaneously), starting from the lowest energy, until the sum of their weights equals the total number of electrons per unit cell, i.e. the discretized equivalent of Eq. (27) is solved at $T = 0$. Then, the step function is replaced by the Fermi function and the Fermi energy is determined from the requirement that:

$$N = \sum_{\mathbf{k}} \sum_{\nu} w(\mathbf{k}, \epsilon_{\nu}(\mathbf{k}) - E_F) , \quad (81)$$

where the weights are given by:

$$w(\mathbf{k}, \epsilon_{\nu}(\mathbf{k}) - E_F) = w(\mathbf{k}) \frac{1}{e^{(\epsilon_{\nu}(\mathbf{k}) - E_F)/k_B T} + 1} . \quad (82)$$

The weights $w(\mathbf{k}, \epsilon_{\nu}(\mathbf{k}) - E_F)$ are stored to be used for later Brillouin zone integrations.

4.4 Representation of the Density and the Potential

The expansion of the charge density n and the potential is very similar to the expansion of the wavefunction. In the interstitial region the two quantities are expanded into three-dimensional plane wave, inside the muffin-tins they are represented by spherical harmonics and radial functions, which are stored on an exponential mesh and in the vacuum they are expanded into two-dimensional plane wave and z -depended functions, which are also given on an exponential mesh. However, the charge density is given by Eq. (1), which contains terms of the form $e^{i(\mathbf{G}-\mathbf{G}')\mathbf{r}}$. Consequently, for a consistent representation the charge density cutoff has to be twice the wavefunction cutoff $G_{\max}$. In Section 4.2.2 we explained, that the potential is also needed up to cutoff of $2G_{\max}$. This leads to a large number of coefficients, that need to be stored. Fortunately, this number can be reduced, if the symmetry of the system is exploited.

Of course, the charge density and the potential possess the lattice symmetry. Therefore, the expansion into plane waves is more general than necessary. The plane waves can be replaced by symmetrized plane waves, the so-called stars. They are defined by:

$$\Phi_s^{3D}(\mathbf{r}) = \frac{1}{N_{op}} \sum_{op} e^{i\mathbf{R}\mathbf{G}(\mathbf{r}-\mathbf{t})}, \quad (83)$$

where $\{\mathbf{R}|\mathbf{t}\}$ are the symmetry operation of the lattice space group; if all the translation vectors $\mathbf{t}$ are zero, the space group is call symmorphic. By this construction all plane waves, that are symmetry equivalent, are combined to form one star. The two-dimensional stars $\Phi_s^{2D}(\mathbf{r})$ are defined in the same way, applying the operations of the two-dimensional space group only.

The same arguments can be applied to the expansion of the $n(V)$ inside the muffin-tins. In this case the relevant symmetry group is the point group of the atom under consideration. Thus, different expansions are used at different atoms in general. The symmetrized

functions are called lattice harmonics and they are linear combinations of spherical harmonics

$$K_\nu(\hat{\mathbf{r}}) = \sum_m c_{\nu,m}^\alpha Y_L(\hat{\mathbf{r}}). \quad (84)$$

The lattice harmonics are real, orthonormal and invariant under the point group operations. Finally, the expansion of the charge density has the form:

$$n(\mathbf{r}) = \begin{cases} \sum_s n_s \Phi_s^{3D}(\mathbf{r}) & \mathbf{r} \in I \\ \sum_\nu n_\nu^\alpha(r) K_\nu(\hat{\mathbf{r}}) & \mathbf{r} \in MT^\alpha \end{cases}. \quad (85)$$

The potential is expanded in exactly the same way.

4.5 Construction of the Electron Density

In this section we will discuss the determination of the charge density from the eigenfunctions. In density functional calculations of an infinite periodic solid, the electron density is given by an integral over the Brillouin zone (cf. Eq. (27)):

$$n(\mathbf{r}) = \frac{1}{V_{BZ}} \int_{BZ} \sum_{\nu, \epsilon_\nu(\mathbf{k}) < E_F} |\psi_\nu(\mathbf{k}, \mathbf{r})|^2 d^3k, \quad (86)$$

where V_{BZ} is the volume of the Brillouin zone, ν is the band index and E_F is the Fermi energy. In spin-polarized calculations the summation includes also the spin-index σ (cf. Section 2.3), while in a non-magnetic calculation a factor “2” has to be added to account for the spin-degeneracy. In the case of film calculations the three-dimensional Brillouin zone is replaced by a two-dimensional Brillouin zone. In both cases integration methods that sample eigenfunctions and the eigenvalues on discrete k-point are used to compute the integrals. These methods transform the integration into a weighted sum over the k-points, where the choice of k-points and their weights depend on the integration method used. These weights depend not only on the k-point, but also on the energy of a band, i.e. on the band (index), because each band contributes to the electron density only if its energy is below the Fermi energy:

$$n(\mathbf{r}) = \sum_{\mathbf{k}} \sum_{\nu} |\psi_\nu(\mathbf{k}, \mathbf{r})|^2 w(\nu, \mathbf{k}). \quad (87)$$

Within the FLAPW method the eigenfunctions are represented in terms of the coefficients of the augmented planewaves:

$$\psi_\nu(\mathbf{k}, \mathbf{r}) = \sum_{\mathbf{G}} c_\nu^{\mathbf{G}}(\mathbf{k}) \varphi_{\mathbf{G}}(\mathbf{k}, \mathbf{r}). \quad (88)$$

Inside the muffin-tin spheres each planewave is coupled to a sum of spherical harmonics and radial functions. Hence, in a sphere μ an eigenfunction is given by:

$$\psi_\nu^\mu(\mathbf{k}, \mathbf{r}) = \sum_{\mathbf{G}} c_\nu^{\mathbf{G}}(\mathbf{k}) \sum_L a_L^{\mu\mathbf{G}}(\mathbf{k}) u_l^\alpha(r) Y_L(\hat{\mathbf{r}}) + b_L^{\mu\mathbf{G}}(\mathbf{k}) \dot{u}_l^\alpha(r) Y_L(\hat{\mathbf{r}}). \quad (89)$$

The $a_L^{\mu\mathbf{G}}(\mathbf{k})$ and $b_L^{\mu\mathbf{G}}(\mathbf{k})$ coefficients can be replaced by band dependent A- and B-coefficients, obtained by performing the contraction over the planewaves:

$$\psi_\nu^\mu(\mathbf{k}, \mathbf{r}) = \sum_L a_{L,\nu}^\mu(\mathbf{k}) u_l^\alpha(r) Y_L(\hat{\mathbf{r}}) + b_{L,\nu}^\mu(\mathbf{k}) \dot{u}_l^\alpha(r) Y_L(\hat{\mathbf{r}}), \quad (90)$$

where

$$A_{L,\nu}^\mu(\mathbf{k}) = \sum_{\mathbf{G}} c_\nu^\mathbf{G}(\mathbf{k}) a_L^\mu \mathbf{G}(\mathbf{k}), \quad B_{L,\nu}^\mu(\mathbf{k}) = \sum_{\mathbf{G}} c_\nu^\mathbf{G}(\mathbf{k}) b_L^\mu \mathbf{G}(\mathbf{k}). \quad (91)$$

4.5.1 “ l -Like” Charge

Since the wavefunctions are expanded into spherical harmonics inside the muffin-tin spheres, they can be split up into contributions with a certain l -character:

$$\psi_\nu^\mu(\mathbf{k}, \mathbf{r}) = \sum_l \psi_{\nu,l}^\mu(\mathbf{k}, \mathbf{r}). \quad (92)$$

The particle density of a certain state depends on the square of the wavefunction. Therefore, it contains cross-terms with a mixture of different l 's:

$$n_\nu^\mu(\mathbf{r}) = \frac{1}{V_{BZ}} \int_{BZ} \sum_l |\psi_{\nu,l}^\mu(\mathbf{k}, \mathbf{r})|^2 + \sum_{l' \neq l} 2 \left(\psi_{\nu,l'}^\mu(\mathbf{k}, \mathbf{r}) \right)^* \psi_{\nu,l}^\mu(\mathbf{k}, \mathbf{r}) d^3 k. \quad (93)$$

If, however, the density is integrated over the muffin-tin, the cross-terms vanish because of the orthogonality of the spherical harmonics. Thus, the total electron density inside a sphere can be written as a sum over contributions with definite l -character:

$$n_\nu^\mu = \sum_l n_{\nu,l}^\mu, \quad n_{\nu,l}^\mu = \frac{1}{V_{BZ}} \int_{BZ} \int_{MT^\mu} |\psi_{\nu,l}^\mu(\mathbf{k}, \mathbf{r})|^2 d^3 r d^3 k, \quad (94)$$

where $n_{\nu,l}^\mu$ is called “ l -like” charge. We can also define a $\mathbf{k}$ -dependent l -like charge by:

$$n_{\nu,l}^\mu(\mathbf{k}) = \int_{MT^\mu} |\psi_{\nu,l}^\mu(\mathbf{k}, \mathbf{r})|^2 d^3 r. \quad (95)$$

Substituting Eq. (90) yields:

$$n_{\nu,l}^\mu(\mathbf{k}) = \sum_{m=-l}^l |A_{L,\nu}^\mu(\mathbf{k})|^2 + |B_{L,\nu}^\mu(\mathbf{k})|^2 \dot{N}_l^\alpha, \quad (96)$$

where

$$\dot{N}_l^\alpha = \int_0^{R_{MT^\alpha}} (\dot{u}_l^\alpha(r))^2 r^2 dr \quad (97)$$

and the orthogonality of the spherical harmonics, the normalization of u_l^α and the orthogonality of u_l^α and $\dot{u}_l^\alpha$ have been used.

4.5.2 Determination of the Optimal Energy Parameter

In order to minimize the linearization error, the energy parameters should be chosen as close to the band energies as possible. However, the band energies $\epsilon_\nu(\mathbf{k})$ depend on $\mathbf{k}$ whereas the energy parameters E_l^α are constants. In addition, the radial functions contribute to the eigenfunctions of different band with different energies. Therefore, deviations between $\epsilon_\nu(\mathbf{k})$ and E_l^α have to be accepted. An optimal choice can be obtained from the requirement, that the energy parameters minimize

$$\int_{BZ} \sum_{\nu, \epsilon_\nu(\mathbf{k}) < E_F} (\epsilon_\nu(\mathbf{k}) - E_l^\alpha)^2 n_{\nu,l}^\mu(\mathbf{k}) d^3 k, \quad (98)$$

which is the quadratic error weighted with the amount of charge that each band contributes to the l-like charge with the l-character of the energy parameter. Setting the derivative $(\partial/\partial E_l^\alpha)$ equal to zero yields the optimal energy parameter:

$$E_l^\alpha = \left(\sum_{\mathbf{k}} \sum_{\nu} \epsilon_{\nu}(\mathbf{k}) n_{\nu,l}^{\mu}(\mathbf{k}) w(\nu, \mathbf{k}) \right) / \left(\sum_{\mathbf{k}} \sum_{\nu} n_{\nu,l}^{\mu}(\mathbf{k}) w(\nu, \mathbf{k}) \right) . \quad (99)$$

4.5.3 Construction of the Electron Density in the Muffin-Tins

Substituting Eq. (90) into Eq. (86) yields the electron density in the muffin-tin spheres:

$$n^{\mu}(\mathbf{r}) = \frac{1}{V_{BZ}} \int_{BZ} \sum_{\nu, \epsilon_{\nu}(\mathbf{k}) < E_F} \sum_{L'} \left(A_{L',\nu}^{\mu}(\mathbf{k}) u_{l'}^{\alpha}(r) + B_{L',\nu}^{\mu}(\mathbf{k}) \dot{u}_{l'}^{\alpha}(r) \right)^* Y_{L'}^*(\hat{\mathbf{r}}) \cdot \sum_L \left(A_{L,\nu}^{\mu}(\mathbf{k}) u_l^{\alpha}(r) + B_{L,\nu}^{\mu}(\mathbf{k}) \dot{u}_l^{\alpha}(r) \right) Y_L(\hat{\mathbf{r}}) d^3 k. \quad (100)$$

The particle density inside the muffin-tins is also expanded into spherical harmonics

$$n^{\mu}(\mathbf{r}) = \sum_L C_L^{\mu}(r) Y_L(\hat{\mathbf{r}}). \quad (101)$$

The coefficients $C_{L''}^{\mu}(r)$ can be determined by multiplying Eq. (100) with $\int d\Omega Y_{L''}(\hat{\mathbf{r}})$:

$$C_{L''}^{\mu}(r) = \frac{1}{V_{BZ}} \int_{BZ} \sum_{\nu, \epsilon_{\nu}(\mathbf{k}) < E_F} \sum_{L'} \left(A_{L',\nu}^{\mu}(\mathbf{k}) u_{l'}^{\alpha}(r) + B_{L',\nu}^{\mu}(\mathbf{k}) \dot{u}_{l'}^{\alpha}(r) \right)^* \sum_L \left(A_{L,\nu}^{\mu}(\mathbf{k}) u_l^{\alpha}(r) + B_{L,\nu}^{\mu}(\mathbf{k}) \dot{u}_l^{\alpha}(r) \right) G_{ll'l''}^{mm'm''} d^3 k, \quad (102)$$

where it has been used, that the Gaunt coefficients are real, i.e.

$$\int Y_{lm} Y_{l'm'}^* Y_{l''m''}^* d\Omega = \int Y_{lm}^* Y_{l'm'} Y_{l''m''} d\Omega . \quad (103)$$

Finally, applying a Brillouin zone integration method yields:

$$C_{L''}^{\mu}(r) = \sum_{l'l} \left(\sum_{\mathbf{k}} \sum_{\nu} \sum_{m'm} \left(A_{L',\nu}^{\mu}(\mathbf{k}) \right)^* A_{L,\nu}^{\mu}(\mathbf{k}) G_{ll'l''}^{mm'm''} w(\nu, \mathbf{k}) \right) u_{l'}^{\alpha}(r) u_l^{\alpha}(r) + \dots A^* B + B^* A + B^* B \dots . \quad (104)$$

4.5.4 Construction of the Electron Density in the Interstitial Region

In the interstitial region the wavefunctions are represented in the following form:

$$\psi_{\nu}(\mathbf{k}, \mathbf{r}) = \sum_{\mathbf{G}} c_{\nu}^{\mathbf{G}}(\mathbf{k}) e^{i(\mathbf{G}+\mathbf{k})\mathbf{r}} . \quad (105)$$

Starting from Eq. (1) the electron density is given by:

$$n(\mathbf{r}) = \frac{1}{V_{BZ}} \int_{BZ} \sum_{\nu, \epsilon_{\nu}(\mathbf{k}) < E_F} \sum_{\mathbf{G}' \mathbf{G}''} \left(c_{\nu}^{\mathbf{G}'}(\mathbf{k}) \right)^* c_{\nu}^{\mathbf{G}''}(\mathbf{k}) d^3 k e^{i(\mathbf{G}'' - \mathbf{G}')\mathbf{r}} . \quad (106)$$

The electron density in the interstitial region is also expanded into planewaves:

$$n(\mathbf{r}) = \sum_{\mathbf{G}} n^{\mathbf{G}} e^{i\mathbf{G}\mathbf{r}} . \quad (107)$$

Hence, the planewave coefficients of the electron density are:

$$n^{\mathbf{G}} = \frac{1}{V_{BZ}} \int_{BZ} \sum_{\nu, \epsilon_{\nu}(\mathbf{k}) < E_F} \sum_{\substack{\mathbf{G}', \mathbf{G}'' \\ \mathbf{G}'' - \mathbf{G}' = \mathbf{G}}} \left(c_{\nu}^{\mathbf{G}'}(\mathbf{k}) \right)^* c_{\nu}^{\mathbf{G}''}(\mathbf{k}) d^3k . \quad (108)$$

Apparently, the planewave cutoff of the particle density has to be twice the cutoff of the wavefunction expansion ($G_{\max}$) to allow an accurate description. The $\mathbf{k}$ and state dependent density

$$n_{\nu}^{\mathbf{G}}(\mathbf{k}) = \sum_{\substack{\mathbf{G}', \mathbf{G}'' \\ \mathbf{G}'' - \mathbf{G}' = \mathbf{G}}} \left(c_{\nu}^{\mathbf{G}'}(\mathbf{k}) \right)^* c_{\nu}^{\mathbf{G}''}(\mathbf{k}) = \sum_{\mathbf{G}'} \left(c_{\nu}^{\mathbf{G}'}(\mathbf{k}) \right)^* c_{\nu}^{(\mathbf{G} + \mathbf{G}')}(\mathbf{k}) \quad (109)$$

is given by a convolution in momentum space. For each coefficient a sum over $\mathbf{G}$ has to be performed. Consequently, the numerical effort put into the determination of $n_{\nu}^{\mathbf{G}}(\mathbf{k})$ scales proportional to the number of G -vectors squared, i.e. proportional to $(G_{\max})^6$. However, $n_{\nu}^{\mathbf{G}}(\mathbf{k})$ can be calculated more efficiently using the fast Fourier transform (FFT). First, $c_{\nu}^{\mathbf{G}}(\mathbf{k})$ is Fourier transformed to real space, where it is squared on a real space mesh yielding $n_{\nu}(\mathbf{k}, \mathbf{r})$, then all states are summed up and finally the resulting particle density is back-transformed to momentum space.

$$c_{\nu}^{\mathbf{G}}(\mathbf{k}) \xrightarrow{FFT} \psi_{\nu}(\mathbf{k}, \mathbf{r}) \xrightarrow{square} n_{\nu}(\mathbf{k}, \mathbf{r}) \xrightarrow{\sum_{\nu}} n(\mathbf{k}, \mathbf{r}) \xrightarrow{FFT^{-1}} n^{\mathbf{G}}(\mathbf{k})$$

With this scheme the numerical effort increases proportional to $(G_{\max})^3 \ln G_{\max}^3$, which is a major improvement for large systems. In a last step the planewaves have to be combined to form the three-dimensional stars.

4.6 Construction of the Coulomb Potential

The Coulomb potential consists of two parts, the Hartree term $V_H(\mathbf{r})$ and the external potential of the nuclei $V_i(\mathbf{r})$:

$$V_c(\mathbf{r}) = V_H(\mathbf{r}) + V_i(\mathbf{r}) . \quad (110)$$

The Hartree potential has to be determined from the charge density via the Poisson equation:

$$\Delta V_H(\mathbf{r}) = 4\pi n(\mathbf{r}) . \quad (111)$$

In real space the solution of Eq. (111) is given by

$$V_H(\mathbf{r}) = \int \frac{4\pi n(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d^3\mathbf{r}' . \quad (112)$$

In reciprocal space, however, the Poisson equation is diagonal, as a result the solution is very simple:

$$V_H(\mathbf{G}) = \frac{4\pi n(\mathbf{G})}{G^2} . \quad (113)$$

Therefore, and because of the representation of the charge density and the potential in the interstitial and vacuum region, the solution of the Poisson equation in reciprocal space appears to be convenient. However, due to the rather localized core and valence states the charge density changes on a very small length scale near the nuclei. Therefore, the planewave expansion of n converges slowly, and a direct use of Eq. (113) is impractical, if not impossible. This difficulty can be circumvented via the pseudocharge method.

4.6.1 The Pseudocharge Method

The pseudocharge method, developed by Weinert²³, is a very elegant technique to calculate the interstitial and vacuum Hartree potential. The underlying idea is to divide the solution of the Poisson equation into two steps. In the first step the true muffin-tin charge is replaced by a convergent pseudocharge density $\tilde{n}$, that leads to the same potential outside the muffin-tins. Then, the interstitial (and vacuum) potential is calculated in reciprocal space. In the second step the muffin-tin potential is determined from the Dirichlet boundary value problem, defined by the exact muffin-tin charge and the interstitial potential on the muffin-tin sphere boundaries. The potential outside the muffin-tin spheres due to a charge distribution inside the sphere is determined completely by its multipole moments q_L :

$$V(\mathbf{r}) = \sum_{l=0}^{\infty} \sum_{m=-l}^l \frac{4\pi}{2l+1} \frac{q_L}{r^{l+1}} Y_L(\hat{\mathbf{r}}) . \quad (114)$$

However, the multipole moments do not define the charge density uniquely. The charge density is given by:

$$n(\mathbf{r}) = n_I(\mathbf{r})\Theta(\mathbf{r} \in I) + \sum_{\alpha} n_{\alpha}(\mathbf{r})\Theta(\mathbf{r} \in MT^{\alpha}) . \quad (115)$$

Of course, in film calculation there is also a vacuum charge, and we will come back to this later. Eq. (115) can be rewritten as

$$n(\mathbf{r}) = n_I(\mathbf{r}) + \sum_{\alpha} [n_{\alpha}(\mathbf{r}) - n_I(\mathbf{r})]\Theta(\mathbf{r} \in MT^{\alpha}) . \quad (116)$$

Thus, the interstitial charge has been extended into the muffin-tin and subtracted there again. The second term in Eq. (116) can now be replaced by a pseudocharge $\tilde{n}^{\alpha}$, that has the same multipole moments (cf.²³ for details). The resultant pseudocharge $\tilde{n}$ is given by

$$\tilde{n}(\mathbf{r}) = n_I(\mathbf{r}) + \sum_{\alpha} \tilde{q}^{\alpha}(\mathbf{r}) ; \quad (117)$$

$\tilde{n}(\mathbf{r})$ is constructed to have a more rapidly converging Fourier expansion than the original charge density $n(\mathbf{r})$. Therefore, the Poisson equation can now be solved using Eq. (113).

Still, the muffin-tin potential V_{MT}^{α} remains to be determined. For this step the exact muffin-tin charge n_{α} has to be used. Since the interstitial potential is already known at this point, the calculation of V_{MT}^{α} constitutes a classical spherically symmetric Dirichlet boundary value problem, which can be solved by the Green's function method⁷⁵:

$$V_{MT}^{\alpha}(\mathbf{r}) = \int_{MT^{\alpha}} n_{\alpha}(\mathbf{r}') G(\mathbf{r}, \mathbf{r}') d^3 r' - \frac{R_{\alpha}^2}{4\pi} \oint_{S^{\alpha}} V_I(\mathbf{r}') \frac{\partial G}{\partial n'} d\Omega' \quad (118)$$

The second integral is over the muffin-tin sphere boundary S^α , and it is necessary to satisfy the boundary conditions. The Green's function is given by:

$$G^\alpha(\mathbf{r}, \mathbf{r}') = 4\pi \sum_{l,m} \frac{Y_L(\hat{\mathbf{r}}') Y_L(\hat{\mathbf{r}})}{2l+1} \frac{r_{<}^l}{r_{>}^{l+1}} \left(1 - \left(\frac{r_{>}}{R_{MT^\alpha}} \right)^{2l+1} \right) \quad (119)$$

where $r_{>} = \max\{|\mathbf{r}|, |\mathbf{r}'|\}$, $r_{<} = \min\{|\mathbf{r}|, |\mathbf{r}'|\}$. Finally, the muffin-tin potential has to be expanded into lattice harmonics $K_\nu(\hat{\mathbf{r}})$:

$$V_{MT}^\alpha(\mathbf{r}) = \sum_\nu V_{MT,\nu}^\alpha(r) K_\nu(\hat{\mathbf{r}}) . \quad (120)$$

The potential of the nuclei $V_i^\alpha(\mathbf{r}) = \frac{eZ_i^\alpha}{|\mathbf{r}|}$ is added to the spherical ($l = 0$) component of the potential $V_{MT,0}^\alpha(r)$.

The muffin-tin potential is computed in the same way for both, bulk and film calculations. Apparently, the interstitial and the vacuum have to be treated differently in the two cases, due to the different boundary conditions and the different representation of the vacuum potential. Therefore, in the next two sections the solution of the Poisson equation will be outlined separately for these cases.

4.6.2 Determination of the Interstitial Coulomb Potential in Bulk Calculations

In the case of bulk calculations we have periodic boundary conditions in three dimensions. Therefore, the solution of the Poisson equation,

$$\mathbf{G}^2 V(\mathbf{G}) = 4\pi \tilde{n}(\mathbf{G}) \quad (121)$$

is very simple. Obviously, this equation can only be solved, if $\tilde{n}(0) = 0$. Since $\tilde{n}(0)$ is the average charge density, this means, that charge neutrality is essential. Still, $V(0)$ remains undetermined by Eq. (121), i.e. one has the freedom to shift the potential by a constant. This is a consequence of the periodic boundary conditions, because they do not fix the reference of the potential. Usually $V(0)$ is chosen to be zero, hence the Coulomb potential in the interstitial region is given by:

$$V_I(\mathbf{r}) = \sum_{\mathbf{G} \neq 0} \frac{4\pi \tilde{n}(\mathbf{G})}{\mathbf{G}^2} e^{i\mathbf{G}\mathbf{r}} = \sum_{s \neq 0} \frac{4\pi \tilde{n}_s}{\mathbf{G}_s^2} \Phi_s^{3D}(\mathbf{r}) , \quad (122)$$

where the first summation is expressed in terms of G -vectors and the second in terms of stars.

4.7 Computation of the Exchange-Correlation Potential

The problem of the determination of the exchange-correlation potential is quite different from the Coulomb potential. On one hand, V_{xc}^σ is a local quantity, i.e. $V_{xc}^\sigma(\mathbf{r})$ depends only on $n_\uparrow(\mathbf{r})$ and $n_\downarrow(\mathbf{r})$ at the same position $\mathbf{r}$. Thus, the muffin-tins, the interstitial and vacuum region can be treated independently. On the other hand, V_{xc}^σ and ϵ_{xc}^σ are non-linear functions of $n_\uparrow$ and $n_\downarrow$. Therefore, V_{xc}^σ and ϵ_{xc}^σ have to be calculated in real space. V_{xc}^σ and ϵ_{xc}^σ are determined in the same way. First, $n_\uparrow$ and $n_\downarrow$ are transformed to real space, where V_{xc}^σ and ϵ_{xc}^σ are calculated. Then, V_{xc}^σ and ϵ_{xc}^σ are back-transformed. Then, V_{xc}^σ is added to the Coulomb potential, yielding the spin-dependent potential $V_\uparrow$ and $V_\downarrow$. ϵ_{xc}^σ is needed for the determination of the total energy.

4.7.1 Calculation of ϵ_{xc}^σ and V_{xc}^σ in the Interstitial Region

In the interstitial region the charge density is expanded into three-dimensional stars with coefficients n_s^σ . Multiplying these by $e^{i\mathbf{R}\mathbf{G}\mathbf{t}}$ yields the plane-wave coefficients n_G^σ . If the space group is symmorphic the star and plane-wave coefficients are identical. However, due to numerical inaccuracy, the calculated coefficients of symmetry equivalent plane-waves are not exactly equal, and the corresponding star coefficient is obtained from the average of the plane-wave coefficients. In the next step a three-dimensional Fast Fourier transform is carried out. Then, the exchange-correlation potential is calculated on a real space mesh $\mathbf{r}_i$. Finally, V_{xc}^σ is back-transformed, and the star coefficients are computed:

$$n_s^\sigma \longrightarrow n_G^\sigma \xrightarrow{FFT} n^\sigma(\mathbf{r}_i) \longrightarrow V_{xc}^\sigma(\mathbf{r}_i) \xrightarrow{FFT^{-1}} V_{xc}^{\sigma,G} \longrightarrow V_{xc}^{\sigma,s}.$$

4.7.2 Calculation of ϵ_{xc}^σ and V_{xc}^σ in the Muffin-Tin Spheres

The muffin-tin charge is expanded into lattice harmonics and radial functions. The radial functions are stored on a discrete real-space mesh. Thus, the transform to real space affects only the angular part. The charge density is calculated on a set of special angular points $\hat{\mathbf{r}}_i = (\theta_i, \phi_i)$. Again, the exchange correlation potential is calculated in real space. Thereafter, the result $V_{xc}^\sigma(\mathbf{r})$ is expanded into spherical harmonics Y_L . The Y_L are orthonormal, therefore the coefficients can be obtained from

$$v_{xc,L}^\sigma(r) = \int Y_L(\hat{\mathbf{r}}) V_{xc}^\sigma(r, \hat{\mathbf{r}}) d\Omega. \quad (123)$$

The choice of the points $\hat{\mathbf{r}}_i = (\theta_i, \phi_i)$, on which $n^\sigma(\mathbf{r})$ and $V_{xc}^\sigma(\mathbf{r})$ are calculated, depends on the integration method, that is used to perform the angular integration. In the current implementation Eq. (123) is computed via a Gauß-Legendre integration and the angular points are chosen such, that the orthonormality condition of the Y_L holds also for the angular mesh $\hat{\mathbf{r}}_i$.

5 The FLAPW Method for Specialized Geometries

5.1 The Film Geometry for Surfaces and Thin Films

Today, the physics of surfaces and films is an field of major interest and investigation. However, surfaces are difficult to treat, because they break the translational symmetry, i.e. there is only the 2-dimensional symmetry parallel to the surface left to be used to reduce the problem, and a semi-infinite problem is left perpendicular to the surface. In our approach surfaces are approximated by thin films, typically 10–15 atomic layers thick. Obviously, this approximation, which is called the thin-slab approximation, can only yield good results if the interaction between the two surfaces of the film is weak enough, so that each of them shows the properties of the surfaces of an ideal semi-infinite crystal. In the case of film calculations space is divided into three distinct regions, the muffin-tins, the interstitial and the vacuum region (cf. Figure 12). The interstitial region now stretches from $-D/2$ to $D/2$ in z-direction, which is defined to be the direction perpendicular to the film. The representation of the wavefunctions inside the muffin-tin spheres remains exactly the same as in the bulk case. Since the periodicity along the z-direction is lost, the

The resulting basis functions have the form

$$\varphi_{\mathbf{G}_{\parallel}G_{\perp}}(\mathbf{k}_{\parallel}, \mathbf{r}) = \{a_{\mathbf{G}_{\parallel}G_{\perp}}(\mathbf{k}_{\parallel})u_{\mathbf{G}_{\parallel}}(\mathbf{k}_{\parallel}, z) + b_{\mathbf{G}_{\parallel}G_{\perp}}(\mathbf{k}_{\parallel})\dot{u}_{\mathbf{G}_{\parallel}}(\mathbf{k}_{\parallel}, z)\} e^{i(\mathbf{G}_{\parallel}+\mathbf{k}_{\parallel})\mathbf{r}_{\parallel}}. \quad (127)$$

The coefficients $a_{\mathbf{G}_{\parallel}G_{\perp}}(\mathbf{k}_{\parallel})$ and $b_{\mathbf{G}_{\parallel}G_{\perp}}(\mathbf{k}_{\parallel})$ are determined in exactly the same way as it is done for the muffin-tins by requiring that the functions are continuous and differentiable at the vacuum boundary. It should be mentioned, that the vacuum basis functions offer less variational freedom than the basis set in the interstitial region does. This can be seen by noting that there are only two functions, $u_{\mathbf{G}_{\parallel}}$ and $\dot{u}_{\mathbf{G}_{\parallel}}$ times the corresponding planar plane-wave, to be matched to all plane-waves of the interstitial region with the same $\mathbf{G}_{\parallel}$. But there are generally far more than two different $G_{\perp}$'s, i.e the number of basis functions in the vacuum region is significantly smaller than in the interstitial region. However, this can be improved rather easily. In Eq. (125) only one energy parameter E_{vac} is used. Instead one can use a whole series of parameters E_{vac}^i to cover an energy region. A possible choice of the energy parameters could be $E_{vac}^i = E_{vac}^{G_{\perp}} = E_{vac} - \frac{\hbar^2}{2m}G_{\perp}^2$, which leads correspondingly to $G_{\perp}$ dependent basis functions $u_{\mathbf{G}_{\parallel}G_{\perp}}(\mathbf{k}_{\parallel}, z)$. For more details see Ref. 76. In general, however, the present approximation is accurate, the energy spectrum of the electrons in the vacuum region is small due to the work-function.

Finally we would like to summarize the basis set used for thin film calculation with the FLAPW method:

$$\varphi_{\mathbf{G}_{\parallel}G_{\perp}}(\mathbf{k}_{\parallel}, \mathbf{r}) = \begin{cases} e^{i(\mathbf{G}_{\parallel}+\mathbf{k}_{\parallel})\mathbf{r}_{\parallel}} e^{iG_{\perp}z} & \text{interstitial} \\ \{a_{\mathbf{G}_{\parallel}G_{\perp}}(\mathbf{k}_{\parallel})u_{\mathbf{G}_{\parallel}}(\mathbf{k}_{\parallel}, z) + b_{\mathbf{G}_{\parallel}G_{\perp}}(\mathbf{k}_{\parallel})\dot{u}_{\mathbf{G}_{\parallel}}(\mathbf{k}_{\parallel}, z)\} e^{i(\mathbf{G}_{\parallel}+\mathbf{k}_{\parallel})\mathbf{r}_{\parallel}} & \text{vacuum} \\ \sum_L a_L^{\mu\mathbf{G}}(\mathbf{k})u_l(r)Y_L(\hat{\mathbf{r}}) + b_L^{\mu\mathbf{G}}(\mathbf{k})\dot{u}_l(r)Y_L(\hat{\mathbf{r}}) & \text{MT}^{\mu} \end{cases} \quad (128)$$

This expansion has been suggested by H. Krakauer, M. Posternak and A. J. Freeman³¹. Correspondingly, the charge density and potential is expanded in the form:

$$n(\mathbf{r}) = \begin{cases} \sum_s n_s \Phi_s^{3D}(\mathbf{r}) & \mathbf{r} \in \text{interstitial region} \\ \sum_s n_s(z) \Phi_s^{2D}(\mathbf{r}) & \mathbf{r} \in \text{vacuum} \\ \sum_{\nu} n_{\nu}^{\mu}(r) K_{\nu}(\hat{\mathbf{r}}) & \mathbf{r} \in \text{MT}^{\mu} \end{cases} \quad (129)$$

and the Hamiltonian and overlap matrix consists now of an additional term (compare to Eq. (55)), the vacuum contribution, paying tribute that the space is now partitioned in three regions

$$H = H_I + H_{MT} + H_V \quad \text{and} \quad S = S_I + S_{MT} + S_V. \quad (130)$$

$n_s(z)\Phi_s^{2D}$ contains important information for the analysis and interpretation of STM topography and spectroscopy results on the basis of the Tersoff-Hamann model⁷⁷ as worked out by Heinze *et al.*⁷⁸.

5.2 The Wire Geometry for Chains, Wires and Tubes

In the FLAPW method for one-dimensional systems⁴⁶, the infinite three-dimensional space is again partitioned into three regions: the muffin-tin spheres around the atoms, the interstitial region between the atoms and within a cylinder along the axis of the wire (z) of

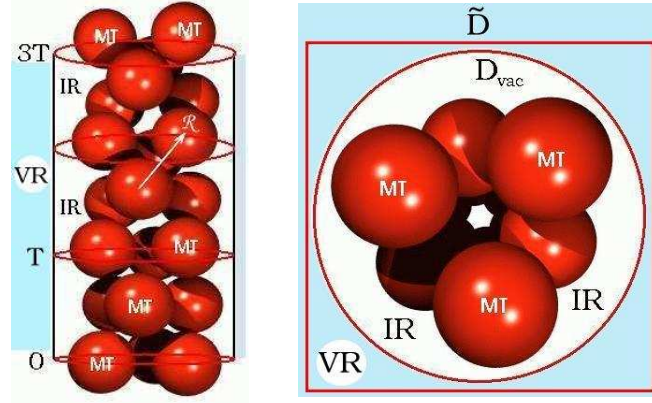


Figure 13. Spatial partitioning of space into muffin-tin spheres (MT), interstitial region (IR) and vacuum region (VR) (shown in blue color) is shown from aside (left) and from the top (right). The vacuum region is the infinite region outside the cylinder with the diameter D_{vac} . In-plane reciprocal vectors $\mathbf{G}_{\parallel}$ are generated in an in-plane square lattice with the lattice constant $\tilde{D} > D_{\text{vac}}$.

the radius R_{vac} . Outside this cylinder there is an infinitely extended vacuum region (VR in Figure 13). From here on we define the z -axis as the axis of one-dimensional translational symmetry. As our method is based on the use of LAPW basis functions,^{19,31,12} the set of reciprocal vectors $\mathbf{G} = (\mathbf{G}_{\parallel}, G_z)$ is generated in a rectangular box, which reflects the translational periodicity of the system in z -direction. The corresponding Bloch number, k_z , lies within the first one-dimensional Brillouin zone. The in-plane reciprocal lattice vectors $\mathbf{G}_{\parallel}$ are generated in an in-plane square lattice with the lattice constant $\tilde{D}$. The vacuum region is an infinite region outside the cylinder with the diameter $D_{\text{vac}} < \tilde{D}$ ($D_{\text{vac}} = 2R_{\text{vac}}$), with the axis along z -direction.

As characteristic for the FLAPW method, optimally adjusted basis functions are used in three different regions of space. In the interstitial region and in the spheres, the usual LAPW basis functions are used. In the vacuum the following representation is used:

$$\varphi_{\mathbf{G}}(k_z, \mathbf{r}) = \sum_m \left(a_m^{\mathbf{G}}(k_z) u_m^{G_z}(k_z, r) + b_m^{\mathbf{G}}(k_z) \dot{u}_m^{G_z}(k_z, r) \right) e^{im\varphi} e^{i(G_z + k_z)z}. \quad (131)$$

The space coordinate $\mathbf{r}$ is written in terms of cylindrical coordinates (r, φ, z) and the summation over m goes up to the angular expansion parameter m_{max} , which ensures that the oscillations of the plane-waves on the cylindrical vacuum boundary continue smoothly to the vacuum side. Since the vacuum potential is rather flat, relativistic effects on the basis functions can safely be ignored, and the cylindrically symmetrical part of the vacuum potential $V_0(r)$ and the vacuum energy parameter E_v , determined in every iteration, enter in solving the radial Schrödinger equation for every pair (m, G_z) giving rise to the vacuum radial basis wavefunctions $u_m^{G_z}(k_z, r)$ and their energy derivatives $\dot{u}_m^{G_z}(k_z, r)$.

The sets of augmentation coefficients a and b both for the MT spheres and the vacuum region are determined such that the basis functions and their spatial derivatives are continuous across the MT spheres, interstitial and vacuum region boundaries. All the basis functions with reciprocal lattice vector $\mathbf{G}$ that fulfill the condition $|k_z + \mathbf{G}| < K_{\text{max}}$ are

included. The corresponding representation of the charge density and potential involves all vectors $\mathbf{G}$ with $|\mathbf{G}| < G_{\max}$. Typically, $G_{\max} \approx 3 \cdot K_{\max}$ in order to describe multiplication of the interstitial potential with the step function. The vacuum parameter $m_{\max}$ is defined in the same manner as $l_{\max}$ in the spheres:¹⁴ $m_{\max} \simeq K_{\max} \cdot R_{\text{vac}}$.

6 Where has the CPU Time Gone?

A thorough understanding where and how much CPU time is spent during a FLAPW calculation is compulsory to judge which problems can be solved with given computers and to develop strategies to overcome possible limitations. Along the discussion of two possible paths to speed-up the calculations put forward in the following two sections, we estimate the CPU time requests for different parts of the FLAPW method as implemented in the FLEUR code³⁰.

6.1 Using Symmetry to Speed-Up Calculations

As surface science is an important field of materials science and condensed matter physics, in this section we compare the very efficient film approach (see Section 5.1) to simulate surfaces in electronic structure calculations with the very popularly used repeated slab model within a super-cell approach, where the unit cell consists of a film of a finite number of layers of a material and empty space describing the vacuum and the decoupling of the films from unit cell to unit cell. To repeat what has been said in Section 5.1, the key fact is, that the film approach reduces the number of LAPW basis functions N_{APW} as compared to the super-cell model.

To illustrate the effect, we choose as reference system a calculation of a 9-layer film of Cu(110) with a single $\mathbf{k}$ -point in a $p(4 \times 2)$ unit cell containing 72 atoms. This system was also chosen in performance tests of the unparallelized WIEN-code^{79,80}, which offers the possibility to gauge different implementations of the same method. Our calculations were performed with equivalent cutoff parameters also on a single node IBM/RS 6000 AIX-SP2 system to facilitate a comparison of the run times. We made no attempt to optimize the code for a certain architecture. In comparison to a repeated slab calculation ($N_{\text{APW}} = 7069$), the film geometry ($\tilde{D} = 30$ a.u.) reduces N_{APW} by approx. 1000 or 14%. This reduces the computational effort in the Hamiltonian setup (scaling with N_{APW}^2) by 27%, in the diagonalization ($\propto N_{\text{APW}}^3$) by 37% (cf. chart in Figure 14). Potential generation (linear with N_{APW}) and charge density construction for the IR (this is done via discrete FFT and scales with $N_{\text{APW}} \ln(N_{\text{APW}})$) are moderately affected.

Our reference system, as well as many other systems of interest, possesses inversion symmetry. In the above calculation, we exploited this fact already, solving a symmetric rather than an hermitian eigenvalue problem, saving time and memory. But a further benefit can be drawn from this symmetry: The augmentation coefficients $P_{L,\mu}^{\mathbf{k},\mathbf{G}}$ ($P = a, b$ cf. Eq. (79)) for an atom μ at a position $\boldsymbol{\tau}_\mu$ are evaluated as

$$P_{L,\mu}^{\mathbf{k},\mathbf{G}} = 4\pi i^l e^{i(\mathbf{k}+\mathbf{G}) \cdot \boldsymbol{\tau}_\mu} f_l(x) Y_L^*(\widehat{\mathbf{k}+\mathbf{G}}), \quad (132)$$

where R_μ is the MT radius and $f_l(x)$ is a linear combination of Bessel functions of argument $x = (k + G)R_\mu$. This time consuming evaluation has to be done for all atoms in the

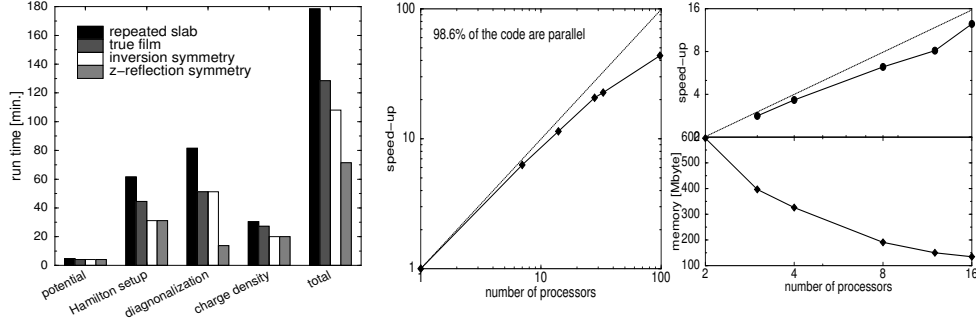


Figure 14. Left: Breakup of the computing time for a calculation of the $p(4 \times 2)$ Cu(110) film. The run-times are given for a repeated slab model, a true film calculation, and the latter with making use of inversion and z-reflection symmetry. The total time of 178 min. per self-consistency cycle should be compared to 174 min. obtained with the optimized WIEN code⁸⁰ at the same computer using standard diagonalization and otherwise similar cutoff parameters. Middle: Speed-up from the parallelization over the $\vec{k}$ -points for a 9-layer film of Fe(100) with 196 $\vec{k}$ -points. Right: Speed-up and memory requirements due to eigenvector parallelization tested on a 9-layer film of $p(4 \times 2)$ Cu(110) with a single $\vec{k}$ -point on a Cray T3E. The memory requirements for the calculation with two processors was extrapolated and a similar value was also found for the unparallelized calculation.

Hamiltonian as well as in the charge density setup. If two equivalent atoms μ and μ' are connected by inversion symmetry, then $-\tau_\mu = \tau_{\mu'}$ and therefore it is easily verified from Eq. (132) that $P_{L,\mu'}^{\mathbf{k},\mathbf{G}} = (-1)^{l+m} (P_{L,-m,\mu}^{\mathbf{k},\mathbf{G}})^*$. This simple relation helps to speed-up both the matrix setups and the charge density generation. In our specific example 44% of the coefficients can be constructed in this way. The actual speed-up gained from this procedure can be seen in the chart of Figure 14.

A film with a mirror plane perpendicular to its surface normal (which shall point in z-direction), i.e. z-reflection symmetry, allows the use of symmetrized basis functions³¹:

$$\chi_{\mathbf{k}_\parallel, \mathbf{G}}^\pm(\mathbf{r}) = \sqrt{2} e^{i(\mathbf{k}_\parallel + \mathbf{G}_\parallel) \cdot \mathbf{r}_\parallel} \begin{cases} \cos(G_z z); (+) \\ \sin(G_z z); (-) \end{cases}, \quad (133)$$

where $\mathbf{G} = (\mathbf{G}_\parallel, G_z)$. Here we exploit the fact, that in film calculations $\mathbf{k} = \mathbf{k}_\parallel$. Therefore, not only the density and potential, but also the basis functions have z-reflection symmetry. This enables a block diagonalization of the eigenvalue problem, decomposing the Hamilton- and overlap matrix into symmetric and antisymmetric blocks. Apart from a small overhead – resulting from sorting the eigenvectors – we expect a speed-up of a factor 4 for a N_{APW}^3 scaling diagonalization routine. From the chart in Figure 14 one finds that this value is almost obtained. In summary, we gained a speed-up of more than a factor 2.5 as compared to a repeated slab calculation where these symmetries were neglected.

Using the same concept for one-dimensional systems, compared to the conventional FLAPW formulation, the 1D-method can be 150 times faster than the supercell approach, and the Hamiltonian construction and the diagonalization part of the 1D code even 270 times faster than that of the bulk super-cell code⁴⁶. As compared to the supercell approach using the film geometry, per self-consistency iteration the 1D-method is then 15 times faster, and the Hamiltonian construction and the diagonalization part of the 1D code is 25 times faster than that of the film super-cell code.

6.2 Parallelization of the Algorithm

Due to the rather complex algorithm of a FLAPW program as compared to a quantum Monte Carlo code for instance, there exists no straight forward parallelization strategy. From the chart in Figure 14 it is evident that there is no single, well-defined part of the program that consumes almost all computer time. Moreover, depending on the actual problem, the computational costs are distributed quite differently: while in large systems most of the time is spent in the diagonalization of the eigenvalue problem, in systems of moderate size other parts of the code might get equally time consuming. But even small systems can become supercomputer applications when the eigenvalue problem has to be solved for a huge number of $\mathbf{k}$ -points, $N_{\mathbf{k}}$, as it is required in many magnetic problems. Applications on workstation clusters require a minimization of communication between the nodes, while on massively parallel machines the memory requirements for the setup of the Hamilton matrix may surpass the resources available on a single processor. This calls for flexible parallelization strategies.

To fulfill adequately the needs of all different kinds of calculations we introduced parallelization on two levels: a coarse grained $\mathbf{k}$ -point parallelization and a fine grained combined eigenvector and eigenvalue parallelization. Depending on the actual values of $N_{\mathbf{k}}$ and the number of processors, N_{pe} , available for the application as well as on the memory-resources on the individual processors, the simultaneous application of both strategies allows a flexible load balancing.

$\mathbf{k}$ -point parallelization

For a calculation of the $p(4 \times 2)$ Cu(110) surface with only one $\mathbf{k}$ -point we find that 92.6% of the time was spent in the $\mathbf{k}$ -dependent part of the code, whereas the rest was used mainly for the potential construction (5.7%). From this we expect that the $\mathbf{k}$ -point parallelization will be a very efficient strategy beyond one $\mathbf{k}$ -point.

In the parallelized version of the FLEUR-code the setup and the potential construction are done on a single processor. All necessary variables are then broadcasted to the other processors and the potential can be read from a file if there is a common file system for all nodes. The eigenvalue problem that has to be solved for all $\mathbf{k}$ -points can now be distributed over the processors and the results (eigenvalues and -vectors) are written to a global direct-access file. If no common file system exists, each node may write the results to a private file and send the eigenvalues back to the node that did the setup and which determines now the Fermi-level. With this result, the occupancy of the eigenvalues can be calculated and the result is sent to all processors. With this information all nodes can calculate a partial charge density (and e.g. forces on the atoms, orbital moments etc.) from the eigenvectors that were calculated in the last step. Up to this point the communication between the nodes was almost negligible, but now the charge densities from all nodes are sent back to and summed up by the first node. Here, the charge density mixing and all necessary output is done that finishes a step of the self-consistency cycle.

Since there is no additional computational effort and a moderate amount of communication between the nodes, the performance is almost ideal. In case of the film containing 72 Cu atoms a parallelization of more than 92% can be expected from the unparallelized calculation. But the performance is also splendid for smaller systems as it is shown in Figure 14. Test was carried out for a 196 $\mathbf{k}$ -point calculation as it occurs typically in the connection with calculations of the magnetic anisotropy calculations where often even a denser sampling of the Brillouin zone is necessary.

Eigenstate parallelization

Due to memory (per node) limitations on massively parallel machines, **k**-point parallelization alone does not suffice to calculate big systems. In this case one distributes the eigenvalue problem for each **k**-point on $N_{\text{pe}}^{\text{ev}}$ nodes.

We found it most useful to adopt a parallel QR-algorithm for the solution of the generalized (symmetric or hermitian) eigenvalue problem that uses the matrices in a column-wise distributed fashion, i.e. a column i of the Hamilton and overlap matrix can be found on the processor with the number $\text{mod}(i, N_{\text{pe}}^{\text{ev}})$. Since our matrices are symmetric or hermitian we calculate only one part of every column, the other part is sent from the other nodes to complete the column. Compared to the unparallelized code, where packed matrices are used, this gives us no big improvement on the use of memory if $N_{\text{pe}}^{\text{ev}} = 2$, but when four processors work on one **k**-point we use only half of the memory per node for the matrices (and they normally use most of the memory in this step). Since the communication is moderate and no additional computational effort arises for the matrix setup, the scaling of this part with the number of processors is almost linear. The parallel QR-algorithm reduces the generalized eigenvalue problem via a Cholesky-factorization (see also Section 2.4) to a normal one and uses a Householder transform to get a tridiagonal matrix. From this matrix the ($\approx 10\%$) lowest eigenvalues are determined and the eigenvector calculation is distributed over the $N_{\text{pe}}^{\text{ev}}$ nodes. Finally, each processor holds approximately the same number of eigenvalue/eigenvector pairs.

For the determination of the charge density there is no conceptual difference whether each node calculates a partial charge density from a subset of **k**-points or from a subset of eigenvalues of a selected **k**-point. Therefore, it is rather easy to implement a parallelized charge-density generator once the complete eigenvectors are available on the nodes. The efficiency of this parallelization strategy for the Cu(110) reference system is shown in the right panel of Figure 14.

The parallelization of FLAPW codes enabled calculations of systems with unprecedented complexity involving unit cells of several hundredth of atoms dealing simultaneously with complex magnetic structures. The strategy of parallelization presented combines the flexibility required to tackle very different problems with transparency in the source-code and an acceptable performance on various parallel computer platforms.

Acknowledgments

This article benefited from discussions and collaborations with Henry Krakauer, Eugene Krasovskii, Philipp Kurz, Yuri Mokrousov, Rositza Pentcheva, Josef Redinger and Mike Weinert, which took place over a long period of time.

References

1. H. Hohenberg and W. Kohn, *Inhomogeneous Electron Gas*, Phys. Rev. **136**, B864 (1964).
2. R. M. Dreizler and J. da Provincia, *Density Functional Methods in Physics* (Plenum, New York, 1985).
3. R. O. Jones and O. Gunnarsson, *The density functional formalism, its applications and prospects*, Rev. Mod. Phys. **61**, 689 (1989).

4. M. Ernzerhof, J. P. Perdew, and K. Burke, *Density Functionals: Where do they come from, why do they work?*, in *Topics in Current Chemistry*, Vol. 180, R.F. Nalewajski Ed. (Springer, Berlin, 1996).
5. A. Zunger and A. J. Freeman, *Ground-state electronic properties of diamond in the local-density formalism*, Phys. Rev. B **15**, 5049 (1977).
6. V. L. Moruzzi, J. F. Janak, and A. R. Williams, *Calculated Electronic Properties of Metals* (Pergamon, New York, 1978).
7. E. Runge and E. K. U. Gross, *Density-Functional Theory for Time-Dependent System*, Phys. Rev. Lett. **52**, 997 (1984).
8. M. Petersilka, U. J. Gossmann, and E. K. U. Gross, *Excitation Energies from Time-Dependent Density-Functional Theory*, Phys. Rev. Lett. **76**, 1212 (1996).
9. J. D. Talman and W. F. Shadwick, *Optimized effective atomic central potential*, Phys. Rev. A **14**, 36 (1976).
10. V. I. Anisimov, A. I. Poteryaev, M. A. Korotin, A. O. Anokhin, and G. Kotliar, *First-principles calculations of the electronic structure and spectra of strongly correlated systems: dynamical mean-field theory*, J. Phys: Condens. Matt. **9**, 7359 (1997).
11. A. I. Lichtenstein and M. I. Katsnelson, *Ab initio calculations of quasiparticle band structure in correlated systems: LDA++ approach*, Phys. Rev. B **57**, 6884 (1998).
12. E. Wimmer, H. Krakauer, M. Weinert, and A. J. Freeman, *Full-potential self-consistent linearized-augmented-plane-wave method for calculating the electronic structure of molecules and surfaces: O₂ molecule*, Phys. Rev. B **24**, 864 (1981).
13. M. Weinert, E. Wimmer, and A. J. Freeman, *Total-energy all-electron density functional method for bulk solids and surfaces*, Phys. Rev. B **26**, 4571 (1982).
14. D. Singh, *Planewaves, Pseudopotentials and the LAPW Method*, (Kluwer Academic Publishers, Boston/Dordrecht/London, 1994).
15. J. C. Slater, *Wave Functions in a Periodic Potential*, Phys. Rev. **51**, 846 (1937).
16. J. C. Slater, *Energy band calculations by the augmented plane wave method*, Advances in Quantum Chemistry **1**, 35 (1964).
17. T. Loucks, *Augmented Plane Wave Method*, (Benjamin, New York, 1967).
18. H. Bross, G. Bohn, G. Meister, W. Schubö, and H. Stöhr, *New version of the modified augmented-plane-wave method*, Phys. Rev. B **2**, 3098 (1970).
19. O. K. Andersen, *Linear methods in band theory*, Phys. Rev. B **12**, 3060 (1975).
20. D. D. Koelling and G. O. Arbman, *Use of energy derivative of the radial solution in an augmented plane wave method: application to copper*, J. Phys. F (Metal Phys.) **5**, 2041 (1975).
21. R. V. Kasowski, *Band Structure of NiS as Calculated Using a Simplified Linear-Combination-of-Muffin-Tin-Orbitals Method*, Phys. Rev. B **8**, 1378 (1973).
22. P. M. Marcus, *Variational Methods in the Computation of Energy Bands*, Int. J. Quantum Chem. Suppl. **1**, 567 (1967).
23. M. Weinert, *Solution of Poisson's equation: beyond Ewald-type methods*, J. Math. Phys. **22**, 2433 (1981).
24. D. R. Hamann, *Semiconductor Charge Densities with Hard-Core and Soft-Core Pseudopotentials*, Phys. Rev. Lett. **42**, 662 (1979).
25. H. J. F. Jansen and A. J. Freeman, *Total-energy full-potential linearized augmented-plane-wave method for bulk solids: Electronic and structural properties of tungsten*, Phys. Rev. B **30**, 561 (1984).

26. S.-H. Wei, H. Krakauer, and M. Weinert, *Linearized augmented-plane-wave calculation of the electronic structure and total energy of tungsten*, Phys. Rev. B **32**, 7792 (1985).
27. S.-H. Wei and H. Krakauer, *Local-density-functional calculation of the pressure-induced-metalization of BaSe and BaTe*, Phys. Rev. Lett. **55**, 1200 (1985).
28. L. F. Mattheiss and D. R. Hamann, *Linear augmented-plane-wave calculation of the structural properties of bulk Cr, Mo, and W*, Phys. Rev. B **33**, 823 (1986).
29. P. Blaha, K. Schwarz, P. Sorantin, and S.B. Trickey, *Full-potential, linearized augmented plane wave programs for crystalline systems*, Comp. Phys. Commun. **59**, 399 (1990).
30. see <http://www.flapw.de>
31. H. Krakauer, M. Posternak, and A. J. Freeman, *Linearized augmented plane-wave method for the electronic band structure of thin films*, Phys. Rev. B **19**, 1706 (1979).
32. D. R. Hamann, L. F. Mattheiss, and H. S. Greenside, *Comparative LCAO-LAPW study of C1 chemisorption on the Ag(001) surface*, Phys. Rev. B **24**, 6151 (1981).
33. E. Wimmer, H. Krakauer, and A. J. Freeman, *Theory of surface electronic structure*, Adv. Electronics Electron Phys. **65**, 337 (1985).
34. In the scalar relativistic approximation (SRA)^{35,81–83} of the full relativistic Kohn–Sham equations the mass-velocity and Darwin terms are included to all orders of $(1/c^2)^n$, where c is the velocity of light, but the spin–orbit interaction is systematically omitted. In this approach the angular momentum quantum number ℓ and the spin quantum number σ remain good quantum numbers.
35. D. D. Koelling and B. N. Harmon, *A technique for relativistic spin-polarised calculations*, J. Phys. C (Solid State Physics) **10**, 3107 (1977).
36. C. Li, A. J. Freeman, H. J. F. Jansen, and C. L. Fu, *Magnetic anisotropy in low-dimensional ferromagnetic systems: Fe monolayers on Ag(001), Au(001), and Pd(001) substrates*, Phys. Rev. B **42**, 5433 (1990).
37. J. M. Soler and A. R. Williams, *Augmented-plane-wave forces*, Phys. Rev. B **42**, 9728 (1990).
38. R. Yu, D. Singh, and H. Krakauer, *All-electron and pseudopotential force calculations using the linearized-augmented-plane-wave method*, Phys. Rev. B **43**, 6411 (1991).
39. S. Blügel, *First Principles Calculations of the Electronic Structures of Magnetic Overlayers on Transition Metal Surfaces*, (Forschungszentrum Jülich, Jül. Report 2197, 1988).
40. D. Singh, *Simultaneous solution of diagonalization and self-consistency problems for transition-metal systems*, Phys. Rev. B **40**, 5428 (1989).
41. G. W. Fernando, *Iterative approaches to electronic structure with augmented bases*, Phys. Rev. B **41**, 903 (1990).
42. R. Wu and A. J. Freeman, *An efficient step-forward way to solve the Schrödinger eigenvalue equation in self-consistent calculations*, Comp. Phys. Comm. **76**, 58 (1993).
43. D. Singh, *Ground-state properties of lanthanum: Treatment of extended-core states*, Phys. Rev. B **43**, 6388 (1991).
44. E. Sjöstedt, L. Nordström, and D. Singh, *An alternative way of linearizing the augmented plane-wave method*, Solid State Commun. **114**, 15 (2004).
45. Ph. Kurz, F. Förster, L. Nordström, G. Bihlmayer, and S. Blügel, *Ab initio treat-*

- ment of noncollinear magnets with the full-potential linearized augmented plane wave method, Phys. Rev. B **69**, 024415 (2004).
46. Y. Mokrousov, G. Bihlmayer, and S. Blügel, *Full-potential linearized augmented plane-wave method for one-dimensional systems: Gold nanowire and iron monowires in a gold tube*, Phys. Rev. B **72**, 045402 (2005).
 47. M. Usuda, N. Hamada, T. Kotani, and M. van Schilfhaarde, *All-electron GW calculation based on the LAPW method: Application to wurtzite ZnO*, Phys. Rev. B **66**, 125101 (2002).
 48. D. Wortmann, H. Ishida, and S. Blügel, *An ab initio Green-function formulation of the transfer matrix: Application to complex bandstructures*, Phys. Rev. B **65**, 165103 (2002).
 49. D. Wortmann, H. Ishida, and S. Blügel, *An embedded Green-function approach to the ballistic electron transport through an interface*, Phys. Rev. B **65**, 165103 (2002).
 50. E. E. Krasovskii, *Augmented-plane-wave approach to scattering of Bloch electrons by an interface*, Phys. Rev. B **70**, 245322 (2004).
 51. M. L. Cohen, *Electronic structure of solids*, Phys. Rep. **110**, 293 (1984).
 52. W. E. Pickett, *Pseudopotential methods in condensed matter applications*, Comp. Phys. Rep. **9**, (1)15 (1989).
 53. H. Akai, M. Akai, S. Blügel, B. Drittler, H. Ebert, K. Terakura, R. Zeller, and P. H. Dederichs, *Theory of Hyperfine Interactions in Metals*, Prog. Theo. Phys. (Suppl) **101**, 11 (1990).
 54. M. Wuttig, Y. Gauthier, and S. Blügel, *Magnetically Driven Buckling and Stability of Ordered Surface Alloys: Cu(100) c(2 × 2) Mn*, Phys. Rev. Lett. **70**, 3619 (1993).
 55. W. Kohn and L.J. Sham, *Self-Consistent Equations Including Exchange and Correlation Effects*, Phys. Rev. **140**, A1133 (1965).
 56. S. Blügel, *First Principles Calculations of the Electronic Structures of Magnetic Overlayers on Transition Metal Surfaces*, (PhD thesis, RWTH Aachen, 1988).
 57. L. Hedin and B. I. Lundqvist, *Explicit local exchange-correlation potentials*, J. Phys. C (Solid State Physics) **4**, 2064 (1971).
 58. S. H. Vosko, L. Wilk, and N. Nusair, *Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis*, Can. J. Phys. **58**, 1200 (1980).
 59. J. P. Perdew, J. A. Chevary, S. H. Vosko, K. A. Jackson, M. R. Pederson, D. J. Singh, and C. Fiolhais, *Atoms, molecules, solids, and surfaces: Applications of the generalized gradient approximation for exchange and correlation*, Phys. Rev. B **46**, 6671 (1992).
 60. J. P. Perdew, K. Burke, and M. Ernzerhof, *Generalized Gradient Approximation Made Simple*, Phys. Rev. Lett. **77**, 3865 (1996).
 61. J. Stoer, *Numerische Mathematik 1*, (Springer-Verlag, Berlin, 1994).
 62. D. J. Chadi and Marvin L. Cohen, *Special Points in the Brillouin Zone*, Phys. Rev. B **8**, 5747 (1973).
 63. S. L. Cunningham, *Special points in the two-dimensional Brillouin zone*, Phys. Rev. B **10**, 4988 (1974).
 64. O. Jepson and O. K. Andersen, *The electronic structure of h.c.p. Ytterbium*, Solid State Commun. **9**, 1763 (1971).
 65. G. Lehmann and M. Taut, *On the Numerical Calculation of the Density of States and*

- Related Properties*, phys. stat. sol. (b) **54**, 469 (1972).
66. P. E. Blöchl, O. Jepsen, and O. K. Andersen, *Improved tetrahedron method for Brillouin-zone integrations*, Phys. Rev. B **49**, 16223 (1994).
 67. R. Pentcheva, *Ab-initio Rechnungen zum Zusammenhang zwischen Magnetismus und Struktur ultradünner Filme*, (Diploma thesis, RWTH Aachen, 1995).
 68. A. R. Williams, J. Kübler, and C. D. Gelatt Jr., *Cohesive properties of metallic compounds: Augmented-Spherical-Wave calculations*, Phys. Rev. B **19**, 6094 (1979).
 69. T. Takeda and J. Kübler, *Linear augmented plane wave method for self-consistent calculations*, J. Phys. F **9**, 661 (1979).
 70. S. Goedecker, *Treatment of semicore states in the linearized augmented-plane-wave method and other linearized electronic-structure methods*, Phys. Rev. B **47**, 9881 (1993).
 71. J. Yu, A. J. Freeman, R. Podloucky, P. Herzig, and P. Weinberger, *Origin of electric-field gradients in high-temperature superconductors: $\text{YBa}_2\text{Cu}_3\text{O}_7$* , Phys. Rev. B **43**, 532 (1991).
 72. E. E. Krasovskii and W. Schattke, *The extended-LAPW-based $\mathbf{k}\cdot\mathbf{p}$ method for complex band structure calculations*, Solid State Comm. **93**, 775 (1995).
 73. E. E. Krasovskii, *Accuracy and convergence properties of the extended linear augmented-plane-wave method*, Phys. Rev. B **56**, 12866 (1997).
 74. G. K. H. Madsen, P. Blaha, K. Schwarz, E. Söstedt, and L. Nordström, *Efficient linearization of the augmented plane-wave method*, Phys. Rev. B **64**, 195134 (2001).
 75. J. D. Jackson, *Klassische Elektrodynamik* (Walter de Gruyter, Berlin, New York, 1983).
 76. W. Ning, C. Kailai, and W. Dingsheng, *Work Function of Transition-Metal Surface with Submonolayer Alkali-Metal Coverage*, Phys. Rev. Lett. **56**, 2759 (1986).
 77. J. Tersoff and D. R. Hamann, *Theory and Application for the Scanning Tunneling Microscope*, Phys. Rev. Lett. **50**, 1998 (1983).
 78. S. Heinze, S. Blügel, R. Pascal, M. Bode, R. Wiesendanger, *Prediction of Corrugation Reversal in STM-images of bcc-(110)-surfaces: $\text{W}(110)$, $\text{Ta}(110)$, $\text{Fe}(110)$* , Phys. Rev. B **58**, 16432 (1998).
 79. P. Blaha, K. Schwarz, P. Dufek, and R. Augustyn, **WIEN95** (Technical University, Vienna, 1995).
 80. M. Petersen, F. Wagner, L. Hufnagel, and M. Scheffler, *Improving the efficiency of FP-LAPW calculations* Comp. Phys. Comm. **126**, 294 (2000).
 81. T. Takeda, *The scalar relativistic approximation*, Z. Physik B **43**, 32 (1978).
 82. J. H. Wood and A. M. Boring, *Improved Pauli Hamiltonian for local-potential problems*, Phys. Rev. B **18**, 2701 (1978).
 83. H. Gollisch and L. Fritsche, *Relativistic one-particle equation for electron states of heavy metals*, phys. stat. sol. (b) **86**, 145 (1978).

The Korringa-Kohn-Rostoker (KKR) Green Function Method

I. Electronic Structure of Periodic Systems

Phivos Mavropoulos¹ and Nikos Papanikolaou²

¹ Institute of Solid State Research
Forschungszentrum Jülich
52425 Jülich, Germany
E-mail: ph.mavropoulos@fz-juelich.de

² Institute of Microelectronics
NCSR Demokritos, GR-15310 Athens, Greece

The Korringa-Kohn-Rostoker (KKR) method for the calculation of the electronic structure of materials is founded on the concepts of the Green function and of multiple-scattering. In this manuscript, after a short introduction to Green functions, we present a description of single-site scattering (including anisotropic potentials) and multiple-scattering theory and the KKR equations. The KKR representation of the Green function and the algebraic Dyson equation are introduced. We then discuss the screened KKR formalism, and its advantages in the numerical effort for the calculation of layered systems. Finally we give a summary of the self-consistency algorithm for the calculation of the electronic structure.

1 Introduction and Historical Survey

The multiple-scattering method of Korringa, Kohn and Rostoker for the calculation of the electronic structure of materials was introduced in 1947 by Korringa¹ and in 1954 by Kohn and Rostoker.² Characteristic of this method is the use of *multiple-scattering theory* for the solution of the Schrödinger equation and the determination of the electron band structure. In such an approach, the scattering properties of each scattering center (atom) are determined in a first step and described by a scattering matrix, while the multiple-scattering by all atoms in the lattice is determined in a second step by demanding that the incident wave at each center is the sum of the outgoing waves from all other centers. In this way, a separation between the potential and geometric properties is achieved.

A further significant development of the KKR scheme came when it was reformulated as a Green function method.^{3,4} Once more separating the single-site scattering problem from the multiple-scattering effects, the method is able to produce the crystal Green function efficiently by relating it to the Green function of free space via the Dyson equation. In a second step the crystal Green function can be used as a reference in order to calculate the Green function of an impurity in the crystal,⁵ again via a Dyson equation. This way of solving the impurity problem is extremely efficient, avoiding the construction of huge supercells which are needed in wavefunction methods.

Chemically disordered alloys can also be treated by the KKR method within the coherent potential approximation (CPA).⁷ In this approximation one defines the Green function of an average crystal medium, determined self-consistently through the condition that the concentration average of the various atom types should not produce any scattering in this medium.

The development of *screened*, or *tight-binding*, KKR was a further breakthrough for the numerical efficiency of the method.⁸ Via a transformation of the reference system remote lattice sites are decoupled, and the principal layer technique allows the calculation time to scale linearly with the number of atoms. This is especially useful for layered systems (surfaces, interfaces, multilayers) and allows the study of, e.g., interlayer exchange coupling or ballistic transport through junctions.

Transport properties have also been addressed within the KKR method, since the computation Green function allows for approaches of the Kubo-Greenwood type (for diffusive transport)⁹ or the Landauer-Büttiker-type (for ballistic transport).¹⁰ Combined with the Boltzmann equation⁶ the method for solving the impurity problem is also ideally suited for the calculation of transport properties in dilute metallic alloys.

In recent years, an attempt for *ab initio* calculations beyond density functional theory (DFT) has led to the development of hybrid theories which combine the local density approximation (LDA) to DFT with dynamical mean-field theory (DMFT). In these “LDA+DMFT” schemes the Green function is a central quantity, since the electron self-energy is evaluated diagrammatically and included in the propagators through a Dyson equation. Therefore the KKR Green function formalism is well suited for further development in this direction, and steps have already been taken to include DMFT in the KKR scheme.¹¹

A short list of successful applications of the KKR method for electronic structure of solids, combined with density functional theory, includes bulk materials¹², surfaces¹³, interfaces and tunnel junctions¹⁴, and impurities in bulk and on surfaces.¹⁵ Spectroscopic properties¹⁶ and transport properties^{17–19} have also been studied within this method. The KKR scheme can incorporate the Dirac equation, whenever relativistic effects become important.²⁰

Since the KKR method is concerned with the propagation of waves through an array of scatterers, its applications are not restricted to the field of the electronic structure of solids. The propagation of, e.g., electromagnetic waves in photonic crystals,²¹ or the propagation of elastic waves through photonic crystals²² can be efficiently described by multiple-scattering theory, and the useful features of KKR, such as the CPA, can be also generalized for these cases.²³

In this manuscript we present the basics of the KKR formalism. After a general introduction to Green functions in Section 2, we give an analysis of the single-site scattering problem in Section 3. We continue with a detailed presentation of multiple-scattering theory (Section 4), where the basic ideas are given independently of a Green function formulation; we then attend to the KKR representation of the Green function in Section 5. Details on the single-site scattering problem by non-spherical potentials are given in Section 6, while the method for total energy calculations is summarized in Section 7. We carry on with the basics of screened KKR in Section 8, and briefly show how to calculate surfaces or interfaces in Section 9. Finally, we discuss the self-consistency procedure, including the idea of complex-energy contour integration, in Section 10. The reader is assumed to know the basics of DFT, of the LDA, and of electronic structure calculations. These are not essential for the multiple-scattering formalism, but become important in Sections 7 and 10. The solution of the impurity problem within the KKR Green function method is described in the manuscript of Peter H. Dederichs (see page 279).

2 Definition and General Properties of the Green Function

In this section we give a brief introduction to Green functions in scattering problems. The purpose of the introduction is to remind of the concepts and formulas which will be used in the subsequent sections. For more details, including mathematical rigor, we refer to the literature, e.g., to the book by Newton²⁴ or by Economou.²⁵

2.1 Time-Dependent Green Functions

Consider an electron under the influence of a potential $V(\vec{r})$. The Hamiltonian is

$$H = -\nabla^2 + V(\vec{r}) = H_{\text{free}} + V(\vec{r}). \quad (1)$$

Here, and in the following, we use atomic units ($\hbar = 1$, $m_e = \frac{1}{2}$, $e = -\sqrt{2}$). The time-dependent Schrödinger equation, determining the time evolution of the electron wavefunction $\psi(t)$, is

$$i\frac{\partial}{\partial t}\psi(t) = H\psi(t) \quad (2)$$

with formal solution

$$\psi(t) = e^{-iHt}\psi(0), \quad (3)$$

and with $\psi(0)$ representing the wavepacket preparation as an initial condition. The quantity $U(t) = e^{-iHt}$ is the time evolution operator in quantum mechanics.

Corresponding to the Schrödinger equation, we define now two propagators, the *retarded Green function* $G^R(t)$ and the *advanced Green function* $G^A(t)$, as

$$\left(i\frac{\partial}{\partial t} - H\right)G^{R,A}(t) = \delta(t) \quad (4)$$

with boundary conditions

$$G^R(t) = 0 \quad t < 0 \quad (5)$$

$$G^A(t) = 0 \quad t > 0. \quad (6)$$

The formal solution of these equations is

$$G^R(t) = \begin{cases} -ie^{-iHt} & t > 0 \\ 0 & t < 0 \end{cases} \quad (7)$$

$$G^A(t) = \begin{cases} 0 & t > 0 \\ ie^{-iHt} & t < 0 \end{cases}. \quad (8)$$

Evidently, the Green functions coincide, up to a factor, with the time evolution operator; $G^R(t)$ is used to propagate the wavefunction forward in time, and $G^A(t)$ to propagate the wavefunction backward in time:

$$\psi(t) = iG^R(t-t')\psi(t') \quad t' < t \quad (9)$$

$$\psi(t) = -iG^A(t-t')\psi(t') \quad t < t'. \quad (10)$$

In the following we shall restrict the discussion to the retarded Green function G^R and drop the index R .

Given a perturbing potential $\Delta V(\vec{r})$ added to a Hamiltonian H_0 , it follows from Eq. (5) that the Green function G_1 corresponding to the new Hamiltonian $H_1 = H_0 + \Delta V$ is related to the Green function G_0 corresponding to H_0 via the Dyson integral equation

$$G_1(t) = G_0(t) + \int_0^t G_0(t-t') \Delta V G_1(t') dt'. \quad (11)$$

Furthermore, it can be proven that an incoming wavepacket ψ_0 , which without the interaction with ΔV would evolve as $\psi_0(t)$, evolves into the wavefunction $\psi(t)$ as

$$\psi(t) = \psi_0(t) + \int_{-\infty}^t G_0(t-t') \Delta V \psi(t') dt'; \quad (12)$$

this is the *Lippmann-Schwinger equation*.

2.2 Energy-Dependent Green Functions

A Fourier transform of the Green function,

$$G(E) = \int_{-\infty}^{\infty} G(t) e^{i(E+i\varepsilon)t} dt \quad (13)$$

results in the formal solution

$$G(E) = (E + i\varepsilon - H)^{-1}. \quad (14)$$

Here, ε is a positive real number, to be taken in the end to the limit $\varepsilon \rightarrow 0$, which ensures convergence of the integral in Eq. (13) for $t \rightarrow \infty$. More generally, one may define the (time-independent) Green function as the resolvent of the time-independent Schrödinger equation, via the operator equation

$$G(E) = (E - H)^{-1} \quad (15)$$

for an arbitrary complex energy E (as long as $(E - H)$ can be inverted). The singularities of $G(E)$ determine the eigenvalue spectrum; in particular, $G(E)$ has poles at the eigenenergies of the bound states, and a branch cut along the energies of the continuous spectrum. For $\text{Im}E > 0$, $G(E)$ is an analytical function of E . In terms of a complete set of eigenfunctions of H , $|\psi_i\rangle$, corresponding to eigenvalues ϵ_i , the following spectral representation can be obtained:

$$G(E) = \sum_i \frac{|\psi_i\rangle\langle\psi_i|}{E - \epsilon_i}. \quad (16)$$

Represented in real space, the above equation becomes

$$G(\vec{r}, \vec{r}'; E) = \sum_i \frac{\psi_i(\vec{r}) \psi_i^*(\vec{r}')}{E - \epsilon_i} \quad (17)$$

representing, in the limit of $\text{Im}E = \varepsilon \rightarrow 0^+$, an outgoing wave at $\vec{r}$ with a source term at $\vec{r}'$. From the above equation one can see that the imaginary part of G is directly related to the spectrally- and space-resolved density of states $n(\vec{r}; E)$ (for real E):

$$n(\vec{r}; E) = -\frac{1}{\pi} \text{Im} G(\vec{r}, \vec{r}; E). \quad (18)$$

This follows from the Dirac identity,

$$\int_{-\infty}^{\infty} \frac{f(x)}{x - x_0 \pm i\varepsilon} dx = P \left(\int_{-\infty}^{\infty} \frac{f(x)}{x - x_0} dx \right) \mp i\pi f(x_0) \quad (19)$$

where P stands for the Cauchy principal part of the integral. One can deduce from Eq. (18) an expression for the spectral density of states,

$$n(E) = -\frac{1}{\pi} \text{Im} \int G(\vec{r}, \vec{r}; E) d^3r = -\frac{1}{\pi} \text{Im} \text{Tr} G(E), \quad (20)$$

where the last step stresses that the trace of the operator $G(E)$ can be taken in any basis, not just in real-space representation. On the other hand, the charge density is found as an integral of $n(\vec{r}; E)$ over the energies up to the Fermi level:^a

$$\rho(\vec{r}) = -\frac{1}{\pi} \text{Im} \int_{-\infty}^{E_F} G(\vec{r}, \vec{r}; E) dE = -\frac{1}{\pi} \text{Im} \int_{-\infty}^{E_F} \text{Tr} (\hat{r} G(E)) dE \quad (21)$$

with $\hat{r} = |\vec{r}\rangle \delta(\vec{r} - \vec{r}') \langle \vec{r}'|$ is the position operator.

In general, from the spectral representation (Eq. (16)) it follows that the expectation value of any physical quantity, represented by an operator $\hat{A}$, can be harvested via the relation

$$\langle A \rangle = -\frac{1}{\pi} \text{Im} \int_{-\infty}^{E_F} \text{Tr} [\hat{A} G(E)] dE. \quad (22)$$

Therefore, the Green function contains all information which is given by the eigenfunctions. If the Green function can be computed, then all physical properties of the system can be found.

2.3 Relation between Perturbed and Unperturbed System. The T -Matrix

The time-dependent Dyson equation (11) has its energy-dependent counterpart. Given from Eq. (15) that for the reference system $G_0^{-1}(E) = E - H_0$, while for the perturbed system $G_1^{-1}(E) = E - (H_0 + \Delta V)$, we conclude that the two Green functions are connected via

$$G_1^{-1}(E) = G_0^{-1}(E) - \Delta V. \quad (23)$$

The Dyson equation can be rewritten in the following forms (which can be directly verified by substitution in Eq. (23)):

$$G_1(E) = [1 - G_0(E) \Delta V]^{-1} G_0(E) \quad (24)$$

$$= G_0(E) [1 - \Delta V G_0(E)]^{-1} \quad (25)$$

$$= G_0(E) + G_0(E) \Delta V G_1(E). \quad (26)$$

The last expression, although not in a closed form, reminds of expression in Eq. (11) and allows for an interpretation of the Dyson equation via multiple-scattering events by the perturbing potential ΔV if we expand G_1 in the right-hand side:

$$G_1(E) = G_0(E) + G_0(E) \Delta V G_0(E) + G_0(E) \Delta V G_0(E) \Delta V G_0(E) + \dots; \quad (27)$$

^aThis is of course valid for fermions at $T = 0$; otherwise a convolution with the Fermi function, or the appropriate distribution function, is needed.

this is the analogue of a Born series expansion for the Green function.

Also the Lippmann-Schwinger equation (12) has a time-independent counterpart. Observing that the Schrödinger equation for the perturbed system can be written as

$$(E - H) |\psi_1\rangle = \Delta V |\psi_1\rangle, \quad (28)$$

we can verify by substitution that the solution $|\psi_1\rangle$ can be written in terms of the unperturbed eigenstates $|\psi_0\rangle$ as

$$|\psi_1\rangle = |\psi_0\rangle + G_0(E) \Delta V |\psi_1\rangle. \quad (29)$$

Expanding $|\psi_1\rangle$ on the right-hand side of this Lippmann-Schwinger equation leads to the Born series

$$|\psi_1\rangle = |\psi_0\rangle + G_0(E) \Delta V |\psi_0\rangle + G_0(E) \Delta V G_0(E) \Delta V |\psi_0\rangle + \dots \quad (30)$$

In the case that E does not belong to the spectrum of the unperturbed Hamiltonian H , $|\psi_0\rangle = 0$ and we have

$$|\psi_1\rangle = G_0(E) \Delta V |\psi_1\rangle. \quad (31)$$

This expression can be used to obtain the discrete spectrum of H_1 .

We now introduce the transition matrix, or T -matrix, of a scattering system. It relates the states $|\psi_1\rangle$ of the perturbed Hamiltonian to the states $|\psi_0\rangle$ of the unperturbed system via the defining equation

$$\Delta V |\psi_1\rangle = T(E) |\psi_0\rangle. \quad (32)$$

In terms of the T -matrix the Lippmann-Schwinger equation (29) is written as

$$|\psi_1\rangle = |\psi_0\rangle + G_0(E) T(E) |\psi_0\rangle, \quad (33)$$

the Dyson equation (26) as

$$G_1(E) = G_0(E) + G_0(E) T(E) G_0(E), \quad (34)$$

while the following relations can be easily verified:

$$T(E) = \Delta V [1 - G_0(E) \Delta V]^{-1} \quad (35)$$

$$= [1 - \Delta V G_0(E)]^{-1} \Delta V \quad (36)$$

$$= \Delta V + \Delta V G_0(E) T(E). \quad (37)$$

Expanding the right-hand side of the last equation leads once more to a Born series:

$$T(E) = \Delta V + \Delta V G_0(E) \Delta V + \Delta V G_0(E) \Delta V G_0(E) \Delta V + \dots, \quad (38)$$

while for small perturbations we are led to the first-order Born approximation, $T \simeq \Delta V$.

Via the T -matrix one can express the change in the integrated density of states, $\Delta N(E)$, between the perturbed and the unperturbed system. From Eq. (20) it follows that the change in the density of states is

$$\Delta n(E) \equiv n_1(E) - n_0(E) = -\frac{1}{\pi} \text{ImTr} [G_1(E) - G_0(E)]. \quad (39)$$

Eq. (15) gives us, on the other hand, the identity $G = \partial(\ln G^{-1})/\partial E$. Combining this with (39) we proceed as

$$\Delta n(E) = -\frac{1}{\pi} \text{Im Tr} \frac{\partial}{\partial E} [\ln G_1^{-1} - \ln G_0^{-1}] = -\frac{1}{\pi} \text{Im Tr} \frac{\partial}{\partial E} \ln [G_0 G_1^{-1}] \quad (40)$$

$$= -\frac{1}{\pi} \text{Im} \frac{\partial}{\partial E} \text{Tr} \ln [1 - G_0 \Delta V] \quad (41)$$

$$= -\frac{1}{\pi} \text{Im} \frac{\partial}{\partial E} \text{Tr} \ln [T^{-1} \Delta V]. \quad (42)$$

In the last two steps we used the Dyson equation (24) and the property (Eq. (35)) of the T -matrix. The integrated change in the density of states is thus

$$\Delta N(E) = \int_{-\infty}^E \Delta n(E') dE' = -\frac{1}{\pi} \text{Im Tr} \ln [T^{-1}(E) \Delta V]_{-\infty}^E. \quad (43)$$

By splitting the logarithm as $\ln(T^{-1}(E) \Delta V) = \ln T^{-1}(E) + \ln \Delta V$ we see that the contribution of ΔV drops out.^b Since the lower integration limit ($-\infty$) is below the spectrum of the Hamiltonian, it also gives no contribution. We are thus left with the result

$$\Delta N(E) = \frac{1}{\pi} \text{Im Tr} \ln T(E). \quad (44)$$

In the following section we will associate the perturbation ΔV with a spherically symmetric potential of finite range in free space. Then we have a single-site scattering problem, and we will denote the T -matrix as t ; in angular-momentum representation it is diagonal and its elements are related to the scattering phase shifts $\delta_l(E)$ via

$$t_l = -\frac{1}{\sqrt{E}} \sin(\delta_l(E)) e^{i\delta_l(E)}.$$

In this special case, Eq. (44) becomes

$$\Delta N(E) = \frac{1}{\pi} \sum_l (2l+1) \delta_l(E). \quad (46)$$

If the scattering potential represents an impurity in a free-electron host, then the integrated difference of density of states up to the Fermi level, $\Delta N(E_F)$, must be equal to the additional charge introduced by the impurity, $Z_{\text{imp}} - Z_{\text{host}}$. This is expressed by the *Friedel sum rule*. Including a factor 2 for the spin degeneracy, the rule reads:

$$Z_{\text{imp}} - Z_{\text{host}} = \frac{2}{\pi} \sum_l (2l+1) \delta_l(E_F). \quad (47)$$

3 Single-Site Scattering

We turn now to the scattering problem of a spherical atomic potential embedded in free space (actually in an environment of constant potential). The reference system is thus a free-electron system (where the Hamiltonian contains only the kinetic energy term, and

^bThis is because ΔV does not depend on the energy; in systems where the perturbation is energy-dependent, e.g., a self-energy, it must be accounted for.

the eigenfunctions are plain waves). The Green function of a free-electron system has the form^{24, 25}

$$g(\vec{r}, \vec{r}'; E) = -\frac{1}{4\pi} \frac{e^{ik|\vec{r}-\vec{r}'|}}{|\vec{r}-\vec{r}'|}. \quad (48)$$

with $k = \sqrt{E}$. In the case of scattering by a central potential, it is useful to work in angular-momentum representation. We therefore represent an incoming plain wave as

$$\psi_{\vec{k}}^{\text{inc}}(\vec{r}) = e^{i\vec{k}\vec{r}} = \sum_L 4\pi i^l j_l(\sqrt{E}r) Y_L(\vec{r}) Y_L(\vec{k}) \quad (49)$$

where j_l is the spherical Bessel function, while Y_L are the real spherical harmonics (see Appendix 10). We use the combined index $L := (l, m)$, where l and m are angular momentum indexes. On the other hand, the free-space Green function (48) can be expanded as:

$$g(\vec{r}, \vec{r}'; E) = \sum_L Y_L(\vec{r}) g_l(r, r'; E) Y_L(\vec{r}') \quad (50)$$

with

$$g_l(r, r'; E) = -i\sqrt{E} j_l(\sqrt{E}r_<) h_l(\sqrt{E}r_>), \quad (51)$$

where $h_l = j_l + i n_l$ are spherical Hankel functions, n_l are spherical Neumann functions,²⁶ while $r_< (r_>)$ is the smaller (bigger) of the radii r and r' . The Bessel functions $j_l(x)$ are finite as $x \rightarrow 0$, while $h_l(x)$ and $n_l(x)$ are diverging as $x \rightarrow 0$.

Suppose now that there is a scattering potential of finite range, of the form

$$V(\vec{r}) = \begin{cases} V(r) & r \leq S \\ 0 & r > S. \end{cases} \quad (52)$$

Then the radial wavefunctions $R_l(r; E)$ satisfy the radial Schrödinger equation

$$\left[-\frac{1}{r} \frac{\partial^2}{\partial r^2} r + \frac{l(l+1)}{r^2} + V(r) - E \right] R_l(r; E) = 0. \quad (53)$$

The asymptotic form of $R_l(r; E)$ for $r \rightarrow \infty$ is

$$R_l(r; E) \rightarrow \frac{A_l}{\sqrt{E}r} \sin \left(\sqrt{E}r - \frac{l\pi}{2} + \delta_l(E) \right) \quad (54)$$

where A_l is independent of r and $\delta_l(E)$ is the phase shift with respect to the wavefunction for vanishing potential.

For $r > S$, $V(r) = 0$ and the general solution of the radial equation is a sum of two linearly independent special solutions:

$$R_l(r; E) = B_l j_l(\sqrt{E}r) + C_l n_l(\sqrt{E}r) \quad (55)$$

where B_l and C_l are constants. Using the asymptotic form of Bessel functions,

$$\lim_{x \rightarrow \infty} j_l(x) = \frac{1}{x} \sin(x - l\pi/2) \quad (56)$$

$$\lim_{x \rightarrow \infty} n_l(x) = -\frac{1}{x} \cos(x - l\pi/2) \quad (57)$$

together with Eqs. (54) and (55), we obtain

$$R_l(r; E) = A_l \left(j_l(\sqrt{E}r) \cos \delta_l - n_l(\sqrt{E}r) \sin \delta_l \right) \quad \text{for } r > S. \quad (58)$$

On the other hand, the Lippmann-Schwinger equation gives

$$R_l(r; E) = j_l(\sqrt{E}r) + \int_0^S g_l(r, r'; E) V(r') R_l(r'; E) r'^2 dr' \quad (59)$$

which, using Eq. (51), yields for $r > S$:

$$R_l(r; E) = j_l(\sqrt{E}r) - i h_l(\sqrt{E}r) \sqrt{E} \int_0^S j_l(\sqrt{E}r') V(r') R_l(r'; E) r'^2 dr'. \quad (60)$$

The integral is just the t -matrix element in angular-momentum representation (see Eq. (32)),

$$t_l(E) = \int_0^S j_l(\sqrt{E}r) V(r) R_l(r; E) r^2 dr, \quad (61)$$

so that we obtain

$$R_l(r; E) = j_l(\sqrt{E}r) - i \sqrt{E} t_l(E) h_l(\sqrt{E}r) \quad (r > S) \quad (62)$$

with the t -matrix related to the phase shift by (cf. (58))

$$t_l(E) = -\frac{1}{\sqrt{E}} \sin \delta_l(E) e^{i\delta_l(E)}. \quad (63)$$

An incoming wave can be expanded in Bessel functions and spherical harmonics as (cf. (49))

$$\psi_{\vec{k}}^{\text{inc}}(\vec{r}) = \sum_L a_{\vec{k}L}^0 j_l(\sqrt{E}r) Y_L(\vec{r}) \quad (\text{for } r > S). \quad (64)$$

The scattered wave can be expanded analogously as:

$$\psi_{\vec{k}}^{\text{sc}}(\vec{r}) = \sum_L a_{\vec{k}L}^{\text{sc}} j_l(\sqrt{E}r) Y_L(\vec{r}) \quad (\text{for } r > S). \quad (65)$$

Within the range S of the potential we have

$$\psi_{\vec{k}}(\vec{r}) = \sum_L a_{\vec{k}L} R_l(r; E) Y_L(\vec{r}) \quad (66)$$

with the boundary conditions of R_l given by (62). From the continuity of the wavefunction at $r = S$, we obtain then

$$a_{\vec{k}L}^{\text{sc}} = -i \sqrt{E} t_l(E) a_{\vec{k}L}^0. \quad (67)$$

Finally, we give without proof the Green function for the scattering problem by a central potential. It can be written as the product of two linearly independent solutions, R_l (regular, i.e., converging at $r \rightarrow 0$) and H_l (irregular, i.e., diverging at $r \rightarrow 0$), of the radial equation:

$$\begin{aligned} G(\vec{r}, \vec{r}'; E) &= -i \sqrt{E} \sum_L R_l(r_{<}; E) H_l(r_{>}; E) Y_L(\vec{r}) Y_L(\vec{r}') \\ &\equiv \sum_L G_l(r, r'; E) Y_L(\vec{r}) Y_L(\vec{r}'). \end{aligned} \quad (68)$$

The boundary conditions of R_l are given by Eq. (62). In order to find the boundary conditions of H_l we use the identity (34) $G_l = g_l + g_l t_l g_l$ with g_l given by (51) and obtain

$$H_l(r; E) = h_l(\sqrt{E}r) \quad (r > S). \quad (69)$$

In practice, we integrate Eq. (53) numerically outwards from the origin to $r = S$ in order to obtain R_l . At $r = S$ the requirement for continuity of the logarithmic derivative yields the t -matrix elements. Analogously, a numerical integration inwards yields the diverging radial wavefunction H_l , given the boundary condition (69) at $r = S$.

4 Multiple-Scattering Theory

In the previous section we discussed the solution of the scattering problem for an isolated scattering potential. In this section we will extend the study to a set of scatterers. At this point we assume the *muffin-tin* (MT) approximation to the scattering potential, according to which the potential is spherically symmetric around each scattering center (atomic site), within a sphere of radius R_{MT} , and constant otherwise; the spheres can be touching each other, but are assumed to be non-overlapping.^c The theory presented here is based on the bookkeeping of the scattering amplitudes in a periodic array of scatterers, so that a secular equation connecting the amplitudes of incoming and outgoing waves emerges. The procedure leads to the KKR equations for the band structure of periodic crystals. Although the procedure is not necessary for the KKR Green function theory, it provides physical insight to the ideas founding multiple-scattering theory, used in many wave-scattering problems even when a Green function is not introduced. The connection to the KKR representation of the Green function will be done in the next section.

Assume a periodic structure of MT potentials, centered at lattice sites $\vec{R}^n$. The constant potential in the interstitial can be taken to be zero. An outgoing wave from site $\vec{R}^n$ can be expanded in incoming waves at site $\vec{R}^{n'}$ by employing a theorem for the transformation of Hankel functions:

$$h_L(\vec{r}' + \vec{R}^{n'} - \vec{R}^n; E) = \frac{i}{\sqrt{E}} \sum_{L'} g_{LL'}^{nn'}(E) j_{L'}(\vec{r}'; E), \quad (70)$$

where we have used the abbreviations

$$j_L(\vec{r}'; E) = j_l(\sqrt{E}r) Y_L(\vec{r}) \quad \text{and} \quad h_L(\vec{r}'; E) = h_l(\sqrt{E}r) Y_L(\vec{r}). \quad (71)$$

The expansion coefficients, also called *structure constants*, are given by

$$g_{LL'}^{nn'}(E) = -(1 - \delta_{nn'}) 4\pi i \sqrt{E} \sum_{L''} i^{l-l'+l''} C_{LL'L''} h_{L''}(\vec{R}^n - \vec{R}^{n'}; E) \quad (72)$$

where the Gaunt coefficients

$$C_{LL'L''} = \int d\Omega Y_L(\vec{r}) Y_{L'}(\vec{r}) Y_{L''}(\vec{r}) \quad (73)$$

^cIn Section 6 we will see how the method can be extended to the general case of non-spherical site-centered potentials (the so-called *full-potential* problem).

have been introduced. The summation in Eq. (73) includes a finite number of terms, since the Gaunt coefficients vanish for $l'' > l' + l$. Using the above expansions, the free-electron Green function can be cast in the form:

$$g(\vec{r} + \vec{R}^n, \vec{r}' + \vec{R}^{n'}; E) = -\frac{1}{4\pi} \frac{e^{i\vec{k} \cdot \vec{r} + \vec{R}^n - \vec{r}' - \vec{R}^{n'}}}{|\vec{r} + \vec{R}^n - \vec{r}' - \vec{R}^{n'}|} \\ = -i\sqrt{E} \sum_L j_L(\vec{r}_{<}; E) h_L(\vec{r}_{>}; E) + \sum_{LL'} j_L(\vec{r}; E) g_{LL'}^{nn'}(E) j_{L'}(\vec{r}'; E). \quad (74)$$

We proceed by considering an *outgoing* wave in the interstitial, after a scattering event by a potential at $\vec{R}^n$. It has the form

$$\psi_{\vec{k}}^{\text{sc}(n)}(\vec{r}) = \sum_L b_{\vec{k}L}^{\text{sc}(n)} h_L(\vec{r}; E). \quad (75)$$

The same wave can be resolved as an *incoming* wave at a different center $\vec{R}^{n'}$ via

$$\psi_{\vec{k}}^{\text{inc}(n')}(\vec{r}') = \sum_L b_{\vec{k}L}^{(n')} h_L(\vec{r}'; E). \quad (76)$$

Thus a wave scattered at site n can be expressed as incoming to site n' :

$$\sum_L b_{\vec{k}L}^{\text{sc}(n)} h_L(\vec{r}' + \vec{R}^{n'} - \vec{R}^n; E) = \sum_L b_{\vec{k}L}^{(n')} h_L(\vec{r}'; E). \quad (77)$$

Using the identity in Eq. (70) we find

$$b_{\vec{k}L}^{(n')} = \frac{i}{\sqrt{E}} \sum_{L'} g_{LL'}^{nn'} b_{\vec{k}L'}^{\text{sc}(n')}. \quad (78)$$

Given the periodicity of the crystal lattice, we employ the Bloch condition according to which the amplitude of a scattered wave at position $\vec{R}^n$ differs from the amplitude of a scattered wave at position $\vec{R}^{n'}$ by a phase factor of $\exp(i\vec{k} \cdot (\vec{R}^n - \vec{R}^{n'}))$. The amplitude of the *total incoming* wave at the scatterer $\vec{R}^{n'}$, originating from all other identical scatterers, can therefore be written as

$$c_{\vec{k}L}^{(n')} = \frac{i}{\sqrt{E}} \sum_{L'} \sum_{n \neq n'} g_{LL'}^{nn'}(E) e^{i\vec{k} \cdot (\vec{R}^n - \vec{R}^{n'})} c_{\vec{k}L'}^{\text{sc}(n')} = \frac{i}{\sqrt{E}} \sum_{L'} g_{LL'}(\vec{k}; E) c_{\vec{k}L'}^{\text{sc}(n')}. \quad (79)$$

The quantities c and c^{sc} are the expansion coefficients for the total incoming and scattered wave, respectively, while

$$g_{LL'}(\vec{k}; E) = \sum_{n \neq n'} g_{LL'}^{nn'}(E) e^{i\vec{k} \cdot (\vec{R}^n - \vec{R}^{n'})} \quad (80)$$

is the Fourier transform of $g_{LL'}^{nn'}(E)$. The KKR structure constants $g_{LL'}(\vec{k}; E)$ (and $g_{LL'}^{nn'}(E)$) depend only on the geometry of the lattice and not on the scattering potential.^d

The total scattered wave from a scatterer at lattice site n' is connected to the total incoming wave at that site via the t -matrix, with elements $t_l(E)$. In the presence of an

^dNote that the calculation of $g_{LL'}(\vec{k}; E)$ demands a cumbersome Ewald summation, because the real-space $g_{LL'}^{nn'}(E)$ does not fall off fast enough in real space. Alternatively, one can use the tight-binding KKR formalism, presented in Section 8.

external incoming wave (from outside the crystal) of amplitude $a_{\vec{k}L}^0$, the total incoming wave at some site consists of the external one plus the waves scattered from all other lattice sites, so that:

$$c_{\vec{k}L}^{\text{sc}(n')} = -i\sqrt{E} t_l(E) \left(a_{\vec{k}L}^0 + c_{\vec{k}L}^{(n')} \right). \quad (81)$$

Combining Eqs. (81) and (79), we arrive at the following system of equations:

$$a_{\vec{k}L}^0 = \sum_{L'} \left(\delta_{LL'} - g_{LL'}(\vec{k}; E) t_{l'}(E) \right) \frac{i}{\sqrt{E}} t_{l'}^{-1}(E) c_{\vec{k}L'}^{\text{sc}(n')}. \quad (82)$$

In the absence of an external incoming wave, these are rewritten as

$$\sum_{L'} \left(\delta_{LL'} - g_{LL'}(\vec{k}; E) t_{l'}(E) \right) c_{\vec{k}L'}^{(n')} = 0. \quad (83)$$

The electronic eigenvalues in a crystal are given by the solutions of (83), i.e., in the absence of an external wave. In order to have non-trivial solutions of system (83), the necessary and sufficient condition is that the determinant vanishes:

$$\text{Det} \left[\delta_{LL'} - g_{LL'}(\vec{k}; E) t_{l'}(E) \right] = 0. \quad (84)$$

This is the KKR secular equation, giving the energy band structure $E(\vec{k})$ of periodic crystals.

The total wavefunction in the interstitial around a scatterer n is given by the sum of the incoming and scattered waves, Eqs. (75) and (76). Using Eqs. (81) and (83) we arrive at the form

$$\psi_{\vec{k}}(\vec{r} + \vec{R}^n) = \sum_L c_{\vec{k}L}^{(n)} \left[j_L(\vec{r}; E) - i\sqrt{E} h_L(\vec{r}; E) \right] \quad r \geq R_{\text{MT}}. \quad (85)$$

Each component L of the wavefunction has thus the form of $R_L(\vec{r}; E) = R_l(r; E) Y_L(\vec{r})$ of Eq. (62) for $r \geq S$. We are led therefore to the following form of the Bloch wavefunction in the crystal:

$$\psi_{\vec{k}}(\vec{r} + \vec{R}^n) = \sum_L c_{\vec{k}L}^{(n)} R_L(\vec{r}; E); \quad (86)$$

Bloch's theorem implies

$$c_{\vec{k}L}^{(n)} = e^{i\vec{k} \cdot \vec{R}^n} c_{\vec{k}L}^{(0)}. \quad (87)$$

The coefficients $c_{\vec{k}L}^{(n)}$ can be viewed as the matrix elements of a unitary transformation between the $\vec{k}$ and (L, n) representations. The orthogonality of the bases implies:

$$\sum_{n,L} c_{\vec{k}L}^{(n)*} c_{\vec{k}'L}^{(n)} = \delta_{\vec{k}\vec{k}'} \quad (88)$$

$$\sum_{\vec{k}} c_{\vec{k}L}^{(n)*} c_{\vec{k}L'}^{(n')} = \delta_{nn'} \delta_{LL'}. \quad (89)$$

5 KKR Representation of the Green Function

We now turn to the description of the KKR Green function method, by introducing the KKR representation of the Green function and the Dyson equation.

Consider the Green function of a periodic array of spherically symmetric, non-overlapping potentials. The potential is given by

$$V(\vec{r} + \vec{R}^n) = V^n(r) \quad (90)$$

and the Green function is defined via

$$(-\nabla^2 + V^n(r) - E) G(\vec{r} + \vec{R}^n, \vec{r}' + \vec{R}^{n'}; E) = -\delta_{nn'} \delta(\vec{r} - \vec{r}') . \quad (91)$$

For $n \neq n'$ the Green function satisfies the homogeneous Schrödinger equation and can be expanded in the regular solutions of the Schrödinger equation $R_L^n(\vec{r}; E)$ and $R_L^{n'}(\vec{r}; E)$. In the muffin-tin approximation (see equation (52)), these are simply given by $R_L^n(\vec{r}; E) = R_l^n(r; E) Y_L(\vec{r})$ (see equation (53)), while in the full-potential case a more cumbersome calculation is needed (see Section 6). For $n = n'$ we actually have the Green function for a central potential as in Eq. (68), but with a boundary condition of back-scattering by all other potentials in the crystal. In the end, the Green function is a sum of the general solution of the homogeneous equation corresponding to Eq. (91), plus a special solution of the inhomogeneous equation (91). In the mixed site-angular-momentum representation, the Green function is finally expanded as:

$$G(\vec{r} + \vec{R}^n, \vec{r}' + \vec{R}^{n'}; E) = -i\sqrt{E} \sum_L R_L^n(\vec{r}_{<}; E) H_L^n(\vec{r}_{>}; E) \delta_{nn'} + \sum_{LL'} R_L^n(\vec{r}; E) G_{LL'}^{nn'}(E) R_L^{n'}(\vec{r}'; E), \quad (92)$$

where we have used the abbreviation $H_L^n(\vec{r}; E) = H_l^n(r; E) Y_L(\vec{r})$ for the irregular solution of the radial equation at the atomic cell n . The coefficients $G_{LL'}^{nn'}(E)$ are called *structural Green functions*, and are at the moment unknown quantities to be calculated by use of the Dyson equation. This, taking free space as the unperturbed system, is:

$$G(\vec{r} + \vec{R}^n, \vec{r}' + \vec{R}^{n'}; E) = g(\vec{r} + \vec{R}^n, \vec{r}' + \vec{R}^{n'}; E) + \sum_{n''} \int d^3 r'' g(\vec{r} + \vec{R}^n, \vec{r}'' + \vec{R}^{n''}; E) V^{n''}(r'') G(\vec{r}'' + \vec{R}^{n''}, \vec{r}' + \vec{R}^{n'}; E) . \quad (93)$$

By substituting the expansions Eqs. (74) and (92) into Eq. (93), one can arrive, after some algebra,⁷ at the following *algebraic Dyson equation* determining the structural Green functions:

$$G_{LL'}^{nn'}(E) = g_{LL'}^{nn'}(E) + \sum_{n'', L''} g_{LL''}^{nn''}(E) t_{l''}^{n''}(E) G_{L''L'}^{n''n'}(E) \quad (94)$$

where the t -matrix enters, instead of the potential which appears in the normal Dyson equation. The physical significance of this equation becomes clearer if we expand the sum on the right-hand side, whence we obtain:

$$G_{LL'}^{nn'} = g_{LL'}^{nn'} + \sum_{n'', L''} g_{LL''}^{nn''} t_{l''}^{n''} g_{L''L'}^{n''n'} + \sum_{n'', L''} \sum_{n''', L'''} g_{LL''}^{nn''} t_{l''}^{n''} g_{L''L'''}^{n''n'''} t_{l'''}^{n'''} g_{L'''L'}^{n'''n'} + \dots , \quad (95)$$

meaning that an the electron can travel from n' to n directly, or after being scattered once by any site, or by two sites, etc.

In practice, the structural Green functions are first calculated in $\vec{k}$ -space using matrix inversion; a subsequent Fourier transform gives us the real-space quantities. We write, then,

$$G_{LL'}(\vec{k}; E) = \sum_{n'} G_{LL'}^{nn'}(E) e^{-i\vec{k} \cdot (\vec{R}^n - \vec{R}^{n'})} \quad (96)$$

(which, due to translational symmetry, is independent of n). The algebraic Dyson equation (94) becomes

$$G_{LL'}(\vec{k}; E) = g_{LL'}(\vec{k}; E) + \sum_{L''} g_{LL''}(\vec{k}; E) t_{l''}(E) G_{L''L'}(\vec{k}; E) \quad (97)$$

(the t -matrix is independent of n , again due to translational symmetry). The structural Green functions $G_{LL'}$ and $g_{LL'}$, and the t -matrix t_l , are considered as matrices in L and L' , and (97) is solved by matrix inversion after a cutoff at some $l = l_{\max}$ for which the t -matrix becomes negligible (usually $l_{\max} = 3$ or 4 suffices). The result is

$$G_{LL'}^{nn'}(E) = \frac{1}{V_{\text{BZ}}} \int_{\text{BZ}} d^3k e^{i\vec{k} \cdot (\vec{R}^n - \vec{R}^{n'})} \left[\left(1 - \mathbf{g}(\vec{k}; E) \mathbf{t}(E) \right)^{-1} \mathbf{g}(\vec{k}; E) \right]_{LL'} \quad (98)$$

where the integral is over the Brillouin zone volume V_{BZ} . For the calculation of the charge density or of the density of states, only the on-site term $n = n'$, $G_{LL'}^{nn}(E)$, is needed.

It is straightforward to generalize the method to the case of more than one atoms per unit cell, say N_{at} . Introducing an index $\mu = 1, \dots, N_{\text{at}}$ to account for the different atom types, and reserving the index n for the periodic lattice positions, an atomic position in the crystal is defined by the lattice vector $\vec{R}^n$ plus the site vector $\vec{\chi}^\mu$ connecting the lattice point to the basis atom:

$$\vec{R}^{n\mu} = \vec{R}^n + \vec{\chi}^\mu. \quad (99)$$

The Fourier transforms are done then with respect to n only, so that we have an expression analogous to Eq. (96),

$$G_{LL'}^{\mu\mu'}(\vec{k}; E) = \sum_{n'} G_{LL'}^{n\mu, n'\mu'}(E) e^{-i\vec{k} \cdot (\vec{R}^n - \vec{R}^{n'})}, \quad (100)$$

and the analogous Fourier transform for the free-space structure constants $g^{n\mu, n'\mu'}$, giving the amplitude of electron propagation from atom μ' at lattice position n' to atom μ at lattice position n . The Dyson equation (98) then takes the form

$$G_{LL'}^{n\mu, n'\mu'}(E) = \frac{1}{V_{\text{BZ}}} \int_{\text{BZ}} d^3k e^{i\vec{k} \cdot (\vec{R}^n - \vec{R}^{n'})} e^{i\vec{k} \cdot (\vec{\chi}^\mu - \vec{\chi}^{\mu'})} \times \left[\left(1 - \mathbf{g}(\vec{k}; E) \mathbf{t}(E) \right)^{-1} \mathbf{g}(\vec{k}; E) \right]_{LL'}^{\mu\mu'}. \quad (101)$$

Here, the t -matrix $\mathbf{t}(E)$ depends on the atom-type μ and on angular-momentum indexes (it is site-diagonal, $(\mathbf{t})_l^{\mu\mu'} = t_l^\mu \delta_{\mu\mu'}$). The structure constants $\mathbf{g}(\vec{k}; E)$ are considered as matrices in both (L, L') and (μ, μ') , and thus the computational effort for the matrix inversion increases as $O(N_{\text{at}}^3)$. A considerable speed-up can be achieved for large systems by using the concept of the screening transformation (see Section 8).

6 Description of the Full-Potential

Although the approximation of spherical potentials around the atomic gives in many cases reasonable results, the correct description of the full anisotropic potential can give significant corrections when details of the electronic structure are required. The full-potential can be especially important in systems of reduced symmetry, such as non-cubic crystals, at surfaces or interfaces, etc.; in particular the calculation of forces on atoms, leading to lattice relaxations, is highly inaccurate without the full-potential. In this section we give a summary of the full-potential treatment within the KKR method, and refer to the literature for details, concerning in particular the convergence properties.²⁷

We begin our discussion on the level of scattering theory by a single anisotropic potential $V(\vec{r})$ of finite maximal radius S , which later will represent the site-centered atomic potential of a crystal. An incoming wave of wavevector $\vec{k}$, scattered by this potential, results in a wave $\psi_{\vec{k}}(\vec{r})$. The dependence on $\vec{k}$ is lifted by an expansion in spherical harmonics:

$$\psi_{\vec{k}}(\vec{r}) = \sum_{L'} 4\pi i^{l'} Y_{L'}(\vec{k}) R_{L'}(\vec{r}; E), \quad (102)$$

where $R_{L'}(\vec{r}; E)$ is the regular (converging at zero) solution of the Schrödinger equation corresponding to an incoming spherical wave of symmetry L' . The Lippmann-Schwinger equation, determining $R_{L'}(\vec{r}; E)$, is

$$R_{L'}(\vec{r}) = j_{l'}(\sqrt{Er}) Y_{L'}(\vec{r}) + \int g(\vec{r}, \vec{r}'; E) V(\vec{r}') R_{L'}(\vec{r}') d^3 r', \quad (103)$$

where the incoming free wave $j_{l'}(\sqrt{Er}) Y_{L'}(\vec{r})$ and the free-electron Green function $g(\vec{r}, \vec{r}'; E)$ appear. While the index L' refers to the incoming wave, the wavefunction $R_{L'}(\vec{r}')$ can be expanded once more²⁸ in spherical harmonics to lift the dependence on the direction of $\vec{r}$ (if the potential is spherically symmetric, $R_{L'}(\vec{r})$ is just $R_{l'}(r) Y_{L'}(\vec{r})$):

$$R_{L'}(\vec{r}) = \sum_L R_{LL'}(r; E) Y_L(\vec{r}). \quad (104)$$

Similarly, the potential $V(\vec{r})$ is expanded in spherical harmonics as

$$V(\vec{r}) = \sum_L V_L(r) Y_L(\vec{r}). \quad (105)$$

Using these expansions, the Lippmann-Schwinger equation (103) can be re-written in the form

$$R_{LL'}(r; E) = j_l(\sqrt{Er}) \delta_{LL'} + \int_0^S g_l(r, r'; E) \sum_{L''} V_{LL''}(r') R_{L''L'}(r'; E) r'^2 dr' \quad (106)$$

where we have introduced the notation

$$V_{LL'}(r) = \sum_{L''} C_{LL'L''} V_{L''}(r). \quad (107)$$

The advantage of Eq. (106) is that it contains only a one-dimensional integral instead of the three-dimensional integral of (103). However, one has now to solve a system of coupled

radial equations, since the Schrödinger equation reads:

$$\sum_{L''} \left[\left(-\frac{1}{r} \frac{\partial^2}{\partial r^2} r + \frac{l(l+1)}{r^2} - E \right) \delta_{LL''} + V_{LL''} \right] R_{L''L'}(r; E) = 0. \quad (108)$$

Obviously, the non-spherical components of the potential mix the angular-momentum channels, resulting in the above coupled equations. Also the t -matrix contains now mixed angular-momentum components. It is found by matching $R_{L''L'}$ to an outgoing free wave at the boundary S , and related to the solution $R_{L''L'}$ by:

$$t_{LL'}(E) = \int_0^S j_l(\sqrt{Er}) \sum_{L''} V_{LL''} R_{L''L'}(r'; E) r'^2 dr' \quad (109)$$

which is analogous to expression in Eq. (61) for the spherical potential. Similar to $R_L(\vec{r}; E)$, the irregular (diverging at the origin) solution of the Schrödinger equation, $H_L(\vec{r}; E)$, is expanded in spherical harmonics as

$$H_{L'}(\vec{r}; E) = \sum_L H_{LL'}(r; E) Y_L(\vec{r}). \quad (110)$$

Instead of solving directly the coupled Schrödinger equations, a different procedure is followed, based on perturbation theory. Since the non-spherical part of the potential is usually weak compared to the spherical part, first the radial solutions $R_l^{\text{sph}}(r; E)$ and $H_l^{\text{sph}}(r; E)$ for the spherical component ($V_{l=0}(r)$) are calculated, as in Section 3 (see Eq. (53)). This is used as a reference system. Then, using also the corresponding spherical Green function

$$G_l^{\text{sph}}(r, r'; E) = -i\sqrt{E} R_l^{\text{sph}}(r < E) H_l^{\text{sph}}(r > E) \quad (111)$$

(see also eq. (68)), the radial wavefunctions of the non-spherical case are connected to $R_l^{\text{sph}}(r; E)$ via the Lippmann-Schwinger equation

$$R_{LL'}(r; E) = R_l^{\text{sph}}(r; E) \delta_{LL'} + \int_0^S G_l^{\text{sph}}(r, r'; E) \sum_{L''} \Delta v_{LL''}(r') R_{L''L'}(r'; E) r'^2 dr' \quad (112)$$

where only the non-spherical part of the potential enters:

$$\Delta v_{LL'}(r) = \sum_{L'' \neq 0} C_{LL'L''} V_{L''}(r). \quad (113)$$

Eq. (112) can be solved iteratively, as a Born series. Practice shows that two or three iterations are enough for most applications. Similar is the solution for the irregular wavefunction $H_{LL'}(\vec{r}; E)$, starting from $H_l^{\text{sph}}(r > E)$.

Within a crystal, the site-centered potentials must be cut off at the boundary of the Wigner-Seitz cells. In order to ensure this, shape functions are used²⁹ and expanded in spherical harmonics. In particular, the shape function $\Theta^n(\vec{r})$ of a site n is defined as

$$\Theta^n(\vec{r}) = \begin{cases} 1 & \text{if } \vec{r} \text{ is in the WS-cell of site } n, \\ 0 & \text{otherwise.} \end{cases} \quad (114)$$

The crystal potential is then given by

$$V^n(\vec{r}) = V(\vec{r} + \vec{R}^n) \Theta^n(\vec{r}). \quad (115)$$

The shape functions are expanded as

$$\Theta^n(\vec{r}) = \sum_L \Theta_L^n(r) Y_L(\vec{r}). \quad (116)$$

The coefficients $\Theta_L^n(r)$ enter in the expansions for the charge density and the potential, in order to ensure the correct cutoff at the Wigner-Seitz boundary. Without going into details, we note that, although the convergence of Eq. (116) is slow, the property of the Gaunt coefficients

$$C_{LL'L''} \neq 0 \text{ only if } |l' - l''| \leq l \leq l' + l'' \quad (117)$$

allows us to take into account the shape functions up to only $4l_{\max}$, if the t -matrix and the Green functions are cut off at $l_{\max}$.

7 Total Energy

Within density functional theory, the total energy of a many-electron system is written as a sum of three terms: the single-particle kinetic energy $T[\rho_s]$, the Hartree energy $E_H[\rho]$, and the exchange and correlation energy $E_{xc}[\rho_s]$ ($s = (+, -)$ is the spin index). The kinetic and the exchange-correlation energy, as well as the total energy, are functionals of the spin density $\rho_s(\vec{r}) := (\rho_+(\vec{r}), \rho_-(\vec{r}))$, while the Hartree energy is a functional of the charge density $\rho = \rho_+ + \rho_-$. We have:

$$E_{\text{tot}}[\rho_+, \rho_-] = T[\rho_+, \rho_-] + E_H[\rho] + E_{xc}[\rho_+, \rho_-]. \quad (118)$$

Given the single-particle energies ϵ_{is} (eigenenergies of the Kohn-Sham equations), the kinetic energy can be written in terms of these and of the effective Kohn-Sham potential $V_s^{\text{eff}}(\vec{r})$ as:

$$T[\rho_+, \rho_-] = \sum_s \left(\sum_i \epsilon_{is} - \int \rho_s(\vec{r}) V_s^{\text{eff}}(\vec{r}) d^3r \right). \quad (119)$$

In this way, the sum of the single-particle energies

$$E_{\text{SP}} := \sum_{is} \epsilon_{is} \quad (120)$$

is singled out and can be thought of a “band energy”, which would be relevant if we had non-interacting electrons in an external potential, while the remaining terms are packed up together in what is called the “double-counting energy terms”, as corrections to the single-particle picture:

$$E_{\text{DC}}[\rho_+, \rho_-] = - \sum_s \int \rho_s(\vec{r}) V_s^{\text{eff}}(\vec{r}) d^3r + E_H[\rho] + E_{xc}[\rho_+, \rho_-]. \quad (121)$$

The total energy is the sum of the two:

$$E_{\text{tot}}[\rho_+, \rho_-] = E_{\text{SP}}[\rho_+, \rho_-] + E_{\text{DC}}[\rho_+, \rho_-]. \quad (122)$$

We proceed to analyze each term separately. The sum of single-particle energies can be written in terms of the spin-dependent density of states $n_s(E)$ as

$$\begin{aligned} E_{\text{SP}} &= \sum_{is} \epsilon_{is} \\ &= \sum_s \int^{E_F} E n_s(E) dE \end{aligned} \quad (123)$$

$$= E_F N - \sum_s \int^{E_F} N_s(E) dE. \quad (124)$$

In the last step we introduced $N_s(E)$ as the integrated density of states up to energy E and used the fact that the total number of electrons per spin is just $N_s = \int_{E_F} n(E) dE$, and $N = N_+ + N_-$. In practice, expression in Eq. (123) can be used for periodic systems. Expression in Eq. (124) is useful for systems with broken periodicity, such as impurities in crystals, where the perturbed charge density converges very slowly with distance from the impurity due to Friedel oscillations. Then, $N_s(E)$ is calculated not by integration of $n_s(E)$, but by using the Friedel sum rule (or its multiple-scattering analogue, Lloyd's formula), which takes into account the Friedel oscillations up to infinity.

The double-counting term includes the electrostatic energy and the exchange-correlation energy. The electrostatic (Hartree) energy depends on the charge densities at each cell n , $\rho_n(\vec{r}) := \rho(\vec{r} + \vec{R}^n)$, and on the nuclear charges Z^n . We have:

$$\begin{aligned} E_H[\rho] &= \sum_{nn'} \int_{\Omega^n} d^3r \int_{\Omega^{n'}} d^3r' \frac{\rho^n(\vec{r}) \rho^{n'}(\vec{r}')}{|\vec{r} + \vec{R}^n - \vec{r}' - \vec{R}^{n'}|} \\ &\quad - 2 \sum_{nn'} Z^{n'} \int_{\Omega^n} d^3r \frac{\rho^n(\vec{r})}{|\vec{r} + \vec{R}^n - \vec{R}^{n'}|} + \sum_n \sum_{n' \neq n} \frac{Z^n Z^{n'}}{|\vec{R}^n - \vec{R}^{n'}|} \end{aligned} \quad (125)$$

where Ω^n is the volume of the atomic cell n . It proves convenient to define the Coulomb potential

$$V_C^n(\vec{r}) = 2 \sum_{n'} \int_{\Omega^{n'}} d^3r' \frac{\rho^{n'}(\vec{r}')}{|\vec{r} + \vec{R}^n - \vec{r}' - \vec{R}^{n'}|} - 2 \sum_{n'} \frac{Z^{n'}}{|\vec{R}^n - \vec{R}^{n'}|}. \quad (126)$$

The Madelung potential $V_M(\vec{R}^n)$ is the Coulomb potential at position $\vec{R}^n$ if we exclude the term $n' = n$ from the second sum of (126):

$$V_M(\vec{R}^n) = 2 \sum_{n'} \int_{\Omega^{n'}} d^3r' \frac{\rho^{n'}(\vec{r}')}{|\vec{r} + \vec{R}^n - \vec{r}' - \vec{R}^{n'}|} - 2 \sum_{n' (\neq n)} \frac{Z^{n'}}{|\vec{R}^n - \vec{R}^{n'}|}. \quad (127)$$

Using the above definitions we re-write the electrostatic energy as

$$E_H[\rho] = \frac{1}{2} \left[\sum_n \int_{\Omega^n} d^3r \rho^n(\vec{r}) V_C^n(\vec{r}) - \sum_n Z^n V_M(\vec{R}^n) \right]. \quad (128)$$

The Coulomb potential and the charge density at site n are expanded in spherical harmonics around $\vec{R}^n$:

$$V_C^n(\vec{r}) = \sum_L V_{C,L}^n(r) Y_L(\vec{r}); \quad (129)$$

$$\rho^n(\vec{r}) = \sum_L \rho_L^n(r) Y_L(\vec{r}). \quad (130)$$

In this way the calculation of the Coulomb potential is reduced to summing up terms containing the moments $\rho_L^n(r)$ of the charge density over all lattice sites. For the higher l terms the corresponding summations converge rapidly, but for low l an Ewald summation is required. The details of this procedure are omitted here. Once the expansion in Eq. (129) of the Coulomb potential is known, the Madelung potential can be calculated using the value of V_C^n at a sphere of radius R around $\vec{R}^n$ and by knowledge of the charge distribution within this sphere. The result can be obtained by solving the corresponding boundary-value problem in electrostatics (the proof is omitted here):

$$V_M(\vec{R}^n) = 2\sqrt{4\pi} \int_0^R \rho_{l=0}^n(r) dr + \frac{2}{R} (Z^n - N^n(R)) + \frac{1}{\sqrt{4\pi}} V_{C,l=0}^n(R), \quad (131)$$

where $N^n(R)$ is the number of electrons within the sphere of radius R .

The effective Kohn-Sham potential at cell n can be written in terms of the Coulomb potential and the exchange-correlation energy:

$$V_{\text{eff},s}^n(\vec{r}) = V_C^n(\vec{r}) + \frac{\delta E_{xc}[\rho_+, \rho_-]}{\delta \rho_s(\vec{r})}. \quad (132)$$

Finally, the exchange-correlation energy within the local density approximation is given by

$$E_{xc}^{\text{LDA}}[\rho] = \int \rho(\vec{r}) \epsilon_{xc}(\rho_+(\vec{r}), \rho_-(\vec{r})) d^3r \quad (133)$$

where $\epsilon_{xc}(\rho_+, \rho_-)$ is the exchange-correlation energy density for a homogeneous electron gas of spin density ρ_s . This is again expanded in spherical harmonics as

$$\epsilon_{xc}(\rho_+^n(\vec{r}), \rho_-^n(\vec{r})) = \sum_L \epsilon_{xc,L}^n(r) Y_L(\vec{r}), \quad (134)$$

and the exchange-correlation energy is given by

$$E_{xc}^{\text{LDA}}[\rho] = \sum_n \sum_{LL'L''} C_{LL'L''} \int_0^{R_c^n} \Theta_L^n(r) \rho_{L'}^n(r) \epsilon_{xc,L''}^n(r) r^2 dr. \quad (135)$$

8 Screened (Tight-Binding) KKR Method

An improvement of the KKR method has been achieved by the so-called *screened* or *tight-binding* KKR formalism,⁸ which allows a considerable reduction of the calculation time for large systems. In particular, while the traditional KKR formalism requires a matrix inversion for the solution of the algebraic Dyson equation with computational effort of $O(N^3)$ (for N different atoms in the unit cell), in the screened KKR method the same

results can be obtained with an effort of, ideally, $O(N)$; this is optimally achieved for layered systems. This is made possible by a transformation of the reference system after which the reference Green function falls off exponentially with distance, thus allowing the inversion of sparse, or even tridiagonal, matrices, which is much faster than a full matrix inversion. Due to the decoupling between distant atomic sites which follows, the corresponding transformation is called *screening transformation* and the method *screened KKR*; due to its formal resemblance to tight-binding theory, the method is also called *tight-binding KKR*.

Three observations lead to the tight-binding KKR formalism:

- Using the Dyson equation one can connect the crystal Green function to any reference system (of the same periodic structure), and not just to the free electron system.
- A reference system of repulsive potentials can be constructed in which there are no states in the energy region of interest (up to 1-2 Ry higher than E_F).
- The structural Green functions of this reference system fall off exponentially in space, so that the corresponding KKR matrix becomes practically sparse or tridiagonal.

In what follows, we shall describe these ideas in more detail.

8.1 Transformation to an Arbitrary Reference System

In the Dyson equation, introduced in Section 2, no formal statement is made about the nature of the reference system. Subsequently, when describing the KKR method, we have chosen a reference system of constant potential as the most natural one, since the Green function and structure constants are then given by simple analytical expressions. Here we show that a choice of another reference system of the same lattice structure as the real system leads to the same form of the algebraic Dyson equation, with the difference of t -matrices between the real and reference system ($t_{LL'}^n(E) - t_{LL'}^{r,n}(E)$) entering in the place of the $t_{LL'}^n(E)$ in Eq. (94).

Consider then a reference system of potentials $V^{r,n}(\vec{r})$ (to be given in detail below) placed at the lattice sites n of the crystal. These are characterized by t -matrices $t^{r,n}(E)$. The structural Green function of this system, $G_{LL'}^{r,nn'}(E)$, are related to the free-space structure constants, $g_{LL'}^{nn'}(E)$ (Eq. (72)), via the algebraic Dyson equation. Adopting boldface symbols to denote matrices (and with $\mathbf{t}^r$ standing for the site-diagonal t -matrix), this is written

$$\mathbf{G}^r(E) = \mathbf{g}(E) + \mathbf{g}(E) \mathbf{t}^r(E) \mathbf{G}^r(E) \quad (136)$$

which is rewritten as

$$(\mathbf{G}^r)^{-1} = \mathbf{g}^{-1} - \mathbf{t}^r. \quad (137)$$

At the same time the structural Green functions of the real system, $G_{LL'}^{nn'}(E)$, are related to the free-space structure constants via the t -matrix $t^n(E)$ of the real system. The corresponding algebraic Dyson equation reads:

$$(\mathbf{G})^{-1} = \mathbf{g}^{-1} - \mathbf{t}. \quad (138)$$

Expressions in Eqs. (136) and (137) lead directly to the algebraic Dyson equation connecting the new reference system to the real system:

$$(\mathbf{G})^{-1} = (\mathbf{G}^r)^{-1} - (\mathbf{t} - \mathbf{t}^r), \quad (139)$$

which can be expanded as

$$\mathbf{G}(E) = \mathbf{G}^r(E) + \mathbf{G}^r(E) \Delta \mathbf{t}(E) \mathbf{G}(E) \quad (140)$$

with $\Delta \mathbf{t}(E) = \mathbf{t}(E) - \mathbf{t}^r(E)$. This equation has the same form as Eq. (94) with $\Delta \mathbf{t}(E)$ in the place of $\mathbf{t}(E)$.

8.2 Choice of a Reference System of Repulsive Potentials

Having shown that the algebraic Dyson equation holds after a transformation to a new reference system, we proceed with the choice of an adequate reference system in which the structural Green functions fall exponentially with distance. Such a system is defined by a collection of repulsive muffin-tin potentials (one around each site n) of the form:

$$V^{r,n}(\vec{r}) = \begin{cases} V_C, & r \leq R_{\text{MT}}^n \\ 0, & \text{otherwise} \end{cases} \quad (141)$$

with R_{MT}^n the muffin-tin radius at site n , and V_C a positive constant, usually chosen to be a few Rydbergs. It is physically expected (and computationally verified) that for such a potential the eigenvalue spectrum starts from an energy E_{bot} somewhat smaller than V_C ; a choice of $V_C = 4$ Ry is adequate to push E_{bot} high above E_F . Schematically, this is shown in Figure 1.

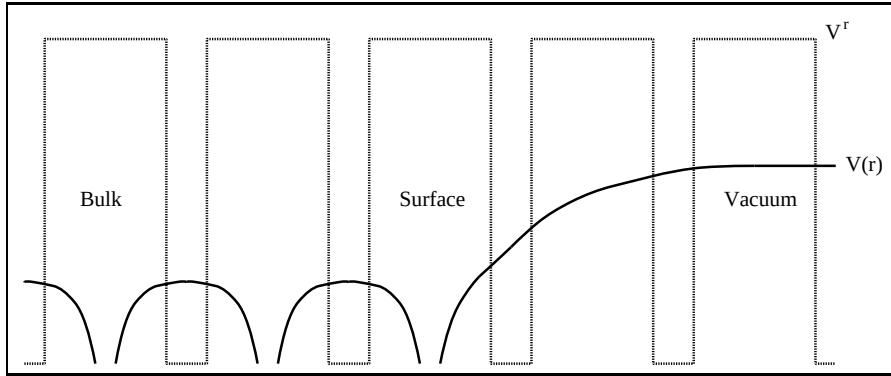


Figure 1. Schematic representation of the repulsive potential $V^r(\vec{r})$ and the crystal potential $V(\vec{r})$ in the bulk, at a surface, and in the vacuum.

In this way, it is ensured that for $E < E_{\text{bot}}$ the Green function of the reference system drops rapidly, and in practice exponentially, with distance; the same is true for the structural Green functions. In order to demonstrate this, we introduce the *partial norm*

$$N_{ll'}(|\vec{R}n - \vec{R}n'|; E) = \frac{|E|^{(l+l')/2}}{(2l+1)!!(2l'+1)!!} \left[\sum_{mm'} |G_{lm,l'm'}^{r,nn'}(E)|^2 \right]^{1/2} \quad (142)$$

with $(2l + 1)!! = (2l + 1)(2l - 1) \dots (3)(1)$. In Figure 2 (left), the partial norms for an fcc lattice are plotted for $l = l'$ as function of the distance $|\mathbf{R}^{nn'}|$ for the choice $E = 0.65$ Ry which is representative for the Fermi energy of Cu. They are compared with the corresponding norms of the Green function g for potential-free space in Figure 2 (right panel).⁸

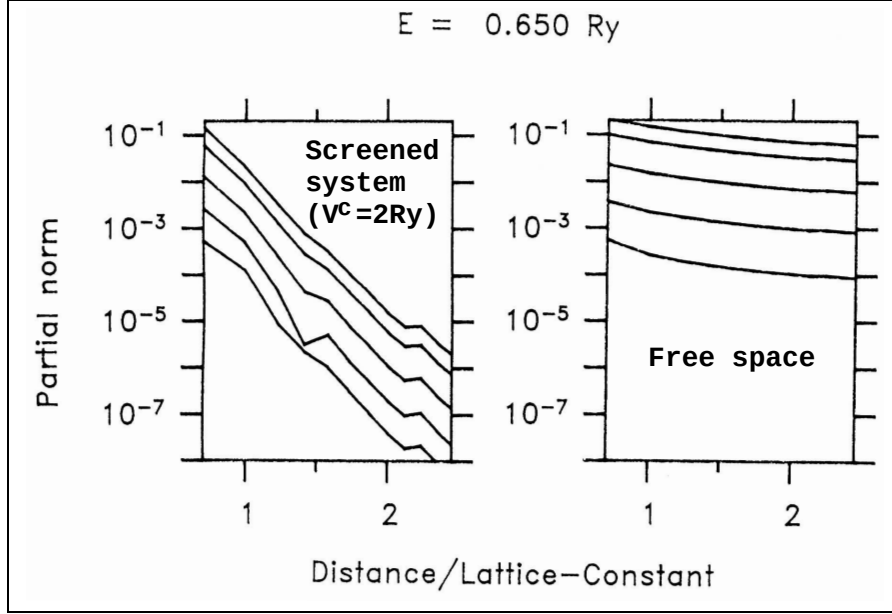


Figure 2. Screened (left panel) and unscreened (right panel) partial norms (Eq. (142)) for $l = l'$ and $E = 0.65$ Ry as function of the distance in units of the lattice constant for an fcc lattice. The results for $l = 0, 1, 2, 3, 4$ are shown from top to bottom.⁸

The rapid decay of these *screened* structural Green functions allows their direct evaluation in real space, with no need of a Brillouin zone integration. Thus the cumbersome calculation of the free-space $g_{LL'}(\vec{k}; E)$, which demands an Ewald summation for convergence, is avoided. Instead, the matrices in equation (136) can be cut off in real space at a finite value of $|\vec{R}^n - \vec{R}^{n'}|$ (two lattice constants are enough for fcc), and solved by direct matrix inversion, yielding $G_{LL'}^{r,nn'}(E)$; it is straightforward to obtain $G_{LL'}^r(\vec{k}; E)$ by a subsequent Fourier transform, since the summand is confined to finite $|\vec{R}^n - \vec{R}^{n'}|$.

8.3 Computational Speed-up for Large Systems

At the end of Section 5 we commented that the computational effort for large systems, with many atoms per unit cell, scales as $O(N_{\text{at}}^3)$ due to the matrix inversion involved in the Dyson equation (see Eq. (101)). We will now see why the screening transformation allows for a considerable acceleration of the calculations.

Using the notation introduced at equations (99–101), we consider the Fourier-transformed structural Green functions of a large system, and the algebraic Dyson equation

after the screening transformation:

$$G_{LL'}^{\mu\mu'}(\vec{k}; E) = \left[\left(1 - \mathbf{G}^r(\vec{k}; E) \Delta \mathbf{t}(E) \right)^{-1} \mathbf{G}^r(\vec{k}; E) \right]_{LL'}^{\mu\mu'}. \quad (143)$$

For large systems, where the interatomic distance in the unit cell is large for several pairs of atoms, the reference structural Green functions $|\mathbf{G}^{r,\mu\mu'}|$ are practically vanishing for the corresponding (μ, μ') and can be neglected. Then the matrix which must be inverted in Eq. (143) is sparse, and one can use algorithms for inversion and multiplication of sparse matrices to speed up the calculation. Moreover, one matrix multiplication in (143) can be avoided by converting the equation as:

$$\mathbf{G} = (1 - \mathbf{G}^r \mathbf{t})^{-1} \mathbf{G}^r = -\mathbf{t}^{-1} + \mathbf{t}^{-1} (\mathbf{t}^{-1} - \mathbf{G}^r)^{-1} \mathbf{t}^{-1}. \quad (144)$$

Since the t -matrix is site-diagonal, the inversion $\mathbf{t}^{-1}$ and the multiplications with $\mathbf{t}^{-1}$ are numerically inexpensive.

This is particularly important in the case of layered systems, which are extending over many monolayers in one direction. Then the indexes μ, μ' are layer indexes, and reference structural Green function matrix takes a block-tridiagonal form, given schematically as

$$\begin{bmatrix} x & x & 0 & 0 & \dots & 0 & 0 & 0 & x \\ x & x & x & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & x & x & x & \dots & 0 & 0 & 0 & 0 \\ 0 & 0 & x & x & \dots & 0 & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & x & x & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & x & x & x & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & x & x & x \\ x & 0 & 0 & 0 & \dots & 0 & 0 & x & x \end{bmatrix}. \quad (145)$$

Each symbol x represents a matrix block, with its size being related to the distance $|\vec{\chi}^\mu - \vec{\chi}^{\mu'}|$ after which $|\mathbf{G}_{LL'}^{r,\mu\mu'}|$ is considered negligible; the x at the upper right and lower left corners are present only if we have a repeated supercell. If only first neighbors would couple, the blocks would be matrices in L and L' only (of size $(l_{\max} + 1)^2 \times (l_{\max} + 1)^2$). For longer-range coupling one can use the principal layer technique by which several layers are combined into a principal layer such that only nearest neighboring principal layers couple. Then one obtains again the form Eq. (144), but now the blocks are submatrices with L and L' indexes and i and i' indexes enumerating the layers within the principal layer. The matrix structure displayed in Eq. (144) is the appropriate one for the multilayer geometry, whereas in slab geometry the lowest left block and the highest right block vanish. For surfaces and interfaces of half-infinite crystals the matrix (Eq. (144)) has infinite range to one or both sides.

In any case the numerical complexity is greatly reduced from $O(N^3)$ to $O(N)$ for N different and $O(\log N)$ for N identical principal layers. Note that, for the calculation of the spin density (and therefore for self-consistent calculations), only the on-site terms of the real-system structural Green function are needed,^e i.e., the block-diagonal part ($\mu' = \mu$)

^e Because $\rho_s(\vec{R}^\mu + \vec{r}) = -\frac{1}{\pi} \text{Im} \int_{-\infty}^{EF} G_s(\vec{R}^\mu + \vec{r}, \vec{R}^\mu + \vec{r}; E) dE$.

of $G_{LL'}^{\mu\mu'}$. This fact is included in the $O(N)$ speed-up; if one wishes to calculate additional, non-diagonal matrix elements of $G_{LL'}^{\mu\mu'}$, e.g. for impurity or transport calculations, the numerical cost is $O(N)$ for each additional element, giving a total of $O(N^3)$ if all N^2 elements are to be calculated.

The treatment of matrices like Eq. (144) is well known in tight-binding surface physics and, for instance for the tight-binding linear-muffin-tin-orbital method,³⁰ Wenzien et al.³¹ have presented an efficient formalism to calculate the Green function of an ideal semi-infinite crystal and the corresponding $\mathbf{k}_{\parallel}$ resolved densities of states.

9 Two-Dimensional Systems: Finite-Thickness Slabs and Half-Infinite Crystals

The extension of the KKR method to the treatment of layered systems, such as surfaces and interfaces, is straightforward, and most efficient within the screened KKR formalism, where $O(N)$ scaling can be achieved (where N is the number of layers) as discussed in Section 8.

When treating a layered system, a surface-adapted geometry is used, in the sense that the two-dimensional periodicity of the atomic layers parallel to the surface (or interface) is exploited while the direction perpendicular to these layers is treated as if these were different atoms in a unit cell. The Fourier transforms are done now within the two-dimensional surface Brillouin zone (SBZ), and the corresponding integration is over all $\vec{k}_{\parallel}$ in the SBZ. Thus, we have the analogue of Eq. (101),

$$G_{LL'}^{n\mu, n'\mu'}(E) = \frac{1}{A_{\text{SBZ}}} \int_{\text{SBZ}} d^2 k_{\parallel} e^{i\vec{k}_{\parallel} \cdot (\vec{R}^n - \vec{R}^{n'})} e^{i\vec{k}_{\parallel} \cdot (\vec{\chi}^{\mu} - \vec{\chi}^{\mu'})} \times \left[\left(1 - \mathbf{G}^r(\vec{k}_{\parallel}; E) \Delta \mathbf{t}(E) \right)^{-1} \mathbf{G}^r(\vec{k}_{\parallel}; E) \right]_{LL'}^{\mu\mu'}. \quad (146)$$

where now $\vec{R}^n$ are in-plane position vectors of the two-dimensional Bravais lattice, while $\vec{\chi}^{\mu}$ are vectors connecting atomic positions in different layers; A_{SBZ} is the area of the SBZ.

In the case of surfaces, the vacuum is described by empty sites, meaning that the lattice structure is continued into the vacuum but no nuclei are positioned there. In this way, the vacuum potential and charge density are calculated within the multiple-scattering formalism on the same footing as the bulk. In practice, three or four monolayers of vacuum sites are enough for the calculation of the electronic structure; Eq. (146) can be cutoff after that.

Depending on the problem, one can choose to use a slab of finite thickness in order to represent a surface or interface, or one can choose to take half-infinite boundary conditions. In the latter case, and starting from a “boundary” layer, the crystal is continued by periodically repeating the potential of this boundary layer to all subsequent layers up to infinity. One is then faced with a problem of inverting an infinite matrix, which due to the screening transformation has a tridiagonal form (see expression (145)), in order to find the Green function in the region of interest. This is done efficiently by the decimation technique,³² which is based on an iterative inversion of matrices of doubled size at each step. In this way the number of layers which are included in the Green function grows exponentially with the number of steps, and the limit of the half-infinite crystal is rapidly achieved.

10 Self-Consistency Algorithm and Energy Contour Integration

We proceed with a short description of the self-consistency algorithm for the calculation of the electronic structure by the KKR method. As in all first-principles schemes, the central quantity is the charge density which is found by solving the Kohn-Sham equations. The steps followed are:

1. Start with an input potential V_s^{in} (s is a spin index used in magnetic systems).
2. Calculate the wavefunctions R_L and H_L and, from these, the t -matrix $t_{LL'}$.
3. Calculate the structure constants of the reference system $g_{LL'}^{\text{ref}}(\vec{k}; E)$ (or the free-space structure constants if the tight-binding formalism is not used).
4. Calculate the t -matrix of the reference system, $t_{LL'}^{\text{ref}}$, and the difference $\Delta t_{LL'} = t_{LL'} - t_{LL'}^{\text{ref}}$.
5. Solve the algebraic Dyson equation by matrix inversion for the structural Green function and integrate over $\vec{k}$ (see Eq. (98) for $n = n'$ and for Δt in the place of t).
6. Calculate the Green function using the structural Green function and R_L and H_L (Eq. (92)). Integrate the Green function from the bottom of the valence band E_b up to E_F using a complex-energy contour (see below) and take the imaginary part to find the valence electron spin density: $\rho_s^v(\vec{r}) = -\frac{1}{\pi} \text{Im} \int_{E_b}^{E_F} G_s(\vec{r}, \vec{r}; E) dE$.
7. Calculate the core-electron wavefunctions and core-electron spin density $\rho_s^c(\vec{r})$; here the multiple-scattering formalism is not needed, because the core wavefunctions are assumed to be highly localized at the atomic sites. Obtain the total spin density $\rho_s = \rho_s^c + \rho_s^v$.
8. Find the output potential V_s^{out} by solving the Poisson equation and adding the exchange-correlation potential. If $V_s^{\text{out}} = V_s^{\text{in}}$ to a reasonable accuracy, exit the cycle, otherwise:
9. Properly mix V_s^{out} with V_s^{in} to obtain a new input potential, and return to step 1.

We now comment on the energy integration for the valence charge density or spin density. According to Eq. (21), the integral to be taken is

$$\rho_s^v(\vec{r}) = -\frac{1}{\pi} \text{Im} \int_{E_b}^{E_F} G_s(\vec{r}, \vec{r}; E) dE . \quad (147)$$

In order to reach the desired accuracy for achieving self-consistency, a large number of integration points are required for the evaluation of Eq. (147), typically between 1000 and 2000. Fortunately, the numerical effort can be strongly reduced by using the analytical properties of the Green function. Since $G(E)$ is analytical on the upper energy half-plane ($\text{Im} E > 0$), the integral in Eq. (147) can be evaluated on a contour starting at the real axis below E_b , continuing at complex $E = E_R + i\Gamma$ (with $E_R = \text{Re} E$), and ending at E_F .³³ The gain comes from the fact that, for larger Γ , the Green function has much less structure than at the real axis. This can be understood by considering that the density of

states corresponding to an eigenvalue ϵ_i is of the form $-\frac{1}{\pi}\text{Im}G(E) = \delta(E - \epsilon_i)$ for real E , while it takes the form $-\frac{1}{\pi}\text{Im}G(E_R + i\Gamma) = \Gamma\pi^{-1}/((E_R - \epsilon_i)^2 + \Gamma^2)$ for complex E , i.e., it is broadened to a Lorentzian function of half-width Γ . As a result, 30 integration points are typically enough when integrating over such a contour. Furthermore, for the energies far from the real axis, fewer k -points are needed in the Brillouin zone integration (Eq. (98)), reducing even more the numerical cost.

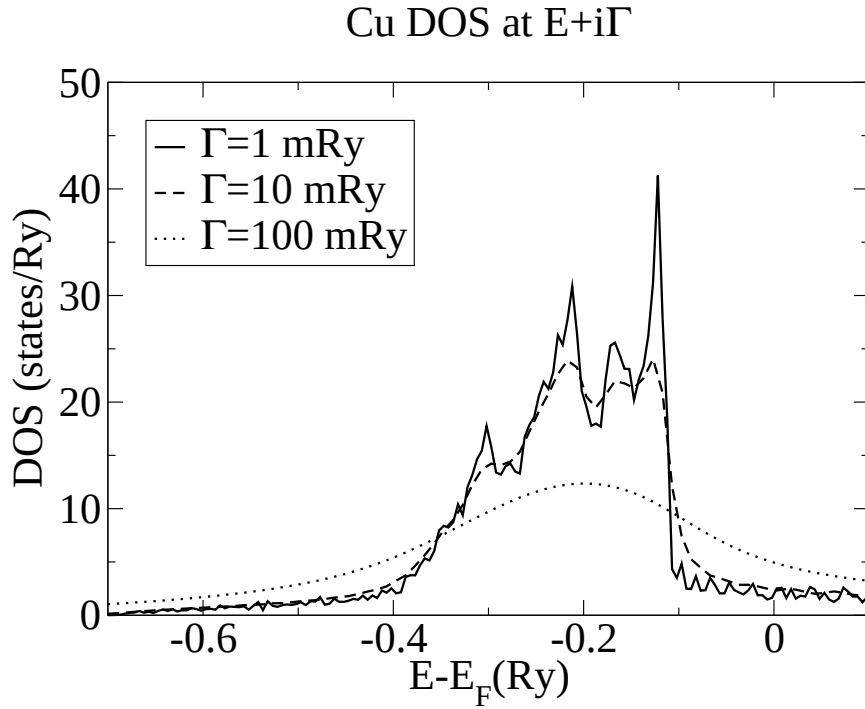


Figure 3. Density of states (DOS) of fcc Cu calculated for $\Gamma = 100$ mRy (typical for self-consistency calculations), 10 mRy, and 1 mRy (typical for density-of-states calculations). The smoothening of the Green function for larger Γ , leading to a drastic decrease of the number of necessary integration points, is evident.

In order to demonstrate the smoothening of the Green function with increasing Γ , we show in Figure 3 the density of states of fcc Cu calculated for $\Gamma = 100$ mRy (typical for self-consistency calculations), 10 mRy, and 1 mRy (typical for density-of-states calculations).

Appendix

Real Spherical Harmonics

The real spherical harmonics Y_{lm} are defined by a unitary transformation of the (usual) complex spherical harmonics Y_{lm}^c :

$$Y_{lm} = \begin{cases} \frac{1}{\sqrt{2}}(Y_{lm}^c + Y_{l,-m}^c) & m > 0 \\ Y_{l,0}^c & m = 0 \\ \frac{-i}{\sqrt{2}}(Y_{l,|m|}^c - Y_{l,-|m|}^c) & m < 0. \end{cases} \quad (148)$$

The complex spherical harmonics are defined as

$$Y_{lm}^c(\theta, \phi) = (-1)^{(m+|m|)/2} \sqrt{\frac{2l+1}{4\pi} \frac{(l-|m|)!}{(l+|m|)!}} P_l^{|m|}(\cos \theta) e^{im\phi} \quad (149)$$

where $P_l^{|m|}(x)$ are the Legendre functions

$$P_l^{|m|}(x) = \frac{1}{2^l l!} (1-x^2)^{|m|/2} \frac{d^{l+|m|}}{dx^{l+|m|}} (x^2-1)^l. \quad (150)$$

References

1. J. Korringa, *Physica* **13**, 392 (1947).
2. W. Kohn and N. Rostoker, *Phys. Rev.* **94**, 1111 (1954).
3. T. H. Dupree, *Ann. Phys. (N. Y.)*, **15**, 63 (1961); J. L. Beeby, *Proc. Roy. Soc. London Ser. A* **302**, 113 (1967); G. J. Morgan, *Proc. Phys. Soc.* **89**, 365 (1966).
4. For a review of the KKR method, see N. Papanikolaou, R. Zeller, and P. H. Dederichs, *J. Phys.: Condens. Matter* **14**, 2799 (2002).
5. R. Zeller and P. H. Dederichs, *Phys. Rev. Lett.* **42**, 1713 (1979).
6. I. Mertig, E. Mrosan, and P. Ziesche, *Multiple Scattering Theory of Point Defects in Metals: Electronic Properties*, Teubner-Texte zur Physik 11, Teubner Verlag Leipzig (1987).
7. For a comprehensive presentation of the KKR and KKR-CPA methods and many of their applications, see: A. Gonis, *Theoretical Materials Science*, Materials Research Society (2000).
8. R. Zeller, P. H. Dederichs, B. Újfalussy, L. Szunyogh, and P. Weinberger, *Phys. Rev. B* **52**, 8807 (1995); K. Wildberger, R. Zeller, and P. H. Dederichs, *Phys. Rev. B* **55**, 10074 (1997).
9. J. Banhart, H. Ebert, P. Weinberger, and J. Voithländer, *Phys. Rev. B* **50**, 2104 (1994); W. H. Butler, *Phys. Rev. B* **31**, 3260 (1995).
10. Ph. Mavropoulos, N. Papanikolaou, and P. H. Dederichs, *Phys. Rev. B* **69**, 125104 (2004).
11. J. Minar, L. Chioncel, A. Perlov, H. Ebert, M. I. Katsnelson, and A. I. Lichtenstein, *Phys. Rev. B* **72**, 045125 (2005).

12. M. Asato, A. Settels, T. Hoshino, T. Asada, S. Blügel, R. Zeller, and P. H. Dederichs, Phys. Rev. B **60**, 5202 (1999).
13. I. Galanakis, G. Bihlmayer, V. Bellini, N. Papanikolaou, R. Zeller, S. Blügel, and P.H. Dederichs, Europhys. Lett. **58**, 751 (2002).
14. M. Freyss, N. Papanikolaou, V. Bellini, R. Zeller, and P. H. Dederichs, Phys. Rev. B **66**, 014445 (2002).
15. B. Nonas, K. Wildberger, R. Zeller, P. H. Dederichs, and B. L. Gyorffy, Phys. Rev. B **57**, 84 (1998).
16. H. Ebert and S. Mankovsky, Phys. Rev. Lett. **90**, 077404 (2003).
17. Ph. Mavropoulos, O. Wunnicke, and P. H. Dederichs, Phys. Rev. B **66**, 024416 (2002).
18. N. Papanikolaou, J. Opitz, P. Zahn, and I. Mertig, Phys. Rev. B **66**, 165441 (2002).
19. A. Vernes, H. Ebert, and J. Banhart, Phys. Rev. B **68**, 134404 (2003).
20. V. Popescu, H. Ebert, B. Nonas, and P. H. Dederichs, Phys. Rev. B **64**, 184407 (2001).
21. N. Stefanou, V. Karathanos, and A. Modinos, J. Phys. Condens. Matter **4**, 7389 (1992).
22. I. E. Psarobas, N. Stefanou, and A. Modinos, Phys. Rev. B **62**, 278 (2000); R. Sainidou, N. Stefanou, and A. Modinos, Phys. Rev. B **69**, 064301 (2004).
23. A. Modinos, V. Yannopapas, and N. Stefanou, Phys. Rev. B **61**, 8099 (2000).
24. R. G. Newton, *Scattering Theory of Waves and Particles*, second edition, Springer (1982); reprint by Dover (2002).
25. E. N. Economou, *Green's functions in Quantum Physics*, Springer Series in Solid-State Sciences 7, Springer Verlag Berlin (1979).
26. G. Arfken, *Mathematical Methods for Physicists*, Academic Press International Edition (1970).
27. R. Zeller, J. Phys. C: Solid State Phys. **20** 2347 (1987).
28. B. Drittler, M. Weinert, R. Zeller, and P. H. Dederichs, Solid State Commun. **79**, 31 (1991).
29. N. Stefanou, H. Akai, and R. Zeller, Comp. Phys. Commun. **60**, 231 (1990); N. Stefanou and R. Zeller, J. Phys.: Condens. Matter **3**, 7599 (1991).
30. O. K. Andersen and O. Jepsen, Phys. Rev. Lett. **53**, 2571 (1984).
31. B. Wenzien, J. Kudrnovský, V. Drchal, and M. Šob, J. Phys.: Condens. Matter **1**, 9893 (1989).
32. J. Kudrnovský, V. Drchal, I. Turek, and P. Weinberger, Phys. Rev. B **50**, 16105 (1994).
33. K. Wildberger, P. Lang, R. Zeller, and P. H. Dederichs, Phys. Rev. B **52**, 11502 (1995).

Molecular Dynamics - Vision and Reality

Godehard Sutmann

John von Neumann Institute for Computing
Central Institute for Applied Mathematics
Forschungszentrum Jülich
52425 Jülich, Germany
E-mail: g.sutmann@fz-juelich.de

“We may regard the present state of the universe as the effect of its past and the cause of its future. An intellect which at any given moment knew all of the forces that animate nature and the mutual positions of the beings that compose it, if this intellect were vast enough to submit the data to analysis, could condense into a single formula the movement of the greatest bodies of the universe and that of the lightest atom; for such an intellect nothing could be uncertain and the future just like the past would be present before its eyes.”¹

Marquis Pierre Simon de Laplace, 1814

1 Introduction

With the age of enlightenment and the development of mathematical tools the vision of computability and predictability of natural phenomena arose among scientists and philosophers. Pierre Simon de Laplace phrased this vision in terms of a controlling, omniscient instance which would be able to look into the future as well as into the past due to the deterministic nature of processes, governed by the solution of differential equations. This omniscient instance, introduced by Laplace was henceforth called the *Laplace demon*. This rational view of Laplace, however, had to be corrected with the advent of chaos theory, starting with the work of Poincaré, which states that minimal changes in initial configurations of nonlinear differential equations might lead to a diverging behavior between solutions. Although the general view of Laplace’s vision is corrected nowadays by chaos theory and quantum mechanics, it expresses two main features of classical mechanics, i.e. (i) determinism of processes and (ii) time reversibility of the fundamental equations. His understanding of nature was one of the first ideas for doing molecular dynamics simulations, i.e. considering an isolated system of particles, the behavior of which is fully determined by the solution of the classical equations of motion

$$\dot{\mathbf{p}}_i = -\frac{\partial \mathcal{H}}{\partial \mathbf{q}_i} \quad , \quad \dot{\mathbf{q}}_i = \frac{\partial \mathcal{H}}{\partial \mathbf{p}_i} \quad (1)$$

where $\mathcal{H}$ is the Hamiltonian of the system and $\mathbf{p}_i$, $\mathbf{q}_i$ are the generalized momenta and coordinates of particle with index i . However, as is shown in Section 3, even for small systems, which are precisely described by initial and boundary conditions, the vision of Laplace is not fulfilled.

Nevertheless, the main line, which governs the idea of *Laplace's demon* is found in modern simulation methodologies and computer simulations based on this principle have become a powerful tool to treat the dynamical behavior of nonlinear many-body systems. The initial conditions as well as boundary conditions enter either from theoretical considerations (limiting laws) or from experimental values. If the physical system is fully characterized, simulations may be an indispensable tool to solve theoretical models beyond certain approximations or to provide additional help for experimentalists to get a deeper view into phenomena or to answer questions which are not possible to treat with current experimental facilities.

Although there are different methods to obtain information about complex systems, particle simulations always require a model for the interaction between system constituents. The model can only be considered as an approximation to reality, but computer simulations should provide a tool to solve this model exactly (within numerical precision). In order to get a connection to real world systems, this model has to be tested against experimental results, i.e. it should reproduce or approximate experimental findings like distribution functions or phase diagrams and theoretical constraints, i.e. it should obey certain fundamental or limiting laws like energy conservation.

Concerning MD simulations the ingredients for a program are basically threefold:

- (i) As already mentioned, a model for the interaction between system constituents (atoms, molecules, surfaces etc.) is needed. Often, it is assumed that particles interact only pairwise, which is exact e.g. for particles with fixed partial charges. This assumption greatly reduces the computational effort and the work to implement the model into the program.
- (ii) An integrator is needed, which propagates particle positions and velocities from time t to $t + \delta t$. It is a finite difference scheme which moves trajectories discretely in time. The time step δt has properly to be chosen to guarantee stability of the integrator, i.e. there should be no drift in the system's energy.
- (iii) A statistical ensemble has to be chosen, where thermodynamic quantities like pressure, temperature or the number of particles are controlled. The natural choice of an ensemble in MD simulations is the microcanonical ensemble (NVE), since the system's Hamiltonian without external potentials is a conserved quantity. Nevertheless, there are extensions to the Hamiltonian which also allow to simulate different statistical ensembles.

These steps essentially define an MD simulation. Having this tool at hand, it is possible to obtain *exact* results within numerical precision. Results are only correct with respect to the model which enters into the simulation and they have to be tested against theoretical predictions and experimental findings. If the simulation results differ from the *real system* properties or are incompatible with *solid* theoretical manifestations, the model has to be refined. This procedure can be understood as an adaptive refinement which leads in the end to an approximation of a model of the *real world* at least for certain properties. The model itself may be constructed from plausible considerations, where parameters are chosen from neutron diffraction or NMR measurements. It may also result from first principle investigations, like quantum *ab initio* calculations. Although the electronic distribution of the particles is calculated very accurately, this type of model building contains also some approximations, since many-body interactions are mostly neglected (this would increase the parameter space in the model calculation enormously). However, it often provides a good starting point for a realistic model.

The problem which is often met in molecular simulations is the existence of a variety

of time scales, which govern the physical system. Since molecular dynamics consists of an integration of differential equations, one often has to take into account the fastest motion in the system by imposing a small time step of integration in order to sample the movements correctly and conserve first integrals. Considering e.g. simulations of DNA, there is a huge gap between time scales which govern certain phenomena, i.e. fast atomic vibrations take place on a sub-picosecond time scale (10^{-14} s), twisting of the molecule on a picosecond scale (10^{-12} s), bending of the whole molecule on a nanosecond scale ($10^{-7} - 10^{10}$ s) and supercoiling of the molecule on a broad time scale band up to seconds ($10^{-6} - 10^0$ s)² (similar range of timescales holds for protein dynamics³). Therefore, *brute-force* calculations have to be done on the most powerful computers at hand, or approximations have to be done, reducing system components in order to coarsen the dynamics of the system. In the second approach it is of course very important to keep the essential ingredients which still reproduce global motions or long time dynamics. Examples for this approach may be found e.g. in Ref. 4. An impressive example for the first approach was given by the group of Peter Kollman, where a small protein (subdomain HP-36 from the villin headpiece) was studied in an all-atom simulation for a micro-second time interval^{5,6}. Although this is still a small system, consisting of about 10000 atoms (600 protein atoms and 3000 water molecules), CRAY T3D and CRAY T3E machines with 256 processors had to be kept busy (continuously) for several months. This huge computational work had to be spent because 5×10^8 steps of integration had to be performed because of the small time step of integration of $\delta t = 2 \times 10^{-15}$ s which was necessary to apply in order to resolve high frequency motions in the system. Although with the development of computer hardware the time scales as well as the system sizes can be extended more and more it is still a current field of research how to model and simulate complex systems on a long time scale⁷.

It is clear that the performance of particle dynamics simulations strongly depends on the computer facilities at hand. The first studies using MD simulation techniques were performed in 1957 by B. J. Alder and T. E. Wainright^{8,9} who simulated systems of hard spheres. In this early simulation, which was run on an IBM-704, up to 500 particles could be simulated, for which 500 collisions per hour could be calculated. Taking into account 200000 collisions for a production run, these simulations lasted for more than two weeks. Nowadays, systems for several million or even billion particles are performed¹⁰⁻¹². These huge systems make parallel computing an indispensable tool. Fortunately, molecular dynamics is in principle possible to parallelize to 100% (not taking into account input/output operations). Therefore, the problem sizes and timescales are usually extended with the inauguration of more powerful parallel systems. Nevertheless it is still challenging to develop programs which are able to scale up to thousands of processors. In principle, new architectures, like the IBM Blue Gene/L^{13,14} should provide a platform for tackling the grand challenge problems, e.g. protein folding. Nevertheless, practical experience tells that parallel programs do not scale good enough to use numbers of processors up to 10000 or more. This makes it necessary to improve parallel algorithms and to think in new algorithms, which overcome different time and length scales^{7,15}.

Classical molecular dynamics methods are nowadays applied to a huge class of problems, e.g. properties of liquids, defects in solids, fracture, surface properties, friction, molecular clusters, polyelectrolytes and biomolecules. Due to the large area of applicability, simulation codes for molecular dynamics were developed by many groups. On the internet homepage of the Collaborative Computational Project No.5 (CCP5)¹⁶ there are a

lot of computer codes assembled for condensed phase dynamics. During the last years several programs were designed for parallel computers. Among them, which are partly available free of charge, are, e.g., Amber/Sander¹⁷, CHARMM¹⁸, NAMD¹⁹, NWCHEM²⁰ and LAMMPS²¹.

2 Models and Methods for Particle Interactions

The part of a simulation where physics come into play is the modeling of interactions between particles. This is the part of a molecular dynamics program which makes the difference between simulating a galaxy or simulating a droplet of water molecules. In general the physical system is determined by its Hamiltonian (or other way around, the Hamiltonian is written down according to the system under consideration). It can be written as *intrinsic* part $\mathcal{H}_0$ and *external* part $\mathcal{H}_1(t)$

$$\mathcal{H} = \mathcal{H}_0 + \mathcal{H}_1(t) \quad (2)$$

where the time-dependence of $\mathcal{H}_1$ indicates that time-varying external fields, e.g. sinusoidal laser beams, may enter into the energetic description. If the external part of the Hamiltonian is omitted then it is clear from classical mechanics that the system Hamiltonian is a conserved quantity.

The intrinsic part of the Hamiltonian can often be written as

$$\begin{aligned} \mathcal{H}_0 = & \frac{1}{2} \mathbf{p}^T \mathbf{M}^{-1} \mathbf{p} + \frac{1}{2} \boldsymbol{\omega}^T \boldsymbol{\theta}^{-1} \boldsymbol{\omega} + \sum_{i=1}^{N-1} \sum_{j=i+1}^N u(r_{ij}) \\ & + \sum_{i=1}^{N-2} \sum_{j=i+1}^{N-1} \sum_{k=j+1}^N u^{(3)}(r_{ij}, r_{ik}, r_{jk}) + \dots \end{aligned} \quad (3)$$

where $\mathbf{p}$ is a 3N-dimensional vector of particle momenta, $\mathbf{M}$ a diagonal mass matrix, $r_{ij} = |\mathbf{r}_i - \mathbf{r}_j|$ the distance between particle pairs and $u(r)$ a pair potential function. Furthermore, if simulating rigid bodies or molecules with fixed atomic distances, $\boldsymbol{\omega}$ is the angular velocity and $\boldsymbol{\theta}$ the tensor of inertia. If not only pair interactions are to be considered, 3-body potentials $u^{(3)}$, or multi-body potentials $u^{(n)}$ can be included into the Hamiltonian. Mainly this is avoided, since it is not easy to model and also it is rather time consuming to evaluate potentials and forces originating from these many-body terms.

Roughly speaking the potential functions can be classified into two main groups, namely short-range and long-range interactions. One can estimate the range of a potential function by considering the leading term of its expansion in powers of r^{-n} . The integral over the leading term gives an estimate of the range of influence, i.e.

$$\int \frac{1}{r^n} dr^d = \begin{cases} \text{finite} & : \text{short range} \\ \infty & : \text{long range} \end{cases} \quad (4)$$

where d is the dimension of the physical space. Therefore, in 3-dimensional space, potentials dropping as r^{-3} , e.g. dipole-dipole interactions, are still long ranged.

There may be different terms contributing to the interaction potential between particles, i.e. there is no universal expression, as one can imagine for first principles calculations. In fact, contributions to interactions depend on the model which is used and this is the result of collecting various contributions into different terms, coarse graining interactions or

imposing constraints, to name a few. Generally one can distinguish between bonded and non-bonded terms, or intra- and inter-molecular terms. The first class denotes all contributions originating between particles which are closely related to each other by constraints or potentials which guaranty defined particles as close neighbors. The second class denotes interactions between particles which can *freely* move, i.e. there are no defined neighbors, but interactions simply depend on distances.

A typical form for a (so-called) force field (e.g. AMBER¹⁷) looks as follows

$$\mathcal{U} = \sum_{\text{bonds}} K_r (r - r_{eq})^2 + \sum_{\text{angles}} K_\theta (\theta - \theta_{eq})^2 + \sum_{\text{dihedrals}} \frac{V_n}{2} [1 + \cos(n\phi - \gamma)] \quad (5)$$

$$+ \sum_{i < j} \left[\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^6} \right] + \sum_{\text{H-bonds}} \left[\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} \right] + \sum_{i < j} \frac{q_i q_j}{r_{ij}}$$

In the following, short- and long-range interaction potentials and methods are briefly described in order to show differences in their algorithmical treatment.

2.1 Short Range Interactions

In Eq. (5) the first three terms are examples for bonded, short range interactions. They consist of bond vibrations (i.e. two molecules oscillating in a potential), angular vibrations (i.e. varying bond angle between three molecules) and dihedral motion (i.e. torsional deformation of four bonded particles). It is clear by definition that these interactions are short ranged. Since the bond partners between molecules are fixed in a simulation (in principle it is possible to include bond breaking in an approximate way²² also in a classical simulation), one can set up fixed lists in the beginning of the simulation, containing all indices of particle-pairs, -triples, -quadruples etc. which exhibit bonded interactions. This omits lengthy checks, which particles take part in an interaction.

The next two terms are examples for short range, non-bonded terms. The first is the famous Lennard-Jones potential, which includes a strongly repelling term, modeling the Pauli exclusion, i.e. hard core repulsion. The attracting term on the other hand models the effect of interactions between induced dipoles, attraction due to fluctuating charge distributions. The other term is sometimes included to give a more realistic description of hydrogen bonds. This term is only evaluated between special particle pairs, i.e. hydrogens and electronegative atoms, which are able to form hydrogen bonds. The constants, A_{ij} , B_{ij} , C_{ij} , D_{ij} are in general different for different atom types or pairs, i.e. they have to be adjusted by proper modeling. Often the constants are evaluated for single atom types and the parameters for cross terms between different atoms are calculated by combination rules.

The Lennard-Jones and the hydrogen bond terms can be written in a different way

$$u(r_{i\alpha,j\beta}) = 4\epsilon_{\alpha\beta} \left(\left(\frac{\sigma_{\alpha\beta}}{r_{i\alpha,j\beta}} \right)^a - \left(\frac{\sigma_{\alpha\beta}}{r_{i\alpha,j\beta}} \right)^b \right) \quad (6)$$

where i, j denotes different particle indices and α, β different types of atoms (for Lennard-Jones e.g. it is $\sigma = (A/B)^{-6}$, $\epsilon = B^2/4A$, $a = 12$ and $b = 6$). The most frequently applied combining rule is the Lorentz-Berthelot formula²³

$$\sigma_{\alpha\beta} = \frac{\sigma_{\alpha\alpha} + \sigma_{\beta\beta}}{2}, \quad \epsilon_{\alpha\beta} = \sqrt{\epsilon_{\alpha\alpha}\epsilon_{\beta\beta}} \quad (7)$$

It is simply noted here that there are also more complicated rules, e.g. the Kong rule²⁴ or the Waldman-Hagler rule²⁵ which sometimes give better results in the calculation of pressure-density profiles of liquids²⁶.

Since these potentials are short ranged one can restrict the evaluation to limited region of space. Usually one introduces a cutoff radius, R_c , beyond which the interactions are set to zero. A common choice for the cutoff radius is given by $R_c = 2.5\sigma$. Since at R_c the interaction is not identical zero, it is clear that particles entering from outside R_c into inside region exhibit a jump in potential and forces. Formally, the force at the cutoff distance ($\mathbf{F}_{ij} = -\nabla u(r_{ij})$) is infinitely large, since the potential exhibits a step due to truncation. This sudden acceleration of particles usually leads to a heating of the system, since the motion is not reversible. Consider, e.g. a two-particle system, where one particle comes from outside and moving with velocity v_0 to an interparticle distance $r = R_c - \varepsilon$, with $\varepsilon \ll 1$. Then it gets an abrupt force contribution, accelerating it, leading to a velocity $v_1 > v_0$. On the other hand, if a particle starts from $R_c - \varepsilon$ with velocity v_1 to leave the sphere with radius R_c , then it still has velocity $v_1 > v_0$. For a many-particle system this means that $\langle v^2 \rangle$ increases due to many crossings and recrossings of the interaction sphere, i.e. the temperature increases.

In order to avoid this statistical effect, one may either introduce smoothing functions, which continuously drop the potential and the forces to zero. The disadvantage with this approach is that there will be a zone of large forces at the cutoff distance if the forces are properly evaluated as derivatives of the potential. Therefore one has to smooth forces in this region, leading however to a non-conservative system. Therefore a different method is most often used, which consists in shifting the whole potential and force by a certain amount, which guarantees that both the potential and the force are exactly zero at the cutoff distance, i.e.

$$u^{(sfp)}(r_{ij}) = \begin{cases} u(r_{ij}) - u(R_c) + (r_{ij} - R_c) F(R_c) & : r_{ij} \leq R_c \\ 0 & : r_{ij} > R_c \end{cases} \quad (8)$$

$$\mathbf{F}^{(sfp)}(\mathbf{r}_{ij}) = \begin{cases} \mathbf{F}(\mathbf{r}_{ij}) - F(R_c) \hat{\mathbf{r}}_{ij} & : r_{ij} \leq R_c \\ 0 & : r_{ij} > R_c \end{cases} \quad (9)$$

This ensures a smooth transition from outside to inside the cutoff region and vice versa.

Since there are only relatively few particles which have to be considered for the interaction with a tagged particle (i.e. those particles within the cutoff range), it would be a computational bottleneck if in any time step all particle pairs would have to be checked whether they lie inside or outside the interaction range. This becomes more and more a problem as the number of particles increases. A way to overcome this bottleneck is to introduce list techniques. The first implementation dates back to the early days of molecular dynamics simulations. In 1967, Verlet introduced a list²⁷, where at a given time step all particle pairs were stored within a range $R_c + R_s$, where R_s is called the skin radius and which serves as a reservoir of particles, in order not to update the list in each time step (which would make the list redundant). Therefore, in a force routine, not all particles have to be tested, whether they are in a range $r_{ij} < R_c$, but only those particle pairs, stored in the list. Since particles are moving during the simulation, it is necessary to update the list from

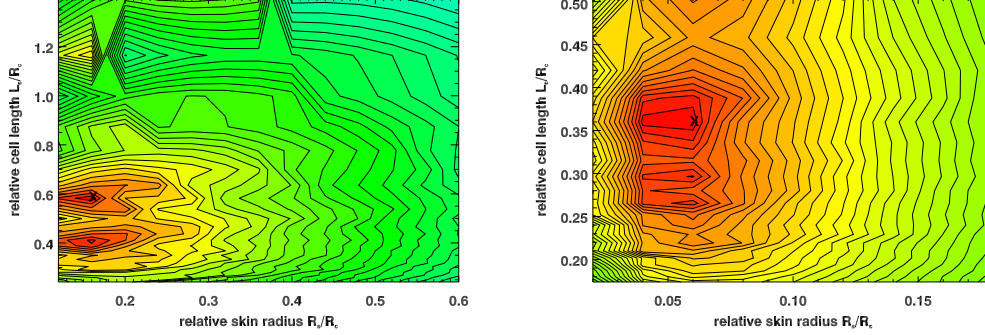


Figure 1. Contour plots of the performance for the combination of linked-cell and Verlet list as a function of the cell length and the size of the skin radius. Crosses mark the positions predicted from an optimization procedure²⁹. Test systems were composed of 4000 Lennard-Jones particles with $R_c = 2.5 \sigma$ at temperature $T = 1.4 \epsilon/k_B$. Left: $\rho = 0.75/\sigma^3$. Right: $\rho = 2.0/\sigma^3$.

time to time. A criterion to update the list could be, e.g.

$$\max_i |\mathbf{r}_i(t) - \mathbf{r}_i(t_{upd})| \geq \frac{R_s}{2} \quad (10)$$

where t_{upd} is the time from the last list update. This ensures that particles cannot move from the outside region into the cutoff sphere without being recognized. This technique, though efficient, has still complexity $\mathcal{O}(N^2)$, since at an update step, *all* particle pairs have to be checked for their mutual distances. Another problem arises when simulating many particles, since the memory requirements are relatively large (size of the list is $4\pi(R_c + R_s)^3 \rho N/3$). There is, of course also the question, how large the skin radius should be chosen. Often, it is chosen as $R_s = 1.5\sigma$. In Ref. 28 it was shown that an optimal choice strongly depends on the number of particles in the system and an optimization procedure was outlined.

An alternative list technique, which scales linearly with the number of particles is the linked-cell method^{30,31}. The linked-cell method starts with subdividing the whole system into cubic cells and sorting all particles into these cells according to their position. The size of the cells, L_c , is chosen to be $L_c \leq L_{Box}/\text{floor}(L_{Box}/R_c)$, where L_{Box} is the length of the simulation box. All particles are then sorted into a list array of length N . The list is organized in a way that particles, belonging to the same cell are linked together, i.e. the entry in the list referring to a particle points directly to the entry of a next particle inside the same cell. A zero entry in the list stops the search in the cell and a next cell is checked for entries. This technique not only has computational complexity of $\mathcal{O}(N)$, since the sorting into the cells and into the N -dimensional array is of $\mathcal{O}(N)$, but also has memory requirements which only grow linearly with the number of particles. These features make this technique very appealing. However, the technique is not well vectorizable and also the addressing of next neighbors in the cells require indirect access (e.g. `i=index(i)`), which may lead to cache misses. In order not to miss any particle pair in the interactions every box has to have a neighbor region in each direction which extends to R_c . In the case, where $L_c \geq R_c$, every cell is surrounded by 26 neighbor cells in three dimensional systems. This gives rise

to the fact that the method gives only efficiency gains if $L_{Box} \geq 4R_c$, i.e. subdividing each box direction into more than 3 cells. In order to approximate the cutoff sphere in a better way by cubic cells, one may reduce the cell size and simultaneously increasing the total number of cells. In an optimization procedure²⁸, it was found that a reduction of cell sizes to $L_c = R_c/2$ or even smaller often gives very much better results.

It is, of course, possible to combine these list techniques, i.e. using the linked-cell technique in the update step of the Verlet list. This reduces the computational complexity of the Verlet list to $\mathcal{O}(N)$ while fully preserving the efficiency of the list technique. It is also possible to model the performance of this list combination and to optimize the length of the cells and the size of the skin radius. Figure 1 shows the result of a parameter study, where the performance of the list was measured as a function of (L_c, R_s) . Also shown is the prediction of parameters coming out of an optimization procedure²⁹.

2.2 Long Range Interactions

Long range interactions essentially require to take all particle pairs into account for a proper treatment of interactions. This may become a problem, if periodic boundary conditions are imposed to the system, i.e. formally simulating an infinite number of particles (no explicit boundaries imply infinite extend of the system). Therefore one has to devise special techniques to treat this situation. On the other hand one also has to apply fast techniques to overcome the inherent $\mathcal{O}(N^2)$ complexity of the problem, since for large numbers of particles this would imply an intractable computational bottleneck. In general one can classify algorithms for long range interactions into the following system:

- Periodic boundary conditions
 - Grid free algorithms, e.g. Ewald summation method^{32–34}
 - Grid based algorithms, e.g. Smoothed Particle Mesh Ewald^{35,36}, Particle-Particle-Particle-Mesh method^{37–39}
- Open boundary conditions
 - Grid free algorithms, e.g. Fast Multipole Method^{40–45} (FMM), Barnes-Hut Tree method^{46,47}
 - Grid based algorithms, e.g. Particle-Particle Particle-Multigrid method⁴⁸ (P³Mg), Particle Mesh Wavelet method⁴⁹ (PMW)

In the following two important members of these classes will be described, the Ewald summation method and the Fast Multipole Method.

2.2.1 Ewald Summation Method

The Ewald summation method originates from crystal physics, where the problem was to determine the Madelung constant⁵⁰, describing a factor for an effective electrostatic energy in a perfect periodic crystal. Considering the electrostatic energy of a system of N particles in a cubic box and imposing periodic boundary conditions, leads to an equivalent problem.

At position $\mathbf{r}_i$ of particle i , the electrostatic potential, $\phi(\mathbf{r}_i)$, can be written down as a lattice sum

$$\phi(\mathbf{r}_i) = \sum_{\mathbf{n}}^{\dagger} \sum_{j=1}^N \frac{q_j}{|\mathbf{r}_{ij} + \mathbf{n}L|} \quad (11)$$

where $\mathbf{n} = (n_x, n_y, n_z)$, $n_\alpha \in \mathbb{Z}$ is a vector along cartesian coordinates and L is the length of the simulation box. The sign " $\dagger$ " means that $i \neq j$ for $|\mathbf{n}| = 0$.

Eq. (11) is conditionally convergent, i.e. the result of the outcome depends on the order of summation. Also the sum extends over infinite number of lattice vectors, which means that one has to modify the procedure in order to get an absolute convergent sum and to get it fast converging. The original method of Ewald consisted in introducing a convergence factor e^{-ns} , which makes the sum absolute convergent; then transforming it into different fast converging terms and then putting s in the convergence factor to zero. The final result of the calculation can be easier understood from a physical picture. If every charge in the system is screened by a counter charge of opposite sign, which is smeared out, then the potential of this composite charge distribution becomes short ranged (it is similar in electrolytic solutions, where ionic charges are screened by counter charges - the result is an exponentially decaying function, the Debye potential⁵¹). In order to compensate for the added charge distribution it has to be subtracted again. The far field of a localized charge distribution is, however, again a Coulomb potential. Therefore this term will be long ranged. There would be nothing gained if one would simply sum up these different terms. The efficiency gain shows up, when one calculates the short range interactions as direct particle-particle contributions in real space, while summing up the long range part of the smeared charge cloud in reciprocal Fourier space. Choosing as the smeared charge distribution a Gaussian charge cloud of half width $1/\alpha$ the corresponding expression for the energy becomes

$$\begin{aligned} \phi(\mathbf{r}_i) = & \sum_{\mathbf{n}}^{\dagger} \sum_{j=1}^N q_j \frac{\text{erfc}(\alpha|\mathbf{r}_{ij} + \mathbf{n}L|)}{|\mathbf{r}_{ij} + \mathbf{n}L|} \\ & + \frac{4\pi}{L^3} \sum_{\mathbf{k} \neq 0} \sum_{j=1}^N \frac{q_j}{|\mathbf{k}|^2} e^{-|\mathbf{k}|^2/4\alpha^2} e^{i\mathbf{k}\mathbf{r}_{ij}} - q_i \frac{2\alpha}{\sqrt{\pi}} \end{aligned} \quad (12)$$

The last term corresponds to a self energy contribution which has to be subtracted, as it is considered in the Fourier part. Eq. (12) is an exact equivalent of Eq. (11), with the difference that it is an absolute converging expression. Therefore nothing would be gained without further approximation. Since the complimentary error function can be approximated for large arguments by a Gaussian function and the k-space parts decreases like a Gaussian, both terms can be approximated by stopping the sums at a certain lattice vector $\mathbf{n}$ and a maximal k -value k_{max} . The choice of parameters depends on the error, $\epsilon = \exp(-p^2)$, which one accepts to tolerate. Setting the error tolerance p and choosing the width of the counter charge distribution, one gets

$$R_c^2 + \frac{\log(R_c)}{\alpha^2} = \frac{1}{\alpha^2} (p^2 - \log(2)) \quad (13)$$

$$k_{max}^2 + 8\alpha^2 \log(k_{max}) = 4\alpha^2 p^2 + \log\left(\frac{4\pi}{L^3}\right) \quad (14)$$

This can be solved iteratively or if one is only interested in an approximate estimate for the error, i.e. neglecting logarithmic terms, one gets

$$R_c = \frac{p}{\alpha} \quad (15)$$

$$k_{max} = 2\alpha p \quad (16)$$

Using this error estimate and furthermore introducing execution times, spent for the real- and reciprocal-space part, it is possible to show that parameters R_c , α and k_{max} can be chosen to get a complexity of $\mathcal{O}(N^{3/2})$ for the Ewald sum^{52,53}. In this case, parameters are

$$\frac{R_c}{L} \approx \sqrt{\frac{\pi}{N^{1/3}}} \quad , \quad \alpha L \approx \frac{L k_{max}}{2\pi} = \sqrt{\pi N^{1/3}} \quad (17)$$

Figure 2 shows the contributions of real- and reciprocal parts in Eq. (12), as a function of the spreading parameter α , where an upper limit in both the real- and reciprocal-contributions was applied. In the real-space part usually one restricts the sum to $|\mathbf{n}| = 0$ and applies a spherical cutoff radius, R_c . For fixed values of R_c and k_{max} there is a broad plateau region, where the two terms add up to a constant value. Within this plateau region, a value for α should be chosen. Often it is chosen according to $\alpha = 5/L$. Also shown is the potential energy of a particle, calculated with the Ewald sum. It is well observed that due to the periodicity of the system the potential energy surface is not radial symmetric, which may cause problems for small numbers of particles in the system.

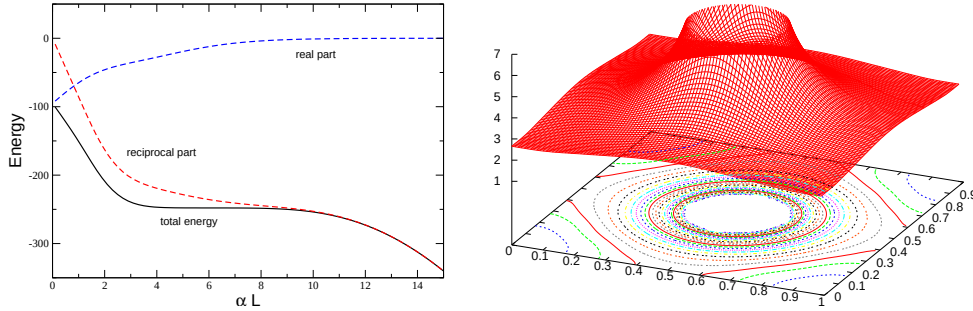


Figure 2. Left: Dependence of the calculated potential on the choice of the scaled inverse width, αL , of the smeared counter charge distribution. Parameters for this test were $N = 152$, $R_c = 0.5 L$ and $k_{max} L / 2\pi = 6$. Right: Surface plot and contours for the electrostatic potential of a charge, located in the center of the simulation volume. Picture shows the xy-plane for $z = L/2$. Parameters were $R_c = 0.25 L$, $\alpha L = 12.2$ and $k_{max} L / 2\pi = 6$.

2.2.2 The Fast Multipole Method

In open geometries there is no lattice summation, but only the sum over all particle pairs in the whole system. The electrostatic energy at a particle's position is therefore simply calculated as

$$\phi(\mathbf{r}_i) = \sum_{j=1}^N \frac{q_j}{|\mathbf{r}_i - \mathbf{r}_j|} \quad (18)$$

Without further approximation this is always an $\mathcal{O}(N^2)$ algorithm since there are $N(N-1)/2$ interactions to consider in the system (here Newton's third law was taken into account). The idea of a multipole method is to group particles which are far away from a tagged particle together and to consider an effective interaction of a particle with this particle group⁵⁴⁻⁵⁶. The physical space is therefore subdivided in a hierarchical way, where the whole system is considered as level 0. Each further level is constructed by dividing the length in each direction by a factor of two. The whole system is therefore subdivided into a hierarchy of boxes where each *parent box* contains eight *children boxes*. This subdivision is performed at maximum until the level, where each particle is located in an individual box. Often it is enough to stop the subdivision already at a lower level.

In the following it is convenient to work in spherical coordinates. The main principle of the method is that the interaction between two particles, located at $\mathbf{r} = r, \theta, \varphi$ and $\mathbf{a} = (a, \alpha, \beta)$ can be written as a multipole expansion⁵⁷

$$\frac{1}{|\mathbf{r} - \mathbf{a}|} = \sum_{l=0}^{\infty} \sum_{m=-l}^l \frac{(l-|m|)!}{(l+|m|)!} \frac{a^l}{r^{l+1}} P_{lm}(\cos \alpha) P_{lm}(\cos \theta) e^{-im(\beta-\varphi)} \quad (19)$$

where $P_{lm}(x)$ are associated Legendre polynomials⁵⁸. This expression requires that $a/r < 1$ and this gives a lower limit for the so-called *well separated* boxes. This makes it necessary to have at least one box between a tagged box and the zone, where contributions can be expanded into multipoles. Defining the operators

$$O_{lm}(\mathbf{a}) = a^l (l-|m|)! P_{lm}(\cos \alpha) e^{-im\beta} \quad (20)$$

$$M_{lm}(\mathbf{r}) = \frac{1}{r^{l+1}} \frac{1}{(l+|m|)!} P_{lm}(\cos \theta) e^{im\varphi} \quad (21)$$

with which Eq. (19) may simply be rewritten in a more compact way, it is possible to write further three operators, which are needed, in a compact scheme, i.e.

1.) a translation operator, which relates the multipole expansion of a point located at $\mathbf{a}$ to a multipole expansion of a point located at $\mathbf{a} + \mathbf{b}$

$$O_{lm}(\mathbf{a} + \mathbf{b}) = \sum_{j=0}^l \sum_{k=-j}^j A_{jk}^{lm}(\mathbf{b}) O_{jk}(\mathbf{a}) \quad , \quad A_{jk}^{lm}(\mathbf{b}) = O_{l-j, m-k}(\mathbf{b}) \quad (22)$$

2.) a transformation operator, which transforms a multipole expansion centered at the origin into a Taylor expansion centered at location $\mathbf{b}$

$$M_{lm}(\mathbf{a} - \mathbf{b}) = \sum_{j=0}^l \sum_{k=-j}^j B_{jk}^{lm}(\mathbf{b}) O_{jk}(\mathbf{a}) \quad , \quad B_{jk}^{lm}(\mathbf{b}) = M_{l+j, m+k}(\mathbf{b}) \quad (23)$$

3.) a translation operator, which translates a Taylor expansion of a point $\mathbf{r}$ about the origin into a Taylor expansion of $\mathbf{r}$ about a point $\mathbf{b}$

$$M_{lm}(\mathbf{r} - \mathbf{b}) = \sum_{j=0}^l \sum_{k=-j}^j C_{jk}^{lm}(\mathbf{b}) M_{jk}(\mathbf{r}) \quad , \quad C_{jk}^{lm}(\mathbf{b}) = A_{jk}^{lm}(\mathbf{b}) \quad (24)$$

The procedure to calculate interactions between particles is then subdivided into five passes. Figure 3 illustrates four of them. The first pass consists of calculating the multipole

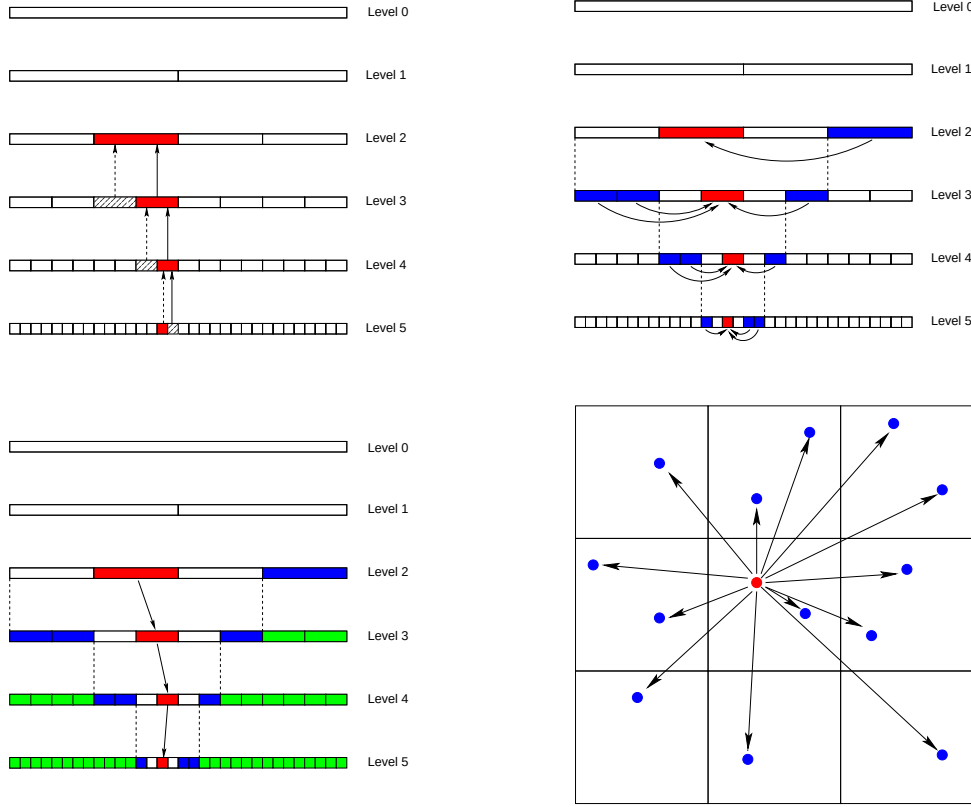


Figure 3. Schematic of different passes in the Fast Multipole Method. Upper left: Pass 1, evaluation of multipole terms in finest subdivision and translating information upwards the tree. Upper right: Pass 2, transforming multipole expansions in well separated boxes into local Taylor expansions. Lower left: Pass 3, transferring multipole expansions downwards the tree, thus collecting information of the whole system, except nearest neighbor boxes. Lower right: Pass 5, direct calculation of particle-particle interactions in local and nearest neighbor boxes.

expansions in the lowest level boxes (finest subdivision). Using the translation operator $O_{lm}(\mathbf{a} + \mathbf{b})$, the multipole expansions are translated into the center of their parent boxes and summed up. This procedure is repeated then subsequently for each level, until level 2 is reached, from where no further information is passed to a coarser level. In pass 2, using operator $M_{lm}(\mathbf{a} - \mathbf{b})$, multipole expansions are translated into Taylor expansions in a box from well separated boxes, whose parent boxes are nearest neighbor boxes. Well separated means, that for all particles in a given box the multipole expansion in a separated box is valid. Since the applicability of Eq. (19) implies $r > a$, well separateness means on level l that boxes should be separated by a distance 2^{-l} . This also explains, why there is no need to transfer information higher than level 2, since from there on it is not possible to have well separated boxes anymore, i.e. multipole expansions are not valid any more. In pass 3, using the operator $M_{lm}(\mathbf{a} - \mathbf{b})$, this information is then translated downwards the tree, so that finally on the finest level all multipole information is known in order to inter-

act individual particles with expansions, originating from all other particles in the system which are located in well separated boxes of the finest level. In pass 4 this interaction between individual particles and multipoles is performed. Finally in pass 5, explicit pair-pair interactions are calculated between particles in a lowest level box and those which are in nearest neighbor boxes, i.e. those boxes which are not called well separated.

It can be shown⁴² that each of the steps performed in this algorithm is of order $\mathcal{O}(N)$, making it an optimal method. Also the error made by this method can be controlled rather reliably⁴⁵. A very conservative error estimate is thereby given as^{54, 42, 59}

$$\left| \phi(r) - \frac{q}{|\mathbf{r} - \mathbf{a}|} \right| \leq \frac{|q|}{r - a} \left(\frac{a}{r} \right)^{p+1} \quad (25)$$

At the current description the evaluation of multipole terms scales as $\mathcal{O}(l_{max}^4)$, when l_{max} is the largest value of l in the multipole expansion, Eq. (19). A faster version which scales as $\mathcal{O}(l_{max}^3)$ and therefore strongly reducing the prefactor of the overall scheme, was proposed in Ref. 43, where multipoles are evaluated in a rotated coordinate frame, which makes it possible to reduce calculations to Legendre polynomials and not requiring associated Legendre polynomials.

Also to mention is that there are approaches to extend the Fast Multipole Method to periodic systems^{60, 61}.

3 The Integrator

The propagation of a classical particle system can be described by the temporal evolution of the phase space variables $(\mathbf{p}, \mathbf{q})$, where the phase space $\Gamma(\mathbf{p}, \mathbf{q}) \in \mathbb{R}^{6N}$ contains all possible combinations of momenta and coordinates of the system. The exact time evolution of the system is thereby given by a flow map

$$\Phi_{\delta t, \mathcal{H}} : \mathbb{R}^{6N} \rightarrow \mathbb{R}^{6N} \quad (26)$$

which means

$$\Phi_{\delta t, \mathcal{H}}(\mathbf{p}(t), \mathbf{q}(t)) = (\mathbf{p}(t) + \delta\mathbf{p}, \mathbf{q}(t) + \delta\mathbf{q}) \quad (27)$$

where

$$\mathbf{p} + \delta\mathbf{p} = \mathbf{p}(t + \delta t) \quad , \quad \mathbf{q} + \delta\mathbf{q} = \mathbf{q}(t + \delta t) \quad (28)$$

For a nonlinear many-body system, the equations of motion cannot be integrated exactly and one has to rely on numerical integration of a certain order. Propagating the coordinates by a constant step size h , a number of different finite difference schemes may be used for the integration. But there are a number of requirements, which have to be fulfilled in order to be useful for molecular dynamics simulations. An integrator, suitable for many-body simulations should fulfill the following requirements:

- Accuracy, i.e. the solution of an analytically solvable test problem should be as close as possible to the numerical one.
- Stability, i.e. very long simulation runs should produce physically relevant trajectories, which are not governed by numerical artifacts

- Conservativity, there should be no drift or divergence in conserved quantities, like energy, momentum or angular momentum
- Reversibility, i.e. it should have the same temporal structure as the underlying equations
- Effectiveness, i.e. it should allow for large time steps without entering instability and should require a minimum of force evaluations, which usually need about 95 % of CPU time per time step
- Symplecticity, i.e. the geometrical structure of the phase space should be conserved

It is obvious that the numerical flow, $\phi_{\delta t, \mathcal{H}}$, of a finite difference scheme will not be fully equivalent to $\Phi_{\delta t, \mathcal{H}}$, but the system dynamics will be described correctly if the items above will be fulfilled.

In the following the mentioned points will be discussed and a number of different integrators will be compared.

3.1 Basic Methods

The most simple integration scheme is the Euler method, which may be constructed by a first order difference approximation to the time derivative of the phase space variables

$$\mathbf{p}_{n+1} = \mathbf{p}_n - \delta t \frac{\partial}{\partial \mathbf{q}} \mathcal{H}(\mathbf{p}_n, \mathbf{q}_n) \quad (29)$$

$$\mathbf{q}_{n+1} = \mathbf{q}_n + \delta t \frac{\partial}{\partial \mathbf{p}} \mathcal{H}(\mathbf{p}_n, \mathbf{q}_n) \quad (30)$$

where δt is the step size of integration. This is equivalent to a Taylor expansion which is truncated after the first derivative. Therefore, it is obvious that it is of first order. Knowing all variables at step n , this scheme has all relevant information to perform the integration. Since only information from one time step is required to do the integration, this scheme is called the one-step explicit Euler scheme. The basic scheme, Eqs. (29,30) may also be written in different forms.

The implicit Euler method

$$\mathbf{p}_{n+1} = \mathbf{p}_n - \delta t \frac{\partial}{\partial \mathbf{q}} \mathcal{H}(\mathbf{p}_{n+1}, \mathbf{q}_{n+1}) \quad (31)$$

$$\mathbf{q}_{n+1} = \mathbf{q}_n + \delta t \frac{\partial}{\partial \mathbf{p}} \mathcal{H}(\mathbf{p}_{n+1}, \mathbf{q}_{n+1}) \quad (32)$$

can only be solved iteratively, since the derivative on the right-hand-side (*rhs*) is evaluated at the coordinate positions on the left-hand-side (*lhs*).

An example for a so-called partitioned Runge-Kutta method is the *velocity implicit method*

$$\mathbf{p}_{n+1} = \mathbf{p}_n - \delta t \frac{\partial}{\partial \mathbf{q}} \mathcal{H}(\mathbf{p}_{n+1}, \mathbf{q}_n) \quad (33)$$

$$\mathbf{q}_{n+1} = \mathbf{q}_n + \delta t \frac{\partial}{\partial \mathbf{p}} \mathcal{H}(\mathbf{p}_{n+1}, \mathbf{q}_n) \quad (34)$$

Since the Hamiltonian usually splits into kinetic $\mathcal{K}$ and potential $\mathcal{U}$ parts, which only depend on one phase space variable, i.e.

$$\mathcal{H}(\mathbf{p}, \mathbf{q}) = \frac{1}{2} \mathbf{p}^T \mathbf{M}^{-1} \mathbf{p} + \mathcal{U}(\mathbf{q}) \quad (35)$$

where $\mathbf{M}^{-1}$ is the inverse of the diagonal mass matrix, this scheme may also be written as

$$\mathbf{p}_{n+1} = \mathbf{p}_n - \delta t \frac{\partial}{\partial \mathbf{q}} \mathcal{U}(\mathbf{q}_n) \quad (36)$$

$$\mathbf{q}_{n+1} = \mathbf{q}_n + \frac{\delta t}{m} \mathbf{p}_{n+1} \quad (37)$$

showing that it is not necessary to solve it iteratively.

Obviously this may be written as a *position implicit method*

$$\mathbf{p}_{n+1} = \mathbf{p}_n - \delta t \frac{\partial}{\partial \mathbf{q}} \mathcal{U}(\mathbf{q}_{n+1}) \quad (38)$$

$$\mathbf{q}_{n+1} = \mathbf{q}_n + \frac{\delta t}{m} \mathbf{p}_n \quad (39)$$

Applying first Eq. (39) and afterwards Eq. (38) also this variant does not require an iterative procedure.

All of these schemes are first order accurate but have different properties, as will be shown below. Before discussing these schemes it will be interesting to show a higher order scheme, which is also based on a Taylor expansion. First write down expansions

$$\mathbf{q}(t + \delta t) = \mathbf{q}(t) + \delta t \dot{\mathbf{q}}(t) + \frac{1}{2} \delta t^2 \ddot{\mathbf{q}}(t) + O(\delta t^3) \quad (40)$$

$$= \mathbf{q}(t) + \frac{\delta t}{m} \mathbf{p}(t) + \frac{1}{2m} \delta t^2 \dot{\mathbf{p}}(t) + O(\delta t^3) \quad (41)$$

$$\mathbf{p}(t + \delta t) = \mathbf{p}(t) + \delta t \dot{\mathbf{p}}(t) + \frac{1}{2} \delta t^2 \ddot{\mathbf{p}}(t) + O(\delta t^3) \quad (42)$$

$$= \mathbf{p}(t) + \frac{\delta t}{2} (\dot{\mathbf{p}}(t) + \dot{\mathbf{p}}(t + \delta t)) + O(\delta t^3) \quad (43)$$

where in Eq. (41), the relation $\dot{\mathbf{q}} = \mathbf{p}/m$ was used and in Eq. (43) a first order Taylor expansion for $\dot{\mathbf{p}}$ was inserted. From these expansions a simple second order, one-step splitting scheme may be written as

$$\mathbf{p}_{n+1/2} = \mathbf{p}_n + \frac{\delta t}{2} \mathbf{F}(\mathbf{q}_n) \quad (44)$$

$$\mathbf{q}_{n+1} = \mathbf{q}_n + \frac{\delta t}{m} \mathbf{p}_{n+1/2} \quad (45)$$

$$\mathbf{p}_{n+1} = \mathbf{p}_{n+1/2} + \frac{\delta t}{2} \mathbf{F}(\mathbf{q}_{n+1}) \quad (46)$$

where the relation $\dot{\mathbf{p}} = -\partial \mathcal{H} / \partial \mathbf{q} = \mathbf{F}$ was used. This scheme is called the *Velocity Verlet* scheme. In a pictorial way it is sometimes described as half-kick, drift, half-kick, since the first step consists in applying forces for half a time step, second step consists in free flight of a particle with momentum $\mathbf{p}_{n+1/2}$ and the last step applies again a force for half a time step. In practice, forces only need to be evaluated once in each time step. After having calculated the new positions, $\mathbf{q}_{n+1}$, forces are calculated for the last integration step. They

are, however, stored to be used in the first integration step as *old* forces in the next time step of the simulation.

This algorithm comes also in another flavor, called the *Position Verlet* scheme. It can be expressed as

$$\mathbf{q}_{n+1/2} = \mathbf{q}_n + \frac{\delta t}{2m} \mathbf{p}_n \quad (47)$$

$$\mathbf{p}_{n+1} = \mathbf{p}_n + \delta t \mathbf{F}(\mathbf{q}_{n+1/2}) \quad (48)$$

$$\mathbf{q}_{n+1} = \mathbf{q}_{n+1/2} + \frac{\delta t}{2m} \mathbf{p}_{n+1} \quad (49)$$

In analogy to the description above this is sometimes described as half-drift, kick, half-drift. Using the relation $\mathbf{p} = \dot{\mathbf{q}}/m$ and expressing this as a first order expansion, it is obvious that $\mathbf{F}(\mathbf{q}_{n+1/2}) = \mathbf{F}((\mathbf{q}_n + \mathbf{q}_{n+1})/2)$ which corresponds to an implicit midpoint rule.

3.2 Stability

Performing simulations of stable many-body systems for long times should produce configurations which are in thermal equilibrium. This means that system properties, e.g. pressure, internal energy, temperature etc. are fluctuating around constant values. To measure these equilibrium properties it should not be relevant where to put the time origin from where configurations are considered to calculate average quantities. This requires that the integrator should propagate phase space variables in such a way that small fluctuations do not lead to a diverging behavior of a system property. This is a kind of minimal requirement in order to simulate any physical system without a domination of numerical artifacts. It is clear, however, that any integration scheme will have its own stability range depending on the step size δt . This is a kind of sampling criterion, i.e. if the step size is too large, in order to resolve details of the energy landscape, an integration scheme may end in instability.

For linear systems it is straight forward to analyze the stability range of a given numerical scheme. Consider e.g. the harmonic oscillator, for which the equations of motion may be written as $\dot{q}(t) = p(t)$ and $\dot{p}(t) = -\omega^2 q(t)$, where ω is the vibrational frequency and it is assumed that it oscillates around the origin. The exact solution of this problem may be written as

$$\begin{pmatrix} \omega q(t) \\ p(t) \end{pmatrix} = \begin{pmatrix} \cos \omega t & \sin \omega t \\ -\sin \omega t & \cos \omega t \end{pmatrix} \begin{pmatrix} \omega q(0) \\ p(0) \end{pmatrix} \quad (50)$$

For a numerical integrator the stepwise solution may be written as

$$\begin{pmatrix} \omega q_{n+1} \\ p_{n+1} \end{pmatrix} = \mathbf{M}(\delta t) \begin{pmatrix} \omega q_n \\ p_n \end{pmatrix} \quad (51)$$

where $\mathbf{M}(\delta t)$ is a propagator matrix. It is obvious that any stable numerical scheme requires eigenvalues $|\lambda(\mathbf{M})| \leq 1$. For $|\lambda| > 1$ the scheme will be unstable and divergent, for $|\lambda| < 1$ it will be stable but will exhibit friction, i.e. will loose energy. Therefore, in view of the conservativity of the scheme, it will be required that $|\lambda(\mathbf{M})| = 1$.

As an example the propagator matrices for the Implicit Euler (IE) and Position Verlet (PV) algorithms are calculated as

$$\mathbf{M}_{IE}(\delta t) = \frac{1}{1 + \omega^2 \delta t^2} \begin{pmatrix} 1 & \omega \delta t \\ -\omega \delta t & 1 \end{pmatrix} \quad (52)$$

$$\mathbf{M}_{PV}(\delta t) = \begin{pmatrix} 1 - \frac{1}{2}\omega^2\delta t^2 & \omega\delta t \left(1 - \frac{1}{4}\omega^2\delta t^2\right) \\ -\omega\delta t & 1 - \frac{1}{2}\omega^2\delta t^2 \end{pmatrix} \quad (53)$$

It is then straight forward to calculate the eigenvalues as roots of the characteristic polynomials. The eigenvalues are then calculated as

$$\lambda_{EE} = 1 \pm i\omega\delta t \quad (54)$$

$$\lambda_{IE} = \frac{1}{1 + \omega^2\delta t^2}(1 \pm i\omega\delta t) \quad (55)$$

$$\lambda_{PV} = \lambda_{VV} = \lambda_{VIE} = \lambda_{PIE} = 1 - \frac{1}{2}\omega^2\delta t^2 \left(1 \pm \sqrt{1 - \frac{4}{\omega^2\delta t^2}}\right) \quad (56)$$

This shows that the absolute values for the Explicit Euler (EE) and the Implicit Euler methods never equals one for $\delta t \neq 0$, i.e. both methods do not produce stable trajectories. This is different for the Position Verlet, the Velocity Verlet (VV), the Position Implicit Euler (PIE) and the Velocity Implicit Euler (VIE), which all have the same eigenvalues. It is found that the range of stability for all of them is in the range $\omega^2\delta t^2 < 2$. For larger values of δt the absolute values of the eigenvalues bifurcates, getting larger and smaller values than one. In Figure 4 the absolute values are shown for all methods and in in Figure 5 the imaginary versus real parts of λ are shown. For EE it is clear that the imaginary part diverges linearly with increase of δt . The eigenvalues of the stable methods are located on a circle until $\omega^2\delta t^2 = 2$. From there one branch diverges to $-\infty$, while the other decreases to zero.

As a numerical example the phase space trajectories of the harmonic oscillator for $\omega = 1$ are shown for the different methods in Figure 6. For the stable methods, results for a time step close to instability is shown. All different methods produce closed, stable orbits, but it is seen on the other hand that they strongly deviate from the exact solution, which is shown for reference. This demonstrates that stability is a necessary, but only a weak criterion for correct results. Numerically correct results are only obtained for much smaller time steps in the range of $\delta t \approx 0.01$. Also shown are the results for EE and IE. Here a very much smaller time step, $\delta t = 0.01$ is chosen. It is seen that the phase space trajectory of EE spirals out while the one of IE spirals in with time, showing the instable or evanescent character of the methods.

Another issue related to stability is the effect of a trajectory perturbation. If initial conditions are slightly perturbed, will a good integrator keep this trajectory close to the reference trajectory? The answer is No and it is even found that the result is not that strong dependent on the integrator. Even for integrators of high order, trajectories will not stay close to each other. The time evolution of the disturbance may be studied similar to the system trajectory. Consider the time evolution for $\Gamma + \delta\Gamma$, where $\Gamma = (\mathbf{p}, \mathbf{q})$ and $\delta\Gamma = (\delta\mathbf{p}, \delta\mathbf{q})$ is a small disturbance. Then

$$\frac{d\Gamma}{dt} = \nabla_{\Gamma} \mathcal{H}(\Gamma) \quad (57)$$

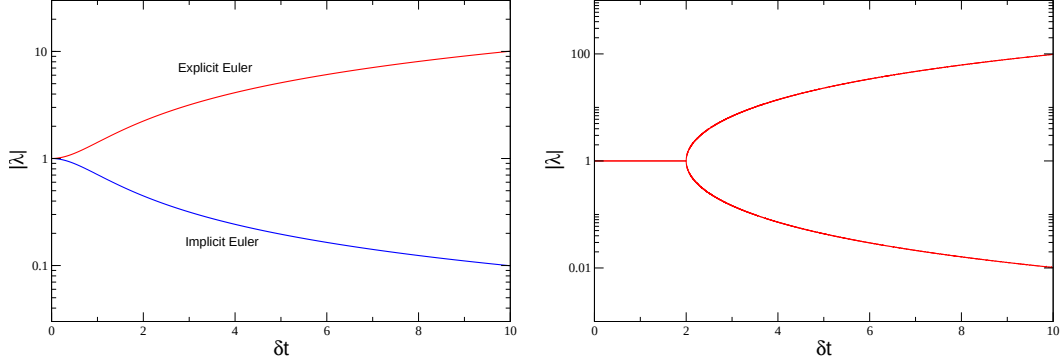


Figure 4. Absolute value of the eigenvalues λ as function of the time step δt . Left: Explicit and implicit Euler method. Right: Velocity and Position Verlet as well as Velocity Implicit and Position implicit Euler method. All methods have the eigenvalues.

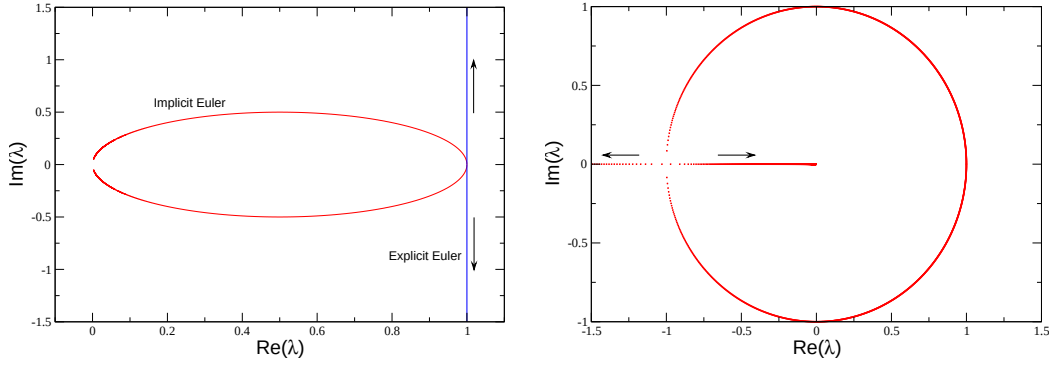


Figure 5. Imaginary versus real part of eigenvalues λ of the propagator matrices. Left: Implicit and Explicit Euler. Right: Velocity and Position Verlet as well as Velocity Implicit and Position implicit Euler method.

Similarly one can write for small $\delta\Gamma$

$$\frac{d}{dt}(\Gamma + \delta\Gamma) = \nabla_{\Gamma} \mathcal{H}(\Gamma + \delta\Gamma) \quad (58)$$

$$= \nabla_{\Gamma} \mathcal{H}(\Gamma) + \nabla_{\Gamma}(\nabla_{\Gamma} \mathcal{H}(\Gamma))\delta\Gamma \quad (59)$$

where the second line is a truncated Taylor series. Comparing terms one simply gets as equation of motion for a perturbation

$$\frac{d\delta\Gamma}{dt} = \nabla_{\Gamma}^2 \mathcal{H}(\Gamma)\delta\Gamma \quad (60)$$

It is found that the disturbance develops exponentially, with a characteristic, system dependent exponent, which is the Ljapunov exponent^{62,63}.

Now consider the following situation where identical starting configurations are taken for two simulations. They will be carried out by different yet exact algorithms, therefore leading formally to the same result. Nevertheless it may happen that different orders of

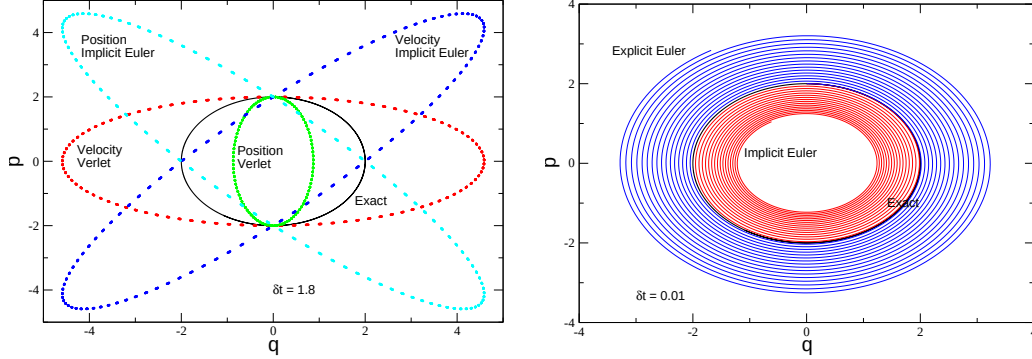


Figure 6. Phase space trajectories for the one-dimensional harmonic oscillator, integrated with the Velocity Implicit Euler, Position Implicit Euler, Velocity Verlet, Position Verlet and integration step size of $\delta t = 1.8$ (left) and the Implicit Euler and Explicit Euler and step size $\delta t = 0.01$ (right).

floating-point operations are used in both algorithms. Due to round off errors, floating-point arithmetic is not necessarily associative, i.e. in general

$$a \hat{\circ} (b \hat{\circ} c) \neq (a \hat{\circ} b) \hat{\circ} c \quad (61)$$

where $\hat{\circ}$ is a floating-point machine operation (+, -, /, *). Therefore, both simulations will be different by round off errors. According to the above discussion, this may be considered as the slightest disturbance of a system trajectory, $\delta\Gamma_{\min}$, and the question is, what effect such a round off error will have. A different method to study difference in system trajectories is the calculation of the difference

$$\gamma_x(t) = \frac{1}{3N} \sum_{i=1}^N \sum_{\alpha=x,y,z} (x(t) - \tilde{x}(t))^2 \quad (62)$$

where N is the number of particles, $x(t)$ a certain property, e.g. the coordinates or momenta, and $\tilde{x}$ the same property of a disturbed trajectory. In Figure 7 results are shown for a system of Lennard-Jones particles, where the disturbance was induced by reversing the order of summation in the force routine, thereby provoking round off errors in the first time step. Shown are results for the coordinates, the velocities and the forces and it is seen that all quantities diverge exponentially from machine accuracy up to a certain behavior at long times, which is shown in the inset. To understand the long time behavior, $\gamma_x(t)$ can be written as average property

$$\gamma_x(t) = \langle (x(t) - x(0) - \tilde{x}(t) + x(0))^2 \rangle \quad (63)$$

$$= \langle |x(t) - x(0)|^2 \rangle + \langle |\tilde{x}(t) - x(0)|^2 \rangle - 2\langle x(t)\tilde{x}(t) \rangle + 2\langle x(0)\tilde{x}(t) \rangle + 2\langle x(t)x(0) \rangle - 2\langle x(0)^2 \rangle \quad (64)$$

In the second equation the first two terms are mean square displacements of x in the two systems (note that $\tilde{x}(0) = x(0)$ since the same starting configurations are used), the next term is a cross correlation between the systems. This will vanish if the systems become independent of each other. The next two systems consist of auto-correlation functions of x

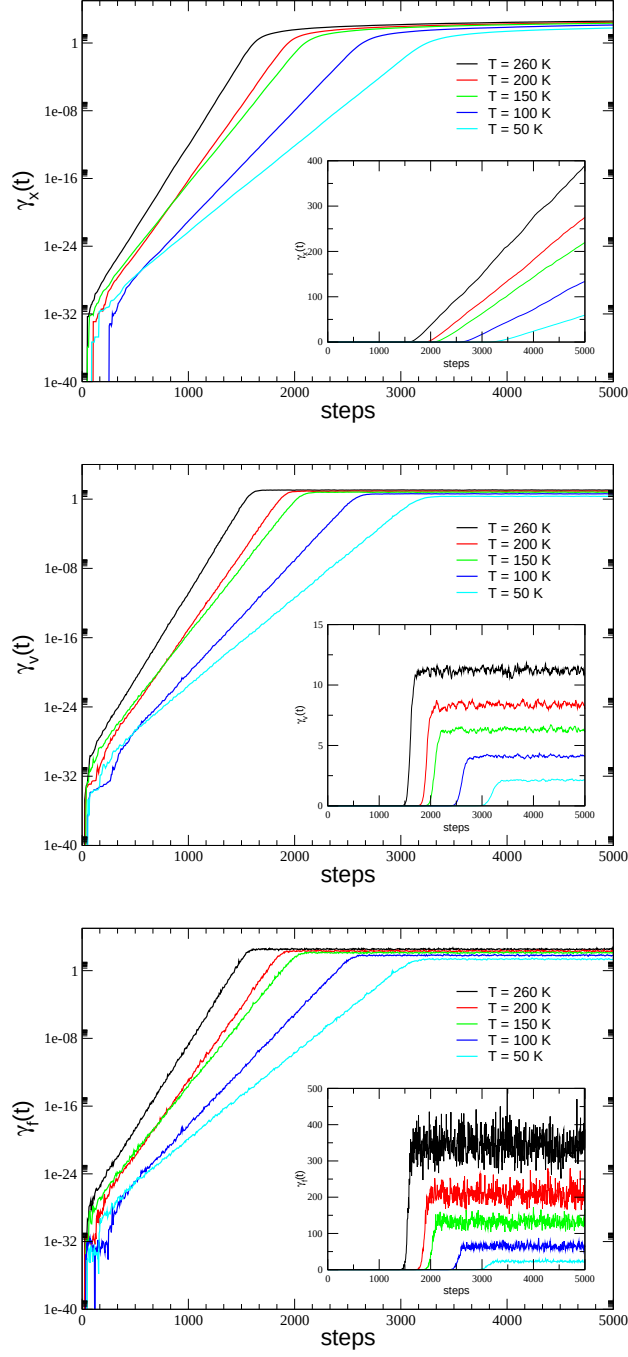


Figure 7. Divergent behavior of trajectories due to round off errors, induced by different summation order in the force routine. From top to bottom: coordinates, velocities, forces. The insets show on a linear scale the long time behavior of the trajectory differences, i.e. when the two systems get decorrelated.

in each system. For long times they will also decrease to zero. Finally, the last term gives a constant offset which does not depend on time. Therefore the long time behavior will be governed for coordinates, momenta and forces by

$$\lim_{t \rightarrow \infty} \gamma_q(t) = 2\langle |\mathbf{q}(t) - \mathbf{q}(0)|^2 \rangle = 12Dt \quad (65)$$

$$\lim_{t \rightarrow \infty} \gamma_p(t) = 2\langle \mathbf{p}(t)^2 \rangle = mk_B T \quad (66)$$

$$\lim_{t \rightarrow \infty} \gamma_f(t) = 2\langle \mathbf{F}(t)^2 \rangle = 2(\nabla \mathcal{W})^2 \quad (67)$$

where D is the diffusion coefficient, T the temperature and $\mathcal{W}$ the potential of mean force.

That the divergent behavior of neighbored trajectories is a system dependent property is shown in Figure 7 where results for Lennard-Jones systems at different temperatures are shown.

In conclusion, the individual trajectories of a physical complex system will end up at different places in phase space when introducing round off errors or small perturbations. Round off errors cannot be avoided with simple floating-point arithmetic (only discrete calculations are able to avoid round off errors; but then the physical problem is transformed into a different space). Since one cannot say anything about a *true* summation order, the location in phase space cannot have an absolute meaning. Therefore, the solution to come out of this dilemma is to interpret the phase space location as a *possible* and *allowed* realization of the system, which makes it necessary, however, to average over a lot of possible realizations.

3.3 Time Reversibility

Considering the classical equations of motion for the case of a time independent Hamiltonian, i.e. no externally applied time-dependent potentials, it becomes clear that the dynamics of conservative particle systems should be time reversible. Since the momentum $\mathbf{p}$ has dimension [length/time], a time transformation will lead to

$$t \rightarrow -t, \quad \mathbf{p} \rightarrow -\mathbf{p} \quad (68)$$

Therefore this transform will lead to

$$\frac{\partial \mathbf{p}}{\partial t} = -\frac{\partial \mathcal{H}}{\partial \mathbf{q}} \xrightarrow{t \rightarrow -t} \frac{\partial(-\mathbf{p})}{\partial(-t)} = -\frac{\partial \mathcal{H}}{\partial \mathbf{q}} \quad (69)$$

$$\frac{\partial \mathbf{q}}{\partial t} = \frac{\partial \mathcal{H}}{\partial \mathbf{p}} = \frac{\mathbf{p}}{m} \xrightarrow{t \rightarrow -t} \frac{\partial(\mathbf{q})}{\partial(-t)} = -\frac{\partial \mathcal{H}}{\partial(-\mathbf{p})} = -\frac{\mathbf{p}}{m} \quad (70)$$

showing the equations of motion are unchanged under a time reverse transformation.

Therefore it is not a qualitative change to calculate the mechanical system towards the past and not towards the future. This means that an integrator should be able to propagate a system trajectory up to a certain time step and from there, by a time reversal transformation $\delta t \rightarrow -\delta t$, should go exactly the same trajectory back to the starting configuration, i.e.

$$\phi_{\delta t}^* = \phi_{-\delta t}^{-1} \stackrel{!}{=} \phi_{\delta t} \quad (71)$$

The map $\phi_{\delta t}^*$ is called the adjoint method⁶⁴ and if Eq. (71) holds, $\phi_{\delta t}$ is symmetric.

The time reversibility can easily be checked by inspection for each integrator. As an example consider the Position Implicit Euler method. The reverse of this method, to move positions and momenta from step $n + 1$ back to step n or by a time reversal would be

$$\begin{array}{ll} \text{Reverse} & \text{Transform } \delta t \rightarrow -\delta t \\ \mathbf{q}_n = \mathbf{q}_{n+1} - \frac{\delta t}{m} \mathbf{p}_n & \mathbf{q}_n = \mathbf{q}_{n+1} - \frac{\delta t}{m} \mathbf{p}_{n+1} \end{array} \quad (72)$$

$$\mathbf{p}_n = \mathbf{p}_{n+1} - \delta t \mathbf{F}(\mathbf{q}_{n+1}) \quad \mathbf{p}_n = \mathbf{p}_{n+1} - \delta t \mathbf{F}(\mathbf{q}_n) \quad (73)$$

As it is seen the method with time reversal is not identical to the inverse map, since the derivatives $\partial_q \mathcal{H}$ and $\partial_p \mathcal{H}$ are evaluated at the wrong time step. Therefore the Position Implicit Euler method is not time reversible. The same is true for all other Euler methods, shown in Section 3.1. However, it is instructive to note that the reverse map corresponds to the Velocity Implicit Euler. Applying a proper time reversal transform $\mathbb{T}$ to the algorithms yields the following

$$\mathbb{T} \{ \phi_{\delta t}^{EE} \} = \phi_{\delta t}^{IE} \quad (74)$$

$$\mathbb{T} \{ \phi_{\delta t}^{IE} \} = \phi_{\delta t}^{EE} \quad (75)$$

$$\mathbb{T} \{ \phi_{\delta t}^{PIE} \} = \phi_{\delta t}^{VIE} \quad (76)$$

$$\mathbb{T} \{ \phi_{\delta t}^{VIE} \} = \phi_{\delta t}^{PIE} \quad (77)$$

Therefore one can simply combine different methods in order to construct a new scheme. In order to propagate the system by a time step δt , apply each scheme for $\delta t/2$. For example consider the combination $(\phi_{\delta t/2}^{EE} \circ \phi_{\delta t/2}^{IE})$, which leads to the following steps

$$\mathbf{q}_{n+1/2} = \mathbf{q}_n + \frac{\delta t}{2m} \mathbf{p}_{n+1/2} \quad (78)$$

$$\mathbf{p}_{n+1/2} = \mathbf{p}_n + \frac{\delta t}{2} \mathbf{F}(\mathbf{q}_{n+1/2}) \quad (79)$$

$$\mathbf{q}_{n+1} = \mathbf{q}_{n+1/2} + \frac{\delta t}{2m} \mathbf{p}_{n+1/2} \quad (80)$$

$$\mathbf{p}_{n+1} = \mathbf{p}_{n+1/2} + \frac{\delta t}{2} \mathbf{F}(\mathbf{q}_{n+1/2}) \quad (81)$$

A simple comparison of terms yields that $\mathbf{q}_{n+1/2} = (\mathbf{q}_n + \mathbf{q}_{n+1})/2$ and $\mathbf{p}_{n+1/2} = (\mathbf{p}_n + \mathbf{p}_{n+1})/2$ which leads to

$$\mathbf{q}_{n+1} = \mathbf{q}_n + \frac{\delta t}{2m} (\mathbf{p}_n + \mathbf{p}_{n+1}) \quad (82)$$

$$\mathbf{p}_{n+1} = \mathbf{p}_n + \frac{\delta t}{2} \mathbf{F}((\mathbf{q}_n + \mathbf{q}_{n+1})/2) \quad (83)$$

which is the Implicit Midpoint Euler scheme (IME).

Combining different types of methods, it comes out that the following relations hold

$$\phi_{\delta t/2}^{EE} \circ \phi_{\delta t/2}^{IE} = \phi_{\delta t}^{IME} \quad (84)$$

$$\phi_{\delta t/2}^{IE} \circ \phi_{\delta t/2}^{EE} = \phi_{\delta t}^{ITE} \quad (85)$$

$$\phi_{\delta t/2}^{PIE} \circ \phi_{\delta t/2}^{VIE} = \phi_{\delta t}^{VV} \quad (86)$$

$$\phi_{\delta t/2}^{VIE} \circ \phi_{\delta t/2}^{PIE} = \phi_{\delta t}^{PV} \quad (87)$$

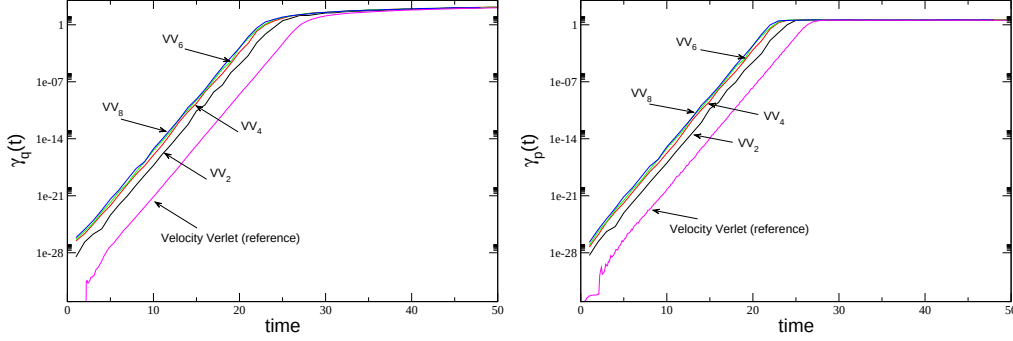


Figure 8. Results for calculations with time reversal transformation, $\delta t \rightarrow -\delta t$ after time t . Results show $\overline{\gamma}_q(t)$ and $\overline{\gamma}_p(t)$, Eq. (88), for the Velocity Verlet integrator (VV2) and higher order composition schemes (VV4, VV6, VV8, cmp. Section 3.4 and the Appendix). As a reference, results are shown for $\gamma_q(t)$ and $\gamma_p(t)$, obtained for two trajectories with different round off errors at $t = 0$.

where $\phi_{\delta t}^{ITE}$ is the Implicit Trapezoidal Euler. This result shows some interesting properties: (i) symmetric methods can be constructed by composition of non-symmetric methods (example for the adjoint method), (ii) since PIE and VIE are both first order methods, while VV is of second order, this shows that composition of methods increases the order of a method.

Now that time reversibility is established in principle for four different methods (VV, PV, IME, ITE), the question is how good time reversibility is achieved in a molecular dynamics computer simulation. In practice one can test this by simulating $M/2$ time steps and then reversing the sign of δt . For a symmetric method, the trajectories should end up exactly in the initial configuration after M time steps. E.g. if particles started from a regular lattice with zero momentum, this should be exactly the final configuration. Here now it is important to consider the effect of round off errors in the simulation which accumulate during the run. It was already discussed that a very small perturbation of the initial configuration may lead to a diverging behavior of two (initially very close) trajectories. If in the time reverse motion of particles the round off errors are not fully compensated, the trajectory will finally end up a distance apart, which corresponds to an initial perturbation of the size of the cumulated round off error. In Figure 8 the results for such a numerical experiment is shown. Calculated is the norm of the trajectory differences

$$\overline{\gamma}_x(t) = \|\mathbf{x}(0) - \mathbf{x}_{\leftarrow}(0|t)\|^2 \quad (88)$$

where $\mathbf{x} = (\mathbf{q}, \mathbf{p})$ and t denotes the time step, from where the trajectory was reversed. Figure 8 shows results for this calculation for a system of 864 Lennard-Jones particles. In addition to the trajectory difference, calculated from time reversal, a *reference* curve is shown, which is the result from a different summation order. This shows that true time reversibility cannot be achieved easily for a trajectory in molecular dynamics, although a true time reversible integrator is used. This problem can be traced back to the round off errors. The largest number of floating-point operations usually is done in the force routine of an MD program. This usually takes about 95% of CPU time of the whole simulation time. Since forces are used to update the velocities, which in turn are needed to propagate

the coordinates, it seems to be clear that the main source of round off errors is to be found in the force loop. Therefore one attempt would be to reduce round off errors in this part of the program. Several techniques were proposed for this task, e.g. up- or down-sorting for force contributions^{65,66}, tree like summation or compensated summation⁶⁷⁻⁷⁴. These methods have in common that they are able to reduce the round off error, but they do not avoid it. Therefore there will be still a slight deviation from the *true* summation, which may act as a slight perturbation for the trajectory.

Another approach, which avoids round off errors is the use of integer arithmetic⁷⁵⁻⁷⁸. In this approach all relevant quantities, e.g. length, time, charge, energy, are scaled in a way that they can be represented by large integer values. The size of the values thus determines the resolution with which differences in lengths etc. can be resolved.

A different approach, which still relies on floating-point arithmetic was introduced by Skeel⁷⁹. If a variable is denoted by x and its upper limit is known, i.e. $\max |x| \leq \bar{x}$, where $\bar{x}$ is a power of 2 and if μ is denoted the target precision, one can construct an equally spaced floating-point grid, with grid spacing $\delta x = 2^{-\mu} \bar{x}$. If ν denotes the number of significant bits in the floating-point representation of a machine (often $\nu = 24$ for single precision, $\nu = 53$ for double precision and $\nu = 113$ for quadruple precision), then use the following procedures to transform a machine number x_i with upper bound $\bar{x}$:

If $\mu \leq \nu - 2$ define

$$\text{round}_x(x_i) = (x_i \overset{\wedge}{+} (0.75 \times 2^{\nu-\mu} \bar{x})) \overset{\wedge}{-} (0.75 \times 2^{\nu-\mu} \bar{x}) \quad (89)$$

If $\mu = \nu - 1$ define

$$\text{round}_x(x_i) = (x_i \overset{\wedge}{+} \text{sign}(x_i) \bar{x}) \overset{\wedge}{-} \text{sign}(x_i) \bar{x} \quad (90)$$

If $\mu = \nu$ define

$$\text{round}_x(x_i) = (x_i \overset{\wedge}{-} \text{sign}(x_i) \bar{x}) \overset{\wedge}{+} \text{sign}(x_i) \bar{x} \quad (91)$$

where $\overset{\wedge}{\circ}$ denotes a machine operation.

E.g. the Velocity Verlet algorithm then reads as

$$\mathbf{p}_{n+1/2} = \mathbf{p}_n + \text{round}_p \left((0.5 \overset{\wedge}{\times} \delta t \overset{\wedge}{\div} m) \overset{\wedge}{\times} \mathbf{F}(\mathbf{q}_n) \right) \quad (92)$$

$$\mathbf{q}_{n+1} = \mathbf{q}_n + \text{round}_q \left((\delta t \overset{\wedge}{\div} m) \overset{\wedge}{\times} \mathbf{p}_n \right) \quad (93)$$

$$\mathbf{p}_{n+1} = \mathbf{p}_{n+1/2} + \text{round}_p \left((0.5 \overset{\wedge}{\times} \delta t \overset{\wedge}{\div} m) \overset{\wedge}{\times} \mathbf{F}(\mathbf{q}_{n+1}) \right) \quad (94)$$

Varying the value of μ it is possible to change the accuracy of floating-point operations gradually while preserving exact machine operations, i.e. avoiding round off errors. This is shown in Figure 9 where the quadratic displacement between trajectories is shown for interval floating-point arithmetic with different values of μ and a reference trajectory with ordinary floating-point arithmetic. In practical applications it will be expensive to apply the round-function in every force evaluation. Therefore, the forces are calculated in the usual way, accepting round off errors. μ has to be chosen then in such a way that the last digits, affected by round off errors in the forces are truncated in the integration step.

The interest of such round off free algorithms is not only an academic one to prove a theoretical property of an integration scheme. Since round off errors accumulate, the calculated trajectory is dominated by round off errors from a certain time on. This does not mean that the trajectory does not carry any information anymore, but it may lead to the fact that correlations between particles over this time interval cannot be seen in a simulation. Avoiding round off errors this should not happen and small effects, which are usually covered by round off errors show up.

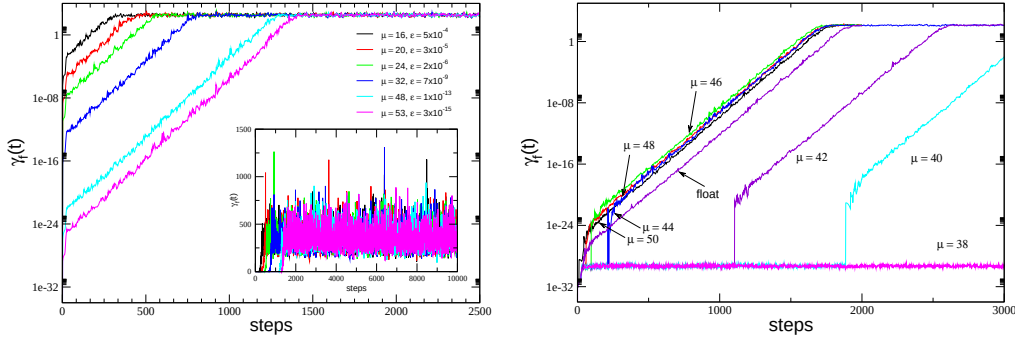


Figure 9. Left: Divergent behavior of system trajectories, calculated with the usual floating-point arithmetic and round off free interval floating-point arithmetic. In both cases a Velocity Verlet integration scheme is used. Right: Result for $\gamma_f(t)$ of floating-point interval arithmetic where trajectories were calculated with different summation sequences.

A numerical test shows that there is no accumulation of numerical errors in such a simulation. Reversing the time step δt after $M/2$ time steps and comparing the two trajectory pieces in the intervals $t \in [1, M/2] \delta t$ and $t \in [M/2 + 1, M] \delta t$ leads to differences in trajectories identical to zero (no fluctuations in position-, velocity- or force-differences)! This shows that complete reversibility is possible.

What about the effect of different summation order? If forces are not subject to floating-point interval arithmetic, there will be round off errors and the result of the forces will depend on the order of summing up individual force contributions. This does not do any harm to the integrator, unless the value of $\text{round}_p \left((0.5 \hat{\times} \delta t \div m) \hat{\times} \mathbf{F}(\mathbf{q}_n) \right)$ lies in different intervals for different summation orders. It is clear that with a wider floating-point interval the probability of such an event becomes smaller and consequently, for smaller μ it is to be expected that round off errors due to summation order show up later than for large μ .

In Figure 9 the quadratic displacement, $\gamma_f(t)$, of forces is shown due to a different order of summation in the force routine. It is found that the divergent behavior of trajectories starts at different times depending on the value of μ , as can be expected from the above discussion. For $\mu = 38$ there are only small fluctuations on the level of machine accuracy. Nevertheless, also this function will start to diverge at larger times. Compared to floating-point interval arithmetic is the result for usual double precision floating-point arithmetic. It is found that for large μ the diverging behavior is even more pronounced (although the Ljapunov exponent is for all cases the same of course). This effect is due to a stronger effect of initial disturbance in the case of interval arithmetic. This effect, induced by different

order of summation, may be avoided by simply using interval arithmetic also in the force routine, i.e. rounding properly every individual force contribution before summing up. This introduces, however, a stronger computational overhead since the round-function has to be applied very much more often (in the case of short range interactions 10-100 times for each particle and time step).

3.4 Accuracy

For an integrator of order $p \geq 1$, the local error may be written as an upper bound⁸⁰

$$\|\Phi_{\delta t, \mathcal{H}} - \phi_{\delta t}\| \leq M \delta t^{p+1} \quad (95)$$

where $M > 0$ is a constant, $\Phi_{\delta t, \mathcal{H}}$ is the exact and $\phi_{\delta t}$ the numerical flow of the system. The global error, i.e. the accumulated error for larger times, is thereby bound for stable methods by⁸⁰

$$\|\Gamma(t_n) - \Gamma_n\| \leq K (e^{Lt_n} - 1) \delta t^p, \quad t_n = n\delta t \quad (96)$$

where $K > 0$ is a constant, $L > 0$ the Lipschitz constant, $\Gamma(t_n) = (\mathbf{p}(t_n), \mathbf{q}(t_n))$ the exact and $\Gamma_n = (\mathbf{p}_n, \mathbf{q}_n)$ the numerically computed trajectory at time t_n . This estimate gives of course not too much information for $Lt_n \gg 1$ unless δt is chosen very small. Nevertheless, qualitatively this estimate shows a similar exponential divergent behavior of numerical and exact solution for a numerical scheme, as was observed in Section 3.2.

A different approach to the error behavior of a numerical scheme is backward error analysis, first mentioned in Ref. 81 in the context of differential equations. The idea is to consider the numerical solution of a given scheme as the exact solution of a modified equation. The comparison of the original and the modified equation then gives qualitative insight into the long time behavior of a given scheme.

It is assumed that the numerical scheme can be expressed as a series of the form

$$\phi_{\delta t}(\Gamma_n) = \Gamma_n + \delta t f(\Gamma) + \delta t^2 g_2(\Gamma) + \delta t^3 g_3(\Gamma) \pm \dots \quad (97)$$

where the g_i are known coefficients and for consistency of the differential equation it must hold

$$f(\Gamma) = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \nabla_p \\ \nabla_q \end{pmatrix} \mathcal{H}(\mathbf{p}, \mathbf{q}) \quad (98)$$

On the other hand it is assumed that there exists a modified differential equation of the form

$$\frac{d}{dt} \tilde{\Gamma} = f(\tilde{\Gamma}) + \delta t f_2(\tilde{\Gamma}) + \delta t^2 f_3(\tilde{\Gamma}) + \dots \quad (99)$$

where $\tilde{\Gamma}$ will be equivalent to the numerically obtained solution. In order to construct the

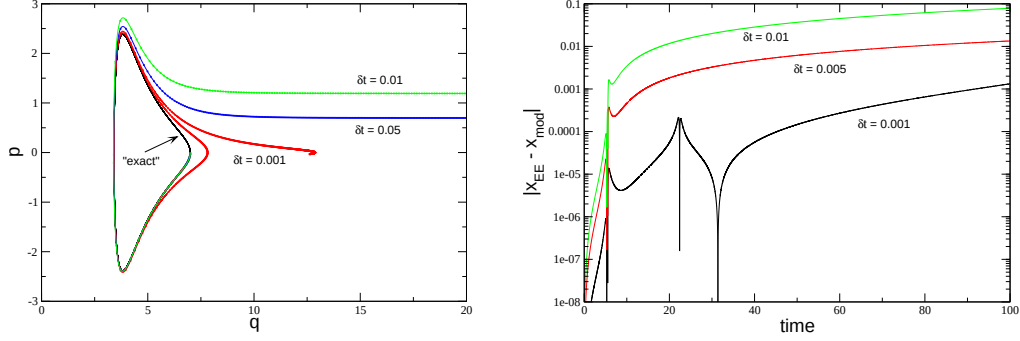


Figure 10. Phase space trajectories of the Lennard-Jones oscillator calculated with the Explicit Euler scheme and different time steps of integration. The *exact* solution (numerical solution of a high order composition scheme with small time step) is shown as a reference - it forms closed orbits. Superimposed to the solutions are results, obtained with a Velocity Verlet scheme, applied to the modified equations, Eqs. (101,102). The right figure shows the differences in coordinates between the calculation with Explicit Euler scheme applied to Lennard-Jones oscillator and Velocity Verlet applied to the modified equation, $|\mathbf{q}_{EE}(t) - \mathbf{q}_{mod}(t)|$.

modified equation, the solution of Eq. (99) is Taylor expanded, i.e.

$$\begin{aligned}
 \tilde{\Gamma}(t + \delta t) = & \tilde{\Gamma}(t) + \delta t \left(f(\tilde{\Gamma}) + \delta t f_2(\tilde{\Gamma}) + \delta t^2 f_3(\tilde{\Gamma}) + \dots \right) \\
 & + \frac{\delta t^2}{2!} \left(f'(\tilde{\Gamma}) + \delta t f_2'(\tilde{\Gamma}) + \dots \right) \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \left(f(\tilde{\Gamma}) + \delta t f_2(\tilde{\Gamma}) + \dots \right) \\
 & + \frac{\delta t^3}{3!} \left\{ \left(f''(\tilde{\Gamma}) + \delta t f_2''(\tilde{\Gamma}) + \dots \right) \left(\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \left(f(\tilde{\Gamma}) + \delta t f_2(\tilde{\Gamma}) + \dots \right) \right)^2 \right. \\
 & \quad + \left(f'(\tilde{\Gamma}) + \delta t f_2'(\tilde{\Gamma}) + \dots \right) \left(\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \left(f'(\tilde{\Gamma}) + \delta t f_2'(\tilde{\Gamma}) + \dots \right) \right) \\
 & \quad \left. \times \left(f(\tilde{\Gamma}) + \delta t f_2(\tilde{\Gamma}) + \dots \right) \right\} \\
 & + \dots
 \end{aligned} \tag{100}$$

The procedure to construct the unknown functions f_i proceeds in analogy to perturbation theory, i.e. coefficients with same powers of δt are collected which leads to a recursive scheme to solve for all unknowns.

To give an example the Lennard-Jones oscillator is considered, i.e. a particle performing stable motions in negative part of a Lennard-Jones potential. As was observed already for the harmonic oscillator, the Explicit Euler method will gain energy during the time, i.e. the particle will increase kinetic energy which finally will lead to an escape of the Lennard-Jones potential well. Solving for the modified equation of the Explicit Euler, one gets as a first correction

$$\dot{\mathbf{q}} = \frac{\partial \mathcal{H}}{\partial \mathbf{p}} + \frac{\delta t}{2} \frac{\partial \mathcal{H}}{\partial \mathbf{q}} \tag{101}$$

$$\dot{\mathbf{p}} = -\frac{\partial \mathcal{H}}{\partial \mathbf{q}} + \frac{\delta t}{2} \mathbf{p} \frac{\partial^2 \mathcal{H}}{\partial \mathbf{p}^2} \tag{102}$$

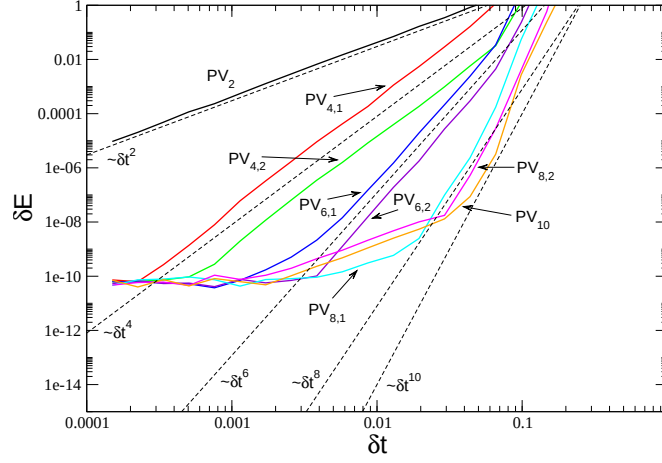


Figure 11. Comparison of symmetric composition high order integrators, based on the Position Verlet scheme. Coefficients of composition methods are given in the Appendix.

Figure 10 shows results for the integration of equations of motion with the Explicit Euler scheme. Different time steps for integration were applied which show a faster escape from a stable orbit with increasing time step. Also plotted in the same figure is the solution of the modified equations with a high order symplectic scheme, which can be considered as *exact* on these time scales. It is found that the trajectories more or less coincide and cannot be distinguished by eye. A more quantitative analysis (Figure 10) shows that for relatively long times the solution is rather well approximated by the modified equation, although with increasing time the differences between solutions become more pronounced. This means that for longer times it would be necessary to include more terms of higher order in δt into the modified equation. It should be mentioned that, in general, the series expansion of the modified equation diverges.

This gives a method to analyze the error. A better way, of course, is to avoid errors. A natural approach would be to increase the order p of the integrator at hand. As was shown already in Section 3.3 in the context of time reversibility, one can compose different methods in order to get (i) symmetric methods and (ii) increase the order of the method. A method for constructing higher order symplectic methods was proposed by Yoshida^{82,83} and Suzuki^{84,85}, where an s -fold composition of symmetric methods was applied. The resulting flow can thus be written as

$$\psi_{\delta t} = \phi_{w_s \delta t} \circ \phi_{w_{s-1} \delta t} \circ \dots \circ \phi_{w_2 \delta t} \circ \phi_{w_1 \delta t} \quad (103)$$

where $\phi_{\delta t}$ is a symmetric basic method, e.g. Velocity Verlet ($\phi_{\delta t} = \phi_{\delta t/2, \mathcal{T}} \circ \phi_{\delta t, \mathcal{U}} \circ \phi_{\delta t/2, \mathcal{T}}$, Eqs. (44-46,86)) or Implicit Midpoint Rule ($\phi_{\delta t} = \phi_{\delta t/2, \mathcal{H}} \circ \phi_{\delta t, \mathcal{H}}^*$, Eq. (84)) and w_i are chosen parameters which determine the width of a substep. For consistency and if the composition method itself is symmetric, it is required that

$$\sum_{i=1}^s w_i = 1 \quad , \quad w_{s+1-k} = w_k \quad , \quad k = 1, \dots, s \quad (104)$$

In order to calculate the coefficients one has the additional condition

$$\sum_{i=1}^s w_i^{p+1} = 0 \quad (105)$$

If the basic method is of order p it can be shown that the composed method is at least of order $p + 1$ ⁶⁴. A symmetric method has an even order since all odd terms in the Taylor expansion cancel. If the composition method is constructed as a symmetric method according to Eq. (104), then $\tilde{\psi}_{\delta t}$ is even (at least) of order $p + 2$. This procedure can be repeated recursively in order to construct symmetric methods of any order. Of course, the number of stages increases each time the order is increased, e.g. starting with a symmetric method of order $p = 2$ as basic method and taking $s = 3$, this will result in a method of order $p = 4$. This order-four method can again be taken as a basic method and one ends up with a resulting order of $p = 6$ and $s = 9$. For even higher methods the number of stages will strongly increase, making this type of high order methods computationally demanding (cmp. Appendix). It is a matter of choice with which number of steps the procedure is started. For $s = 3$, one gets the scheme of Yoshida⁸², corresponding to

$$w_1 = w_3 = \frac{1}{2 - 2^{1/(p+1)}} \quad , \quad w_2 = -\frac{2^{1/(p+1)}}{2 - 2^{1/(p+1)}} \quad (106)$$

For $s = 5$ one gets the scheme of Suzuki⁸⁴, corresponding to

$$w_1 = w_2 = w_4 = w_5 = \frac{1}{4 - 4^{1/(p+1)}} \quad , \quad w_3 = -\frac{4^{1/(p+1)}}{4 - 4^{1/(p+1)}} \quad (107)$$

There are also refined composition schemes, which do not need that large number of substeps. They are also based on Eq. (103) with symmetric basic methods like the Velocity Verlet. References and coefficients are given in the Appendix. Figure 11 shows the error energy fluctuation

$$\delta E = \sqrt{\frac{1}{N_t} \sum_{k=1}^{N_t} \left(1 - \frac{E_k}{E_0}\right)^2} \quad (108)$$

for a test case of a system of Lennard-Jones particles, where E_k is the total energy in the system at time step k and N_t the number of MD steps. As a basic method the Position Verlet scheme was applied. It is found that all integration schemes obey the prescribed order. Due to numerical round off errors the methods saturate for very small time steps.

4 Conclusions

The old vision of Laplace as it was sketched in Section 1 even cannot be met in computer simulations, where all initial conditions are known *exactly*. Slight disturbances, which may even be introduced by simple round off errors, as was shown by different orders of summation, lead to Lyapunov instabilities of close trajectories. Only in the case, where really discrete lattice maps are introduced, reversibility can be achieved and no effects due to round off errors will be seen. Nevertheless, also in such a discrete system, there will be Lyapunov instabilities if disturbing the initial conditions in the slightest way. These

instabilities, however, have no dramatic consequences for the overall stability of trajectories of the system. The system develops into different parts of accessible phase space but if symplectic, time reversible integrators are used to solve the equations of motions, the geometrical structure of phase space is conserved. Physically this behavior can be interpreted as interference of a system with an uncontrollable, slight disturbance, which could happen within an experimental setup e.g. by thermal fluctuations or impurities etc.. This also should not drive a system out of equilibrium. Actually the fact that a disturbance develops an exponential drift of two nearby trajectories is often used to impose artificial initial conditions, i.e. start a simulation from an ordered regular grid and impose a uniform velocity distribution corresponding to a desired temperature. Equilibration will then lead a loss of memory to the initial conditions and to a proper statistical distribution of physical quantities. From that point of view an individual trajectory of particles is of no principal importance. It can only be considered as a possible, allowed realization of the system and lots of different configurations have to be obtained to get a proper averaging over different trajectories.

Appendix

Composition Methods of Order p

As an example a composition is done with a Velocity Verlet integrator as basic method, i.e.

$$\phi_{\delta t, \mathcal{H}} = \phi_{\delta t/2, \mathcal{U}} \circ \phi_{\delta t, \mathcal{T}} \circ \phi_{\delta t/2, \mathcal{U}}$$

Then the composition scheme is performed in a step wise fashion. In the program an s -stage composition scheme is then implemented as follows (here for the Velocity Verlet integrator)

```

do is = 1, s
  do i = 1, n_part
    do k = 1, 3
      v(k,i) = v(k,i) + 0.5*w(is)*h*f(k,i)/m(i)
      x(k,i) = x(k,i) + w(is)*h*v(k,i)
    enddo
  enddo

  ... calculate forces ...

  do i = 1, n_part
    do k = 1, 3
      v(k,i) = v(k,i) + 0.5*w(is)*h*f(k,i)/m(i)
    enddo
  enddo
enddo

```

In the following, coefficients for composition methods, based on symmetric schemes, with different number of stages, s and of order p are given. For all cases, the symmetry relation

$$w_{s+1-k} = w_k \quad , \quad k = 1, \dots, s$$

holds. Therefore only coefficients up to $[s/2] + 1$ are shown.

stages	order	coefficients	reference
3	4	$w_1 = 1.351207191959657771818$ $w_2 = -1.702414403875838200264$	Yoshida ⁸²
5	4	$w_1 = w_2 = 0.414490771794375711944$ $w_3 = -0.657963093139439791912$	Suzuki ⁸⁴
7	6	$w_1 = 0.78451361047755726382$ $w_2 = 0.23557321335935813368$ $w_3 = -1.17767998417887100695$ $w_4 = 1 - 2 \sum_{s=1}^3 w_s$	Yoshida ⁸²
9	6	$w_1 = 0.39216144400731413928$ $w_2 = 0.33259913678935943860$ $w_3 = -0.70624617255763935981$ $w_4 = 0.08221359629355080023$ $w_5 = 1 - 2 \sum_{s=1}^4 w_s$	Kahan & Li ⁸⁶
15	8	$w_1 = 0.74167036435061295345$ $w_2 = -0.40910082580003159400$ $w_3 = 0.19075471029623837995$ $w_4 = -0.57386247111608226666$ $w_5 = 0.29906418130365592384$ $w_6 = 0.33462491824529818378$ $w_7 = 0.31529309239676659663$ $w_8 = 1 - 2 \sum_{s=1}^7 w_s$	McLachlan ⁸⁷
17	8	$w_1 = 0.13020248308889008087881763$ $w_2 = 0.56116298177510838456196441$ $w_3 = -0.38947496264484728640807860$ $w_4 = 0.15884190655515560089621075$ $w_5 = -0.39590389413323757733623154$ $w_6 = 0.18453964097831570709183254$ $w_7 = 0.25837438768632204729397911$ $w_8 = 0.29501172360931029887096624$ $w_8 = 1 - 2 \sum_{s=1}^8 w_s$	Kahan & Li ⁸⁶
33	10	$w_1 = 0.09040619368607278492161150$ $w_2 = 0.53591815953030120213784983$ $w_3 = 0.35123257547493978187517736$	Kahan & Li ⁸⁶

$$\begin{aligned}
w_4 &= -0.31116802097815835426086544 \\
w_5 &= -0.52556314194263510431065549 \\
w_6 &= 0.14447909410225247647345695 \\
w_7 &= 0.02983588609748235818064083 \\
w_8 &= 0.17786179923739805133592238 \\
w_9 &= 0.09826906939341637652532377 \\
w_{10} &= 0.46179986210411860873242126 \\
w_{11} &= -0.33377845599881851314531820 \\
w_{12} &= 0.07095684836524793621031152 \\
w_{13} &= 0.23666960070126868771909819 \\
w_{14} &= -0.49725977950660985445028388 \\
w_{15} &= -0.30399616617237257346546356 \\
w_{16} &= 0.05246957188100069574521612 \\
w_{17} &= 1 - 2 \sum_{s=1}^{16} w_s
\end{aligned}$$

References

1. P. S. de Laplace. Théorie analytique des probabilités. In *Oeuvres Complètes de Laplace*, volume VII., Paris, 1820. Gauthier-Villars.
2. T. Schlick. *Molecular Modeling and Simulation*. Springer, New York, 2002.
3. C. L. Brooks III, M. Karplus, and B. M. Pettitt. Proteins: A theoretical perspective of dynamics, structure and thermodynamics. In *Advances in Chemical Physics*, volume LXXI, New York, 1988. John Wiley.
4. J. Baschnagel, K. Binder, P. Doruker, A. A. Gusev, O. Hahn, K. Kremer, W. L. Matice, F. Müller-Plathe, M. Murat, W. Paul, S. Santos, U. W. Suter, and V. Tries. Bridging the gap between atomistic and coarse-grained models of polymers: Status and perspectives. *Adv. Polym. Sci.*, 152:41, 2000.
5. Y. Duan, L. Wang, and P. A. Kollman. The early stage of folding of villin headpiece subdomain observed in 200-nanosecond fully solvated molecular dynamics simulation. *Proc. Natl. Acad. Sci. USA*, 95:9897, 1998.
6. Y. Duan and P. A. Kollman. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science*, 282:740, 1998.
7. In P. Nielaba, M. Mareschal, and C. Ciccotti, editors, *Bridging the Time Scales: Molecular Simulations for the next Decade*, Heidelberg, 2002. Springer.
8. B. J. Alder and T. E. Wainwright. Phase transition for a hard sphere system. *J. Chem. Phys.*, 27:1208–1209, 1957.
9. B. J. Alder and T. E. Wainwright. Studies in molecular dynamics. I. General method. *J. Chem. Phys.*, 31:459, 1959.
10. K. Kadau, T. C. Germann, P. S. Lomdahl, and B. L. Holian. Microscopic view of structural phase transitions induced by shock waves. *Science*, 296:1681–1684, 2002.
11. K. Kadau, T. C. Germann, and P. S. Lomdahl. Large-scale molecular-dynamics simulation of 19 billion particles. *Int. J. Mod. Phys. C*, 15:193, 2004.
12. V. Springel, S. D. M. White, A. Jenkins, C. S. Frenk, N. Yoshida, L. Gao, J. Navarro, R. Thacker, D. Croton, J. Helly, J. A. Peacock, S. Cole, P. Thomas, H. Couchman,

- A. Evrard, J. Colberg, and F. Pearce. Simulations of the formation, evolution and clustering of galaxies and quasars. *Nature*, 435:629–636, 2005.
13. <http://www.fz-juelich.de/zam/ibm-bgl>.
 14. <http://www.research.ibm.com/bluegene/>.
 15. G. Lu and E. Kaxiras. Overview of multiscale simulations of materials. In M. Rieth and W. Schommers, editors, *Handbook of Theoretical and Computational Nanotechnology*, volume X, pages 1–33. American Scientific Publishers, 2005.
 16. <http://www.ccp5.ac.uk/>.
 17. <http://amber.scripps.edu/>.
 18. <http://www.charmm.org/>.
 19. <http://www.ks.uiuc.edu/Research/namd/>.
 20. <http://www.emsl.pnl.gov/docs/nwchem/nwchem.html>.
 21. <http://www.cs.sandia.gov/~sjplimp/lammps.html>.
 22. S. Walbran and A. A. Kornyshev. Proton transport in polarizable water. *J. Chem. Phys.*, 114:10039–10048, 2001.
 23. G. C. Maitland, M. Rigby, E. B. Smith, and W. A. Wakeham. *Intermolecular Forces*. Clarendon, Oxford, 1981.
 24. J. Kong. Combining rules for intermolecular potential parameters. II. Rules for the Lennard-Jones (12-6) potential and the Morse potential. *J. Chem. Phys.*, 59:2464–2467, 1973.
 25. M. Waldman and A. T. Hagler. New combining rules for rare gas van der Waals parameters. *J. Comp. Chem.*, 14:1077, 1993.
 26. J. Delhommelle and P. Milli . Inadequacy of the Lorentz-Bertelot combining rules for accurate predictions of equilibrium properties by molecular simulation. *Molec. Phys.*, 99:619–625, 2001.
 27. L. Verlet. Computer experiments on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules. *Phys. Rev.*, 159:98, 1967.
 28. G. Sutmann and V. Stegailov. Optimization of neighbor list techniques for molecular dynamics simulations. *J. Mol. Liq.* (in press), 2005.
 29. G. Sutmann and V. Stegailov (to be published).
 30. R. W. Hockney. The potential calculation and some applications. *Meth. Comput. Phys.*, 9:136–211, 1970.
 31. R. W. Hockney, S. P. Goel, and J. W. Eastwood. Quite high-resolution computer models of a plasma. *J. Comp. Phys.*, 14:148, 1974.
 32. P. Ewald. Die berechnung optischer und elektrostatischer Gitterpotentiale. *Ann. Phys.*, 64:253, 1921.
 33. S. W. de Leeuw, J. M. Perram, and E. R. Smith. Simulation of electrostatic systems in periodic boundary conditions. I. Lattice sums and dielectric constants. *Proc. R. Soc. London*, A373:27, 1980.
 34. S. W. de Leeuw, J. M. Perram, and E. R. Smith. Simulation of electrostatic systems in periodic boundary conditions. II. Equivalence of boundary conditions. *Proc. R. Soc. London*, A373:57, 1980.
 35. T. Darden, D. York, and L. Pedersen. A NlogN method for Ewald sums in large systems. *J. Chem. Phys.*, 98:10089, 1993.
 36. U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen. A smooth particle mesh Ewald method. *J. Chem. Phys.*, 103:8577, 1995.

37. R. W. Hockney and J. W. Eastwood. *Computer simulation using particles*. McGraw-Hill, New York, 1981.
38. M. Deserno and C. Holm. How to mesh up Ewald sums. I. a theoretical and numerical comparison of various particle mesh routines. *J. Chem. Phys.*, 109:7678, 1998.
39. M. Deserno and C. Holm. How to mesh up Ewald sums. II. an accurate error estimate for the P3M algorithm. *J. Chem. Phys.*, 109:7694, 1998.
40. L. Greengard and V. Rokhlin. A fast algorithm for particle simulations. *J. Comp. Phys.*, 73:325, 1987.
41. H. Cheng, L. Greengard, and V. Rokhlin. A fast adaptive multipole algorithm in three dimensions. *J. Comp. Phys.*, 155:468–498, 1999.
42. C. A. White and M. Head-Gordon. Derivation and efficient implementation of the fast multipole method. *J. Chem. Phys.*, 101:6593–6605, 1994.
43. C. A. White and M. Head-Gordon. Rotating around the quartic angular momentum barrier in fast multipole method calculations. *J. Chem. Phys.*, 105:5061–5067, 1996.
44. C. A. White and M. Head-Gordon. Fractional tiers in fast multipole method calculations. *Chem. Phys. Lett.*, 257:647–650, 1996.
45. H. Dachsel. An improved implementation of the fast multipole method. In *Proceedings of the 4th MATHMOD Vienna*, Vienna, 2003.
46. J. E. Barnes and P. Hut. A hierarchical $O(N \log N)$ force calculation algorithm. *Nature*, 324:446, 1986.
47. S. Pfalzner and P. Gibbon. *Many Body Tree Methods in Physics*. Cambridge University Press, New York, 1996.
48. G. Sutmann and B. Steffen. A particle-particle particle-multigrid method for long-range interactions in molecular simulations. *Comp. Phys. Comm.*, 169:343–346, 2005.
49. G. Sutmann and S. Wärow. Wavelet based summation of long-range interactions in molecular systems. (in preparation).
50. N. W. Ashcroft and N. D. Mermin. *Solid State Physics*. Saunders College Publishing, Fort Worth, 1976.
51. R. A. Robinson and R. H. Stokes. *Electrolyte Solutions*. Butterworth, London, 1965.
52. J. W. Perram, H. G. Petersen, and S. W. de Leeuw. An algorithm for the simulation of condensed matter which grows as the $3/2$ power of the number of particles. *Molec. Phys.*, 65:875–893, 1988.
53. D. Fincham. Optimisation of the Ewald sum for large systems. *Molec. Sim.*, 13:1–9, 1994.
54. L. Greengard. *The rapid evaluation of potential fields in particle systems*. MIT press, Cambridge, 1988.
55. L. Greengard. The numerical solution of the N-body problem. *Computers in Physics*, 4:142–152, 1990.
56. L. Greengard. Fast algorithms for classical physics. *Science*, 265:909–914, 1994.
57. J. D. Jackson. *Classical Electrodynamics*. Wiley, New York, 1983.
58. M. Abramowitz and I. Stegun. *Handbook of Mathematical Functions*. Dover Publ. Inc., New York, 1972.
59. R. K. Beatson and L. Greengard. A short course on fast multipole methods. In M. Ainsworth, J. Levesley, W.A. Light, and M. Marletta, editors, *Wavelets, Multilevel Methods and Elliptic PDEs*, pages 1–37. Oxford University Press, 1997.

60. C. G. Lambert, T. A. Darden, and J. A. Board. A multipole-based algorithm for efficient calculation of forces and potentials in macroscopic periodic assemblies of particles. *J. Comp. Phys.*, 126:274–285, 1996.
61. M. Challacombe, C. A. White, and M. Head-Gordon. Periodic boundary conditions and the fast multipole method. *J. Chem. Phys.*, 107:10131–10140, 1997.
62. B. V. Chirikov. A universal instability of many-dimensional oscillator systems. *Phys. Rep.*, 52:264–379, 1979.
63. F. Calvo. Largest Lyapunov exponent in molecular systems: Linear molecules and application to nitrogen clusters. *Phys. Rev. E*, 58:5643–5649, 1998.
64. E. Hairer, Ch. Lubich, and G. Wanner. *Geometric Numerical Integration. Structure-Preserving Algorithms for Ordinary Differential Equations*. Springer, Heidelberg, 2002.
65. J. H. Wilkinson. Error analysis of floating point computation. *Numer. Math.*, 2:319–340, 1960.
66. D. J. Higham. The accuracy of floating point summation. *SIAM J. Sci. Comput.*, 14:783–799, 1993.
67. W. Kahan. Further remarks on reducing truncation errors. *Comm. ACM*, 8:40, 1965.
68. O. Møller. Quasi double-precision in floating point addition. *BIT*, 5:37–50 and 251–255, 1965.
69. W. Kahan. A survey of error analysis. In *Proc. IFIP congress, Ljubljana, Information Processing*, volume 71, pages 1214–1239, Amsterdam, 1972. North-Holland.
70. W. Kahan. Implementation of algorithms. Technical Report Tech. Rep. 20, Dept. Comp. Science, Univ. California, Berkeley, 1973.
71. M. Jankowski, A. Smoktunowicz, and H. Wosniaskowski. A note on floating-point summation of very many terms. *J. Inform. Process. Cybern.*, 19:435–440, 1983.
72. M. Jankowski and H. Wosniaskowski. The accurate solution of certain continuous problems using only single precision arithmetic. *BIT*, 25:635–651, 1985.
73. A. Kielbasinski. Summation algorithm with corrections and some of its applications. *Math. Stos.*, 1:22–41, 1973.
74. A. Neumaier. Rundungsfehleranalyse einiger Verfahren zur Summation endlicher Summen. *Z. Angew. Math. Mech.*, 54:39–51, 1974.
75. D. Levesque and W. T. Ashurst. Long time behavior of the velocity autocorrelation function for a fluid of soft repulsive particles. *Phys. Rev. Lett.*, 33:277–280, 1974.
76. C. Scovel. On symplectic lattice maps. *Phys. Lett. A*, 159:396–400, 1991.
77. D. J. Earn and S. Tremaine. Exact numerical studies of Hamiltonian maps: iterating without round off error. *Physica D*, 56:1–22, 1992.
78. D. Levesque and L. Verlet. Molecular dynamics and time reversibility. *J. Stat. Phys.*, 72:519–537, 1993.
79. R. D. Skeel. Symplectic integration with floating-point arithmetic and other approximations. *Appl. Numer. Math.*, 29:3–18, 1999.
80. B. Leimkuhler and S. Reich. *Simulating Hamiltonian Dynamics*. Cambridge University Press, Cambridge, 2004.
81. R. F. Warming and B. J. Hyett. The modified equation approach to the stability and accuracy analysis of finite difference methods. *J. Comp. Phys.*, 14:159–179, 1974.
82. H. Yoshida. Construction of higher order symplectic integrators. *Phys. Lett. A*, 150:262–268, 1990.

- 83. H. Yoshida. Recent progress in the theory and application of symplectic integrators. *Celestial Mech. Dynam. Astronom.*, 56:27–43, 1993.
- 84. M. Suzuki. Fractal decomposition of exponential operators with applications to many-body theories and Monte Carlo simulations. *Phys. Lett. A*, 146:319–323, 1990.
- 85. M. Suzuki. General theory of higher-order decomposition of exponential operators and symplectic integrators. *Phys. Lett. A*, 165:387–395, 1992.
- 86. W. Kahan and R.-C. Li. Composition constants for raising the orders of unconventional schemes for ordinary differential equations. *Math. Comput.*, 66:1089–1099, 1997.
- 87. R. McLachlan. On the numerical integration of ordinary differential equations by symmetric composition methods. *SIAM J. Sci. Comp.*, 16:151–168, 1995.

An Introduction to *Ab Initio* Molecular Dynamics Simulations

Dominik Marx

Chair of Theoretical Chemistry
Ruhr-Universität Bochum
Universitätsstr. 150, 44780 Bochum, Germany
E-mail: dominik.marx@theochem.rub.de

An introduction into the basic ideas of *ab initio* molecular dynamics methods is given. These methods unify molecular dynamics simulations and electronic structure calculations in the sense that the interactions, i.e. the forces used to propagate the classical nuclei, are obtained from concurrent or “on the fly” electronic structure calculations. Several such molecular dynamics schemes are discussed which arise from a sequence of approximations to the fully coupled Schrödinger equation for electrons and nuclei. Special attention is devoted to Car–Parrinello methods being characterized by a fictitious dynamics of the electronic degrees of freedom.

1 Introduction

Classical molecular dynamics using “predefined potentials”, either based on empirical data or on independent electronic structure calculations, is well established as a powerful tool to investigate many-body condensed matter systems. The broadness, diversity, and level of sophistication of this technique is documented in several monographs as well as proceedings of conferences and scientific schools^{4, 47, 118, 88, 26, 22, 72, 267, 220}. At the very heart of any molecular dynamics scheme is the question of how to describe – that is in practice how to approximate – the interatomic interactions. The traditional route followed in molecular dynamics is to determine these potentials in advance. Typically, the full interaction is broken up into two-body, three-body and many-body contributions, long-range and short-range terms etc., which have to be represented by suitable functional forms, see Section 2 of Ref. 108 for a detailed account. After decades of intense research, very elaborate interaction models including the non-trivial aspect to represent them analytically were devised^{108, 226, 243}.

Despite overwhelming success – which will however not be praised in these Lecture Notes – the need to devise a “fixed model potential” implies serious drawbacks, see the introduction sections of several earlier reviews^{222, 201} for a more complete digression on these aspects. Among the most delicate ones are systems where (i) many different atom or molecule types give rise to a myriad of different interatomic interactions that have to be parameterized and / or (ii) the electronic structure and thus the bonding pattern changes qualitatively in the course of the simulation. These systems can be called “chemically complex”.

The reign of traditional molecular dynamics *and* electronic structure methods was greatly extended by the family of techniques that is called here “*ab initio* molecular dynamics”, see Ref. 179 for a comprehensive account. Other names that are currently in use are for instance first principles, on-the-fly, direct, extended Lagrangian, quantum chemical, Hellmann–Feynman, potential-free, or just quantum molecular dynamics among others.

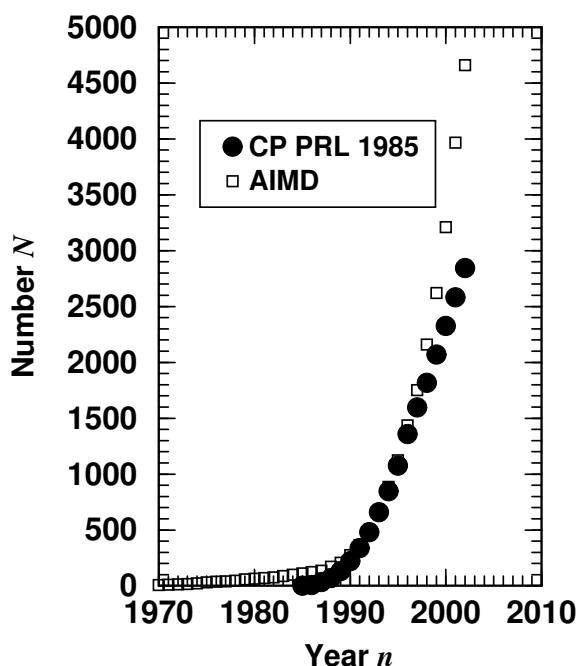


Figure 1. Publication and citation analysis up to the year 2002. Squares: number of publications N which appeared up to the year n containing the keyword “ab initio molecular dynamics” (or synonyms such as “first principles MD”, “Car–Parrinello simulations” etc.) in title, abstract or keyword list. Circles: number of publications N which appeared up to the year n citing the 1985 paper by Car and Parrinello³⁸ (including misspellings of the bibliographic reference). Self-citations and self-papers are excluded, i.e. citations of Ref. 38 in their own papers and papers co-authored by R. Car and / or M. Parrinello are *not* considered in the respective statistics. The analysis is based on the CAPLUS (“Chemical Abstracts Plus”), INSPEC (“Physics Abstracts”), and SCI (“Science Citation Index”) data bases at STN International. Updated statistics as of 05.05.2003 based on data reported in Refs. 178, 179, 181.

The basic idea underlying every *ab initio* molecular dynamics method is to compute the forces acting on the nuclei from electronic structure calculations that are performed “on-the-fly” as the molecular dynamics trajectory is generated. In this way, the electronic variables are not integrated out beforehand, but are considered as active degrees of freedom. This implies that, given a suitable approximate solution of the many-electron problem, also “chemically complex” systems can be handled by molecular dynamics. But this also implies that the approximation is shifted from the level of selecting the model potential to the level of selecting a particular approximation for solving the Schrödinger equation.

Applications of *ab initio* molecular dynamics are particularly widespread in materials science and chemistry, where the aforementioned difficulties (i) and (ii) are particularly severe. The power of this novel technique lead to an explosion of the activity in this field in terms of the number of published papers. The locus can be located in the late-eighties, see the squares in Figure 1 that can be interpreted as a measure of the activity in the area of *ab initio* molecular dynamics. As a matter of fact the time evolution of the number of citations of a particular paper, the one by Car and Parrinello from 1985 entitled “Unified Approach

for Molecular Dynamics and Density-Functional Theory”³⁸, parallels the trend in the entire field, see the circles in Figure 1. Thus, the resonance that the Car and Parrinello paper evoked and the popularity of the entire field go hand in hand in the last decade. Incidentally, the 1985 paper by Car and Parrinello is the last one included in the section “Trends and Prospects” in the reprint collection of “key papers” from the field of atomistic computer simulations⁴⁷. That the entire field of *ab initio* molecular dynamics has grown mature is also evidenced by a separate PACS classification number (71.15.Pd “Electronic Structure: Molecular dynamics calculations (Car-Parrinello) and other numerical simulations”) that was introduced in 1996 into the *Physics and Astronomy Classification Scheme*²¹¹.

Despite its obvious advantages, it is evident that a price has to be paid for putting molecular dynamics on *ab initio* grounds: the correlation lengths and relaxation times that are accessible are much smaller than what is affordable in the framework of standard molecular dynamics. More recently, this caveat got counterbalanced by the ever increasing power of available computing resources, in particular massively parallel platforms¹⁴¹, which shifted many problems in the physical sciences right into the realm of *ab initio* molecular dynamics. Another appealing feature of standard molecular dynamics is less evident, namely the “experimental aspect of playing with the potential”. Thus, tracing back the properties of a given system to a simple physical picture or mechanism is much harder in *ab initio* molecular dynamics. On the other hand it has the power to eventually put phenomena onto a firm basis in terms of the underlying electronic structure and chemical bonding pattern. Most importantly, however, is the fact that new phenomena, which were not foreseen before starting the simulation, can simply happen if necessary. This gives *ab initio* molecular dynamics a truly predictive power.

Ab initio molecular dynamics can also be viewed from another corner, namely from the field of classical trajectory calculations^{270,228}. In this approach, which has its origin in gas phase molecular dynamics, a *global* potential energy surface is constructed in a first step either empirically or based on electronic structure calculations. In a second step, the dynamical evolution of the nuclei is generated by using classical mechanics, quantum mechanics or semi / quasiclassical approximations of various sorts. In the case of using classical mechanics to describe the dynamics – the focus of the present overview – the limiting step for large systems is the first one, why so? There are $3N - 6$ internal degrees of freedom that span the global potential energy surface of an unconstrained N -body system. Using for simplicity 10 discretization points per coordinate implies that of the order of 10^{3N-6} electronic structure calculations are needed in order to map such a global potential energy surface. Thus, the computational workload for the first step grows roughly like $\sim 10^N$ with increasing system size. This is what might be called the “dimensionality bottleneck” of calculations that rely on *global* potential energy surfaces, see for instance the discussion on p. 420 in Ref. 109.

What is needed in *ab initio* molecular dynamics instead? Suppose that a useful trajectory consists of about 10^M molecular dynamics steps, i.e. 10^M electronic structure calculations are needed to generate one trajectory. Furthermore, it is assumed that 10^n independent trajectories are necessary in order to average over different initial conditions so that 10^{M+n} *ab initio* molecular dynamics steps are required in total. Finally, it is assumed that each single-point electronic structure calculation needed to devise the global potential energy surface and one *ab initio* molecular dynamics time step requires roughly the same amount of CPU time. Based on this truly simplistic order of magnitude estimate,

the advantage of *ab initio* molecular dynamics vs. calculations relying on the computation of a global potential energy surface amounts to about $10^{3N-6-M-n}$. The crucial point is that for a given statistical accuracy (that is for M and n fixed and independent of N) and for a given electronic structure method, the computational advantage of “on-the-fly” approaches grows like $\sim 10^N$ with system size.

Of course, considerable progress has been achieved in trajectory calculations by carefully selecting the discretization points and reducing their number, choosing sophisticated representations and internal coordinates, exploiting symmetry etc. but basically the scaling $\sim 10^N$ with the number of nuclei remains a problem. Other strategies consist for instance in reducing the number of active degrees of freedom by constraining certain internal coordinates, representing less important ones by a (harmonic) bath or friction, or building up the global potential energy surface in terms of few-body fragments. All these approaches, however, invoke approximations beyond the ones of the electronic structure method itself. Finally, it is evident that the computational advantage of the “on-the-fly” approaches diminish as more and more trajectories are needed for a given (small) system. For instance extensive averaging over many different initial conditions is required in order to calculate quantitatively scattering or reactive cross sections. Summarizing this discussion, it can be concluded that *ab initio* molecular dynamics is the method of choice to investigate large and “chemically complex” systems.

A host of review articles dealing with *ab initio* molecular dynamics appeared since the early nineties^{222,91,201,193,92,59,265,97,195,225,178,179,268,161,112,87,181,180,142,6} and the interested reader is referred to them for a variety of complementary viewpoints. The present introduction into the basic methodology is a more focussed and updated version of the Lecture Notes¹⁷⁹ “Ab Initio Molecular Dynamics: Theory and Implementation” with special emphasis on the versatile CPMD program package¹³⁹ going back to the *NIC Winter School 2000* on “Modern Methods and Algorithms of Quantum Chemistry”. The present Notes are intended to establish the basis for more specialized Lectures during the *NIC Winter School 2006* “Computational Nanoscience: Do it Yourself!”. The presentation starts from the Schrödinger equation. Classical, Ehrenfest, Born–Oppenheimer, and Car–Parrinello molecular dynamics are “derived” from the time-dependent mean-field approach that is obtained after separating the nuclear and electronic degrees of freedom. The most extensive discussion is related to the features of the basic Car–Parrinello approach but all three *ab initio* approaches to molecular dynamics are contrasted and partly compared. The important issue of how to obtain the correct forces in these schemes is discussed in some depth. The most popular electronic structure theories implemented within *ab initio* molecular dynamics, density functional theory in the first place but also the Hartree–Fock approach, are sketched. Some attention is also given to another important ingredient in *ab initio* molecular dynamics, the choice of the basis set. In addition to the CPMD package¹³⁹ an increasing number of other powerful codes able to perform *ab initio* molecular dynamics simulations is available today (for instance ABINIT¹, CASTEP⁴², CP-PAW⁵², CP2k⁵¹, Dacapo⁵⁴, fhi98md⁷⁷, NWChem¹⁸⁹, PINY²¹³, PWscf/Quantum-ESPRESSO²¹⁸, S/PHI/nX²⁴⁰, or VASP²⁷⁸ among others) which are partly based on very similar techniques as those discussed here.

2 Theoretical Foundations

2.1 Deriving Classical Molecular Dynamics

The starting point of the following discussion is non-relativistic quantum mechanics as formalized via the time-dependent Schrödinger equation

$$i\hbar \frac{\partial}{\partial t} \Phi(\{\mathbf{r}_i\}, \{\mathbf{R}_I\}; t) = \mathcal{H} \Phi(\{\mathbf{r}_i\}, \{\mathbf{R}_I\}; t) \quad (1)$$

in its position representation in conjunction with the standard Hamiltonian

$$\begin{aligned} \mathcal{H} &= - \sum_I \frac{\hbar^2}{2M_I} \nabla_I^2 - \sum_i \frac{\hbar^2}{2m_e} \nabla_i^2 + \sum_{i < j} \frac{e^2}{|\mathbf{r}_i - \mathbf{r}_j|} - \sum_{I,i} \frac{e^2 Z_I}{|\mathbf{R}_I - \mathbf{r}_i|} + \sum_{I < J} \frac{e^2 Z_I Z_J}{|\mathbf{R}_I - \mathbf{R}_J|} \\ &= - \sum_I \frac{\hbar^2}{2M_I} \nabla_I^2 - \sum_i \frac{\hbar^2}{2m_e} \nabla_i^2 + V_{\text{n-e}}(\{\mathbf{r}_i\}, \{\mathbf{R}_I\}) \\ &= - \sum_I \frac{\hbar^2}{2M_I} \nabla_I^2 + \mathcal{H}_e(\{\mathbf{r}_i\}, \{\mathbf{R}_I\}) \end{aligned} \quad (2)$$

for the electronic $\{\mathbf{r}_i\}$ and nuclear $\{\mathbf{R}_I\}$ degrees of freedom. The more convenient atomic units (a.u.) will be introduced at a later stage for reasons that will soon become clear. Thus, only the bare electron–electron, electron–nuclear, and nuclear–nuclear Coulomb interactions are taken into account.

The goal of this section is to derive classical molecular dynamics^{4,118,88,220} starting from Schrödinger’s wave equation and following the elegant route of Tully^{271,272}. To this end, the nuclear and electronic contributions to the total wave function $\Phi(\{\mathbf{r}_i\}, \{\mathbf{R}_I\}; t)$, which depends on *both* the nuclear and electronic coordinates, *have* to be separated. The simplest possible form is a product ansatz

$$\Phi(\{\mathbf{r}_i\}, \{\mathbf{R}_I\}; t) \approx \Psi(\{\mathbf{r}_i\}; t) \chi(\{\mathbf{R}_I\}; t) \exp \left[\frac{i}{\hbar} \int_{t_0}^t dt' \tilde{E}_e(t') \right], \quad (3)$$

where the nuclear and electronic wave functions are separately normalized to unity at every instant of time, i.e. $\langle \chi | \chi \rangle = 1$ and $\langle \Psi | \Psi \rangle = 1$, respectively. In addition, a convenient phase factor

$$\tilde{E}_e = \int d\mathbf{r} d\mathbf{R} \Psi^*(\{\mathbf{r}_i\}; t) \chi^*(\{\mathbf{R}_I\}; t) \mathcal{H}_e \Psi(\{\mathbf{r}_i\}; t) \chi(\{\mathbf{R}_I\}; t) \quad (4)$$

was introduced at this stage such that the final equations will look nice; $\int d\mathbf{r} d\mathbf{R}$ refers to the integration over all $i = 1, \dots$ and $I = 1, \dots$ variables $\{\mathbf{r}_i\}$ and $\{\mathbf{R}_I\}$, respectively. It is mentioned in passing that this approximation is called a one-determinant or single-configuration ansatz for the *total* wave function, which at the end must lead to a mean-field description of the coupled dynamics. Note also that this product ansatz (excluding the phase factor) differs from the Born–Oppenheimer ansatz^{155,162,65} for separating the fast and slow variables

$$\Phi_{\text{BO}}(\{\mathbf{r}_i\}, \{\mathbf{R}_I\}; t) = \sum_{k=0}^{\infty} \tilde{\Psi}_k(\{\mathbf{r}_i\}, \{\mathbf{R}_I\}) \tilde{\chi}_k(\{\mathbf{R}_I\}; t) \quad (5)$$

even in its one-determinant limit, where only a single electronic state k (evaluated for the nuclear configuration $\{\mathbf{R}_I\}$) is included in the expansion.

Inserting the separation ansatz Eq. (3) into Eqs. (1)–(2) yields (after multiplying from the left by $\langle\Psi|$ and $\langle\chi|$ and imposing energy conservation $d\langle\mathcal{H}\rangle/dt \equiv 0$) the following relations

$$i\hbar\frac{\partial\Psi}{\partial t} = -\sum_i\frac{\hbar^2}{2m_e}\nabla_i^2\Psi + \left\{\int d\mathbf{R}\chi^*(\{\mathbf{R}_I\};t)V_{n-e}(\{\mathbf{r}_i\},\{\mathbf{R}_I\})\chi(\{\mathbf{R}_I\};t)\right\}\Psi \quad (6)$$

$$i\hbar\frac{\partial\chi}{\partial t} = -\sum_I\frac{\hbar^2}{2M_I}\nabla_I^2\chi + \left\{\int d\mathbf{r}\Psi^*(\{\mathbf{r}_i\};t)\mathcal{H}_e(\{\mathbf{r}_i\},\{\mathbf{R}_I\})\Psi(\{\mathbf{r}_i\};t)\right\}\chi \quad (7)$$

This set of coupled equations defines the basis of the time-dependent self-consistent field (TDSCF) method introduced as early as 1930 by Dirac⁶², see also Ref. 59. Both electrons and nuclei move quantum-mechanically in time-dependent effective potentials (or self-consistently obtained average fields) obtained from appropriate averages (quantum mechanical expectation values $\langle\dots\rangle$) over the other class of degrees of freedom (by using the nuclear and electronic wave functions, respectively). Thus, the single-determinant ansatz Eq. (3) produces, as already anticipated, a mean-field description of the coupled nuclear-electronic quantum dynamics. This is the price to pay for the simplest possible separation of electronic and nuclear variables.

The next step in the derivation of classical molecular dynamics is the task to approximate the nuclei as classical point particles. How can this be achieved in the framework of the TDSCF approach, given one quantum-mechanical wave equation describing all nuclei? A well-known route to extract classical mechanics from quantum mechanics in general starts with rewriting the corresponding wave function

$$\chi(\{\mathbf{R}_I\};t) = A(\{\mathbf{R}_I\};t) \exp[iS(\{\mathbf{R}_I\};t)/\hbar] \quad (8)$$

in terms of an amplitude factor A and a phase S which are both considered to be real and $A > 0$ in this polar representation, see for instance Refs. 63, 184, 224. After transforming the nuclear wave function in Eq. (7) accordingly and after separating the real and imaginary parts, the TDSCF equation for the nuclei

$$\frac{\partial S}{\partial t} + \sum_I\frac{1}{2M_I}(\nabla_I S)^2 + \int d\mathbf{r}\Psi^*\mathcal{H}_e\Psi = \hbar^2\sum_I\frac{1}{2M_I}\frac{\nabla_I^2 A}{A} \quad (9)$$

$$\frac{\partial A}{\partial t} + \sum_I\frac{1}{M_I}(\nabla_I A)(\nabla_I S) + \sum_I\frac{1}{2M_I}A(\nabla_I^2 S) = 0 \quad (10)$$

is (exactly) re-expressed in terms of the new variables A and S . This so-called “quantum fluid dynamical representation” Eqs. (9)–(10) can actually be used to solve the time-dependent Schrödinger equation⁶¹. The relation for A , Eq. (10), can be rewritten as a continuity equation^{63, 184, 224} with the help of the identification of the nuclear density $|\chi|^2 \equiv A^2$ as directly obtained from the definition Eq. (8). This continuity equation is independent of $\hbar$ and ensures locally the conservation of the particle probability $|\chi|^2$ associated to the nuclei in the presence of a flux.

More important for the present purpose is a more detailed discussion of the relation for S , Eq. (9). This equation contains one term that depends on $\hbar$, a contribution that vanishes

if the classical limit

$$\frac{\partial S}{\partial t} + \sum_I \frac{1}{2M_I} (\nabla_I S)^2 + \int d\mathbf{r} \Psi^* \mathcal{H}_e \Psi = 0 \quad (11)$$

is taken as $\hbar \rightarrow 0$; an expansion in terms of $\hbar$ would lead to a hierarchy of semiclassical methods^{184, 111}. The resulting equation is now isomorphic to equations of motion in the Hamilton–Jacobi formulation^{104, 227}

$$\frac{\partial S}{\partial t} + \mathcal{H}(\{\mathbf{R}_I\}, \{\nabla_I S\}) = 0 \quad (12)$$

of classical mechanics with the classical Hamilton function

$$\mathcal{H}(\{\mathbf{R}_I\}, \{\mathbf{P}_I\}) = T(\{\mathbf{P}_I\}) + V(\{\mathbf{R}_I\}) \quad (13)$$

defined in terms of (generalized) coordinates $\{\mathbf{R}_I\}$ and their conjugate momenta $\{\mathbf{P}_I\}$. With the help of the connecting transformation

$$\mathbf{P}_I \equiv \nabla_I S \quad (14)$$

the Newtonian equation of motion $\dot{\mathbf{P}}_I = -\nabla_I V(\{\mathbf{R}_I\})$ corresponding to Eq. (11)

$$\frac{d\mathbf{P}_I}{dt} = -\nabla_I \int d\mathbf{r} \Psi^* \mathcal{H}_e \Psi \quad \text{or}$$

$$M_I \ddot{\mathbf{R}}_I(t) = -\nabla_I \int d\mathbf{r} \Psi^* \mathcal{H}_e \Psi \quad (15)$$

$$= -\nabla_I V_e^E(\{\mathbf{R}_I(t)\}) \quad (16)$$

can be read off. Thus, the nuclei move according to classical mechanics in an effective potential V_e^E due to the electrons. This potential is a function of only the nuclear positions at time t as a result of averaging $\mathcal{H}_e$ over the electronic degrees of freedom, i.e. computing its quantum expectation value $\langle \Psi | \mathcal{H}_e | \Psi \rangle$, while keeping the nuclear positions fixed at their instantaneous values $\{\mathbf{R}_I(t)\}$.

However, the nuclear wave function still occurs in the TDSCF equation for the electronic degrees of freedom and has to be replaced by the positions of the nuclei for consistency. In this case the classical reduction can be achieved simply by replacing the nuclear density $|\chi(\{\mathbf{R}_I\}; t)|^2$ in Eq. (6) in the limit $\hbar \rightarrow 0$ by a product of delta functions $\prod_I \delta(\mathbf{R}_I - \mathbf{R}_I(t))$ centered at the instantaneous positions $\{\mathbf{R}_I(t)\}$ of the classical nuclei as given by Eq. (15). This yields e.g. for the position operator

$$\int d\mathbf{R} \chi^*(\{\mathbf{R}_I\}; t) \mathbf{R}_I \chi(\{\mathbf{R}_I\}; t) \xrightarrow{\hbar \rightarrow 0} \mathbf{R}_I(t) \quad (17)$$

the required expectation value. This classical limit leads to a time–dependent wave equation for the electrons

$$\begin{aligned} i\hbar \frac{\partial \Psi}{\partial t} &= - \sum_i \frac{\hbar^2}{2m_e} \nabla_i^2 \Psi + V_{n-e}(\{\mathbf{r}_i\}, \{\mathbf{R}_I(t)\}) \Psi \\ &= \mathcal{H}_e(\{\mathbf{r}_i\}, \{\mathbf{R}_I(t)\}) \Psi(\{\mathbf{r}_i\}, \{\mathbf{R}_I\}; t) \end{aligned} \quad (18)$$

which evolve self–consistently as the classical nuclei are propagated via Eq. (15). Note that now $\mathcal{H}_e$ and thus Ψ depend *parametrically* on the classical nuclear *positions* $\{\mathbf{R}_I(t)\}$

at time t through $V_{n-e}(\{\mathbf{r}_i\}, \{\mathbf{R}_I(t)\})$. This means that feedback between the classical and quantum degrees of freedom is incorporated in both directions (at variance with the “classical path” or Mott non-SCF approach to dynamics^{271,272}).

The approach relying on solving Eq. (15) together with Eq. (18) is sometimes called “Ehrenfest molecular dynamics” in honor of Ehrenfest who was the first to address the question^a of how Newtonian classical dynamics can be derived from Schrödinger’s wave equation⁶⁹. In the present case this leads to a hybrid or mixed approach because only the nuclei are forced to behave like classical particles, whereas the electrons are still treated as quantum objects.

Although the TDSCF approach underlying Ehrenfest molecular dynamics clearly is a mean-field theory, transitions between electronic states are included in this scheme. This can be made evident by expanding the *electronic* wave function Ψ (as opposed to the *total* wave function Φ according to Eq. (5)) in terms of many electronic states or determinants Ψ_k

$$\Psi(\{\mathbf{r}_i\}, \{\mathbf{R}_I\}; t) = \sum_{k=0}^{\infty} c_k(t) \Psi_k(\{\mathbf{r}_i\}; \{\mathbf{R}_I\}) \quad (19)$$

with complex coefficients $\{c_k(t)\}$. In this case, the coefficients $\{|c_k(t)|^2\}$ (with $\sum_k |c_k(t)|^2 \equiv 1$) describe explicitly the time evolution of the populations (occupations) of the different states $\{k\}$ whereas interferences are included via the $\{c_k^* c_l\}_{k \neq l}$ contributions. One possible choice for the basis functions $\{\Psi_k\}$ is the adiabatic basis obtained from solving the time-independent electronic Schrödinger equation

$$\mathcal{H}_e(\{\mathbf{r}_i\}; \{\mathbf{R}_I\}) \Psi_k = E_k(\{\mathbf{R}_I\}) \Psi_k(\{\mathbf{r}_i\}; \{\mathbf{R}_I\}) , \quad (20)$$

where $\{\mathbf{R}_I\}$ are the instantaneous nuclear positions at time t according to Eq. (15). The actual equations of motion in terms of the expansion coefficients $\{c_k\}$ are presented in Section 2.2.

At this stage a further simplification can be invoked by restricting the total electronic wave function Ψ to be the ground state wave function Ψ_0 of $\mathcal{H}_e$ at each instant of time according to Eq. (20) and $|c_0(t)|^2 \equiv 1$ in Eq. (19). This should be a good approximation if the energy difference between Ψ_0 and the first excited state Ψ_1 is everywhere large compared to the thermal energy $k_B T$, roughly speaking. In this limit the nuclei move according to Eq. (15) on a single potential energy surface

$$V_e^E = \int d\mathbf{r} \Psi_0^* \mathcal{H}_e \Psi_0 \equiv E_0(\{\mathbf{R}_I\}) \quad (21)$$

that can be computed by solving the *time-independent* electronic Schrödinger equation Eq. (20)

$$\mathcal{H}_e \Psi_0 = E_0 \Psi_0 , \quad (22)$$

for the ground state only. This leads to the identification $V_e^E \equiv E_0$ via Eq. (21), i.e. in this limit the Ehrenfest potential is identical to the ground-state Born–Oppenheimer potential.

^a The opening statement of Ehrenfest’s famous 1927 paper⁶⁹ reads:

“Es ist wünschenswert, die folgende Frage möglichst elementar beantworten zu können: *Welcher Rückblick ergibt sich vom Standpunkt der Quantenmechanik auf die Newtonschen Grundgleichungen der klassischen Mechanik?*”

As a consequence of this observation, it is conceivable to decouple the task of generating the nuclear dynamics from the task of computing the potential energy surface. In a first step E_0 is computed for many nuclear configurations by solving Eq. (22). In a second step, these data points are fitted to an analytical functional form to yield a global potential energy surface²²⁶, from which the gradients can be obtained analytically. In a third step, the Newtonian equation of motion Eq. (16) is solved on this surface for many different initial conditions, producing a “swarm” of classical trajectories. This is, in a nutshell, the basis of *classical trajectory calculations* on global potential energy surfaces^{270, 228}.

As already alluded to in the general introduction, such approaches suffer severely from the “dimensionality bottleneck” as the number of active nuclear degrees of freedom increases. One traditional way out of this dilemma is to approximate the global potential energy surface

$$\begin{aligned} V_e^E \approx V_e^{\text{approx}}(\{\mathbf{R}_I\}) &= \sum_{I=1}^N v_1(\mathbf{R}_I) + \sum_{I<J}^N v_2(\mathbf{R}_I, \mathbf{R}_J) \\ &+ \sum_{I<J<K}^N v_3(\mathbf{R}_I, \mathbf{R}_J, \mathbf{R}_K) + \cdots \end{aligned} \quad (23)$$

in terms of a truncated expansion of many-body contributions^{108, 4, 118}. At this stage, the electronic degrees of freedom are replaced by interaction potentials $\{v_n\}$ and are not featured as explicit degrees of freedom in the equations of motion. Thus, the mixed quantum / classical problem is reduced to purely classical mechanics, once the $\{v_n\}$ are determined. *Classical molecular dynamics*

$$M_I \ddot{\mathbf{R}}_I(t) = -\nabla_I V_e^{\text{approx}}(\{\mathbf{R}_I(t)\}) \quad (24)$$

relies crucially on this idea, where typically only two-body v_2 or three-body v_3 interactions are taken into account^{4, 118}, although more sophisticated models to include non-additive interactions such as polarization exist. This amounts to a dramatic simplification and removes the dimensionality bottleneck as the global potential surface is constructed from a manageable sum of additive few-body contributions — at the price of introducing a drastic approximation and of basically excluding chemical transformations from the realm of simulations.

As a result of this derivation, the essential assumptions underlying classical molecular dynamics become transparent: the electrons follow adiabatically the classical nuclear motion and can be integrated out so that the nuclei evolve on a single Born–Oppenheimer potential energy surface (typically but not necessarily given by the electronic ground state), which is in general approximated in terms of few-body interactions.

Actually, classical molecular dynamics for *many*-body systems is only made possible by somehow decomposing the global potential energy. In order to illustrate this point consider the simulation of $N = 500$ Argon atoms in the liquid phase⁷⁰ where the interactions can faithfully be described by additive two-body terms, i.e. $V_e^{\text{approx}}(\{\mathbf{R}_I\}) \approx \sum_{I<J}^N v_2(|\mathbf{R}_I - \mathbf{R}_J|)$. Thus, the determination of the pair potential v_2 from *ab initio* electronic structure calculations amounts to computing and fitting a one-dimensional function. The corresponding task to determine a global potential energy surface amounts to doing that in about 10^{1500} dimensions, which is simply impossible (and on top of that not necessary for Nobel gases!).

2.2 Ehrenfest Molecular Dynamics

A way out of the dimensionality bottleneck other than to approximate the global potential energy surface Eq. (23) or to reduce the number of active degrees of freedom is to take seriously the classical nuclei approximation to the TDSCF equations, Eqs. (15) and (18). This amounts to computing the Ehrenfest force by actually solving numerically

$$\begin{aligned} M_I \ddot{\mathbf{R}}_I(t) &= -\nabla_I \int d\mathbf{r} \Psi^* \mathcal{H}_e \Psi \\ &= -\nabla_I \langle \Psi | \mathcal{H}_e | \Psi \rangle \end{aligned} \quad (25)$$

$$\begin{aligned} &= -\nabla_I \langle \mathcal{H}_e \rangle \\ &= -\nabla_I V_e^E \\ i\hbar \frac{\partial \Psi}{\partial t} &= \left[-\sum_i \frac{\hbar^2}{2m_e} \nabla_i^2 + V_{n-e}(\{\mathbf{r}_i\}, \{\mathbf{R}_I(t)\}) \right] \Psi \\ &= \mathcal{H}_e \Psi \end{aligned} \quad (26)$$

the coupled set of equations simultaneously. Thereby, the *a priori* construction of any type of potential energy surface is avoided from the outset by solving the time-dependent electronic Schrödinger equation “on-the-fly”. This allows one to compute the force from $\nabla_I \langle \mathcal{H}_e \rangle$ for each configuration $\{\mathbf{R}_I(t)\}$ generated by molecular dynamics; see Section 2.5 for the issue of using the so-called “Hellmann–Feynman forces” instead.

The corresponding equations of motion in terms of the adiabatic basis Eq. (20) and the time-dependent expansion coefficients Eq. (19) read^{271, 272, 65}

$$M_I \ddot{\mathbf{R}}_I(t) = -\sum_k |c_k(t)|^2 \nabla_I E_k - \sum_{k,l} c_k^* c_l (E_k - E_l) \mathbf{d}_I^{kl} \quad (27)$$

$$i\hbar \dot{c}_k(t) = c_k(t) E_k - i\hbar \sum_{I,l} c_l(t) \dot{\mathbf{R}}_I \mathbf{d}_I^{kl}, \quad (28)$$

where the coupling terms are given by

$$\mathbf{d}_I^{kl}(\{\mathbf{R}_I(t)\}) = \int d\mathbf{r} \Psi_k^* \nabla_I \Psi_l \quad (29)$$

with the property $\mathbf{d}_I^{kk} \equiv \mathbf{0}$. The Ehrenfest approach is thus seen to include rigorously non-adiabatic transitions between different electronic states Ψ_k and Ψ_l within the framework of classical nuclear motion and the *mean-field* (TDSCF) approximation to the electronic structure, see e.g. Refs. 271, 272, 65 for reviews and for instance Ref. 223 for an implementation in terms of time-dependent density functional theory.

The restriction to one electronic state in the expansion Eq. (19), which is in most cases the ground state Ψ_0 , leads to

$$M_I \ddot{\mathbf{R}}_I(t) = -\nabla_I \langle \Psi_0 | \mathcal{H}_e | \Psi_0 \rangle \quad (30)$$

$$i\hbar \frac{\partial \Psi_0}{\partial t} = \mathcal{H}_e \Psi_0 \quad (31)$$

as a special case of Eqs. (25)–(26); note that $\mathcal{H}_e$ is time-dependent via the nuclear coordinates $\{\mathbf{R}_I(t)\}$. A point worth mentioning here is that the propagation of the wave function

is unitary, i.e. the wave function preserves its norm and the set of orbitals used to build up the wave function will stay orthonormal, see Section 2.6.

Ehrenfest molecular dynamics is certainly the oldest approach to “on-the-fly” molecular dynamics and is typically used for collision- and scattering-type problems^{55, 270, 185, 56}. Traditionally, the Ehrenfest approach to electron dynamics has not been in widespread use for systems with many active degrees of freedom typical for condensed matter problems. More recently, however, its usage in conjunction with time-dependent density functional theory to describe the electronic subsystem gained a lot of attention, see e.g. Refs. 233, 12, 83, 258, 223, 24, 25, 251 for various such implementations.

2.3 Born–Oppenheimer Molecular Dynamics

An alternative approach to include the electronic structure in molecular dynamics simulations consists in straightforwardly solving the *static* electronic structure problem in each molecular dynamics step given the set of *fixed* nuclear positions at that instance of time. Thus, the electronic structure part is reduced to solving a *time-independent* quantum problem, e.g. by solving the time-independent Schrödinger equation, concurrently to propagating the nuclei via classical molecular dynamics. Thus, the time-dependence of the electronic structure is a consequence of nuclear motion, and not intrinsic as in Ehrenfest molecular dynamics. The resulting Born–Oppenheimer molecular dynamics method is defined by

$$M_I \ddot{\mathbf{R}}_I(t) = -\nabla_I \min_{\Psi_0} \{ \langle \Psi_0 | \mathcal{H}_e | \Psi_0 \rangle \} \quad (32)$$

$$E_0 \Psi_0 = \mathcal{H}_e \Psi_0 \quad (33)$$

for the electronic ground state. A deep difference with respect to Ehrenfest dynamics concerning the nuclear equation of motion is that the minimum of $\langle \mathcal{H}_e \rangle$ has to be reached in each Born–Oppenheimer molecular dynamics step according to Eq. (32). In Ehrenfest dynamics, on the other hand, a wave function that minimized $\langle \mathcal{H}_e \rangle$ initially will also stay in its respective minimum as the nuclei move according to Eq. (30)!

A natural and straightforward extension¹²⁵ of ground-state Born–Oppenheimer dynamics is to apply the same scheme to any excited electronic state Ψ_k without considering any interferences. In particular, this means that also the “diagonal correction terms”¹⁵⁵

$$D_I^{kk}(\{\mathbf{R}_I(t)\}) = - \int d\mathbf{r} \Psi_k^* \nabla_I^2 \Psi_k \quad (34)$$

are *always* neglected; the inclusion of such terms is discussed for instance in Refs. 271, 272. These terms renormalize the Born–Oppenheimer or “clamped nuclei” potential energy surface E_k of a given state Ψ_k (which might also be the ground state Ψ_0) and lead to the so-called “adiabatic potential energy surface” of that state¹⁵⁵. Whence, Born–Oppenheimer molecular dynamics should not be called “adiabatic molecular dynamics”, as is sometime done.

It is useful for the sake of later reference to formulate the Born–Oppenheimer equations of motion for the special case of effective one-particle Hamiltonians. This might be the Hartree–Fock approximation defined to be the variational minimum of the energy expectation value $\langle \Psi_0 | \mathcal{H}_e | \Psi_0 \rangle$ given a single Slater determinant $\Psi_0 = \det\{\psi_i\}$ subject

to the constraint that the one-particle orbitals ψ_i are orthonormal $\langle \psi_i | \psi_j \rangle = \delta_{ij}$. The corresponding constraint minimization of the total energy with respect to the orbitals

$$\min_{\{\psi_i\}} \left\{ \langle \Psi_0 | \mathcal{H}_e | \Psi_0 \rangle \right\} \Big|_{\{\langle \psi_i | \psi_j \rangle = \delta_{ij}\}} \quad (35)$$

can be cast into Lagrange’s formalism

$$\mathcal{L} = - \langle \Psi_0 | \mathcal{H}_e | \Psi_0 \rangle + \sum_{i,j} \Lambda_{ij} (\langle \psi_i | \psi_j \rangle - \delta_{ij}) \quad (36)$$

where Λ_{ij} are the associated Lagrangian multipliers. Unconstrained variation of this Lagrangian with respect to the orbitals

$$\frac{\delta \mathcal{L}}{\delta \psi_i^*} \stackrel{!}{=} 0 \quad (37)$$

leads to the well-known Hartree–Fock equations

$$\mathcal{H}_e^{\text{HF}} \psi_i = \sum_j \Lambda_{ij} \psi_j \quad (38)$$

as derived in standard text books^{252, 183}; the diagonal canonical form $\mathcal{H}_e^{\text{HF}} \psi_i = \epsilon_i \psi_i$ is obtained after a unitary transformation and $\mathcal{H}_e^{\text{HF}}$ denotes the effective one-particle Hamiltonian, see Section 2.7 for more details. The equations of motion corresponding to Eqs. (32)–(33) read

$$M_I \ddot{\mathbf{R}}_I(t) = - \nabla_I \min_{\{\psi_i\}} \{ \langle \Psi_0 | \mathcal{H}_e^{\text{HF}} | \Psi_0 \rangle \} \quad (39)$$

$$0 = - \mathcal{H}_e^{\text{HF}} \psi_i + \sum_j \Lambda_{ij} \psi_j \quad (40)$$

for the Hartree–Fock case. A similar set of equations is obtained if Hohenberg–Kohn–Sham density functional theory^{194, 66, 176} is used, where $\mathcal{H}_e^{\text{HF}}$ has to be replaced by the Kohn–Sham effective one-particle Hamiltonian H_e^{KS} , see Section 2.7 for more details. Instead of diagonalizing the one-particle Hamiltonian an alternative but equivalent approach consists in directly performing the constraint minimization according to Eq. (35) via non-linear optimization techniques.

Early applications of Born–Oppenheimer molecular dynamics were performed in the framework of a semiempirical approximation to the electronic structure problem^{281, 282}. But only a few years later an *ab initio* approach was implemented within the Hartree–Fock approximation¹⁶⁶. Born–Oppenheimer dynamics started to become popular in the early nineties with the availability of more efficient electronic structure codes in conjunction with sufficient computer power to solve “interesting problems”, see for instance the compilation of such studies in Table 1 in a recent overview article³². More recently, a revival of these activities with greatly improved algorithms to perform Born–Oppenheimer simulations is observed^{278, 159, 160, 277, 137}.

Undoubtedly, the breakthrough of Hohenberg–Kohn–Sham density functional theory in the realm of chemistry – which took place around the same time – also helped a lot by greatly improving the “price / performance ratio” of the electronic structure part, see e.g. Refs. 289, 246, 188. A third and possibly the crucial reason that boosted the field of *ab initio* molecular dynamics was the pioneering introduction of the Car–Parrinello approach³⁸,

see also Figure 1. This technique opened novel avenues to treat large-scale problems via *ab initio* molecular dynamics and catalyzed the entire field by making “interesting calculations” possible, see also the closing section on applications.

2.4 Car–Parrinello Molecular Dynamics

2.4.1 Motivation

A non-obvious approach to cut down the computational expenses of molecular dynamics which includes the electrons in a single state was proposed by Car and Parrinello in 1985³⁸. In retrospect it can be considered to combine the advantages of both Ehrenfest and Born–Oppenheimer molecular dynamics. In Ehrenfest dynamics the time scale and thus the time step to integrate Eqs. (30) and (31) simultaneously is dictated by the intrinsic dynamics of the electrons. Since electronic motion is much faster than nuclear motion, the largest possible time step is that which allows to integrate the electronic equations of motion. Contrary to that, there is no electron dynamics whatsoever involved in solving the Born–Oppenheimer Eqs. (32)–(33), i.e. they can be integrated on the time scale given by nuclear motion. However, this means that the electronic structure problem has to be solved self-consistently at each molecular dynamics step, whereas this is avoided in Ehrenfest dynamics due to the possibility to propagate the wave function by applying the Hamiltonian to an initial wave function (obtained e.g. by one self-consistent diagonalization).

From an algorithmic point of view the main task achieved in ground-state Ehrenfest dynamics is simply to keep the wave function automatically minimized as the nuclei are propagated. This, however, might be achieved – in principle – by another sort of deterministic dynamics than first-order Schrödinger dynamics. In summary, the “Best of all Worlds Method” should (i) integrate the equations of motion on the (long) time scale set by the nuclear motion but nevertheless (ii) take intrinsically advantage of the smooth time-evolution of the dynamically evolving electronic subsystem as much as possible. The second point allows to circumvent explicit diagonalization or minimization to solve the electronic structure problem for the next molecular dynamics step. Car–Parrinello molecular dynamics is an efficient method to satisfy requirement (ii) in a numerically stable fashion and makes an acceptable compromise concerning the length of the time step (i).

2.4.2 Car–Parrinello Lagrangian and Equations of Motion

The basic idea of the Car–Parrinello approach can be viewed to exploit the quantum-mechanical adiabatic time-scale separation of fast electronic and slow nuclear motion by transforming that into classical-mechanical adiabatic energy-scale separation in the framework of dynamical systems theory. In order to achieve this goal the two-component quantum / classical problem is mapped onto a two-component purely classical problem with two separate energy scales at the expense of loosing the explicit time-dependence of the quantum subsystem dynamics. Furthermore, the central quantity, the energy of the electronic subsystem $\langle \Psi_0 | \mathcal{H}_e | \Psi_0 \rangle$ evaluated with some wave function Ψ_0 , is certainly a function of the nuclear positions $\{\mathbf{R}_I\}$. But at the same time it *can* be considered to be a functional of the wave function Ψ_0 and thus of a set of one-particle orbitals $\{\psi_i\}$ (or in general of other functions such as two-particle geminals) used to build up this wave function (being for instance a Slater determinant $\Psi_0 = \det\{\psi_i\}$ or a combination thereof).

Now, in classical mechanics the force on the nuclei is obtained from the derivative of a Lagrangian with respect to the nuclear positions. This suggests that a functional derivative with respect to the orbitals, which are interpreted as classical fields, might yield the force on the orbitals, given a suitable Lagrangian. In addition, possible constraints within the set of orbitals have to be imposed, such as e.g. orthonormality (or generalized orthonormality conditions that include an overlap matrix).

Car and Parrinello postulated the following class of Lagrangians³⁸

$$\mathcal{L}_{\text{CP}} = \underbrace{\sum_I \frac{1}{2} M_I \dot{\mathbf{R}}_I^2}_{\text{kinetic energy}} + \underbrace{\sum_i \mu \langle \dot{\psi}_i | \dot{\psi}_i \rangle}_{\text{potential energy}} - \underbrace{\langle \Psi_0 | \mathcal{H}_e | \Psi_0 \rangle}_{\text{potential energy}} + \underbrace{\text{constraints}}_{\text{orthonormality}} \quad (41)$$

to serve this purpose; note that sometimes the prefactor $\mu_i/2$ is used and orbital occupation numbers $f_i = 0, 1, 2$ are introduced. The corresponding Newtonian equations of motion are obtained from the associated Euler–Lagrange equations

$$\frac{d}{dt} \frac{\partial \mathcal{L}}{\partial \dot{\mathbf{R}}_I} = \frac{\partial \mathcal{L}}{\partial \mathbf{R}_I} \quad (42)$$

$$\frac{d}{dt} \frac{\delta \mathcal{L}}{\delta \dot{\psi}_i^*} = \frac{\delta \mathcal{L}}{\delta \psi_i^*} \quad (43)$$

like in classical mechanics, but here for both the nuclear positions and the orbitals; note $\psi_i^* = \langle \psi_i |$ and that the constraints are holonomic¹⁰⁴. Following this route of ideas, generic Car–Parrinello equations of motion are found to be of the form

$$M_I \ddot{\mathbf{R}}_I(t) = - \frac{\partial}{\partial \mathbf{R}_I} \langle \Psi_0 | \mathcal{H}_e | \Psi_0 \rangle + \frac{\partial}{\partial \mathbf{R}_I} \{ \text{constraints} \} \quad (44)$$

$$\mu \ddot{\psi}_i(t) = - \frac{\delta}{\delta \psi_i^*} \langle \Psi_0 | \mathcal{H}_e | \Psi_0 \rangle + \frac{\delta}{\delta \psi_i^*} \{ \text{constraints} \} \quad (45)$$

where μ are the “fictitious masses” or inertia parameters assigned to the orbital degrees of freedom; the units of the mass parameter μ are energy times a squared time for reasons of dimensionality. Note that the constraints within the total wave function lead to “constraint forces” in the equations of motion. Note also that these constraints

$$\text{constraints} = \text{constraints}(\{\psi_i\}, \{\mathbf{R}_I\}) \quad (46)$$

might be a function of both the set of orbitals $\{\psi_i\}$ and the nuclear positions $\{\mathbf{R}_I\}$. These dependencies have to be taken into account properly in deriving the Car–Parrinello equations following from Eq. (41) using Eqs. (42)–(43), see Section 2.5 for a general discussion and see e.g. Ref. 163 for a case with an additional dependence of the wave function constraint on nuclear positions.

According to the Car–Parrinello equations of motion, the nuclei evolve in time at a certain (instantaneous) physical temperature $\propto \sum_I M_I \dot{\mathbf{R}}_I^2$, whereas a “fictitious temperature” $\propto \sum_i \mu \langle \dot{\psi}_i | \dot{\psi}_i \rangle$ is associated to the electronic degrees of freedom. In this terminology, “low electronic temperature” or “cold electrons” means that the electronic subsystem is close to its instantaneous minimum energy $\min_{\{\psi_i\}} \langle \Psi_0 | \mathcal{H}_e | \Psi_0 \rangle$, i.e. close to the exact Born–Oppenheimer surface. Thus, a ground–state wave function optimized for the initial configuration of the nuclei will stay close to its ground state also during time evolution if it is kept at a sufficiently low temperature.

The remaining task is to separate in practice nuclear and electronic motion such that the fast electronic subsystem stays cold also for long times but still follows the slow nuclear motion adiabatically (or instantaneously). Simultaneously, the nuclei are nevertheless kept at a much higher temperature. This can be achieved in nonlinear classical dynamics via decoupling of the two subsystems and (quasi-) adiabatic time evolution. This is possible if the power spectra stemming from both dynamics do not have substantial overlap in the frequency domain so that energy transfer from the “hot nuclei” to the “cold electrons” becomes practically impossible on the relevant time scales. This amounts in other words to imposing and maintaining a metastability condition in a complex dynamical system for sufficiently long times. How and to which extent this is possible in practice was analyzed in detail in a pioneering technical investigation based on well-controlled model systems¹⁹⁷ (see also Refs. 40, 198, 254 and Sects. 3.2 and 3.3 in Ref. 222). Later the adiabaticity issue has been investigated with more mathematical rigor³³ and in terms of a generalization to a second level of adiabatic decoupling¹⁷⁷.

2.4.3 Why Does the Car–Parrinello Method Work ?

In order to shed light on the title question, the dynamics generated by the Car–Parrinello Lagrangian Eq. (41) is analyzed¹⁹⁷ in more detail invoking a “classical dynamics perspective” of a simple model system (eight silicon atoms forming a periodic diamond lattice, local density approximation to density functional theory, normconserving pseudopotentials for core electrons, plane wave basis for valence orbitals, 0.3 fs time step with $\mu = 300$ a.u., in total 20 000 time steps or 6.3 ps), for full details see Ref. 197); a concise presentation of similar ideas can be found in Ref. 40. For this system the vibrational density of states or power spectrum of the electronic degrees of freedom, i.e. the Fourier transform of the statistically averaged velocity autocorrelation function of the classical fields

$$f(\omega) = \int_0^\infty dt \cos(\omega t) \sum_i \langle \dot{\psi}_i; t | \dot{\psi}_i; 0 \rangle \quad (47)$$

is compared to the highest-frequency phonon mode $\omega_n^{\max}$ of the nuclear subsystem in Figure 2. From this figure it is evident that for the chosen parameters the nuclear and electronic subsystems are dynamically separated: their power spectra do not overlap so that energy transfer from the hot to the cold subsystem is expected to be prohibitively slow, see Section 3.3 in Ref. 222 for a similar argument.

This is indeed the case as can be verified in Figure 3 where the conserved energy E_{cons} , physical total energy E_{phys} , electronic energy V_e , and fictitious kinetic energy of the electrons T_e

$$E_{\text{cons}} = \sum_i \mu \langle \dot{\psi}_i | \dot{\psi}_i \rangle + \sum_I \frac{1}{2} M_I \dot{\mathbf{R}}_I^2 + \langle \Psi_0 | \mathcal{H}_e | \Psi_0 \rangle \quad (48)$$

$$E_{\text{phys}} = \sum_I \frac{1}{2} M_I \dot{\mathbf{R}}_I^2 + \langle \Psi_0 | \mathcal{H}_e | \Psi_0 \rangle = E_{\text{cons}} - T_e \quad (49)$$

$$V_e = \langle \Psi_0 | \mathcal{H}_e | \Psi_0 \rangle \quad (50)$$

$$T_e = \sum_i \mu \langle \dot{\psi}_i | \dot{\psi}_i \rangle \quad (51)$$

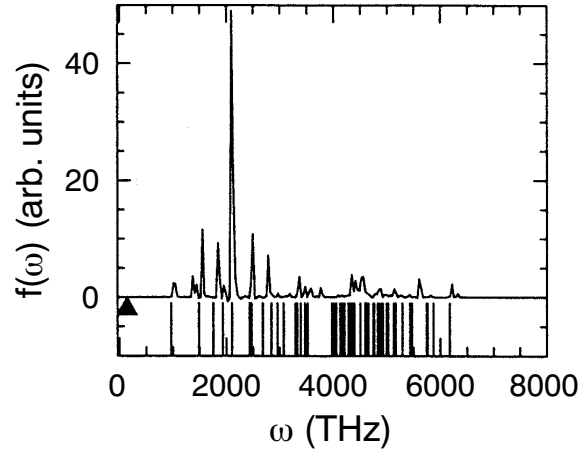


Figure 2. Vibrational density of states Eq. (47) (continuous spectrum in upper part) and harmonic approximation thereof Eq. (52) (stick spectrum in lower part) of the electronic degrees of freedom compared to the highest-frequency phonon mode $\omega_n^{\max}$ (triangle) for a model system; for further details see text. Adapted from Ref. 197.

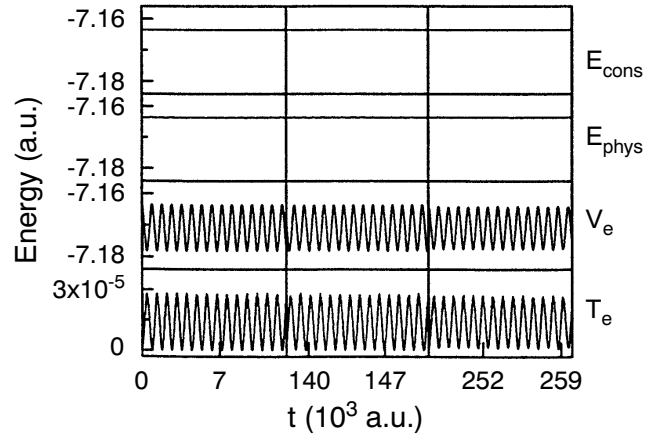


Figure 3. Various energies Eqs. (48)–(51) for a model system propagated via Car–Parrinello molecular dynamics for at short (up to 300 fs), intermediate, and long times (up to 6.3 ps); for further details see text. Adapted from Ref. 197.

are shown for the same system as a function of time. First of all, there should be a conserved energy quantity according to classical dynamics since the constraints are holonomic¹⁰⁴. Indeed “the Hamiltonian” or conserved energy E_{cons} is a constant of motion (with relative variations smaller than 10^{-6} and with no drift), which serves as an extremely sensitive check of the molecular dynamics algorithm. Contrary to that the electronic energy V_e displays a simple oscillation pattern due to the simplicity of the phonon modes.

Most importantly, the fictitious kinetic energy of the electrons T_e is found to perform

bound oscillations around a *constant*, i.e. the electrons “do not heat up” systematically in the presence of the hot nuclei; note that T_e is a measure for deviations from the exact Born–Oppenheimer surface. Closer inspection shows actually two time scales of oscillations: the one visible in Figure 3 stems from the drag exerted by the moving nuclei on the electrons and is the mirror image of the V_e fluctuations. Superimposed on top of that (not shown, but see Figure 4(b)) are small–amplitude high frequency oscillations intrinsic to the fictitious electron dynamics with a period of only a fraction of the visible mode. These oscillations are actually instrumental for the stability of the Car–Parrinello dynamics, *vide infra*. But already the visible variations are three orders of magnitude smaller than the physically meaningful oscillations of V_e . As a result, E_{phys} defined as $E_{\text{cons}} - T_e$ or equivalently as the sum of the nuclear kinetic energy and the electronic total energy (which serves as the potential energy for the nuclei) is essentially constant on the relevant energy and time scales. Thus, it behaves approximately like the strictly conserved total energy in classical molecular dynamics (with only nuclei as dynamical degrees of freedom) or in Born–Oppenheimer molecular dynamics (with fully optimized electronic degrees of freedom) and is therefore often denoted as the “physical total energy”. This implies that the resulting physically significant dynamics of the nuclei yields an excellent approximation to microcanonical dynamics (and assuming ergodicity to the microcanonical ensemble). Note that a different explanation was advocated in Ref. 200 (see also Ref. 201, in particular Section VIII.B and C), which was however revised in Ref. 40. A discussion similar in spirit to the one outlined here¹⁹⁷ is provided in Ref. 222, see in particular Section 3.2 and 3.3.

Given the adiabatic separation and the stability of the propagation, the central question remains if the forces acting on the nuclei are actually the “correct” ones in Car–Parrinello molecular dynamics. As a reference serve the forces obtained from full self–consistent minimizations of the electronic energy $\min_{\{\psi_i\}} \langle \Psi_0 | \mathcal{H}_e | \Psi_0 \rangle$ at each time step, i.e. Born–Oppenheimer molecular dynamics with extremely well converged wave functions. This is indeed the case as demonstrated in Figure 4(a): the physically meaningful dynamics of the x –component of the force acting on one silicon atom in the model system obtained from stable Car–Parrinello fictitious dynamics propagation of the electrons and from iterative minimizations of the electronic energy are extremely close.

Better resolution of one oscillation period in (b) reveals that the gross deviations are also oscillatory but that they are four orders of magnitudes smaller than the physical variations of the force resolved in Figure 4(a). These correspond to the “large–amplitude” oscillations of T_e visible in Figure 3 due to the drag of the nuclei exerted on the quasi–adiabatically following electrons having a *finite* dynamical mass μ . Note that the inertia of the electrons also dampens artificially the nuclear motion (typically on a few–percent scale, see Section V.C.2 in Ref. 28 for an analysis and a renormalization correction of M_I) but decreases as the fictitious mass approaches the adiabatic limit $\mu \rightarrow 0$. Superimposed on the gross variation in (b) are again high–frequency *bound oscillatory small–amplitude* fluctuations like for T_e . They lead on *physically relevant time scales* (i.e. those visible in Figure 4(a)) to “averaged forces” that are very close to the exact ground–state Born–Oppenheimer forces. This feature is an important ingredient in the derivation of adiabatic dynamics^{197, 177}.

In conclusion, the Car–Parrinello force can be said to deviate at most instants of time from the exact Born–Oppenheimer force. However, this does not disturb the physical time

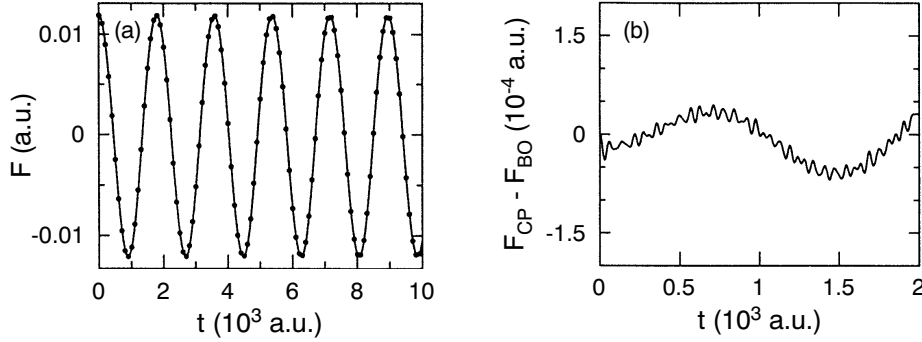


Figure 4. (a) Comparison of the x -component of the force acting on one atom of a model system obtained from Car-Parrinello (solid line) and well-converged Born-Oppenheimer (dots) molecular dynamics. (b) Enlarged view of the difference between Car-Parrinello and Born-Oppenheimer forces; for further details see text. Adapted from Ref. 197.

evolution due to (i) the smallness and boundedness of this difference *and* (ii) the intrinsic averaging effect of small-amplitude high-frequency oscillations within a few molecular dynamics time steps, i.e. on the sub-femtosecond time scale which is irrelevant for *nuclear* dynamics.

2.4.4 How to Control Adiabaticity ?

An important question is under which circumstances the adiabatic separation can be achieved, and how it can be controlled. A simple harmonic analysis of the frequency spectrum of the orbital classical fields close to the minimum defining the ground state yields¹⁹⁷

$$\omega_{ij} = \left(\frac{2(\epsilon_i - \epsilon_j)}{\mu} \right)^{1/2}, \quad (52)$$

where ϵ_j and ϵ_i are the eigenvalues of occupied and unoccupied orbitals, respectively; see Eq. (26) in Ref. 197 for the case where both orbitals are occupied ones. It can be seen from Figure 2 that the harmonic approximation works faithfully as compared to the exact spectrum; see Ref. 199 and Section IV.A in Ref. 201 for a more general analysis of the associated equations of motion. Since this is in particular true for the lowest frequency $\omega_e^{\min}$, the handy analytic estimate for the lowest possible electronic frequency

$$\omega_e^{\min} \propto \left(\frac{E_{\text{gap}}}{\mu} \right)^{1/2}, \quad (53)$$

shows that this frequency increases like the square root of the electronic energy difference E_{gap} between the lowest unoccupied and the highest occupied orbital. On the other hand it increases similarly for a decreasing fictitious mass parameter μ .

In order to guarantee the adiabatic separation, the frequency difference $\omega_e^{\min} - \omega_n^{\max}$ should be large, see Section 3.3 in Ref. 222 for a similar argument. But both the highest phonon frequency $\omega_n^{\max}$ and the energy gap E_{gap} are quantities that are dictated by the

physics of the system. Whence, the only parameter in our hands to control adiabatic separation is the fictitious mass, which is therefore also called “adiabaticity parameter”. However, decreasing μ not only shifts the electronic spectrum upwards on the frequency scale, but also stretches the entire frequency spectrum according to Eq. (52). This leads to an increase of the maximum frequency according to

$$\omega_e^{\max} \propto \left(\frac{E_{\text{cut}}}{\mu} \right)^{1/2}, \quad (54)$$

where E_{cut} is the largest kinetic energy in an expansion of the wave function in terms of a plane wave basis set.

At this place a limitation to decrease μ arbitrarily kicks in due to the maximum length of the molecular dynamics time step $\Delta t^{\max}$ that can be used. The time step is inversely proportional to the highest frequency in the system, which is $\omega_e^{\max}$ and thus the relation

$$\Delta t^{\max} \propto \left(\frac{\mu}{E_{\text{cut}}} \right)^{1/2} \quad (55)$$

governs the largest time step that is possible. As a consequence, Car–Parrinello simulators have to find their way between Scylla and Charybdis and have to make a compromise on the control parameter μ ; typical values for large-gap systems are $\mu = 500\text{--}1500$ a.u. together with a time step of about 5–10 a.u. (0.12–0.24 fs). Recently, an algorithm was devised that optimizes μ during a particular simulation given a fixed accuracy criterion³⁴. Note that a poor man’s way to keep the time step large and still increase μ in order to satisfy adiabaticity is to choose heavier nuclear masses. That depresses the largest phonon or vibrational frequency $\omega_n^{\max}$ of the nuclei (at the cost of renormalizing all *dynamical* quantities in the sense of classical isotope effects).

Up to this point the entire discussion of the stability and adiabaticity issues was based on model systems, approximate and mostly qualitative in nature. But recently it was actually proven³³ that the deviation or the absolute error Δ_μ of the Car–Parrinello trajectory relative to the trajectory obtained on the exact Born–Oppenheimer potential energy surface is controlled by μ :

Theorem 1 iv.): There are constants $C > 0$ and $\mu^* > 0$ such that

$$\Delta_\mu = |\mathbf{R}^\mu(t) - \mathbf{R}^0(t)| + ||\psi^\mu; t\rangle - |\psi^0; t\rangle| \leq C\mu^{1/2}, \quad 0 \leq t \leq T \quad (56)$$

and the fictitious kinetic energy satisfies

$$T_e = \mu \langle \dot{\psi}^\mu; t | \dot{\psi}^\mu; t \rangle \leq C\mu, \quad 0 \leq t \leq T \quad (57)$$

for all values of the parameter μ satisfying $0 < \mu \leq \mu^*$, where up to time $T > 0$ there exists a unique nuclear trajectory on the exact Born–Oppenheimer surface with $\omega_e^{\min} > 0$ for $0 \leq t \leq T$, i.e. there is “always” a finite electronic excitation gap. Here, the superscript μ or 0 indicates that the trajectory was obtained via Car–Parrinello molecular dynamics using a finite mass μ or via dynamics on the exact Born–Oppenheimer surface, respectively. Note that not only the nuclear trajectory is shown to be close to the correct one, but also the wave function is proven to stay close to the fully converged one up to time T . Furthermore, it was also investigated what happens if the initial wave function at $t = 0$ is not the minimum of the electronic energy $\langle \mathcal{H}_e \rangle$ but trapped in an excited state. In this case it is found that the propagated wave function will keep on oscillating at $t > 0$

also for $\mu \rightarrow 0$ and not even time averages converge to any of the eigenstates. Note that this does not preclude Car–Parrinello molecular dynamics in excited states, which is possible given a properly “minimizable” expression for the electronic energy, see e.g. Refs. 125, 86. However, this finding might have crucial implications for electronic level-crossing situations⁶⁵.

What happens if the electronic gap is very small or even vanishes $E_{\text{gap}} \rightarrow 0$ as is the case for metallic systems? In this limit, all the above–given arguments break down due to the occurrence of zero–frequency electronic modes in the power spectrum according to Eq. (53), which necessarily overlap with the phonon spectrum. Following an idea of Sprik²⁴² applied in a classical context it was shown that the coupling of separate Nosé–Hoover thermostats^{4,118,88} to the nuclear and electronic subsystem can maintain adiabaticity by counterbalancing the energy flow from ions to electrons so that the electrons stay “cool”²⁷; see Ref. 84 for a similar idea to restore adiabaticity and Refs. 263, 187, 29 for analyses and improvements. Although this method is demonstrated to work in practice¹⁹⁶, this *ad hoc* cure is not entirely satisfactory from both a theoretical and practical point of view so that the well–controlled Born–Oppenheimer approach is recommended for strongly metallic systems. An additional advantage for metallic systems is that the latter is also better suited to sample many $\mathbf{k}$ –points allows easily for fractional occupation numbers^{194,66,176}, and can handle efficiently the so–called charge sloshing problem²⁰¹.

2.4.5 The Quantum Chemistry Viewpoint

In order to understand Car–Parrinello molecular dynamics also from the “quantum chemistry perspective”, it is useful to formulate it for the special case of the Hartree–Fock approximation using

$$\mathcal{L}_{\text{CP}} = \sum_I \frac{1}{2} M_I \dot{\mathbf{R}}_I^2 + \sum_i \mu \langle \dot{\psi}_i | \dot{\psi}_i \rangle - \langle \Psi_0 | \mathcal{H}_e^{\text{HF}} | \Psi_0 \rangle + \sum_{i,j} \Lambda_{ij} (\langle \psi_i | \psi_j \rangle - \delta_{ij}) . \quad (58)$$

The resulting equations of motion

$$M_I \ddot{\mathbf{R}}_I(t) = -\nabla_I \langle \Psi_0 | \mathcal{H}_e^{\text{HF}} | \Psi_0 \rangle \quad (59)$$

$$\mu \ddot{\psi}_i(t) = -\mathcal{H}_e^{\text{HF}} \psi_i + \sum_j \Lambda_{ij} \psi_j \quad (60)$$

are very close to those obtained for Born–Oppenheimer molecular dynamics Eqs. (39)–(40) except for (i) no need to minimize the electronic total energy expression and (ii) featuring the additional fictitious kinetic energy term associated to the orbital degrees of freedom. It is suggestive to argue that both sets of equations become identical if the term $|\mu \ddot{\psi}_i(t)|$ is small at any time t compared to the physically relevant forces on the right–hand–side of both Eq. (59) and Eq. (60). This term being zero (or small) means that one is at (or close to) the minimum of the electronic energy $\langle \Psi_0 | \mathcal{H}_e^{\text{HF}} | \Psi_0 \rangle$ since time derivatives of the orbitals $\{\psi_i\}$ can be considered as variations of Ψ_0 and thus of the expectation value $\langle \mathcal{H}_e^{\text{HF}} \rangle$ itself. In other words, no forces act on the wave function if $\mu \ddot{\psi}_i \equiv 0$. In conclusion, the Car–Parrinello equations are expected to produce the correct dynamics and thus physical

trajectories in the microcanonical ensemble in this idealized limit. But if $|\mu\ddot{\psi}_i(t)|$ is small for all i , this also implies that the associated kinetic energy $T_e = \sum_i \mu \langle \dot{\psi}_i | \dot{\psi}_i \rangle$ is small, which connects these more qualitative arguments with the previous discussion¹⁹⁷.

At this stage, it is also interesting to compare the structure of the Lagrangian Eq. (58) and the Euler–Lagrange equation Eq. (43) for Car–Parrinello dynamics to the analogues Eqs. (36) and (37), respectively, used to derive “Hartree–Fock statics”. The former reduce to the latter if the dynamical aspect and the associated time evolution is neglected, that is in the limit that the nuclear and electronic momenta are absent or constant. Thus, the Car–Parrinello ansatz, namely Eq. (41) together with Eqs. (42)–(43), can also be viewed as a prescription to derive a new class of “dynamical *ab initio* methods” in very general terms.

2.4.6 The Simulated Annealing and Optimization Viewpoints

In the discussion given above, Car–Parrinello molecular dynamics was motivated by “combining” the positive features of both Ehrenfest and Born–Oppenheimer molecular dynamics as much as possible. Looked at from another side, the Car–Parrinello method can also be considered as an ingenious way to perform *global* optimizations (minimizations) of nonlinear functions, here $\langle \Psi_0 | \mathcal{H}_e | \Psi_0 \rangle$, in a high–dimensional parameter space including complicated constraints. The optimization parameters are those used to represent the total wave function Ψ_0 in terms of simpler functions, for instance expansion coefficients of the orbitals in terms of Gaussians or plane waves, see e.g. Refs. 242, 173, 288, 253 for applications of the same idea in other fields.

Keeping the nuclei frozen for a moment, one could start this optimization procedure from a “random wave function” which certainly does not minimize the electronic energy. Thus, its fictitious kinetic energy is high, the electronic degrees of freedom are “hot”. This energy, however, can be extracted from the system by systematically cooling it to lower and lower temperatures. This can be achieved in an elegant way by adding a non–conservative damping term to the electronic Car–Parrinello equation of motion Eq. (45)

$$\mu\ddot{\psi}_i(t) = -\frac{\delta}{\delta\psi_i^*} \langle \Psi_0 | \mathcal{H}_e | \Psi_0 \rangle + \frac{\delta}{\delta\psi_i^*} \{\text{constraints}\} - \gamma_e \mu \dot{\psi}_i, \quad (61)$$

where $\gamma_e \geq 0$ is a friction constant that governs the rate of energy dissipation²⁵⁶; alternatively, dissipation can be enforced in a discrete fashion by reducing the velocities by multiplying them with a constant factor < 1 . Note that this deterministic and dynamical method is very similar in spirit to simulated annealing¹⁵² invented in the framework of the stochastic Monte Carlo approach in the canonical ensemble. If the energy dissipation is done slowly, the wave function will find its way down to the minimum of the energy. At the end, an intricate global optimization has been performed!

If the nuclei are allowed to move according to Eq. (44) in the presence of another damping term a combined or simultaneous optimization of both electrons and nuclei can be achieved, which amounts to a “global geometry optimization”. This perspective is stressed in more detail in the review Ref. 91 and an implementation of such ideas within the CADPAC quantum chemistry code is described in Ref. 287. This operational mode of Car–Parrinello molecular dynamics is related to other optimization techniques where it is aimed to optimize simultaneously both the structure of the nuclear skeleton and the electronic structure. This is achieved by considering the nuclear coordinates and the expansion

coefficients of the orbitals as variation parameters on the same footing^{21, 130, 253}. But Car–Parrinello molecular dynamics is more than that because even if the nuclei continuously move according to Newtonian dynamics at finite temperature an initially optimized wave function will stay optimal along the nuclear trajectory.

2.4.7 The Extended Lagrangian Viewpoint

There is still another way to look at the Car–Parrinello method, namely in the light of so-called “extended Lagrangians” or “extended system dynamics”⁵, see e.g. Refs. 48, 4, 118, 244, 88 for introductions. The basic idea is to couple additional degrees of freedom to the Lagrangian of interest, thereby “extending” it by increasing the dimensionality of phase space. These degrees of freedom are treated like classical particle coordinates, i.e. they are in general characterized by “positions”, “momenta”, “masses”, “interactions” and a “coupling term” to the particle’s positions and momenta. In order to distinguish them from the physical degrees of freedom, they are often called “fictitious degrees of freedom”.

The corresponding equations of motion follow from the Euler–Lagrange equations and yield a microcanonical ensemble in the extended phase space where the Hamiltonian of the extended system is strictly conserved. In other words, the Hamiltonian of the physical (sub-) system is no more (strictly) conserved, and the produced ensemble is no more the microcanonical one. Any extended system dynamics is constructed such that time-averages taken in that part of phase space that is associated to the physical degrees of freedom (obtained from a partial trace over the fictitious degrees of freedom) are physically meaningful. Of course, dynamics and thermodynamics of the system are affected by adding fictitious degrees of freedom, the classic examples being temperature and pressure control by thermostats and barostats.

In the case of Car–Parrinello molecular dynamics, the basic Lagrangian for Newtonian dynamics of the nuclei is actually extended by classical *fields* $\{\psi_i(\mathbf{r})\}$, i.e. functions instead of coordinates, which represent the quantum wave function. Thus, vector products or absolute values have to be generalized to scalar products and norms of the fields. In addition, the “positions” of these fields $\{\psi_i\}$ actually have a physical meaning, contrary to their momenta $\{\dot{\psi}_i\}$.

2.5 What about Hellmann–Feynman Forces ?

An important ingredient in all dynamics methods is the efficient calculation of the forces acting on the nuclei, see Eqs. (30), (32), and (44). The straightforward numerical evaluation of the derivative

$$\mathbf{F}_I = -\nabla_I \langle \Psi_0 | \mathcal{H}_e | \Psi_0 \rangle \quad (62)$$

in terms of a finite-difference approximation of the total electronic energy is both too costly and too inaccurate for dynamical simulations. What happens if the gradients are evaluated analytically? In addition to the derivative of the Hamiltonian itself

$$\begin{aligned} \nabla_I \langle \Psi_0 | \mathcal{H}_e | \Psi_0 \rangle &= \langle \Psi_0 | \nabla_I \mathcal{H}_e | \Psi_0 \rangle \\ &+ \langle \nabla_I \Psi_0 | \mathcal{H}_e | \Psi_0 \rangle + \langle \Psi_0 | \mathcal{H}_e | \nabla_I \Psi_0 \rangle \end{aligned} \quad (63)$$

there are in general also contributions from variations of the wave function $\sim \nabla_I \Psi_0$. In general means here that these contributions vanish exactly

$$\mathbf{F}_I^{\text{HFT}} = - \langle \Psi_0 | \nabla_I \mathcal{H}_e | \Psi_0 \rangle \quad (64)$$

if the wave function is an exact eigenfunction (or stationary state wave function) of the particular Hamiltonian under consideration. This is the content of the often-cited Hellmann–Feynman Theorem^{133, 76, 168}, which is also valid for many variational wave functions (e.g. the Hartree–Fock wave function) provided that *complete basis sets* are used. If this is not the case, which has to be assumed for numerical calculations, the additional terms have to be evaluated explicitly.

In order to proceed a Slater determinant $\Psi_0 = \det\{\psi_i\}$ of one-particle orbitals ψ_i , which themselves are expanded

$$\psi_i = \sum_{\nu} c_{i\nu} f_{\nu}(\mathbf{r}; \{\mathbf{R}_I\}) \quad (65)$$

in terms of a linear combination of basis functions $\{f_{\nu}\}$, is used in conjunction with an effective one-particle Hamiltonian (such as e.g. in Hartree–Fock or Kohn–Sham theories). The basis functions might depend explicitly on the nuclear positions (in the case of basis functions with origin such as atom-centered orbitals), whereas the expansion coefficients always carry an implicit dependence. This means that from the outset two sorts of forces are expected

$$\nabla_I \psi_i = \sum_{\nu} (\nabla_I c_{i\nu}) f_{\nu}(\mathbf{r}; \{\mathbf{R}_I\}) + \sum_{\nu} c_{i\nu} (\nabla_I f_{\nu}(\mathbf{r}; \{\mathbf{R}_I\})) \quad (66)$$

in addition to the Hellmann–Feynman force Eq. (64).

Using such a linear expansion Eq. (65), the force contributions stemming from the nuclear gradients of the wave function in Eq. (63) can be disentangled into two terms. The first one is called “incomplete-basis-set correction” (IBS) in solid state theory^{21, 247, 73} and corresponds to the “wave function force”²¹⁵ or “Pulay force” in quantum chemistry^{215, 216}. It contains the nuclear gradients of the basis functions

$$\mathbf{F}_I^{\text{IBS}} = - \sum_{i\nu\mu} (\langle \nabla_I f_{\nu} | \mathcal{H}_e^{\text{NSC}} - \epsilon_i | f_{\mu} \rangle + \langle f_{\nu} | \mathcal{H}_e^{\text{NSC}} - \epsilon_i | \nabla_I f_{\mu} \rangle) \quad (67)$$

and the (in practice non-self-consistent) effective one-particle Hamiltonian^{21, 247}. The second term leads to the so-called “non-self-consistency correction” (NSC) of the force^{21, 247}

$$\mathbf{F}_I^{\text{NSC}} = - \int d\mathbf{r} (\nabla_I n) (V^{\text{SCF}} - V^{\text{NSC}}) \quad (68)$$

and is governed by the difference between the self-consistent (“exact”) potential or field V^{SCF} and its non-self-consistent (or approximate) counterpart V^{NSC} associated to $\mathcal{H}_e^{\text{NSC}}$; $n(\mathbf{r})$ is the charge density. In summary, the total force needed in *ab initio* molecular dynamics simulations

$$\mathbf{F}_I = \mathbf{F}_I^{\text{HFT}} + \mathbf{F}_I^{\text{IBS}} + \mathbf{F}_I^{\text{NSC}} \quad (69)$$

comprises in general three qualitatively different terms; see the tutorial article Ref. 73 for a further discussion of core vs. valence states and the effect of pseudopotentials. Assuming that self-consistency is exactly satisfied (which is *never* going to be the case in numerical

calculations), the force $\mathbf{F}_I^{\text{NSC}}$ vanishes and $\mathcal{H}_e^{\text{SCF}}$ has to be used to evaluate $\mathbf{F}_I^{\text{IBS}}$. The Pulay contribution vanishes in the limit of using a complete basis set (which is also not possible to achieve in actual calculations).

The most obvious simplification arises if the wave function is expanded in terms of originless basis functions such as plane waves, see Eq. (100). In this case the Pulay force vanishes exactly, which applies of course to all *ab initio* molecular dynamics schemes (i.e. Ehrenfest, Born–Oppenheimer, and Car–Parrinello) using that particular basis set. This statement is true for calculations where the number of plane waves is fixed. If the number of plane waves changes, such as in (constant pressure) calculations with varying cell volume / shape where the energy cutoff is strictly fixed instead, Pulay stress contributions crop up^{89, 105, 275, 85, 82}. If basis sets with origin are used instead of plane waves Pulay forces arise always and have to be included explicitly in force calculations, see e.g. Refs. 28, 170, 171 for such methods. Another interesting simplification of the same origin is noted in passing: there is no basis set superposition error (BSSE)³⁵ in plane wave–based electronic structure calculations.

A non–obvious and more delicate term in the context of *ab initio* molecular dynamics is the one stemming from non–self–consistency Eq. (68). This term vanishes only if the wave function Ψ_0 is an eigenfunction of the Hamiltonian *within the subspace spanned by the finite basis set used*. This demands less than the Hellmann–Feynman theorem where Ψ_0 has to be an exact eigenfunction of the Hamiltonian and a complete basis set has to be used in turn. In terms of electronic structure calculations complete self–consistency (within a given incomplete basis set) has to be reached in order that $\mathbf{F}_I^{\text{NSC}}$ vanishes. Thus, in numerical calculations the NSC term can be made arbitrarily small by optimizing the effective Hamiltonian and by determining its eigenfunctions to very high accuracy, but it can never be suppressed completely.

The crucial point is, however, that in Car–Parrinello as well as in Ehrenfest molecular dynamics it is not the minimized expectation value of the electronic Hamiltonian, i.e. $\min_{\Psi_0} \{ \langle \Psi_0 | \mathcal{H}_e | \Psi_0 \rangle \}$, that yields the consistent forces. What is merely needed is to evaluate the expression $\langle \Psi_0 | \mathcal{H}_e | \Psi_0 \rangle$ with the Hamiltonian and the associated wave function available at a certain time step, compare Eq. (32) to Eq. (44) or Eq. (30). In other words, it is not required (concerning the present discussion of the contributions to the force!) that the expectation value of the electronic Hamiltonian is actually completely minimized for the nuclear configuration at that time step. Whence, full self–consistency is not required for this purpose in the case of Car–Parrinello (and Ehrenfest) molecular dynamics. As a consequence, the non–self–consistency correction to the force $\mathbf{F}_I^{\text{NSC}}$ Eq. (68) is irrelevant in Car–Parrinello (and Ehrenfest) simulations.

In Born–Oppenheimer molecular dynamics, on the other hand, the expectation value of the Hamiltonian has to be minimized for each nuclear configuration before taking the gradient to obtain the consistent force! In this scheme there is (independently from the issue of Pulay forces) *always* the non–vanishing contribution of the non–self–consistency force, which is unknown by its very definition (if it were known, the problem was solved, see Eq. (68)). It is noted in passing that there are estimation schemes available that correct *approximately* for this systematic error in Born–Oppenheimer dynamics and lead to significant time–savings, see e.g. Ref. 159.

Heuristically one could also argue that within Car–Parrinello dynamics the non–vanishing non–self–consistency force is kept under control or counterbalanced by the non–

vanishing “mass times acceleration term” $\mu\ddot{\psi}_i(t) \approx 0$, which is small but not identical to zero and oscillatory. This is sufficient to keep the propagation stable, whereas $\mu\ddot{\psi}_i(t) \equiv 0$, i.e. an extremely tight minimization $\min_{\Psi_0} \{\langle \Psi_0 | \mathcal{H}_e | \Psi_0 \rangle\}$, is required by its very definition in order to make the Born–Oppenheimer approach stable, compare again Eq. (60) to Eq. (40). Thus, also from this perspective it becomes clear that the fictitious kinetic energy of the electrons and thus their fictitious temperature is a measure for the departure from the exact Born–Oppenheimer surface during Car–Parrinello dynamics.

Finally, the present discussion shows that nowhere in these force derivations was *made use of* the Hellmann–Feynman theorem as is sometimes stated. Actually, it is known for a long time that this theorem is quite useless for numerical electronic structure calculations, see e.g. Refs. 215, 21, 216 and references therein. Rather *it turns out* that in the case of Car–Parrinello calculations using a plane wave basis the resulting relation for the force, namely Eq. (64), looks like the one obtained by simply invoking the Hellmann–Feynman theorem at the outset.

It is interesting to recall that the Hellmann–Feynman theorem as applied to a non-eigenfunction of a Hamiltonian yields only a first-order perturbative estimate of the exact force^{133, 168}. The same argument applies to *ab initio* molecular dynamics calculations where possible force corrections according to Eqs. (67) and (68) are neglected without justification. Furthermore, such simulations can of course not strictly conserve the total Hamiltonian E_{cons} Eq. (48). Finally, it should be stressed that possible contributions to the force in the nuclear equation of motion Eq. (44) due to *position-dependent* wave function *constraints* have to be evaluated following the same procedure. This leads to similar “correction terms” to the force, see e.g. Ref. 163 for such a case.

2.6 Which Method to Choose ?

Presumably the most important question for practical applications is which *ab initio* molecular dynamics method is the most efficient in terms of computer time given a specific problem. An *a priori* advantage of both the Ehrenfest and Car–Parrinello schemes over Born–Oppenheimer molecular dynamics is that no diagonalization of the Hamiltonian (or the equivalent minimization of an energy functional) is necessary, except at the very first step in order to obtain the initial wave function. The difference is, however, that the Ehrenfest time–evolution according to the time–dependent Schrödinger equation Eq. (26) conforms to a unitary propagation^{156, 167, 157}

$$\Psi(t_0 + \Delta t) = \exp[-i\mathcal{H}_e(t_0)\Delta t/\hbar] \Psi(t_0) \quad (70)$$

$$\begin{aligned} \Psi(t_0 + m \Delta t) &= \exp[-i\mathcal{H}_e(t_0 + (m-1)\Delta t) \Delta t/\hbar] \\ &\times \dots \\ &\times \exp[-i\mathcal{H}_e(t_0 + 2\Delta t) \Delta t/\hbar] \\ &\times \exp[-i\mathcal{H}_e(t_0 + \Delta t) \Delta t/\hbar] \\ &\times \exp[-i\mathcal{H}_e(t_0) \Delta t/\hbar] \Psi(t_0) \end{aligned} \quad (71)$$

$$\Psi(t_0 + t^{\text{max}}) \xrightarrow{\Delta t \rightarrow 0} \mathsf{T} \exp \left[-\frac{i}{\hbar} \int_{t_0}^{t_0 + t^{\text{max}}} dt \mathcal{H}_e(t) \right] \Psi(t_0) \quad (72)$$

for infinitesimally short times given by the time step $\Delta t = t^{\text{max}}/m$; here T is the time–ordering operator and $\mathcal{H}_e(t)$ is the Hamiltonian (which is *implicitly* time–dependent via

the positions $\{\mathbf{R}_I(t)\}$ evaluated at time t using e.g. split operator techniques⁷⁴. Thus, the wave function Ψ will conserve its norm and in particular orbitals used to expand it will stay orthonormal, see e.g. Ref. 258. In Car–Parrinello molecular dynamics, on the contrary, the orthonormality has to be imposed brute force by Lagrange multipliers, which amounts to an additional orthogonalization at each molecular dynamics step. If this is not properly done, the orbitals will become non-orthogonal and the wave function unnormalized, see e.g. Section III.C.1 in Ref. 201.

But this theoretical disadvantage of Car–Parrinello vs. Ehrenfest dynamics is in reality more than compensated by the possibility to use a much larger time step in order to propagate the electronic (and thus nuclear) degrees of freedom in the former scheme. In both approaches, there is the time scale inherent to the nuclear motion τ_n and the one stemming from the electronic dynamics τ_e . The first one can be estimated by considering the highest phonon or vibrational frequency and amounts to the order of $\tau_n \sim 10^{-14}$ s (or 0.01 ps or 10 fs, assuming a maximum frequency of about 4000 cm⁻¹). This time scale depends only on the physics of the problem under consideration and yields an upper limit for the time step $\Delta t^{\max}$ that can be used in order to integrate the equations of motion, e.g. $\Delta t^{\max} \approx \tau_n/10$.

The fastest electronic motion in Ehrenfest dynamics can be estimated within a plane wave expansion by $\omega_e^E \sim E_{\text{cut}}$, where E_{cut} is the maximum kinetic energy included in the expansion. A realistic estimate for reasonable basis sets is $\tau_e^E \sim 10^{-16}$ s, which leads to $\tau_e^E \approx \tau_n/100$. The analogous relation for Car–Parrinello dynamics reads however $\omega_e^{\text{CP}} \sim (E_{\text{cut}}/\mu)^{1/2}$ according to the analysis in Section 2.4, see Eq. (54). Thus, in addition to reducing ω_e^{CP} by introducing a finite electron mass μ , the maximum electronic frequency increases much more slowly in Car–Parrinello than in Ehrenfest molecular dynamics with increasing basis set size. An estimate for the same basis set and a typical fictitious mass yields about $\tau_e^{\text{CP}} \sim 10^{-15}$ s or $\tau_e^{\text{CP}} \approx \tau_n/10$. According to this simple estimate, the time step can be about one order of magnitude larger if Car–Parrinello second-order fictitious-time electron dynamics is used instead of Ehrenfest first-order real-time electron dynamics.

The time scale and thus time step problem inherent to Ehrenfest dynamics prompted some attempts to cure it. In Ref. 83 the equations of motion of electrons and nuclei were integrated using two different time steps, the one of the nuclei being 20-times as large as the electronic one. The powerful technology of multiple-time step integration theory^{262,264} could also be applied in order to ameliorate the time scale disparity²⁴⁴. A different approach borrowed from plasma simulations consists in decreasing the nuclear masses so that their time evolution is artificially speeded up²⁵⁸. As a result, the *nuclear* dynamics is fictitious (in the presence of real-time electron dynamics!) and has to be rescaled to the proper mass ratio after the simulation.

In both Ehrenfest and Car–Parrinello schemes the explicitly treated electron dynamics limits the largest time step that can be used in order to integrate simultaneously the coupled equations of motion for nuclei and electrons. This limitation does of course not exist in Born–Oppenheimer dynamics since there is no explicit electron dynamics so that the maximum time step is simply given by the one intrinsic to nuclear motion, i.e. $\tau_e^{\text{BO}} \approx \tau_n$. This is formally an order of magnitude advantage with respect to Car–Parrinello dynamics.

Do these back-of-the-envelope estimates have anything to do with reality? Fortunately, several state-of-the-art studies are reported in the literature for physically similar

systems where all three molecular dynamics schemes have been employed. Ehrenfest simulations^{233,83} of a dilute $K_x \cdot (KCl)_{1-x}$ melt were performed using a time step of 0.012–0.024 fs. In comparison, a time step as large as 0.4 fs could be used to produce a stable Car–Parrinello simulation of electrons in liquid ammonia^{57,58}. Since the physics of these systems has a similar nature — “unbound electrons” dissolved in liquid condensed matter (localizing as F -centers, polarons, bipolarons, etc.) — the time step difference of about a factor of ten confirms the crude estimate given above. In a Born–Oppenheimer simulation²³⁷ of again $K_x \cdot (KCl)_{1-x}$ but up to a higher concentration of unbound electrons the time step used was 0.5 fs.

The time-scale advantage of Born–Oppenheimer vs. Car–Parrinello dynamics becomes more evident if the nuclear dynamics becomes fairly slow, such as in liquid sodium¹⁵⁸ or selenium¹⁵¹ where a time step of 3 fs was used. This establishes the above-mentioned order of magnitude advantage of Born–Oppenheimer vs. Car–Parrinello dynamics in advantageous cases. However, it has to be taken into account that in simulations¹⁵¹ with such a large time step dynamical information is limited to about 10 THz, which corresponds to frequencies below roughly 500 cm^{-1} . In order to resolve vibrations in molecular systems with stiff covalent bonds the time step has to be decreased to less than a femtosecond (see the estimate given above) also in Born–Oppenheimer dynamics.

The comparison of the overall performance of Car–Parrinello and Born–Oppenheimer molecular dynamics in terms of computer time is a delicate issue. For instance it depends crucially on the choice made concerning the accuracy of the conservation of the energy E_{cons} as defined in Eq. (48). Thus, this issue is to some extent subject of “personal taste” as to what is considered to be a “sufficiently accurate” energy conservation. In addition, this comparison might lead to different conclusions as a function of system size. In order to nevertheless shed light on this point, microcanonical simulations of 8 silicon atoms were performed with various parameters using Car–Parrinello and Born–Oppenheimer molecular dynamics as implemented in the CPMD package¹³⁹. This large-gap system was initially extremely well equilibrated and the runs were extended to 8 ps (and a few to 12 ps with no noticeable difference) at a temperature of about 360–370 K (with ± 80 K root-mean-square fluctuations). The wave function was expanded up to $E_{\text{cut}} = 10 \text{ Ry}$ at the Γ -point of a simple cubic supercell and LDA was used to describe the interactions. In both cases the velocity Verlet scheme was used to integrate the equations of motion. It is noted in passing that also the velocity Verlet algorithm²⁶³ allows for stable integration of the equations of motion contrary to the statements in Ref. 222 (see Section 3.4 and Figs. 4–5).

In Car–Parrinello molecular dynamics two different time steps were used, 5 a.u. and 10 a.u. (corresponding to about 0.24 fs), in conjunction with a fictitious electron mass of $\mu = 400 \text{ a.u.}$; this mass parameter is certainly not optimized and thus the time step could be increased furthermore. Also the largest time step lead to perfect adiabaticity (similar to the one documented in Figure 3), i.e. E_{phys} Eq. (49) and T_e Eq. (51) did not show a systematic drift relative to the energy scale set by the variations of V_e Eq. (50). Within Born–Oppenheimer molecular dynamics the minimization of the energy functional was done using the highly efficient DIIS (direct inversion in the iterative subspace) scheme using 10 “history vectors”. In this case, the time step was either 10 a.u. or 100 a.u. and three convergence criteria were used; note that the large time step corresponding to 2.4 fs is already at the limit to be used to investigate typical *molecular* systems (with frequencies up to $3\text{--}4000 \text{ cm}^{-1}$). The convergence criterion is based on the largest element of the wave

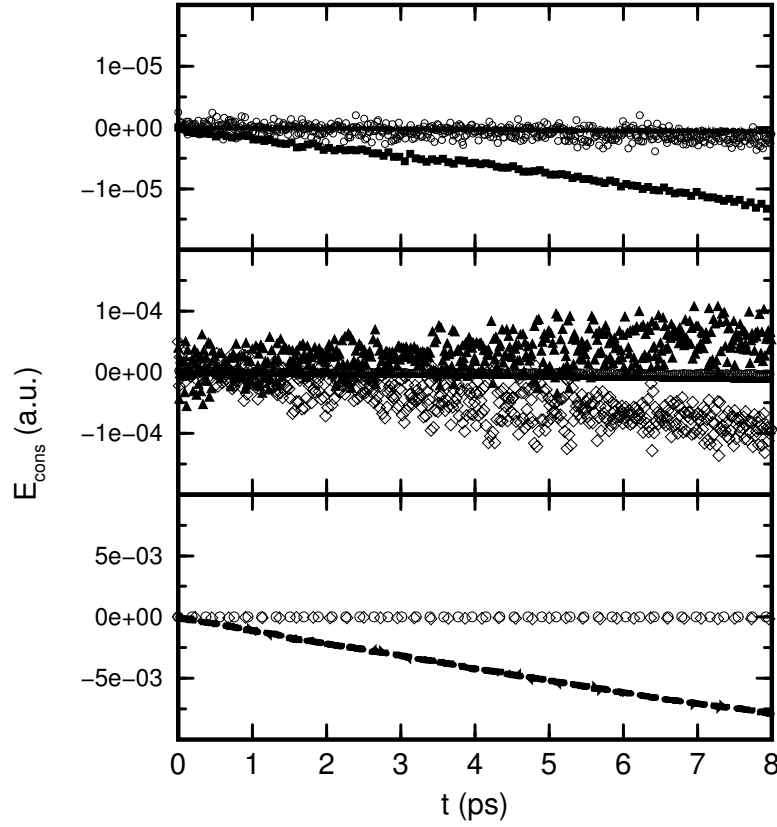


Figure 5. Conserved energy E_{cons} defined in Eq. (48) from Car–Parrinello (CP) and Born–Oppenheimer (BO) molecular dynamics simulations of a model system for various time steps and convergence criteria using the CPMD package¹³⁹; see text for further details and Table 1 for the corresponding timings. Top: solid line: CP, 5 a.u.; open circles: CP, 10 a.u.; filled squares: BO, 10 a.u., 10^{-6} . Middle: open circles: CP, 10 a.u.; filled squares: BO, 10 a.u., 10^{-6} ; filled triangles: BO, 100 a.u., 10^{-6} ; open diamonds: BO, 100 a.u., 10^{-5} . Bottom: open circles: CP, 10 a.u.; open diamonds: BO, 100 a.u., 10^{-5} ; dashed line: BO, 100 a.u., 10^{-4} .

function gradient which was required to be smaller than 10^{-6} , 10^{-5} or 10^{-4} a.u.; note that the resulting energy convergence shows roughly a quadratic dependence on this criterion.

The outcome of this comparison is shown in Figure 5 in terms of the time evolution of the conserved energy E_{cons} Eq. (48) on energy scales that cover more than three orders of magnitude in absolute accuracy. Within the present comparison ultimate energy stability was obtained using Car–Parrinello molecular dynamics with the shortest time step of 5 a.u., which conserves the energy of the total system to about 6×10^{-8} a.u. per picosecond, see solid line in Figure 5(top). Increasing the time step to 10 a.u. leads to an energy conservation of about 3×10^{-7} a.u./ps and much larger energy fluctuations, see open circles

Table 1. Timings in CPU seconds and energy conservation in a.u. / ps for Car–Parrinello (CP) and Born–Oppenheimer (BO) molecular dynamics simulations of a model system for 1 ps of trajectory on an IBM RS6000 / model 390 (Power2) workstation using the CPMD package¹³⁹; see Figure 5 for corresponding energy plots.

Method	Time step (a.u.)	Convergence (a.u.)	Conservation (a.u./ps)	Time (s)
CP	5	—	6×10^{-8}	3230
CP	7	—	1×10^{-7}	2310
CP	10	—	3×10^{-7}	1610
BO	10	10^{-6}	1×10^{-6}	16590
BO	50	10^{-6}	1×10^{-6}	4130
BO	100	10^{-6}	6×10^{-6}	2250
BO	100	10^{-5}	1×10^{-5}	1660
BO	100	10^{-4}	1×10^{-3}	1060

in Figure 5(top). The computer time needed in order to generate one picosecond of Car–Parrinello trajectory increases – to a good approximation – linearly with the increasing time step, see Table 1. The most stable Born–Oppenheimer run was performed with a time step of 10 a.u. and a convergence of 10^{-6} . This leads to an energy conservation of about 1×10^{-6} a.u./ps, see filled squares in Figure 5(top).

As the maximum time step in Born–Oppenheimer dynamics is only related to the time scale associated to nuclear motion it could be increased from 10 to 100 a.u. while keeping the convergence at the same tight limit of 10^{-6} . This worsens the energy conservation slightly (to about 6×10^{-6} a.u./ps), whereas the energy *fluctuations* increase dramatically, see filled triangles in Figure 5(middle) and note the change of scale compared to Figure 5(top). The overall gain is an acceleration of the Born–Oppenheimer simulation by a factor of about seven to eight, see Table 1. In the Born–Oppenheimer scheme, the computer time needed for a fixed amount of simulated physical time decreases only sublinearly with increasing time step since the initial guess for the iterative minimization degrades in quality as the time step is made larger. Further savings of computer time can be easily achieved by decreasing the quality of the wave function convergence from 10^{-6} to 10^{-5} and finally to 10^{-4} , see Table 1. This is unfortunately tied to a significant decrease of the energy conservation from 6×10^{-6} a.u./ps at 10^{-6} (filled triangles) to about 1×10^{-3} a.u./ps at 10^{-4} (dashed line) using the same 100 a.u. time step, see Figure 5(bottom) but note the change of scale compared to Figure 5(middle).

In conclusion, Born–Oppenheimer molecular dynamics can be made as fast as (or even faster than) Car–Parrinello molecular dynamics (as measured by the amount of CPU time spent per picosecond) at the expense of sacrificing accuracy in terms of energy conservation. In the “classical molecular dynamics community” there is a general consensus that this conservation law should be taken seriously being a measure of the numerical quality of the simulation. In the “quantum chemistry and total energy communities” this issue is typically of less concern. There, it is rather the quality of the convergence of the wave function or energy (as achieved in every individual molecular dynamics step) that is believed to be crucial in order to gauge the quality of a particular simulation.

Finally, it is worth commenting in this particular section on a paper entitled “A comparison of Car–Parrinello and Born–Oppenheimer generalized valence bond molecular dynamics”⁹⁴. In this paper one (computationally expensive) term in the nuclear equations of motion is neglected^{269,178}. It is well known that using a basis set with origin, such as

Gaussians $f_\nu^G(\mathbf{r}; \{\mathbf{R}_I\})$ centered at the nuclei, see Eq. (99), produces various Pulay forces, see Section 2.5. In particular a linear expansion Eq. (65) or Eq. (97) based on such orbitals introduces a position dependence into the orthogonality constraint

$$\langle \psi_i | \psi_j \rangle = \sum_{\nu\mu} c_{i\nu}^* c_{j\mu} \underbrace{\langle f_\nu^G | f_\mu^G \rangle}_{S_{\nu\mu}} = \delta_{ij} \quad (73)$$

that is hidden in the overlap matrix $S_{\nu\mu}(\{\mathbf{R}_I\})$ which involves the basis functions. According to Eq. (44) this term produces a constraint force of the type

$$\sum_{ij} \Lambda_{ij} \sum_{\nu\mu} c_{i\nu}^* c_{j\mu} \frac{\partial}{\partial \mathbf{R}_I} S_{\nu\mu}(\{\mathbf{R}_I\}) \quad (74)$$

in the correct Car–Parrinello equation of motion *for the nuclei* similar to the one contained in the electronic equation of motion Eq. (45). This term has to be included in order to yield exact Car–Parrinello trajectories and thus energy conservation, see e.g. Eq. (37) in Ref. 163 for a similar situation. In the case of Born–Oppenheimer molecular dynamics, on the contrary, this term is always absent in the nuclear equation of motion, see Eq. (32). Thus, the particular implementation⁹⁴ underlying the comparison between Car–Parrinello and Born–Oppenheimer molecular dynamics is an approximate one from the outset concerning the Car–Parrinello part; it can be argued that this was justified in the early papers^{125, 126} where the basic feasibility of both the Hartree–Fock– and generalized valence bond–based Car–Parrinello molecular dynamics techniques was demonstrated¹²⁹. Most importantly, this approximation implies that the energy E_{cons} Eq. (48) *cannot be rigorously conserved* in this particular version of Car–Parrinello molecular dynamics. However, energy conservation of E_{cons} was used in Ref. 94 to compare the efficiency and accuracy of these two approaches to GVB *ab initio* molecular dynamics (using DIIS for the Born–Oppenheimer simulations as done in the above–given comparison). Thus, the final conclusion that for “... approaches that utilize non–space–fixed bases to describe the electronic wave function, Born–Oppenheimer AIMD is the method of choice, both in terms of accuracy and speed”⁹⁴ cannot be drawn from this specific comparison for the reasons outlined above (independently of the particular basis set or electronic structure method used).

The toy system investigated here (see Figure 5 and Table 1), i.e. 8 silicon atoms in a periodic supercell, is for the purpose of comparing different approaches to *ab initio* molecular dynamics quite similar to the system used in Ref. 94, i.e. clusters of 4 or 6 sodium atoms (in addition, qualitatively identical results were reported in Section 4 for silicon clusters). Thus, it is admissible to compare the energy conservations reported in Figs. 1 and 2 of Ref. 94 to the ones depicted here in Figure 5 noting that the longest simulations reported in Ref. 94 reached only 1 ps. It should be stressed that the energy conservation seen in Figure 5(top) is routinely achieved in Car–Parrinello molecular dynamics simulations.

2.7 Electronic Structure Methods

2.7.1 Introduction

Up to this point, the electronic structure method to calculate the *ab initio* forces $\nabla_I \langle \Psi | \mathcal{H}_e | \Psi \rangle$ was not specified in detail. It is immediately clear that *ab initio* molecular dynamics is not tied to any particular approach, although very accurate techniques

are of course prohibitively expensive. It is also evident that the strength or weakness of a particular *ab initio* molecular dynamics scheme is intimately connected to the strength or weakness of the chosen electronic structure method. Over the years a variety of different approaches such as density functional^{38,285,13,201,158,14}, Hartree–Fock^{166,109,79,175,125,128,144,132}, generalized valence bond (GVB)^{126,127,93–95}, complete active space SCF (CASSCF)^{235,236}, full configuration interaction (FCI)¹⁷², semiempirical^{281,282,37,78,41,280,124} or other approximate^{202,191,232,192,67,68,8} methods were combined with molecular dynamics, and this list is certainly incomplete.

The focus of the present review clearly is Car–Parrinello molecular dynamics in conjunction with Hohenberg–Kohn–Sham density functional theory^{138,153}. In the following, only those parts of density functional theory are presented that impact directly on *ab initio* molecular dynamics. For a deeper presentation and in particular for a discussion of the assumptions and limitations of this approach (both conceptually and in practice) the reader is referred to the existing excellent literature^{247,147,194,66,176}. For simplicity, the formulae are presented for the spin–unpolarized or restricted special case.

Following the exposition of density functional theory, the fundamentals of Hartree–Fock theory, which is often considered to be the basis of quantum chemistry, are introduced for the same special case. Finally, a glimpse is given at post Hartree–Fock methods. Again, an extensive text–book literature exists for these wave function–based approaches to electronic structure calculations^{252,183}. The very useful connection between the density–based and wave function–based methods goes back to Löwdin’s work in the mid fifties and is e.g. worked out in Chapt. 2.5 of Ref. 194, where Hartree–Fock theory is formulated in density–matrix language.

2.7.2 Density Functional Theory

The total ground–state energy of the interacting system of electrons with classical nuclei fixed at positions $\{\mathbf{R}_I\}$ can be obtained

$$\min_{\Psi_0} \{\langle \Psi_0 | \mathcal{H}_e | \Psi_0 \rangle\} = \min_{\{\phi_i\}} E^{\text{KS}}[\{\phi_i\}]$$

as the minimum of the Kohn–Sham energy^{138,153}

$$E^{\text{KS}}[\{\phi_i\}] = T_s[\{\phi_i\}] + \int d\mathbf{r} V_{\text{ext}}(\mathbf{r}) n(\mathbf{r}) + \frac{1}{2} \int d\mathbf{r} V_{\text{H}}(\mathbf{r}) n(\mathbf{r}) + E_{\text{xc}}[n], \quad (75)$$

which is an explicit functional of the set of auxiliary functions $\{\phi_i(\mathbf{r})\}$ that satisfy the orthonormality relation $\langle \phi_i | \phi_j \rangle = \delta_{ij}$. This is a dramatic simplification since the minimization with respect to all possible *many–body* wave functions $\{\Psi\}$ is replaced by a minimization with respect to a set of orthonormal one–particle functions, the Kohn–Sham orbitals $\{\phi_i\}$. The associated electronic one–body density or charge density

$$n(\mathbf{r}) = \sum_i^{\text{occ}} f_i |\phi_i(\mathbf{r})|^2 \quad (76)$$

is obtained from a single Slater determinant built from the occupied orbitals, where $\{f_i\}$ are integer occupation numbers.

The first term in the Kohn–Sham functional Eq. (75) is the kinetic energy of a non-interacting reference system

$$T_s[\{\phi_i\}] = \sum_i^{\text{occ}} f_i \left\langle \phi_i \left| -\frac{1}{2} \nabla^2 \right| \phi_i \right\rangle \quad (77)$$

consisting of the same number of electrons exposed to the same external potential as in the fully interacting system. The second term comes from the fixed external potential

$$V_{\text{ext}}(\mathbf{r}) = - \sum_I \frac{Z_I}{|\mathbf{R}_I - \mathbf{r}|} + \sum_{I < J} \frac{Z_I Z_J}{|\mathbf{R}_I - \mathbf{R}_J|} \quad (78)$$

in which the electrons move, which comprises the Coulomb interactions between electrons and nuclei and in the definition used here also the internuclear Coulomb interactions; this term changes in the first place if core electrons are replaced by pseudopotentials, see for instance Ref. 179 for a detailed discussion. The third term is the Hartree energy, i.e. the classical electrostatic energy of two charge clouds which stem from the electronic density and is obtained from the Hartree potential

$$V_H(\mathbf{r}) = \int d\mathbf{r}' \frac{n(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} , \quad (79)$$

which in turn is related to the density via

$$\nabla^2 V_H(\mathbf{r}) = -4\pi n(\mathbf{r}) \quad (80)$$

Poisson's equation. The last contribution in the Kohn–Sham functional, the exchange–correlation functional $E_{\text{xc}}[n]$, is the most intricate contribution to the total electronic energy. The electronic exchange and correlation effects are lumped together and basically define this functional as the remainder between the exact energy and its Kohn–Sham decomposition in terms of the three previous contributions.

The minimum of the Kohn–Sham functional is obtained by varying the energy functional Eq. (75) for a fixed number of electrons with respect to the density Eq. (76) or with respect to the orbitals subject to the orthonormality constraint, see e.g. the discussion following Eq. (35) for a similar variational procedure. This leads to the Kohn–Sham equations

$$\left\{ -\frac{1}{2} \nabla^2 + V_{\text{ext}}(\mathbf{r}) + V_H(\mathbf{r}) + \frac{\delta E_{\text{xc}}[n]}{\delta n(\mathbf{r})} \right\} \phi_i(\mathbf{r}) = \sum_j \Lambda_{ij} \phi_j(\mathbf{r}) \quad (81)$$

$$\left\{ -\frac{1}{2} \nabla^2 + V^{\text{KS}}(\mathbf{r}) \right\} \phi_i(\mathbf{r}) = \sum_j \Lambda_{ij} \phi_j(\mathbf{r}) \quad (82)$$

$$H_e^{\text{KS}} \phi_i(\mathbf{r}) = \sum_j \Lambda_{ij} \phi_j(\mathbf{r}) , \quad (83)$$

which are one–electron equations involving an effective *one–particle* Hamiltonian H_e^{KS} with the local potential V^{KS} . Note that H_e^{KS} nevertheless embodies the electronic *many–body* effects by virtue of the exchange–correlation potential

$$\frac{\delta E_{\text{xc}}[n]}{\delta n(\mathbf{r})} = V_{\text{xc}}(\mathbf{r}) . \quad (84)$$

A unitary transformation within the space of the occupied orbitals leads to the canonical form

$$H_e^{\text{KS}} \phi_i = \epsilon_i \phi_i \quad (85)$$

of the Kohn–Sham equations, where $\{\epsilon_i\}$ are the eigenvalues. In conventional static density functional or “band structure” calculations this set of equations has to be solved self-consistently in order to yield the density, the orbitals and the Kohn–Sham potential for the electronic ground state²¹². The corresponding total energy Eq. (75) can be written as

$$E^{\text{KS}} = \sum_i \epsilon_i - \frac{1}{2} \int d\mathbf{r} V_H(\mathbf{r}) n(\mathbf{r}) + E_{\text{xc}}[n] - \int d\mathbf{r} \frac{\delta E_{\text{xc}}[n]}{\delta n(\mathbf{r})} n(\mathbf{r}) , \quad (86)$$

where the sum over Kohn–Sham eigenvalues is the so-called “band structure energy”.

Thus, Eqs. (81)–(83) together with Eqs. (39)–(40) define Born–Oppenheimer molecular dynamics within Kohn–Sham density functional theory, see e.g. Refs. 96, 257, 250, 13, 285, 201, 14, 158, 159 for such implementations. The functional derivative of the Kohn–Sham functional with respect to the orbitals, the Kohn–Sham force acting on the orbitals, can be expressed as

$$\frac{\delta E^{\text{KS}}}{\delta \phi_i^*} = f_i H_e^{\text{KS}} \phi_i , \quad (87)$$

which makes clear the connection to Car–Parrinello molecular dynamics, see Eq. (45). Thus, Eqs. (59)–(60) have to be solved with the effective one-particle Hamiltonian in the Kohn–Sham formulation Eqs. (81)–(83). In the case of Ehrenfest dynamics presented in Section 2.2, which will not be discussed in further detail at this stage, the Runge–Gross time-dependent generalization of density functional theory¹¹⁰ has to be invoked instead, see e.g. Refs. 83, 258, 223.

Crucial to any application of density functional theory is the approximation of the unknown exchange and correlation functional. A discussion focussed on the utilization of suitable functionals in the framework of *ab initio* molecular dynamics is for instance given in Ref. 245. Those exchange–correlation functionals belong to the class of the “Generalized Gradient Approximation”

$$E_{\text{xc}}^{\text{GGA}}[n] = \int d\mathbf{r} n(\mathbf{r}) \varepsilon_{\text{xc}}^{\text{GGA}}(n(\mathbf{r}); \nabla n(\mathbf{r})) , \quad (88)$$

where the unknown functional is approximated by an integral over a function that depends only on the density and its gradient at a given point in space, see Ref. 206 and references therein. The combined exchange–correlation function is typically split up into two additive terms ε_x and ε_c for exchange and correlation, respectively. In the simplest case it is the exchange and correlation energy density $\varepsilon_{\text{xc}}^{\text{LDA}}(n)$ of an interacting but homogeneous electron gas at the density given by the “local” density $n(\mathbf{r})$ at space-point $\mathbf{r}$ in the inhomogeneous system. This simple but astonishingly powerful approximation¹⁴⁷ is the famous local density approximation LDA¹⁵³ (or local spin density LSD in the spin-polarized case¹⁶), and a host of different parameterizations exist in the literature^{194, 66, 176}. The self-interaction correction²⁰³ SIC as applied to LDA was critically assessed for molecules in Ref. 98 with a disappointing outcome.

A significant improvement of the accuracy was achieved by introducing the gradient of the density as indicated in Eq. (88) beyond the well-known straightforward gradient

expansions. These so-called GGAs (also denoted as “gradient corrected” or “semilocal” functionals) extended the applicability of density functional calculation to the realm of chemistry, see e.g. Refs. 204, 18, 164, 206, 205, 207 for a few “popular functionals” and Refs. 146, 71, 238, 148, 49, 248, 249 for extensive tests on molecules, hydrogen-bonded complexes, solids, and surfaces.

and Refs. 146, 71, 238, 148 for extensive tests on molecules, complexes, and solids, respectively.

Another considerable advance was the successful introduction of “hybrid functionals”^{19,20} that include to some extent “exact exchange”¹⁰³ in addition to a standard GGA. Although such functionals can certainly be implemented within a plane wave approach^{113,43,179}, they are prohibitively time-consuming as explained in Ref. 179. A more promising route in this respect are those functionals that include higher-order powers of the gradient (or the local kinetic energy density) in the sense of a generalized gradient expansion beyond the first term. Promising results could be achieved by including Laplacian or local kinetic energy terms^{214,80,81,279,255,31}, but at this stage a sound judgment concerning their “prize / performance ratio” has to await further scrutinizing tests. The “optimized potential method” (OPM) or “optimized effective potentials” (OEP) are another route to include “exact exchange” within density functional theory, see e.g. Section 13.6 in Ref. 245 or Ref. 107 for overviews. Here, the exchange–correlation functional $E_{xc}^{OPM} = E_{xc}[\{\phi_i\}]$ depends on the individual orbitals instead of only on the density or its derivatives.

2.7.3 Hartree–Fock Theory

Hartree–Fock theory is derived by invoking the variational principle in a restricted space of wave functions. The antisymmetric ground-state electronic wave function is approximated by a single Slater determinant $\Psi_0 = \det\{\psi_i\}$ which is constructed from a set of one-particle spin orbitals $\{\psi_i\}$ required to be mutually orthonormal $\langle\psi_i|\psi_j\rangle = \delta_{ij}$. The corresponding variational minimum of the total electronic energy $\mathcal{H}_e$ defined in Eq. (2)

$$\begin{aligned} E^{\text{HF}}[\{\psi_i\}] = & \sum_i \int d\mathbf{r} \psi_i^*(\mathbf{r}) \left[-\frac{1}{2}\nabla^2 + V_{\text{ext}}(\mathbf{r}) \right] \psi_i(\mathbf{r}) \\ & + \frac{1}{2} \sum_{ij} \int \int d\mathbf{r} d\mathbf{r}' \psi_i^*(\mathbf{r}) \psi_j^*(\mathbf{r}') \frac{1}{|\mathbf{r} - \mathbf{r}'|} \psi_i(\mathbf{r}) \psi_j(\mathbf{r}') \\ & + \frac{1}{2} \sum_{ij} \int \int d\mathbf{r} d\mathbf{r}' \psi_i^*(\mathbf{r}) \psi_j^*(\mathbf{r}') \frac{1}{|\mathbf{r} - \mathbf{r}'|} \psi_j(\mathbf{r}) \psi_i(\mathbf{r}') \end{aligned} \quad (89)$$

yields the lowest energy and the “best” wave function within a one-determinant ansatz; the external Coulomb potential V_{ext} was already defined in Eq. (78). Carrying out the constraint minimization within this ansatz (see Eq. (36) in Section 2.3 for a sketch) leads

to

$$\left\{ -\frac{1}{2}\nabla^2 + V_{\text{ext}}(\mathbf{r}) + \sum_j \mathcal{J}_j(\mathbf{r}) - \sum_j \mathcal{K}_j(\mathbf{r}) \right\} \psi_i(\mathbf{r}) = \sum_j \Lambda_{ij} \psi_j(\mathbf{r}) \quad (90)$$

$$\left\{ -\frac{1}{2}\nabla^2 + V^{\text{HF}}(\mathbf{r}) \right\} \psi_i(\mathbf{r}) = \sum_j \Lambda_{ij} \psi_j(\mathbf{r}) \quad (91)$$

$$H_e^{\text{HF}} \psi_i(\mathbf{r}) = \sum_j \Lambda_{ij} \psi_j(\mathbf{r}) \quad (92)$$

the Hartree–Fock integro–differential equations. In analogy to the Kohn–Sham equations Eqs. (81)–(83) these are effective one–particle equations that involve an effective one–particle Hamiltonian H_e^{HF} , the (Hartree–) Fock operator. The set of canonical orbitals

$$H_e^{\text{HF}} \psi_i = \epsilon_i \psi_i \quad (93)$$

is obtained similarly to Eq. (85). The Coulomb operator

$$\mathcal{J}_j(\mathbf{r}) \psi_i(\mathbf{r}) = \left[\int d\mathbf{r}' \psi_j^*(\mathbf{r}') \frac{1}{|\mathbf{r} - \mathbf{r}'|} \psi_j(\mathbf{r}') \right] \psi_i(\mathbf{r}) \quad (94)$$

and the exchange operator

$$\mathcal{K}_j(\mathbf{r}) \psi_i(\mathbf{r}) = \left[\int d\mathbf{r}' \psi_j^*(\mathbf{r}') \frac{1}{|\mathbf{r} - \mathbf{r}'|} \psi_i(\mathbf{r}') \right] \psi_j(\mathbf{r}) \quad (95)$$

are most easily defined via their action on a particular orbital ψ_i . It is found that upon acting on orbital $\psi_i(\mathbf{r})$ the exchange operator for the j –th state “exchanges” $\psi_j(\mathbf{r}') \rightarrow \psi_i(\mathbf{r}')$ in the kernel as well as replaces $\psi_i(\mathbf{r}) \rightarrow \psi_j(\mathbf{r})$ in its argument, compare to the Coulomb operator. Thus, $\mathcal{K}$ is a non–local operator as its action on a function ψ_i at point $\mathbf{r}$ in space requires the evaluation and thus the knowledge of that function throughout all space by virtue of $\int d\mathbf{r}' \psi_i(\mathbf{r}')$... the required integration. In this sense the exchange operator does not possess a simple classical interpretation like the Coulomb operator $\mathcal{C}$, which is the counterpart of the Hartree potential V_H in Kohn–Sham theory. The exchange operator vanishes exactly if the antisymmetrization requirement of the wave function is relaxed, i.e. only the Coulomb contribution survives if a Hartree product is used to represent the wave function.

The force acting on the orbitals is defined

$$\frac{\delta E^{\text{HF}}}{\delta \psi_i^*} = H_e^{\text{HF}} \psi_i \quad (96)$$

similarly to Eq. (87). At this stage, the various *ab initio* molecular dynamics schemes based on Hartree–Fock theory are defined, see Eqs. (39)–(40) for Born–Oppenheimer molecular dynamics and Eqs. (59)–(60) for Car–Parrinello molecular dynamics. In the case of Ehrenfest molecular dynamics the time–dependent Hartree–Fock formalism⁶² has to be invoked instead.

2.7.4 Post Hartree–Fock Theories

Although post Hartree–Fock methods have a very unfavorable scaling of the computational cost as the number of electrons increases, a few case studies were performed with such correlated quantum chemistry techniques. For instance *ab initio* molecular dynamics was combined with GVB^{126,127,93–95}, CASSCF^{235,236}, MP2²³⁰, as well as FCI¹⁷² approaches, see also references therein. It is noted in passing that Car–Parrinello molecular dynamics can only be implemented straightforwardly if energy and wave function are “consistent”. This is not the case in perturbation theories such as e.g. the widely used Møller–Plesset approach¹³¹: within standard MP2 the energy is correct to second order, whereas the wave function is the one given by the uncorrelated HF reference. As a result, the derivative of the MP2 energy with respect to the wave function Eq. (96) does not yield the correct force on the HF wave function in the sense of fictitious dynamics. Such problems are of course absent from the Born–Oppenheimer approach to sample configuration space, see e.g. Ref. 150, 145, 11 for MP2, density functional, and multireference CI *ab initio* Monte Carlo schemes or Ref. 230 for a recent Born–Oppenheimer molecular dynamics simulation using MP2 energies and forces.

It should be kept in mind that the rapidly growing workload of post HF calculations, although extremely powerful in principle, limits the number of explicitly treated electrons to only a few. The rapid development of correlated electronic structure methods that scale linearly with the number of electrons will certainly broaden the range of applicability of this class of techniques in the near future.

2.8 Basis Sets

2.8.1 Gaussians and Slater Functions

Having selected a specific electronic structure method the next choice is related to which basis set to use in order to represent the orbitals ψ_i in terms of simple analytic functions f_ν with well-known properties. In general a *linear* combination of such basis functions

$$\psi_i(\mathbf{r}) = \sum_{\nu} c_{i\nu} f_{\nu}(\mathbf{r}; \{\mathbf{R}_I\}) \quad (97)$$

is used, which represents exactly any reasonable function in the limit of using a complete set of basis functions. In quantum chemistry, Slater–type basis functions (STOs)

$$f_{\mathbf{m}}^S(\mathbf{r}) = N_{\mathbf{m}}^S r_x^{m_x} r_y^{m_y} r_z^{m_z} \exp[-\zeta_{\mathbf{m}}|\mathbf{r}|] \quad (98)$$

with an exponentially decaying radial part and Gaussian–type basis functions (GTOs)

$$f_{\mathbf{m}}^G(\mathbf{r}) = N_{\mathbf{m}}^G r_x^{m_x} r_y^{m_y} r_z^{m_z} \exp[-\alpha_{\mathbf{m}}r^2] \quad (99)$$

have received widespread use, see e.g. Ref. 131 for a concise overview–type presentation. Here, $N_{\mathbf{m}}$, $\zeta_{\mathbf{m}}$ and $\alpha_{\mathbf{m}}$ are constants that are typically kept fixed during a molecular electronic structure calculation so that only the orbital expansion coefficients $c_{i\nu}$ need to be optimized. In addition, fixed linear combinations of the above–given “primitive” basis functions can be used for a given angular momentum channel $\mathbf{m}$, which defines the “contracted” basis sets.

The Slater or Gaussian basis functions are in general centered at the positions of the nuclei, i.e. $\mathbf{r} \rightarrow \mathbf{r} - \mathbf{R}_I$ in Eq. (98)–(99), which leads to the linear combination of *atomic*

orbitals (LCAO) ansatz to solve differential equations algebraically. Furthermore, their derivatives as well as the resulting matrix elements are efficiently obtained by differentiation and integration in real-space. However, Pulay forces (see Section 2.5) will result for such basis functions that are fixed at atoms (or bonds) if the atoms are allowed to move, either in geometry optimization or molecular dynamics schemes. This disadvantage can be circumvented by using *freely* floating Gaussians that are distributed in space²⁴¹, which form an originless basis set since it is localized but not atom-fixed.

A first generation of methods using Gaussian basis functions in the context of *ab initio* molecular dynamics has been proposed roughly in the early to mid nineties in the sense of interfacing existing electronic structure codes with a driver for molecular dynamics, see for instance Refs. 109, 79, 175, 125–128, 93, 144, 132, 172, 94, 95, 235, 236. More recently, a second generation of such approaches that do more explicitly take advantage of the dynamical evolution of the electronic degrees of freedom, which is conceptually at the very root of the efficiency of the original Car–Parrinello algorithm, is developed in the framework of Car–Parrinello^{229, 230, 221, 136}, Born–Oppenheimer^{137, 277} and Ehrenfest¹⁶⁹ dynamics schemes.

2.8.2 Plane Waves

A vastly different approach has its roots in solid-state theory. Here, the ubiquitous periodicity of the underlying lattice produces a periodic potential and thus imposes the same periodicity on the density (implying Bloch’s Theorem, Born–von Karman periodic boundary conditions etc., see e.g. Chapt. 8 in Ref. 9). This heavily suggests to use plane waves as the generic basis set in order to expand the periodic part of the orbitals. Plane waves are defined as

$$f_{\mathbf{G}}^{\text{PW}}(\mathbf{r}) = N \exp[i\mathbf{G}\mathbf{r}] \quad , \quad (100)$$

where the normalization is simply given by $N = 1/\sqrt{\Omega}$; Ω is the volume of the periodic (super-) cell. Since plane waves form a complete and orthonormal set of functions they can be used to expand orbitals according to Eq. (97), where the labeling ν is simply given by the vector $\mathbf{G}$ in reciprocal space / G -space (including only those $\mathbf{G}$ -vectors that satisfy the particular periodic boundary conditions). The total electronic energy is found to have a particularly simple form when expressed in plane waves¹⁴³.

It is important to observe that plane waves are originless functions, i.e. they do *not* depend on the positions of the nuclei $\{\mathbf{R}_I\}$. This implies that the Pulay forces Eq. (67) vanish exactly even within a *finite* basis (and using a fixed number of plane waves, see the discussion related to “Pulay stress” in Section 2.5), which tremendously facilitates force calculations. This also implies that plane waves are a very unbiased basis set in that they are “delocalized” in space and do not “favor” certain atoms or regions over others, i.e. they can be considered as an ultimately “balanced basis set” in the language of quantum chemistry. Thus, the only way to improve the quality of the basis is to increase the “energy cutoff” E_{cut} , i.e. to increase the largest $|\mathbf{G}|$ -vector that is included in the finite expansion Eq. (97). This blind approach is vastly different from the traditional procedures in quantum chemistry that are needed in order to produce reliable basis sets¹³¹. Another appealing feature is that derivatives in real-space are simply multiplications in G -space, and both spaces can be efficiently connected via Fast Fourier Transforms (FFTs). Thus, one can

easily evaluate operators in that space in which they are diagonal, see for instance the flow charts in Ref. 179.

According to the well-known “No Free Lunch Theorem” there cannot be only advantages connected to using plane waves. The first point is that the pseudopotential approximation is intimately connected to using plane waves, why so? A plane wave basis is basically a lattice-symmetry-adapted three-dimensional Fourier decomposition of the orbitals. This means that increasingly large Fourier components are needed in order to resolve structures in real space on decreasingly small distance scales. But already orbitals of first row atoms feature quite strong and rapid oscillations close to the nuclei due to the Pauli principle, which enforces a nodal structure onto the wave function by imposing orthogonality of the orbitals. However, most of chemistry is ruled by the valence electrons, whereas the core electrons are essentially inert. In practice, this means that the innermost electrons can be taken out of explicit calculations. Instead they are represented by a smooth and nodeless effective potential, the so-called pseudopotential^{134, 135, 209, 210, 50}, see for instance Refs. 212, 239, 90 for reviews in the context of “solid state theory” and Refs. 53, 64 for pseudopotentials as used in “quantum chemistry”. The resulting pseudo wave function is made as smooth as possible close to the nuclear core region. This also means that properties that depend crucially on the wave function close to the core cannot be obtained straightforwardly from such calculations. In the field of plane wave calculations the introduction of “soft” norm-conserving *ab initio* pseudopotentials was a breakthrough both conceptually¹¹⁹ and in practice¹⁰. Another important contribution, especially for transition metals, was the introduction of the so-called ultrasoft pseudopotentials by Vanderbilt²⁷⁶. This approach leads to the powerful technique of plane wave pseudopotential electronic structure calculations in the framework of density functional theory^{143, 212}. Within this particular framework the issue of pseudopotentials is elaborated in more detail in Ref. 179.

Another severe shortcoming of plane waves is the backside of the medal of being an unbiased basis set: there is no way to shuffle more basis functions into regions in space where they are more needed than in other regions. This is particularly bad for systems with strong inhomogeneities. Such examples are all-electron calculations or the inclusion of semi-core states, a few heavy atoms in a sea of light atoms, and (semi-) finite systems such as surfaces or molecules with a large vacuum region in order to allow the long-range Coulomb interactions to decay. This is often referred to as the multiple length scale deficiency of plane wave calculations.

2.8.3 Generalized Plane Waves

An extremely appealing and elegant generalization of the plane wave concept^{114, 115} consists in defining them in curved ξ -space

$$f_{\mathbf{G}}^{\text{GPW}}(\xi) = N \det^{1/2} J \exp[i\mathbf{G} \cdot \mathbf{r}(\xi)] \quad (101)$$

$$\det J = \left| \frac{\partial \mathbf{r}^i}{\partial \xi^j} \right| ,$$

where $\det J$ is the Jacobian of the transformation from Cartesian to curvilinear coordinates $\mathbf{r} \rightarrow \xi(\mathbf{r})$ with $\xi = (\xi^1, \xi^2, \xi^3)$ and $N = 1/\sqrt{\Omega}$ as for regular plane waves. These functions are orthonormal, form a complete basis set, can be used for $\mathbf{k}$ -point sampling after replacing $\mathbf{G}$ by $\mathbf{G} + \mathbf{k}$ in Eq. (101), are originless (but nevertheless localized) so that

Pulay forces are absent, can be manipulated via efficient FFT techniques, and reduce to standard plane waves in the special case of an Euclidean space $\xi(\mathbf{r}) = \mathbf{r}$. Thus, they can be used equally well like plane waves in linear expansions of the sort Eq. (65) underlying most of electronic structure calculations. The Jacobian of the transformation is related to the Riemannian metric tensor

$$g_{ij} = \sum_{k=1}^3 \frac{\partial \xi^k}{\partial r^i} \frac{\partial \xi^k}{\partial r^j}$$

$$\det J = \det^{-1/2} \{g_{ij}\} \quad (102)$$

which defines the metric of the ξ -space. The metric and thus the curvilinear coordinate system itself is considered as a variational parameter in the original fully adaptive-coordinate approach^{114,115}, see also Refs. 60, 120–123. Thus, a uniform grid in curved Riemannian space is non-uniform or distorted when viewed in flat Euclidean space (where $g_{ij} = \delta_{ij}$) such that the density of grid points (or the “local” cutoff energy of the expansion in terms of G -vectors) is highest in regions close to the nuclei and lowest in vacuum regions, see Figure 2 in Ref. 120.

Concerning actual calculations, this means that a lower number of generalized plane waves than standard plane waves are needed in order to achieve a given accuracy¹¹⁴, see Figure 1 in Ref. 120. This allows even for all-electron approaches to electronic structure calculations where plane waves fail^{186,217}. More recently, the distortion of the metric was frozen spherically around atoms by introducing deformation functions^{116,117}, which leads to a concept closely connected to non-uniform atom-centered meshes in real-space methods¹⁸⁶, see below. In such non-fully-adaptive approaches using *predefined* coordinate transformations attention has to be given to Pulay force contributions which have to be evaluated explicitly^{116,186}.

2.8.4 Wavelets

Similar to using generalized plane waves is the idea to exploit the powerful multiscale-properties of wavelets. Since this approach requires an extensive introductory discussion (see e.g. Ref. 100 for a gentle introduction) and since it seems still quite far from being used in large-scale electronic structure calculations the interested reader is referred to original papers^{46,283,290,273,99} and review articles^{7,102}. Wavelet-based methods allow intrinsically to exploit multiple length scales without introducing Pulay forces and can be efficiently handled by fast wavelet transforms. In addition, they are also a powerful route to linear scaling or “order- N ” methods^{190,101} as first demonstrated in Ref. 99 with the calculation of the Hartree potential for an all-electron uranium dimer.

2.8.5 Mixed and Augmented Basis Sets

Localized Gaussian basis functions on the one hand and plane waves on the other hand are certainly two extreme cases. There has been a tremendous effort to combine such localized and originless basis functions in order to exploit their mutual strengths. This resulted in a rich collection of mixed and augmented basis sets with very specific implementation requirements. This topic will not be covered here and the interested reader is referred to Refs. 28, 274, 219, 170, 171, 277 and references given therein for some recent implementations used in conjunction with *ab initio* molecular dynamics.

2.8.6 Wannier Functions

An alternative to the plane wave basis set in the framework of periodic calculations in solid-state theory are Wannier functions, see for instance Section 10 in Ref. 9. These functions are formally obtained from a unitary transformation of the Bloch orbitals and have the advantage that they can be exponentially localized under certain circumstances. The so-called maximally localized generalized Wannier functions¹⁸² are the periodic analogues of Boys' localized orbitals defined for isolated systems. Recently the usefulness of Wannier functions for numerical purposes was advocated by several groups, see Refs. 154, 75, 182, 2, 165 and references given therein.

2.8.7 Real Space Grids

A quite different approach is to leave conventional basis set approaches altogether and to resort to real-space methods¹⁷ where continuous space is replaced by a discrete space $\mathbf{r} \rightarrow r_{\mathbf{p}}$. This entails that the derivative operator or the entire energy expression has to be discretized in some way. The high-order central-finite difference approach leads to the expression

$$-\frac{1}{2}\nabla^2\psi_i(\mathbf{r}) \stackrel{h \rightarrow 0}{=} -\frac{1}{2} \left[\sum_{n_x=-N}^N C_{n_x} \psi_i(r_{p_x} + n_x h, r_{p_y}, r_{p_z}) \right. \\ \left. + \sum_{n_y=-N}^N C_{n_y} \psi_i(r_{p_x}, r_{p_y} + n_y h, r_{p_z}) \right. \\ \left. + \sum_{n_z=-N}^N C_{n_z} \psi_i(r_{p_x}, r_{p_y}, r_{p_z} + n_z h) \right] + \mathcal{O}(h^{2N+2}) \quad (103)$$

for the Laplacian which is correct up to the order h^{2N+2} . Here, h is the uniform grid spacing and $\{C_{\mathbf{n}}\}$ are known expansion coefficients that depend on the selected order⁴⁴. Within this scheme, not only the grid spacing h but also the order are disposable parameters that can be optimized for a particular calculation. Note that the discretization points in continuous space can also be considered to constitute a sort of “finite basis set” – despite different statements in the literature – and that the “infinite basis set limit” is reached as $h \rightarrow 0$ for N fixed. A variation on the theme are Mehrstellen schemes where the discretization of the entire differential equation and not only of the derivative operator is optimized³⁶.

The first real-space approach devised for *ab initio* molecular dynamics was based on the lowest-order finite-difference approximation in conjunction with a equally-spaced cubic mesh in real space³⁹. A variety of other implementations of more sophisticated real-space methods followed and include e.g. non-uniform meshes, multigrid acceleration, different discretization techniques, and finite-element methods^{286, 23, 15, 44, 45, 259, 260, 186, 261, 3, 231}, see Ref. 17 for a review. Among the chief advantages of the real-space methods is that linear scaling approaches^{190, 101} can be implemented in a natural way and that the multiple-length scale problem can be coped with by adapting the grid. However, the extension to such non-uniform meshes induces the (in)famous Pulay forces (see Section 2.5) if the mesh moves as the nuclei move.

3 Summary and Outlook

The essentials of what is now called standard *ab initio* molecular dynamics have been reviewed in this Keynote Lecture whereas more advanced approaches are covered in other Lectures at the *NIC Winter School 2006* “Computational Nanoscience: Do it Yourself!”. The notion “standard” implies here the adiabatic propagation in the Born–Oppenheimer ground–state, the usage of classical nuclei and the generation of trajectories in the micro-canonical ensemble. This is a mature computer simulation technique by now and a growing number of powerful computer codes able to perform *ab initio* molecular dynamics simulations in this spirit is available such as for instance the ABINIT¹, CASTEP⁴², CPMD¹³⁹ CP-PAW⁵², CP2K⁵¹, Dacapo⁵⁴, fhi98md⁷⁷, NWChem¹⁸⁹, PINY²¹³, PWscf/Quantum-ESPRESSO²¹⁸, S/PHI/nX²⁴⁰, or VASP²⁷⁸ packages among others. It is not difficult to predict that the expanding family of *ab initio* molecular dynamics techniques, which might differ in the underlying electronic structure theory, the propagation scheme, the basis set or other technical aspects, will have an ever increasing impact on the simulation of “chemically complex” molecular systems.

References

1. ABINIT: Ref. 106; distributed under the terms of the GNU General Public License; see <http://www.abinit.org/>.
2. M. Albrecht, A. Shukla, M. Dolg, P. Fulde, and H. Stoll, Chem. Phys. Lett. **285**, 174 (1998).
3. M. M. G. Alemany, M. Jain, L. Kronik, and J. R. Chelikowsky, Phys. Rev. B **69**, 075101 (2004).
4. M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids* (Clarendon Press, Oxford, 1987; reprinted 1990).
5. H. C. Andersen, J. Chem. Phys. **72**, 2384 (1980).
6. W. Andreoni, D. Marx, and M. Sprik, *Parrinello Festschrift: From Physics via Chemistry to Biology* (Special Issue), ChemPhysChem Volume **6**, Issue 9, Pages 1671–1947 (September 12, 2005).
7. T. A. Arias, Rev. Mod. Phys. **71**, 267 (1999).
8. E. Artacho, D. Sánchez–Portal, P. Ordejón, A. Garcia, and J. M. Soler, phys. stat. sol. (b) **215**, 809 (1999).
9. N. W. Ashcroft and N. D. Mermin, *Solid State Physics* (Saunders College Publishing, Philadelphia, 1976).
10. G. B. Bachelet, D. R. Hamann, and M. Schlüter, Phys. Rev. B **26**, 4199 (1982).
11. P. Bandyopadhyay, S. Ten–no, and S. Iwata, J. Phys. Chem. A **103**, 6442 (1999).
12. R. N. Barnett, U. Landman, and A. Nitzan, J. Chem. Phys. **89**, 2242 (1988).
13. R. N. Barnett, U. Landman, A. Nitzan, and G. Rajagopal, J. Chem. Phys. **94**, 608 (1991).
14. R. N. Barnett and U. Landman, Phys. Rev. B **48**, 2081 (1993).
15. S. Baroni and P. Giannozzi, Europhys. Lett. **17**, 547 (1992).
16. U. von Barth and L. Hedin, J. Phys. C **5**, 1629 (1972).
17. T. L. Beck, Rev. Mod. Phys. **72**, 1041 (2000).

18. A. D. Becke, Phys. Rev. A **38**, 3098 (1988).
19. A. D. Becke, J. Chem. Phys. **98**, 1372 (1993).
20. A. D. Becke, J. Chem. Phys. **98**, 5648 (1993).
21. P. Bendt and A. Zunger, Phys. Rev. Lett. **50**, 1684 (1983).
22. B. J. Berne, G. Ciccotti, and D. F. Coker (Eds.), *Classical and Quantum Dynamics in Condensed Phase Simulations* (World Scientific, Singapore, 1998).
23. J. Bernholc, J.-Y. Yi, and D. J. Sullivan, Faraday Discuss. Chem. Soc. **92**, 217 (1991).
24. K. Yabana and G. F. Bertsch, Phys. Rev. B **54**, 4484 (1996).
25. K. Yabana and G. F. Bertsch, Int. J. Quantum Chem. **75**, 55 (1999).
26. K. Binder and G. Ciccotti (Eds.), *Monte Carlo and Molecular Dynamics of Condensed Matter Systems* (Italian Physical Society SIF, Bologna 1996).
27. P. E. Blöchl and M. Parrinello, Phys. Rev. B **45**, 9413 (1992).
28. P. E. Blöchl, Phys. Rev. B **50**, 17 953 (1994).
29. P. E. Blöchl, Phys. Rev. B **65**, 104 303 (2002).
30. M. Bockstedte, A. Kley, J. Neugebauer, and M. Scheffler, Comput. Phys. Comm. **107**, 187 (1997).
31. A. D. Boese and J. M. L. Martin, J. Chem. Phys. **121**, 3405 (2004).
32. K. Bolton, W. L. Hase, and G. H. Peslherbe, in *Modern Methods for Multidimensional Dynamics Computations in Chemistry*, ed. D. L. Thompson (World Scientific, Singapore, 1998).
33. F. A. Bornemann and C. Schütte, Numer. Math. **78**, 359 (1998).
34. F. A. Bornemann and C. Schütte, Numer. Math. **83**, 179 (1999).
35. S. F. Boys and F. Bernardi, Mol. Phys. **19**, 553 (1970).
36. E. L. Briggs, D. J. Sullivan, and J. Bernholc, Phys. Rev. B **52**, R5471 (1995).
37. J. Broughton and F. Kahn, Phys. Rev. B **40**, 12 098 (1989).
38. R. Car and M. Parrinello, Phys. Rev. Lett. **55**, 2471 (1985).
39. R. Car and M. Parrinello, Solid State Commun. **62**, 403 (1987).
40. R. Car, M. Parrinello, and M. Payne, J. Phys.: Condens. Matter **3**, 9539 (1991).
41. A. Caro, S. Ramos de Debiaggi, and M. Victoria, Phys. Rev. B **41**, 913 (1990).
42. CASTEP: Ref. 201; see
<http://www.tcm.phy.cam.ac.uk/castep/>.
43. S. Chawla and G. A. Voth, J. Chem. Phys. **108**, 4697 (1998).
44. J. R. Chelikowsky, N. Troullier, and Y. Saad, Phys. Rev. Lett. **72**, 1240 (1994).
45. J. R. Chelikowsky, N. Troullier, K. Wu, and Y. Saad, Phys. Rev. B **50**, 11 355 (1994).
46. K. Cho, T. A. Arias, J. D. Joannopoulos, and P. K. Lam, Phys. Rev. Lett. **71**, 1808 (1993).
47. G. Ciccotti, D. Frenkel, and I. R. McDonald, *Simulation of Liquids and Solids* (North-Holland, Amsterdam 1987).
48. G. Ciccotti and J. P. Ryckaert, Comput. Phys. Rep. **4**, 345 (1987).
49. A. J. Cohen and N. C. Handy, Chem. Phys. Lett. **316**, 160 (2000).
50. M. H. Cohen and V. Heine, Phys. Rev. **122**, 1821 (1961).
51. CP2k: “A General Program to Perform Molecular Dynamics Simulations”; CP2k developers group under the terms of the GNU General Public License; see
<http://cp2k.berlios.de/index.html>.
52. CP-PAW: P. E. Blöchl, IBM Zurich Research Laboratory, Ref. 28; see

- <http://www.pt.tu-clausthal.de/~paw/>.
53. T. R. Cundari, M. T. Benson, M. L. Lutz, and S. O. Sommerer, in *Reviews in Computational Chemistry Vol. 8*, p. 145, eds. K. B. Lipkowitz and D. B. Boyd (VCH, New York, 1996).
 54. **Dacapo**: An ab initio molecular dynamics code, based on ultra-soft pseudopotentials; see <http://dcwww.camp.dtu.dk/campos/Dacapo/>.
 55. J. B. Delos, W. R. Thorson, and S. K. Knudson, *Phys. Rev. A* **6**, 709 (1972); see also following two papers.
 56. J. B. Delos, *Rev. Mod. Phys.* **53**, 287 (1981).
 57. Z. Deng, G. J. Martyna, and M. L. Klein, *Phys. Rev. Lett.* **68**, 2496 (1992).
 58. Z. Deng, G. J. Martyna, and M. L. Klein, *Phys. Rev. Lett.* **71**, 267 (1993).
 59. E. Deumens, A. Diz, R. Longo, and Y. Öhrn, *Rev. Mod. Phys.* **66**, 917 (1994).
 60. A. Devenyi, K. Cho, T. A. Arias, and J. D. Joannopoulos, *Phys. Rev. B* **49**, 13 373 (1994).
 61. B. K. Dey, A. Askar, and H. Rabitz, *J. Chem. Phys.* **109**, 8770 (1998).
 62. P. A. M. Dirac, *Proc. Cambridge Phil. Soc.* **26**, 376 (1930).
 63. P. A. M. Dirac, *The Principles of Quantum Mechanics* (Oxford University Press, Oxford, 1947; third edition); see in particular Chapt. V-§31–32.
 64. M. Dolg and H. Stoll, in *Handbook on the Physics and Chemistry of Rare Earths Vol. 22*, p. 607, eds. K. A. Gschneidner Jr. and L. Eyring (Elsevier, Amsterdam, 1996).
 65. N. L. Doltsinis and D. Marx, *J. Theor. Comput. Chem.* **1**, 319 (2002); see <http://www.theochem.rub.de/go/cprev.html>.
 66. R. M. Dreizler and E. K. U. Gross, *Density-Functional Theory* (Springer, Berlin, 1990).
 67. B. I. Dunlap, *Int. J. Quantum Chem.* **69**, 317 (1998).
 68. B. I. Dunlap and R. W. Warren, *Adv. Quantum Chem.* **33**, 167 (1998).
 69. P. Ehrenfest, *Z. Phys.* **45**, 455 (1927).
 70. E. Ermakova, J. Solca, H. Huber, and D. Marx, *Chem. Phys. Lett.* **246**, 204 (1995).
 71. M. Ernzerhof and G. E. Scuseria, *J. Chem. Phys.* **110**, 5029 (1999).
 72. R. Esser, P. Grassberger, J. Grotendorst, and M. Lewerenz (Eds.), *Molecular Dynamics on Parallel Computers* (World Scientific, Singapore, 2000).
 73. M. Fähnle, C. Elsässer, and H. Krimmel, *phys. stat. sol. (b)* **191**, 9 (1995).
 74. M. D. Feit, J. A. Fleck, Jr., and A. Steiger, *J. Comput. Phys.* **47**, 412 (1982).
 75. P. Fernandez, A. Dal Corso, A. Baldereschi, and F. Mauri, *Phys. Rev. B* **55**, R1909 (1997).
 76. R. P. Feynman, *Phys. Rev.* **56**, 340 (1939); see in particular equation (2).
 77. **FHI98md**: Ref. 30; see <http://www.fhi-berlin.mpg.de/th/fhimd/>.
 78. M. J. Field, *Chem. Phys. Lett.* **172**, 83 (1990).
 79. M. J. Field, *J. Phys. Chem.* **95**, 5104 (1991).
 80. M. Filatov and W. Thiel, *Mol. Phys.* **91**, 847 (1997).
 81. M. Filatov and W. Thiel, *Phys. Rev. A* **57**, 189 (1998).
 82. P. Focher, Ph.D. Thesis: *First-principle studies of structural phase transformations* (Scuola Internazionale Superiore di Studi Avanzati (SISSA), Trieste, 1994).

83. E. S. Fois, A. Selloni, M. Parrinello, and R. Car, *J. Phys. Chem.* **92**, 3268 (1988).
84. E. S. Fois, J. I. Penman, and P. A. Madden, *J. Chem. Phys.* **98**, 6361 (1993).
85. G. P. Francis and M. C. Payne, *J. Phys.: Condens. Matter* **2**, 4395 (1990).
86. I. Frank, J. Hutter, D. Marx, and M. Parrinello, *J. Chem. Phys.* **108**, 4060 (1998).
87. I. Frank, *Angew. Chem. Int. Ed.* **42**, 1569 (2003); *Angew. Chem.* **115**, 1607 (2003).
88. D. Frenkel and B. Smit, *Understanding Molecular Simulation – From Algorithms to Applications* (Academic Press, San Diego, 1996).
89. S. Froyen and M. L. Cohen, *J. Phys. C* **19**, 2623 (1986).
90. M. Fuchs and M. Scheffler, *Comput. Phys. Commun.* **119**, 67 (1999).
91. G. Galli and M. Parrinello, in *Computer Simulations in Materials Science*, p. 282, eds. M. Meyer and V. Pontikis (Kluwer, Dordrecht, 1991).
92. G. Galli and A. Pasquarello, in *Computer Simulation in Chemical Physics*, eds. M. P. Allen and D. J. Tildesley (Kluwer, Dordrecht, 1993).
93. D. A. Gibson and E. A. Carter, *J. Phys. Chem.* **97**, 13 429 (1993).
94. D. A. Gibson, I. V. Ionova, and E. A. Carter, *Chem. Phys. Lett.* **240**, 261 (1995).
95. D. A. Gibson and E. A. Carter, *Mol. Phys.* **89**, 1265 (1996).
96. M. J. Gillan, *J. Phys.: Condens. Matter* **1**, 689 (1989).
97. M. J. Gillan, *Contemp. Phys.* **38**, 115 (1997).
98. S. Goedecker and C. J. Umrigar, *Phys. Rev. A* **55**, 1765 (1997).
99. S. Goedecker and O. V. Ivanov, *Solid State Commun.* **105**, 665 (1998).
100. S. Goedecker and O. V. Ivanov, *Computers in Physics* **12**, 548 (1998).
101. S. Goedecker, *Rev. Mod. Phys.* **71**, 1085 (1999).
102. S. Goedecker and C. Chauvin, *J. Theor. Comput. Chem.* **2**, 483 (2003).
103. A. Görling, *Phys. Rev. B* **53**, 7024 (1996).
104. H. Goldstein, C. P. Poole, and J. L. Safko, *Classical Mechanics*, 3rd Edition (Addison–Wesley, San Francisco, 2002); for errors see <http://astro.physics.sc.edu/goldstein/>.
105. P. Gomes Dacosta, O. H. Nielsen, and K. Kunc, *J. Phys. C* **19**, 3163 (1986).
106. X. Gonze, G. M. Rignanese, M. Verstraete, J. M. Beuken, Y. Pouillon, R. Caracas, F. Jollet, M. Torrent, G. Zerah, M. Mikami, P. Ghosez, M. Veithen, J. Y. Raty, V. Olevanov, F. Bruneval, L. Reining, R. Godby, G. Onida, D. R. Hamann, and D. C. Allan, *Z. Krist.* **220**, 558 (2005).
107. T. Grabo, E. K. U. Gross, and M. Lüders, *Orbital Functionals in Density Functional Theory: The Optimized Effective Potential Method*; see <http://psi-k.dl.ac.uk/newsletters/News16/Highlight16.pdf>.
108. C. G. Gray and K. E. Gubbins, *Theory of Molecular Fluids Vol. 1* (Oxford, Clarendon, 1984).
109. J. C. Greer, R. Ahlrichs, and I. V. Hertel, *Z. Phys. D* **18**, 413 (1991); see in particular Section 3.
110. E. K. U. Gross and W. Kohn, *Adv. Quant. Chem.* **21**, 255 (1990).
111. F. Grossmann, *Comments At. Mol. Phys.* **34**, 141 (1999).
112. J. Grotendorst, D. Marx, and A. Muramatsu (Editors), *Quantum Simulations of Complex Many–Body Systems: From Theory to Algorithms*, (John von Neumann Institute for Computing, Forschungszentrum Jülich 2002); Hardcover Version: ISBN 3-00-009057-6, Electronic Version: <http://www.theochem.rub.de/go/cprev.html>, Audio–Visual Lec-

- ture Notes: <http://www.theochem.rub.de/go/cprev.html>.
113. F. Gygi and A. Baldereschi, Phys. Rev. B **34**, 4405 (1986).
 114. F. Gygi, Europhys. Lett. **19**, 617 (1992).
 115. F. Gygi, Phys. Rev. B **48**, 11 692 (1993).
 116. F. Gygi, Phys. Rev. B **51**, 11 190 (1995).
 117. F. Gygi and G. Galli, Phys. Rev. B **52**, R2229 (1995).
 118. R. Haberlandt, S. Fritzsche, G. Peinel, and K. Heinzinger, *Molekulardynamik – Grundlagen und Anwendungen* (Vieweg Verlag, Braunschweig, 1995).
 119. D. R. Hamann, M. Schlüter, and C. Chiang, Phys. Rev. Lett. **43**, 1494 (1979).
 120. D. R. Hamann, Phys. Rev. B **51**, 7337 (1995).
 121. D. R. Hamann, Phys. Rev. B **51**, 9508 (1995).
 122. D. R. Hamann, Phys. Rev. B **54**, 1568 (1996).
 123. D. R. Hamann, Phys. Rev. B **56**, 14 979 (1997).
 124. S. Hammes-Schiffer and H. C. Andersen, J. Chem. Phys. **99**, 523 (1993).
 125. B. Hartke and E. A. Carter, Chem. Phys. Lett. **189**, 358 (1992).
 126. B. Hartke and E. A. Carter, J. Chem. Phys. **97**, 6569 (1992).
 127. B. Hartke and E. A. Carter, Chem. Phys. Lett. **216**, 324 (1993).
 128. B. Hartke, D. A. Gibson, and E. A. Carter, Int. J. Quantum Chem. **45**, 59 (1993).
 129. B. Hartke, private communication.
 130. M. Head-Gordon and J. A. Pople, J. Phys. Chem. **92**, 3063 (1988).
 131. W. J. Hehre, L. Radom, P. v. R. Schleyer, and J. A. Pople, *Ab Initio Molecular Orbital Theory* (John Wiley & Sons, New York, 1986).
 132. A. Heidenreich and J. Sauer, Z. Phys. D **35**, 279 (1995).
 133. H. Hellmann, Z. Phys. **85**, 180 (1933), see equation (3b); note that Hellmann already connects the “Hellmann–Feynman theorem” to first-order perturbation theory for cases where it does not hold rigorously, for instance when the wave function used to evaluate the expectation value is not the exact one.
 134. H. Hellmann, J. Chem. Phys. **3**, 61 (1935).
 135. H. Hellmann and W. Kassatotschkin, J. Chem. Phys. **4**, 324 (1936).
 136. J. M. Herbert and M. Head-Gordon, J. Chem. Phys. **121**, 11 542 (2004).
 137. J. M. Herbert and M. Head-Gordon, Phys. Chem. Chem. Phys. **7**, 3269 (2005).
 138. P. Hohenberg and W. Kohn, Phys. Rev. **136**, B864 (1964).
 139. J. Hutter *et al.*, *CPMD: Car–Parrinello Molecular Dynamics: An Ab Initio Electronic Structure and Molecular Dynamics Program*; IBM Zurich Research Laboratory (1990–2006) and Max–Planck–Institut für Festkörperforschung (1997–2001); for downloads see <http://www.cpmid.org/>.
 140. J. Hutter, M. E. Tuckerman, and M. Parrinello, J. Chem. Phys. **102**, 859 (1995).
 141. J. Hutter and A. Curioni, ChemPhysChem **6**, 1788 (2005).
 142. R. Iftimie, P. Minari, and M. E. Tuckerman, Proc. Natl. Acad. Sci. U.S.A. **102**, 6654 (2005).
 143. J. Ihm, A. Zunger, and M. L. Cohen, J. Phys. C **12**, 4409 (1979).
 144. J. Jellinek, V. Bonačič–Koutecký, P. Fantucci, and M. Wiechert, J. Chem. Phys. **101**, 10 092 (1994).
 145. J. Jellinek, S. Srinivas, and P. Fantucci, Chem. Phys. Lett. **288**, 705 (1998).
 146. B. G. Johnson, P. M. W. Gill, and J. A. Pople, J. Chem. Phys. **98**, 5612 (1993).

147. R. O. Jones and O. Gunnarsson, *Rev. Mod. Phys.* **61**, 689 (1989).
148. Y.-M. Juan and E. Kaxiras, *Phys. Rev. B* **48**, 14 944 (1993).
149. R. A. Kendall, E. Apra, D. E. Bernholdt, E. J. Bylaska, M. Dupuis, G. I. Fann, R. J. Harrison, J. L. Ju, J. A. Nichols, J. Nieplocha, T. P. Straatsma, T. L. Windus, and A. T. Wong, *Comput. Phys. Commun.* **128**, 260 (2000).
150. V. Keshari and Y. Ishikawa, *Chem. Phys. Lett.* **218**, 406 (1994).
151. F. Kirchhoff, G. Kresse, and M. J. Gillan, *Phys. Rev. B* **57**, 10 482 (1998).
152. S. Kirkpatrick, C. D. Gelatt, Jr., and M. P. Vecchi, *Science* **220**, 671 (1983).
153. W. Kohn and L. J. Sham, *Phys. Rev.* **140**, A1133 (1965).
154. W. Kohn, *Chem. Phys. Lett.* **208**, 167 (1993).
155. W. Kołos, *Adv. Quant. Chem.* **5**, 99 (1970).
156. R. Kosloff, *J. Phys. Chem.* **92**, 2087 (1988).
157. R. Kosloff, *Annu. Rev. Phys. Chem.* **45**, 145 (1994).
158. G. Kresse and J. Hafner, *Phys. Rev. B* **47**, 558 (1993).
159. G. Kresse and J. Furthmüller, *Phys. Rev. B* **54**, 11 169 (1996).
160. G. Kresse and J. Furthmüller, *Comput. Mater. Sci* **6**, 15 (1996).
161. G. Kresse, *J. Non-Cryst. Solids* **312**, 52 (2002).
162. W. Kutzelnigg, *Mol. Phys.* **90**, 909 (1997).
163. K. Laasonen, A. Pasquarello, R. Car, C. Lee, and D. Vanderbilt, *Phys. Rev. B* **47**, 10 142 (1993).
164. C. Lee, W. Yang, and R. G. Parr, *Phys. Rev. B* **37**, 785 (1988).
165. Y.-S. Lee, M. Buongiorno Nardelli, and N. Marzari, *Phys. Rev. Lett.* **95**, 076804 (2005).
166. C. Leforestier, *J. Chem. Phys.* **68**, 4406 (1978).
167. C. Leforestier, R. H. Bisseling, C. Cerjan, M. D. Feit, R. Friesner, A. Guldberg, A. Hammerich, G. Jolicard, W. Karrlein, H.-D. Meyer, N. Lipkin, O. Roncero, and R. Kosloff, *J. Comp. Phys.* **94**, 59 (1991).
168. I. N. Levine, *Quantum Chemistry* (Allyn and Bacon, Boston, 1983).
169. X. Li, J. C. Tully, H. B. Schlegel, and M. J. Frisch, *J. Chem. Phys.* **123**, 084 106 (2005).
170. G. Lippert, J. Hutter, and M. Parrinello, *Molec. Phys.* **92**, 477 (1997).
171. G. Lippert, J. Hutter, and M. Parrinello, *Theor. Chem. Acc.* **103**, 124 (1999).
172. Z. Liu, L. E. Carter, and E. A. Carter, *J. Phys. Chem.* **99**, 4355 (1995).
173. H. Löwen, P. A. Madden, and J.-P. Hansen, *Phys. Rev. Lett.* **68**, 1081 (1992); see also Ref. 174.
174. H. Löwen and I. D'amico, *J. Phys.: Condens. Matter* **9**, 8879 (1997).
175. S. A. Maluendes and M. Dupuis, *Int. J. Quantum Chem.* **42**, 1327 (1992).
176. R. M. Martin, *Electronic Structure: Basic Theory and Practical Methods* (Cambridge University Press, Cambridge, 2004); for additional information, errata, web resources etc. see <http://electronicstructure.org/>.
177. D. Marx, M. E. Tuckerman, and G. J. Martyna, *Comput. Phys. Commun.* **118**, 166 (1999); the misprinted definition of the fictitious normal mode masses given in Eq. (2.51) is corrected in Ref. 179. However, the correct definition was implemented in the CPMD package¹³⁹ so that all data reported in the paper are unaffected.
178. D. Marx, *Nachr. Chem. Tech. Lab.* **47**, 186 (1999).
179. D. Marx and J. Hutter, *Ab Initio Molecular Dynamics: Theory and Implementation*, in

- Modern Methods and Algorithms of Quantum Chemistry*, pp. 301–449 (first edition, paperback, ISBN 3-00-005618-1) or pp. 329–477 (second edition, hardcover, ISBN 3-00-005834-6), Editor: J. Grotendorst (John von Neumann Institute for Computing, Forschungszentrum Jülich 2000); see <http://www.theochem.rub.de/go/cprev.html>.
180. D. Marx, Wasser, Eis und Protonen: Mit Quantensimulationen zum molekularen Verständnis von Wasserstoffbrücken, Protonentransfer und Phasenübergängen, *Physik Journal* **3**(5), 33 (2004).
 181. D. Marx, Theoretical Chemistry in the 21st Century: The “Virtual Lab”, in *Proceedings of the “Idea-Finding Symposium: Frankfurt Institute for Advanced Studies”*, pp. 139–153, Editors: W. Greiner and J. Reinhardt (EP Systema Bt., Debrecen, 2004).
 182. N. Marzari and D. Vanderbilt, *Phys. Rev. B* **56**, 12 847 (1997).
 183. R. McWeeny, *Methods of Molecular Quantum Mechanics* (Academic Press, London, 1992).
 184. A. Messiah, *Quantum Mechanics* (North-Holland Publishing Company, Amsterdam, 1964); see in particular Chapter VI.I.4 in Volume I.
 185. H.-D. Meyer and W. H. Miller, *J. Chem. Phys.* **70**, 3214 (1979).
 186. N. A. Modine, G. Zumbach, and E. Kaxiras, *Phys. Rev. B* **55**, 10 289 (1997).
 187. T. Morishita and S. Nosé, *Phys. Rev. B* **59**, 15126 (1999).
 188. Nobel Prize in Chemistry awarded to Walter Kohn “for his development of the density-functional theory” and to John Pople “for his development of computational methods in quantum chemistry” in 1998: W. Kohn, *Rev. Mod. Phys.* **71**, 1253 (1999); J. A. Pople, *Angew. Chem. Int. Ed.* **38**, 1894 (1999); *Angew. Chem.* **111**, 2014 (1999); see <http://nobelprize.org/chemistry/laureates/1998/>.
 189. NWChem: Ref. 149; developed and distributed by Pacific Northwest National Laboratory, USA; see <http://www.emsl.pnl.gov/docs/nwchem/>.
 190. P. Ordejón, *Comp. Mater. Sci.* **12**, 157 (1998).
 191. J. Ortega, J. P. Lewis, and O. F. Sankey, *Phys. Rev. B* **50**, 10 516 (1994).
 192. J. Ortega, J. P. Lewis, and O. F. Sankey, *J. Chem. Phys.* **106**, 3696 (1997).
 193. M. Palumbo, L. Reining, and P. Ballone, *J. Phys. IV (Paris)* **3**:(C7), 1955 (1993).
 194. R. G. Parr and W. Yang, *Density-Functional Theory of Atoms and Molecules* (Oxford University Press, Oxford, 1989).
 195. M. Parrinello, *Solid State Commun.* **102**, 107 (1997).
 196. A. Pasquarello, K. Laasonen, R. Car, Ch. Lee, and D. Vanderbilt, *Phys. Rev. Lett.* **69**, 1982 (1992).
 197. G. Pastore, E. Smargiassi, and F. Buda, *Phys. Rev. A* **44**, 6334 (1991).
 198. G. Pastore, in *Monte Carlo and Molecular Dynamics of Condensed Matter Systems*, Chapt. 24, p. 635, eds. K. Binder and G. Ciccotti (Italian Physical Society SIF, Bologna, 1996).
 199. M. C. Payne, J. D. Joannopoulos, D. C. Allan, M. P. Teter, and D. Vanderbilt, *Phys. Rev. Lett.* **56**, 2656 (1986).
 200. M. C. Payne, *J. Phys.: Condens. Matter* **1**, 2199 (1989).
 201. M. C. Payne, M. P. Teter, D. C. Allan, T. A. Arias, and J. D. Joannopoulos, *Rev. Mod. Phys.* **64**, 1045 (1992).
 202. M. Pearson, E. Smargiassi, and P. Madden, *J. Phys. Condens. Matter* **5**, 3221 (1993).

203. J. P. Perdew and A. Zunger, Phys. Rev. B **23**, 5048 (1981).
204. J. P. Perdew, Phys. Rev. B **33**, 8822 (1986); Erratum: Phys. Rev. B **34**, 7406 (1986).
205. J. P. Perdew and Y. Wang, Phys. Rev. B **45**, 13 244 (1992).
206. J. P. Perdew, J. A. Chevary, S. H. Vosko, K. A. Jackson, M. R. Pederson, D. J. Singh, and C. Fiolhais, Phys. Rev. B **46**, 6671 (1992); Erratum: Phys. Rev. B **48**, 4978 (1993).
207. J. P. Perdew, K. Burke, M. Ernzerhof, Phys. Rev. Lett. **77**, 3865 (1996); Erratum: Phys. Rev. Lett. **78**, 1396 (1997); see Refs. 208,291 for revised versions of the “PBE” functional.
208. J. P. Perdew, K. Burke, M. Ernzerhof, Phys. Rev. Lett. **80**, 891 (1998).
209. J. C. Phillips, Phys. Rev. **112**, 685 (1958).
210. J. C. Phillips and L. Kleinman, Phys. Rev. **116**, 287 (1959).
211. Physics and Astronomy Classification Scheme (PACS); see <http://publish.aps.org/PACS/>.
212. W. E. Pickett, Comput. Phys. Rep. **9**, 115 (1989).
213. PINY: *The PINY_MD(c) Simulation Package*; Principle Authors: G. J. Martyna and M. E. Tuckerman; Other Authors: D. A. Yarne, S. O. Samuelson, A. L. Hughes, Y. Liu, Z. Zhu, M. Diraison, K. Pihakari; see Ref. 266 for the method; see http://homepages.nyu.edu/~mt33/PINY_MD/PINY.html.
214. E. I. Proynov, S. Sirois, and D. R. Salahub, Int. J. Quantum Chem. **64**, 427 (1997).
215. P. Pulay, Molec. Phys. **17**, 197 (1969).
216. P. Pulay, Adv. Chem. Phys. **69**, 241 (1987).
217. A. Putrino and G. B. Bachelet, in *Advances in Computational Materials Science II*, eds. V. Fiorentini and F. Meloni (SIF, Bologna, 1998).
218. PWscf: *Plane-Wave Self-Consistent Field is a set of programs for electronic structure calculations within Density-Functional Theory and Density-Functional Perturbation Theory, using a Plane-Wave basis set and pseudopotentials*; PWscf is released under the GNU General Public License; see <http://www.pwscf.org/>.
219. M. Qiu, X.-Y. Zhou, M. Jiang, P.-L. Cao, and Z. Zeng, Phys. Lett. A **245**, 430 (1998).
220. D. C. Rapaport, *The Art of Molecular Dynamics Simulation* (Cambridge University Press, Cambridge, 2001).
221. C. Raynaud, L. Maron, J.-P. Daudey, and F. Jolibois, Phys. Chem. Chem. Phys. **6**, 4226 (2004).
222. D. K. Remler and P. A. Madden, Molec. Phys. **70**, 921 (1990).
223. U. Saalman and R. Schmidt, Z. Phys. D **38**, 153 (1996).
224. J. J. Sakurai, *Modern Quantum Mechanics* (Addison-Wesley Publishing Company, Redwood City, 1985); see in particular Chapter 2.4.
225. E. Sandré and A. Pasturel, Mol. Simul. **20**, 63 (1997).
226. G. C. Schatz, Rev. Mod. Phys. **61**, 669 (1989).
227. F. Scheck, *Mechanics: From Newton’s Laws to Deterministic Chaos*, 4th Edition (Springer Verlag, Berlin, 2005).
228. R. Schinke, *Photodissociation Dynamics* (Cambridge University Press, Cambridge, 1995).
229. H. B. Schlegel, J. M. Millam, S. S. Iyengar, G. A. Voth, A. D. Daniels, G. E. Scuseria, and M. J. Frisch, J. Chem. Phys. **114**, 9758 (2001).

230. H. B. Schlegel, S. S. Iyengar,, X. Li, J. M. Millam, G. A. Voth, G. E. Scuseria, and M. J. Frisch, *J. Chem. Phys.* **117**, 8694 (2002).
231. R. Schmid, *J. Comput. Chem.* **25**, 799 (2004).
232. G. Seifert, D. Porezag, and Th. Frauenheim, *Int. J. Quantum Chem.* **58**, 185 (1996).
233. A. Selloni, P. Carnevali, R. Car, and M. Parrinello, *Phys. Rev. Lett.* **59**, 823 (1987).
234. A. J. R. da Silva, H.-Y. Cheng, D. A. Gibson, K. L. Sørge, Z. Liu, and E. A. Carter, *Spectrochim. Acta A* **53**, 1285 (1997).
235. A. J. R. da Silva, M. R. Radeke, and E. A. Carter, *Surf. Sci. Lett.* **381**, L628 (1997).
236. A. J. R. da Silva, J. W. Pang, E. A. Carter, and D. Neuhauser, *J. Phys. Chem. A* **102**, 881 (1998).
237. P. L. Silvestrelli, A. Alavi, M. Parrinello, and D. Frenkel, *Europhys. Lett.* **33**, 551 (1996).
238. F. Sim, A. St.-Amant, I. Papai, and D. R. Salahub, *J. Am. Chem. Soc.* **114**, 4391 (1992).
239. D. J. Singh, *Planewaves, Pseudopotentials and the LAPW Method* (Kluwer, Dordrecht, 1994).
240. S/PHI/nX (or SFHingX): written by S. Boeck, J. Neugebauer *et al.*; see <http://www.sfhingx.de/>.
241. M. Sprik and M. L. Klein, *J. Chem. Phys.* **89**, 1592 (1988); Erratum: *J. Chem. Phys.* **90**, 7614 (1989).
242. M. Sprik, *J. Phys. Chem.* **95**, 2283 (1991).
243. M. Sprik, in *Computer Simulation in Chemical Physics*, eds. M. P. Allen and D. J. Tildesley (Kluwer, Dordrecht, 1993).
244. M. Sprik, in *Monte Carlo and Molecular Dynamics of Condensed Matter Systems*, Chapt. 2, p. 43, eds. K. Binder and G. Ciccotti (Italian Physical Society SIF, Bologna, 1996).
245. M. Sprik, in *Classical and Quantum Dynamics in Condensed Phase Simulations*, Chapt. 13, eds. B. J. Berne, G. Ciccotti, and D. F. Coker (World Scientific, Singapore, 1998).
246. M. Springborg (Ed.), *Density-Functional Methods in Chemistry and Materials Science*, (John Wiley & Sons, New York, 1997).
247. G. P. Srivastava and D. Weaire, *Adv. Phys.* **36**, 463 (1987).
248. V. N. Staroverov, G. E. Scuseria, J. Tao, and J. P. Perdew, *J. Chem. Phys.* **119**, 12 129 (2003); Erratum: *J. Chem. Phys.* **121**, 11 507 (2004).
249. V. N. Staroverov, G. E. Scuseria, J. Tao, and J. P. Perdew, *Phys. Rev. B* **69**, 075 102 (2004).
250. I. Štich, R. Car, M. Parrinello, and S. Baroni, *Phys. Rev. B* **39**, 4997 (1989).
251. W. Stier and O. V. Prezhdo, *J. Phys. Chem. B* **106**, 8047 (2002).
252. A. Szabo and N. S. Ostlund, *Modern Quantum Chemistry – Introduction to Advanced Electronic Structure Theory* (McGraw-Hill Publishing Company, New York, 1989).
253. S. Tanaka, *J. Chem. Phys.* **100**, 7416 (1994).
254. P. Tangney and S. Scandolo, *J. Chem. Phys.* **116**, 14 (2002).
255. J. Tao, J. P. Perdew, V. N. Staroverov, and G. E. Scuseria, *Phys. Rev. Lett.* **91**, 146 401 (2003).
256. F. Tassone, F. Mauri, and R. Car, *Phys. Rev. B* **50**, 10 561 (1994).
257. M. P. Teter, M. C. Payne, and D. C. Allen, *Phys. Rev. B* **40**, 12 255 (1989).

258. J. Theilhaber, Phys. Rev. B **46**, 12 990 (1992).
259. E. Tsuchida and M. Tsukada, Solid State Commun. **94**, 5 (1995).
260. E. Tsuchida and M. Tsukada, Phys. Rev. B **52**, 5573 (1995).
261. E. Tsuchida and M. Tsukada, J. Phys. Soc. Japan **67**, 3844 (1998).
262. M. Tuckerman, B. J. Berne, and G. J. Martyna, J. Chem. Phys. **97**, 1990 (1992).
263. M. E. Tuckerman and M. Parrinello, J. Chem. Phys. **101**, 1302 (1994); see also Ref. 140.
264. M. E. Tuckerman and M. Parrinello, J. Chem. Phys. **101**, 1316 (1994).
265. M. E. Tuckerman, P. J. Ungar, T. von Rosenvinge, and M. L. Klein, J. Phys. Chem. **100**, 12 878 (1996).
266. M. E. Tuckerman, D. A. Yarne, S. O. Samuelson, A. L. Hughes, and G. J. Martyna, Comput. Phys. Commun. **128**, 333 (2000).
267. M. E. Tuckerman and G. J. Martyna, J. Phys. Chem. B **104**, 159 (2000); Additions and Corrections: J. Phys. Chem. B **105**, 7598 (2001).
268. M. E. Tuckerman, J. Phys.: Condens. Matter **14**, R1297 (2002).
269. M. E. Tuckerman, private communication.
270. J. C. Tully, in *Modern Theoretical Chemistry: Dynamics of Molecular Collisions*, Part B, p. 217, ed. W. H. Miller (Plenum Press, New York, 1976).
271. J. C. Tully, in *Modern Methods for Multidimensional Dynamics Computations in Chemistry*, ed. D. L. Thompson (World Scientific, Singapore, 1998).
272. J. C. Tully, in *Classical and Quantum Dynamics in Condensed Phase Simulations*, Chapt. 21, p. 489, eds. B. J. Berne, G. Ciccotti, and D. F. Coker (World Scientific, Singapore, 1998).
273. C. J. Tymczak and X.-Q. Wang, Phys. Rev. Lett. **78**, 3654 (1997).
274. K. Uehara, M. Ishitobi, T. Oda, and Y. Hiwatari, Mol. Simul. **18**, 385 (1997); Mol. Simul. **19**, 75 (1997).
275. D. Vanderbilt, Phys. Rev. Lett. **59**, 1456 (1987).
276. D. Vanderbilt, Phys. Rev. B **41**, 7892 (1990).
277. J. VandeVondele, M. Krack, F. Mohamed, M. Parrinello, T. Chassaing, and J. Hutter, Comput. Phys. Commun. **167**, 103 (2005).
278. VASP: *Vienna Ab-initio Simulation Package*^{159,160}; see <http://cms.mpi.univie.ac.at/vasp/>.
279. T. Van Voorhis and G. Scuseria, J. Chem. Phys. **109**, 400 (1998).
280. D. S. Wallace, A. M. Stoneham, W. Hayes, A. J. Fisher, and A. H. Harker, J. Phys.: Condens. Matter **3**, 3879 (1991).
281. I. S. Y. Wang and M. Karplus, J. Amer. Chem. Soc. **95**, 8160 (1973).
282. A. Warshel and M. Karplus, Chem. Phys. Lett. **32**, 11 (1975).
283. S. Wei and M. Y. Chou, Phys. Rev. Lett. **76**, 2650 (1996).
284. R. M. Wentzcovitch, Phys. Rev. B **44**, 2358 (1991).
285. R. M. Wentzcovitch and J. L. Martins, Solid State Commun. **78**, 831 (1991).
286. S. R. White, J. W. Wilkins, and M. P. Teter, Phys. Rev. B **39**, 5819 (1989).
287. A. Willetts and N. C. Handy, Chem. Phys. Lett. **227**, 194 (1994).
288. M. Wilson and P. M. Madden, J. Phys.: Condens. Matter **5**, 2687 (1993).
289. E. Wimmer, Science **269**, 1397 (1995).
290. K. Yamaguchi and T. Mukoyama, J. Phys. B **29**, 4059 (1996).
291. Y. Zhang and W. Yang, Phys. Rev. Lett. **80**, 890 (1998).

Beyond Hartree-Fock: MP2 and Coupled-Cluster Methods for Large Systems

Christof Hättig

Institute of Nanotechnology
Forschungszentrum Karlsruhe
P.O. Box 3640, 76021 Karlsruhe, Germany
E-mail: christof.haettig@int.fzk.de

This chapter reviews briefly the theory of second order Møller-Plesset perturbation theory (MP2), the approximate coupled-cluster singles-and-doubles method CC2 and closely related *ab initio* methods for calculations on excited states which treat the electron-electron interaction correctly through second order. It is shown how the computational costs (CPU time and storage requirements) can be reduced for these methods by orders of magnitudes using the resolution-of-the-identity approximation for electron repulsion integrals. The demands for the auxiliary basis sets are discussed and it is shown how with optimized basis sets the errors due to this approximation become insignificant compared to the usual orbital basis set error. Furthermore, a parallel implementation of RI-MP2 and RI-CC2 in the TURBOMOLE program package for distributed memory architectures (as e.g. clusters of Linux boxes) is presented and some prototypical applications are reviewed.

1 Introduction

During the last year many wavefunction based correlation methods have been proposed for applications on extended molecular systems. Most of them are based on the so-called local correlation approach¹⁻⁹, and/or on an extensive screening of small and often long ranging contributions to the correlation energy^{4,10,11}. Some approaches introduce empirical parameters or rely on a balance between certain contributions which in practice might or might not be given¹²⁻¹⁵. For most of these approaches it is not yet clear to which extent they can be developed in the near future into competitive methods for extended systems. In particular, if the reduction of the computational costs (compared to more traditional implementations of quantum chemical methods) relies on a screening in the atomic orbital (AO) basis set, calculations on extended systems are often only possible with rather small basis sets which cannot supply the accuracy expected from a correlated *ab initio* method. Even though, usually only explored for electronic ground states, most of these approaches could in principle also be generalized to excited states. But for larger molecules, calculations for excited states employ often so-called response methods and the parameterization of the excited state used in these methods hampers the application of the local correlation and related approaches.¹⁶⁻¹⁸

We will therefore in the following not go into the details of these approaches, but restrict the following discussion to the underlying electronic structure methods, which are usually single-reference coupled-cluster (CC) and, in particular, for larger systems Møller-Plesset perturbation theory through second order (MP2) or related methods for excited states. The implementation of the latter methods has during the last decade improved dramatically by combining them with the so-called resolution-of-the-identity (RI) approximation for the four-index electron repulsion integrals (ERIs) with optimized auxiliary

basis sets. Even without any further approximations are these methods today applicable to systems with up to 100 or more atoms. Since the RI approximation depends little on the electronic structure of the investigated system it does not diminish the applicability of the underlying electronic structure methods. It is also compatible and can be combined with the above mentioned screening based approaches to reduce further the computational costs.^{19,20} Thus, it can be expected that these two aspects, the treatment of the electron correlation through second order and the RI approximation for ERIs will remain important ingredients also in future correlated wavefunction based methods for extended systems.

In the following the theory of wavefunction based ab initio methods that treat the electron-electron interaction correctly through second order is briefly reviewed. The emphasis will be on methods for excited states which can be related to the approximate coupled-cluster singles-and-doubles model CC2, an approximation to the coupled-cluster singles-and-doubles method (CCSD). In Section 3 it is shown how the computational costs for these methods can be reduced drastically by using the RI approximation and disc space bottlenecks for these methods can be resolved by an doubles amplitudes-direct implementation. A recent parallel implementation for distributed memory architectures is presented in Section 4 and some example applications with RI-MP2 and RI-CC2 are reviewed in Secs. 5 and 6.

2 Second Order Methods for Ground and Excited States

Second order Møller-Plesset perturbation theory is a conceptually simple and technically the most simplest ab initio^a correlation method. It can be derived by expanding the solution of the electronic Schrödinger equation as a Taylor series in the fluctuation potential (vide infra). This can be done either in the framework of configuration interaction theory or using the single-reference coupled-cluster ansatz for the wavefunction.²¹ We will take here the latter starting point to have a close connection to coupled-cluster response and related methods for excited states. In the coupled-cluster ansatz the wavefunction is parameterized as

$$|\text{CC}\rangle = \exp(\hat{T})|\text{HF}\rangle \quad (1)$$

with the cluster operator defined as

$$\hat{T} = \hat{T}_1 + \hat{T}_2 + \hat{T}_3 + \dots \quad (2)$$

where

$$\hat{T}_1 = \sum_{\mu_1} t_{\mu_1} \hat{\tau}_{\mu_1} = \sum_{ai} t_a^i \hat{\tau}_a^i, \quad \hat{T}_2 = \sum_{\mu_2} t_{\mu_2} \hat{\tau}_{\mu_2} = \sum_{aibj} t_{ab}^{ij} \hat{\tau}_{ab}^{ij}, \quad \dots \quad (3)$$

The coefficients t_{μ_i} are called cluster amplitudes and the excitation operators $\hat{\tau}_{\mu_i}$ generate all possible single, double, and higher excited determinants if applied on the ground state Hartree-Fock (HF) determinant $|\text{HF}\rangle$. Here and in the following, we use the convention that indices $i, j, \dots$ denote occupied, $a, b, \dots$ virtual, and $p, q, \dots$ arbitrary molecular orbitals (MOs).

^a Here and in the following we use “ab initio” for electronic structure methods which are systematically improvable in the sense that they are members of a hierarchy which converges to the exact solution of the electronic Schrödinger equations, i.e. the full configuration interaction (Full CI) limit.

Inserting the ansatz (1) into the electronic Schrödinger equation and multiplying from the left with $\exp(-\hat{T})$ one gets

$$\exp(-\hat{T})\hat{H}\exp(\hat{T})|\text{HF}\rangle = E|\text{HF}\rangle. \quad (4)$$

Projecting the above form of the Schrödinger equation onto the HF determinant and a projection manifold of (suitable linear combinations of) excited determinants one obtains an expression for the ground state energy

$$E = \langle \text{HF} | \exp(-\hat{T})\hat{H}\exp(\hat{T}) | \text{HF} \rangle = \langle \text{HF} | \hat{H} \exp(\hat{T}) | \text{HF} \rangle, \quad (5)$$

and the cluster equations

$$0 = \langle \mu_i | \exp(-\hat{T})\hat{H}\exp(\hat{T}) | \text{HF} \rangle, \quad (6)$$

which determine the amplitudes t_{μ_i} . Since we have not yet made any approximation, the above equations still give the exact ground state solution of the electronic Schrödinger equation. Truncating the cluster operator (2) after the single ($\hat{T}_1$) and double ($\hat{T}_2$) excitations gives the coupled-cluster singles-and-doubles (CCSD) method, truncating it after $\hat{T}_3$ the CCSDT method, and so on.^b

Expressions for Møller-Plesset perturbation theory are found by splitting the Hamiltonian into the Fock operator $\hat{F}$ as zeroth-order and the electron-electron fluctuation potential as first-order contribution to the Hamiltonian

$$\hat{H}^{(0)} = \hat{F}, \quad \hat{H}^{(1)} = \hat{\Phi} = \hat{H} - \hat{F}, \quad (7)$$

and expanding Eqs. (5) and (6) in orders of the fluctuation potential. If the Brillouin-Theorem is fulfilled and $\langle \hat{H}^{(1)} | \text{HF} \rangle = 0$, i.e. for a closed-shell or an unrestricted open-shell Hartree-Fock (UHF) reference, the MP2 energy is obtained as

$$E_{\text{MP2}} = \langle \text{HF} | \hat{\Phi} \hat{T}_2^{(1)} | \text{HF} \rangle = \sum_{aibj} t_{ab}^{ij} \langle \text{HF} | \hat{\Phi}_{ij}^{ab} \rangle \quad (8)$$

with

$$0 = \langle \hat{\Phi}_{ij}^{ab} | [\hat{F}, \hat{T}_2^{(1)}] + \hat{\Phi} | \text{HF} \rangle \Leftrightarrow t_{ab}^{ij} = \frac{\langle \hat{\Phi}_{ij}^{ab} | \text{HF} \rangle}{\epsilon_i - \epsilon_a + \epsilon_j - \epsilon_b} \quad (9)$$

where we assumed canonical molecular orbitals and ϵ_p are the orbital energies.

Møller-Plesset perturbation theory can not straightforwardly be applied to excited states, since wavefunctions for excited states usually require a multi-reference treatment. For reviews on multi-reference many-body perturbation theory and its application on electronically excited states see e.g. Refs. 22, 23. Correlated second order methods for the calculation of excitation energies based on a single-reference treatment for electronic ground states can, however, be derived within the framework of coupled-cluster response theory. The idea behind response theory is to study a system exposed to time-dependent external (e.g. electric) fields and to derive from the response of the wavefunction or density the frequency-dependent properties of the system – for example polarizabilities and hyperpolarizabilities. The latter properties have singularities whenever a frequency of a field

^bSimilar as in configuration interaction theory, a truncation after single excitations (CCS) does not give a useful method for the calculation of ground state energies. As follows from the Brillouin theorem $\langle \hat{H}^{(1)} | \text{HF} \rangle = 0$, the cluster equations have then for a closed-shell or an unrestricted open-shell reference determinant the trivial solution $t_a^i = 0$ and the CCS energy becomes equal the HF energy.

becomes equal to the excitation energy of an allowed transition in the system. Thus, from the poles of frequency-dependent properties one can identify the excitation energies.

Consider a quantum mechanical system described in the unperturbed limit by the time-independent Hamiltonian^c $\hat{H}^{(0)}$ which is perturbed by a time-dependent potential:

$$\hat{H}(t, \epsilon) = \hat{H}^{(0)} + \hat{V}(t, \epsilon) . \quad (10)$$

We assume that the perturbation $\hat{V}$ can be expanded as a sum over monochromatic Fourier components

$$\hat{V}(t, \epsilon) = \sum_j \hat{V}_j \epsilon_j e^{-i\omega_j t} , \quad (11)$$

where $\hat{V}_j$ are hermitian, time-independent one-electron operators (e.g. for an electric field the dipole operator), t the time and ϵ_j are the amplitudes of the associated field strengths. Then the full time-dependent wavefunction of the system, i.e. the solution to the time-dependent Schrödinger equation, can be expanded as a power series in the field strengths as

$$\Psi(t) = \underbrace{\left[\Psi^{(0)} + \sum_j \Psi_j^{(1)}(\omega_j) \epsilon_j e^{-i\omega_j t} + \dots \right]}_{\text{phase-isolated wavefunction } \tilde{\Psi}} e^{-i \int_{t_0}^t dt' \langle \tilde{\Psi}(t) | \hat{H}(t', \epsilon) - i \frac{\partial}{\partial t'} | \tilde{\Psi}(t) \rangle} , \quad (12)$$

and an expectation value for an operator $\hat{\mu}$ as

$$\langle \mu \rangle(t) = \langle \tilde{\Psi}(t) | \hat{\mu} | \tilde{\Psi}(t) \rangle = \mu^{(0)} + \sum_j \langle \langle \mu; V_j \rangle \rangle_{\omega_j} \epsilon_j e^{-i\omega_j t} + \dots \quad (13)$$

For detailed reviews of modern response theory and its implementation for approximate wavefunction methods the interested reader is referred to Refs. 24–29. The important point for the calculation of excitation energies is that the poles in the response functions $\langle \langle \mu; V \rangle \rangle_{\omega}$ occur when ω becomes equal to an eigenvalue of the stability matrix of the employed electronic structure method for the unperturbed system. The stability matrix contains the derivatives of the residuals of the equations which determine the wavefunction parameters with respect to these parameters. For Hartree-Fock, multi-configurational self-consistent field (MC-SCF), density functional theory (DFT), configuration interaction (CI) or other methods which are variational in the sense that the wavefunction parameters are determined by minimization of the energy, the stability matrix is the so-called electronic Hesse matrix – the matrix of the second derivatives of the energy with respect to the wavefunction parameters. For coupled-cluster methods the cluster amplitudes are determined by the cluster equations (6). Arranging the residuals in a vector function

$$\Omega_{\mu_i}(t_{\nu_i}) = \langle \mu_i | \exp(-\hat{T}) \hat{H} \exp(\hat{T}) | \text{HF} \rangle , \quad (14)$$

the stability matrix is given by the Jacobi matrix

$$\mathbf{A}_{\mu_i \nu_j} = \left. \frac{d\Omega_{\mu_i}}{dt_{\nu_j}} \right|_{\epsilon=0} = \langle \mu_i | \exp(-\hat{T}) [\hat{H}, \hat{\tau}_{\nu_j}] \exp(\hat{T}) | \text{HF} \rangle , \quad (15)$$

^cNote that $\hat{H}^{(0)}$ includes here the fluctuation potential in difference to Eq. (7), where the fluctuation potential $\hat{\Phi}$ has been the perturbation.

where $|\epsilon=0$ indicates that the derivatives are taken for the unperturbed system, i.e. at zero field strengths. In configuration interaction theory the stability matrix becomes the matrix representation of the reduced Hamiltonian $\hat{H} - E_0$ (where E_0 is the ground state energy) in the space orthogonal to the electronic ground state.^d In coupled-cluster theory this matrix representation is obtained in a similarity transformed basis^e.

In this way excitation energies can in principle be derived for any electronic structure method. However, to obtain physical meaningful and accurate results, the method has to fulfill certain requirements. For example from the equations for the amplitudes in MP2, Eq. (9), one obtains a Jacobi matrix which gives only excitation energies corresponding to double excitations and these would be equal to the orbital energy differences in the denominator of the amplitudes. The two most important requirements are firstly, that there must be a one-to-one correspondence between the parameters of the wavefunction and at least the investigated part of the spectrum of the Hamiltonian. This requires methods which determine the time-dependent variables by a single set of equations, as e.g. time-dependent Hartree-Fock (HF-SCF), density functional theory (DFT) or multi-configuration self-consistent field (MC-SCF, CAS-SCF, or RAS-SCF), but not a time-dependent configuration interaction (CI) treatment on top of a time-dependent HF-SCF calculation. For this reason the coefficients of the Hartree-Fock orbitals are also above in Eqs. (14) and (15) not considered as parameters of the time-dependent wavefunction, since this second set of variables in the time-dependent problem would lead to a second set of eigenvalues corresponding to single excited states, additionally to the one obtained from the parameterization through the singles cluster amplitudes. Instead, the time-dependent wavefunction is in coupled-cluster response theory usually constructed using the (time-independent) orbitals of the unperturbed system with time-dependent cluster amplitudes. Secondly, to obtain accurate results the stability matrix must also provide an accurate approximation of the those blocks of the Hamiltonian which are most important for the investigated states. For single excitations these are the singles-singles block $\mathbf{A}_{\mu_1\nu_1}$ and the off-diagonal blocks $\mathbf{A}_{\mu_2\nu_1}$ and $\mathbf{A}_{\mu_1\nu_2}$ next to it. With the usual single-reference coupled-cluster methods these blocks are described most accurately and therefore the excitation energies for single excitation dominated transitions are obtained with the highest accuracy, while excitation energies for double and higher excitations are usually considerably less accurate.

Already at the coupled-cluster singles (CCS) level (which for excitation energies is – in contrast to ground state calculations – not equivalent to Hartree-Fock, but to configuration interaction singles (CIS)), excitation energies for states dominated by single replacements of one spin-orbital in the Hartree-Fock reference determinant are obtained correctly through first order in the electron-electron interaction.

A second order method for excited states which accounts for the above requirements and takes over the accuracy of MP2 to excited states dominated by single excitations can be derived by approximating the cluster equations to lowest order in the fluctuation potential. But in difference to the derivation of MP2 in Eqs. (7) – (9) we allow in the Hamiltonian for an additional one-electron perturbation

$$\hat{H}(t) = \hat{F} + \hat{\Phi} + \hat{V}(t), \quad (16)$$

^dIn connection with CI and propagator methods (approximate) matrix representations of $\hat{H} - E_0$ are often also referred to as secular matrix.

^e $\langle \mu_i | \exp(-\hat{T})$ for the bra and $\exp(\hat{T}) | \mu_j \rangle$ for the ket states, where $|\mu_j\rangle = \hat{\tau}_{\mu_j} |\text{HF}\rangle$; for further details see e.g. Ref. 21

which can induce transitions to single excitations and has, as necessary in CC response theory, not been included in the Hartree-Fock calculation. Because of the latter, single excitation amplitudes contribute now to the cluster operator already in zeroth order in the fluctuation potential, $\hat{\Phi}$, and in first order $\hat{T}_1$ and $\hat{T}_2$ both contribute to the wavefunction. Approximating the equations that determine these amplitudes to second (singles) and first order (doubles) one obtains the equations for the approximate coupled-cluster model CC2^{30,31}:

$$0 = \langle_i^a | [\hat{H}, \hat{T}_2] + \hat{\hat{H}} | \text{HF} \rangle, \quad (17)$$

$$0 = \langle_{ij}^{ab} | [\hat{F}, \hat{T}_2] + \hat{\hat{H}} | \text{HF} \rangle, \quad (18)$$

where a similarity transformed Hamiltonian $\hat{\hat{H}} = \exp(-\hat{T}_1) \hat{H} \exp(\hat{T}_1)$ has been introduced to obtain a compact notation. In difference to MP2 the equations for CC2 have to be solved iteratively because of the coupling introduced by $\hat{T}_1$. The ground state energy obtained from CC2

$$E_{\text{CC2}} = \langle \text{HF} | \hat{\Phi} (\hat{T}_2 + \frac{1}{2} \hat{T}_1 \hat{T}_1) | \text{HF} \rangle, \quad (19)$$

is, as for MP2, (only) correct through second order in the fluctuation potential^f, but it leads to a Jacobi matrix with the singles-singles block $\mathbf{A}_{\mu_1 \nu_1}$ correct through second order and the off-diagonal blocks $\mathbf{A}_{\mu_1 \nu_2}$ and $\mathbf{A}_{\mu_2 \nu_1}$ correct through first-order in the fluctuation potential, while the doubles-doubles $\mathbf{A}_{\mu_2 \nu_2}$ block is approximated by the zeroth-order term:

$$\mathbf{A}^{\text{CC2}} = \begin{pmatrix} \langle_i^a | [\hat{\hat{H}} + [\hat{H}, \hat{T}_2]], \hat{\tau}_k^c | \text{HF} \rangle & \langle_i^a | [\hat{\hat{H}}, \hat{\tau}_{kl}^{cd}] | \text{HF} \rangle \\ \langle_{ij}^{ab} | [\hat{\hat{H}}, \hat{\tau}_k^c] | \text{HF} \rangle & \langle_{ij}^{ab} | [\hat{\hat{F}}, \hat{\tau}_{kl}^{cd}] | \text{HF} \rangle \end{pmatrix}. \quad (20)$$

CC2 is the computational simplest iterative coupled-cluster model which gives single excitation energies which are correct through second order. Through the similarity transformed Hamiltonian $\hat{\hat{H}} = \exp(-\hat{T}_1) \hat{H} \exp(\hat{T}_1)$ the Jacobi matrix in Eq. (20) includes, however, also some higher-order terms, since for the unperturbed system the single excitation amplitudes t_{μ_1} contribute only in second- and higher orders to the ground state wavefunction.^g Excluding these terms and replacing the doubles amplitudes by the first-order amplitudes, Eq. (9), from which the MP2 energy is calculated, one obtains the Jacobi matrix of the CIS(D_∞) approximation³⁵, an iterative variant of the perturbative doubles correction³⁶ CIS(D) to CIS (or CCS):

$$\mathbf{A}^{\text{CIS(D}_\infty\text{)}} = \begin{pmatrix} \langle_i^a | [\hat{\hat{H}} + [\hat{H}, \hat{T}_2]], \hat{\tau}_k^c | \text{HF} \rangle & \langle_i^a | [\hat{\hat{H}}, \hat{\tau}_{kl}^{cd}] | \text{HF} \rangle \\ \langle_{ij}^{ab} | [\hat{\hat{H}}, \hat{\tau}_k^c] | \text{HF} \rangle & \langle_{ij}^{ab} | [\hat{\hat{F}}, \hat{\tau}_{kl}^{cd}] | \text{HF} \rangle \end{pmatrix}. \quad (21)$$

This Jacobian contains the minimal number of terms required to obtain the excitation energies for single replacement dominated transitions correct through second order. However,

^fTherefore, CC2 does in general not describe ground state energies, structures, or properties more accurately than MP2. Its advantage upon MP2 is that, combined with coupled-cluster response theory, it can (in contrast to the latter) be applied successfully to excited states.

^gWe assume here that the Brillouin theorem is fulfilled and thus the occupied/virtual block of the Fock matrix vanishes. This holds for closed-shell and unrestricted open-shell Hartree-Fock reference states. For a discussion of additional terms that need to be accounted for in restricted open-shell SCF based calculations we refer e.g. to Refs. 32–34.

it is not possible to construct a coupled-cluster model which leads exactly to such a Jacobi matrix.

The computational savings of CIS(D_∞) compared to CC2 are rather limited³⁵ and CC2 has, as a member of the hierarchy of coupled-cluster methods CCS, CC2, CCSD, CC3, CCSDT, ... certain conceptual advantages. The Jacobi matrix of the CIS(D_∞) approximation may, however, be used as starting point to derive the perturbative doubles correction CIS(D) to the CIS (or CCS) excitation energies³⁵:

$$\omega^{(D)} = \sum_{\mu_1 \nu_1} E_{\mu_1}^{\text{CIS}} \left[\mathbf{A}_{\mu_1 \nu_1}^{\text{CIS(D}_\infty\text{)}} - \mathbf{A}_{\mu_1 \nu_1}^{\text{CIS}} + \sum_{\kappa_2} \frac{\mathbf{A}_{\mu_1 \kappa_2}^{\text{CIS(D}_\infty\text{)}} \mathbf{A}_{\kappa_2 \nu_1}^{\text{CIS(D}_\infty\text{)}}}{\omega^{\text{CIS}} - \epsilon_{\kappa_2}} \right] E_{\nu_1}^{\text{CIS}} \quad (22)$$

or

$$\omega^{\text{CIS(D)}} = \omega^{\text{CIS}} + \omega^{(D)} = \sum_{\mu_1 \nu_1} E_{\mu_1}^{\text{CIS}} \left[\mathbf{A}_{\mu_1 \nu_1}^{\text{CIS(D}_\infty\text{)}} + \sum_{\kappa_2} \frac{\mathbf{A}_{\mu_1 \kappa_2}^{\text{CIS(D}_\infty\text{)}} \mathbf{A}_{\kappa_2 \nu_1}^{\text{CIS(D}_\infty\text{)}}}{\omega^{\text{CIS}} - \epsilon_{\kappa_2}} \right] E_{\nu_1}^{\text{CIS}} \quad (23)$$

where ϵ_{κ_2} contains the orbital energy difference for a double excitation, $\epsilon_{ab}^{ij} = \epsilon_a - \epsilon_i + \epsilon_b - \epsilon_j$.

Another second order method for excited states which is related to CC2 and CIS(D) is the so-called algebraic diagrammatic construction through second order, ADC(2).^{37,38} The secular matrix of ADC(2) is just the symmetric part of $\mathbf{A}^{\text{CIS(D}_\infty\text{)}}$:

$$\mathbf{A}^{\text{ADC(2)}} = \frac{1}{2} \mathbf{A}^{\text{CIS(D}_\infty\text{)}} + \frac{1}{2} \left(\mathbf{A}^{\text{CIS(D}_\infty\text{)}} \right)^\dagger, \quad (24)$$

which leads to some conceptual and also computational simplifications e.g. in the calculation of derivatives (gradients!) since the left and right eigenvectors of a symmetric matrix are identical, while for the non-symmetric Jacobi matrices of CC2 and CIS(D_∞) left and right eigenvectors differ. Both eigenvectors are needed for the calculation of derivatives. Other second order methods for excited states are the second order polarization propagator approach,^{39,40} SOPPA and the perturbative doubles correction,⁴¹ RPA(D), to time-dependent Hartree-Fock, which for excitation energies is also known as the random phase approximation (RPA). The latter method can also be understood as a non-iterative approximation to SOPPA, similar as CIS(D) is a non-iterative approximation to CIS(D_∞). The relation of RPA(D) and SOPPA to the single-reference coupled-cluster response methods is somewhat more difficult, since these methods are members of a different hierarchy of methods (with RPA (TDHF) as first-order model) which is related to the so-called orbital-optimized coupled-cluster (OCC) methods^{42,43}. Therefore, these methods will not be discussed in detail in the following, but we note that the same concepts (doubles amplitude-direct formulation and RI-approximation) can be applied to reduce also for these the computational costs to the same extent as for CC2, ADC(2), CIS(D_∞), and CIS(D).

2.1 Doubles Amplitude-Direct Formulation of Second Order Methods

An important feature of second order methods or approximate doubles methods, as one might also call them, is that an explicit storage (in RAM or on disk) of complete sets of double excitation amplitudes can be avoided similar as the storage of triples amplitudes is avoided in the approximate triples methods CCSD(T), CCSDT-1, CCSDR(3), or CC3.⁴⁴⁻⁴⁷

This is important for applications on large molecules since similar as for the approximate triples methods the storage of the amplitudes would prohibit large-scale applications simply by a storage space or I/O bottleneck.

For example, the MP2 energy can be calculated without storing the double excitation amplitudes using the following scheme^h:

```

do i = 1, nocc
  do j = i, nocc
    do a = 1, nvirt
      do b = b, nvirt
         $t_{ab}^{ij} = (ia|jb)/(\epsilon_i - \epsilon_a + \epsilon_j - \epsilon_b)$ 
         $E_{MP2} = E_{MP2} + (2 - \delta_{ij})\{2(ia|jb) - (ia|jb)\}t_{ab}^{ij}$ 
      end do
    end do
  end do
end do

```

In a similar way also the equations for the doubles amplitudes in CC2 can – for given singles amplitudes t_a^i – immediately be inverted to

$$t_{ab}^{ij} = (ai|\tilde{b}j)/(\epsilon_i - \epsilon_a + \epsilon_j - \epsilon_b) \quad (25)$$

where the similarity transformation with $\exp(\hat{T}_1)$ has been included in the AO-to-MO transformation for the modified two-electron integrals

$$(ai|\tilde{b}j) = \sum_{\alpha} \Lambda_{\alpha a}^p \sum_{\beta} \Lambda_{\beta i}^h \sum_{\gamma} \Lambda_{\gamma b}^p \sum_{\delta} \Lambda_{\delta j}^h (\alpha\beta|\gamma\delta) \quad (26)$$

with $\Lambda_{\alpha a}^p = C_{\alpha a} - \sum_k C_{\alpha k} t_a^k$ and $\Lambda_{\alpha i}^h = C_{\alpha i} + \sum_c C_{\alpha c} t_c^i$. Inserting Eq. (25) into the equation for the singles amplitudes, Eq. (17), gives a set of effective equations for the CC2 singles amplitudes, which reference the doubles amplitudes t_{ab}^{ij} only as intermediates, which can be calculated and contracted with one- and two-electron integrals “on-the-fly” without storing a complete set of these amplitudes on disk:

```

do i = 1, nocc
  do j = 1, nocc
    do a = 1, nvirt
      do b = 1, nvirt
         $t_{ab}^{ij} = (ai|\tilde{b}j)/(\epsilon_i - \epsilon_a + \epsilon_j - \epsilon_b)$ 
         $\Omega_{ci} = \Omega_{ci} + \sum_{abj} (2t_{ab}^{ij} - t_{ba}^{ij})(jb|\tilde{c}a)$ 
         $\Omega_{ak} = \Omega_{ak} - \sum_{bij} (2t_{ab}^{ij} - t_{ba}^{ij})(jb|\tilde{i}k)$ 
         $\vdots$ 
      end do
    end do
  end do
end do

```

^hThe explicit formulas given here and below are for a closed-shell restricted Hartree-Fock reference determinant.

To avoid the storage of doubles amplitudes is even more important for excited states, since in this case else doubles contributions to eigen- or trial-vectors would have to be stored for several simultaneously solved eigenvalues and a number of iterations. An explicit reference to the doubles part of eigen- or trial-vectors during the solution of the eigenproblem can for the approximate doubles methods be removed by exploiting the particular structure of the Jacobi or secular matrices of these methods, in which the doubles-doubles block is in the canonical orbital basis diagonal with the diagonal elements equal to SCF orbital energy differences:

$$\begin{pmatrix} \mathbf{A}_{\mu_1\nu_1} & \mathbf{A}_{\mu_1\nu_2} \\ \mathbf{A}_{\mu_2\nu_1} & \delta_{\mu_2\nu_2}\epsilon_{\nu_2} \end{pmatrix} \begin{pmatrix} E_{\nu_1} \\ E_{\nu_2} \end{pmatrix} = \omega \begin{pmatrix} E_{\nu_1} \\ E_{\nu_2} \end{pmatrix}. \quad (27)$$

The doubles part of the eigenvectors is thus related to the singles part and the eigenvalue through the equation

$$E_{\mu_2} = \frac{\sum_{\nu_1} \mathbf{A}_{\mu_2\nu_1} E_{\nu_1}}{\omega - \epsilon_{\mu_2}}. \quad (28)$$

which allows to partition the linear eigenvalue problem in the space of singles and doubles replacements as an effective eigenvalue problem in the space of only the single excitations:

$$\sum_{\nu_1} \left[\mathbf{A}_{\mu_1\nu_1} + \sum_{\kappa_2} \frac{\mathbf{A}_{\mu_1\kappa_2} \mathbf{A}_{\kappa_2\nu_1}}{\omega - \epsilon_{\kappa_2}} \right] E_{\nu_1} = \sum_{\nu_1} \mathbf{A}_{\mu_1\nu_1}^{eff}(\omega) E_{\nu_1} = \omega E_{\mu_1}. \quad (29)$$

The last equation is, however, in difference to Eq. (27) a nonlinear eigenvalue problem because the effective Jacobi matrix $\mathbf{A}_{\mu_1\nu_1}^{eff}(\omega)$ depends on the eigenvalue ω , which is itself first known when the equation has been solved. But with iterative techniques this eigenvalue problem can be solved almost as efficiently as the original linear eigenvalue problem and the elimination of the need to store the doubles part of solution or trial-vectors more than compensates this complication.⁴⁸

To apply these iterative techniques for the solution of large-scale eigenvalue problems one needs to implement matrix vector products of the form

$$\sigma_{\mu_1}(\omega, b_{\nu_1}) = \sum_{\nu_1} \mathbf{A}_{\mu_1\nu_1}^{eff}(\omega) b_{\nu_1} = \sum_{\nu_1} \mathbf{A}_{\mu_1\nu_1} b_{\nu_1} + \sum_{\kappa_2} \mathbf{A}_{\mu_1\kappa_2} \frac{\sum_{\nu_1} \mathbf{A}_{\kappa_2\nu_1} b_{\nu_1}}{\omega - \epsilon_{\kappa_2}}. \quad (30)$$

Note the similarity of the quotient in the last term with the expression in Eq. (28). For CC2 this term becomes

$$b_{ab}^{ij} = \frac{1}{\epsilon_{iajb}} \sum_{ck} \mathbf{A}_{iajb,ck} b_c^k = \frac{\sum_{ck} \langle ij | [\hat{H}, \hat{\tau}_c^k] | \text{HF} \rangle b_c^k}{\epsilon_i - \epsilon_a + \epsilon_j - \epsilon_b + \omega} = \frac{2(a\bar{i}|bj) - (b\bar{i}|aj)}{\epsilon_i - \epsilon_a + \epsilon_j - \epsilon_b + \omega}, \quad (31)$$

with the modified MO electron repulsion integrals

$$(a\bar{i}|bj) = \hat{P}_{ab}^{ij} \sum_{\alpha\beta} \left(\bar{\Lambda}_{\alpha a}^p \Lambda_{\beta i}^h + \Lambda_{\alpha a}^p \bar{\Lambda}_{\beta i}^h \right) \sum_{\gamma\delta} \Lambda_{\gamma b}^p \Lambda_{\delta j}^h (\alpha\beta|\gamma\delta), \quad (32)$$

where $\bar{\Lambda}_{\alpha a}^p = -\sum_k C_{\alpha k} b_a^k$, $\bar{\Lambda}_{\alpha i}^h = +\sum_c C_{\alpha c} b_i^c$ and $\hat{P}_{ab}^{ij}$ is a symmetrization operator defined through $\hat{P}_{ab}^{ij} f_{ia,jb} = f_{ia,jb} + f_{jb,ia}$. The linear transformation in Eq. (30) can thus be calculated using a similar algorithm as for the residual of the ground state cluster

equations without storing any doubles vectors:

```

do i = 1, nocc
  do j = 1, nocc
    do a = 1, nvirt
      do b = 1, nvirt
         $b_{ab}^{ij} = (ai|bj) / (\epsilon_i - \epsilon_a + \epsilon_j - \epsilon_b + \omega)$ 
         $\sigma_{ci} = \sigma_{ci} + \sum_{abj} (2b_{ab}^{ij} - b_{ba}^{ij}) (jb|ca)$ 
         $\vdots$ 
         $t_{ab}^{ij} = (ai|bj) / (\epsilon_i - \epsilon_a + \epsilon_j - \epsilon_b)$ 
         $\sigma_{ai} = \sigma_{ai} + \sum_{bj} (2t_{ab}^{ij} - t_{ba}^{ij}) \sum_{ck} [2(jb|kc) - (jc|kb)] b_{ck}$ 
      end do
    end do
  end do
end do

```

The fact that the doubles amplitudes of CC2 are determined by the singles amplitudes through Eqs. (25) and (26) and reduce for $t_{\mu_1} \rightarrow t_{\mu_1}^{(1)} = 0$ to the first-order amplitudes of MP2, opens a simple possibility to implement CIS(D_∞) and CIS(D) as approximations to CC2. Considering the effective Jacobi matrix, Eq. (29), as a functional of the singles amplitudes $\mathbf{A}^{eff}(t_{\mu_1}, \omega)$ one obtains the connection:

$$\begin{aligned}
\text{CC2} & : \quad \sum_{\nu_1} \mathbf{A}_{\mu_1 \nu_1}^{eff}(t_{\kappa_1}^{\text{CC2}}, \omega) E_{\nu_1} = \omega E_{\mu_1} \\
\text{CIS(D}_{\infty}) & : \quad \sum_{\nu_1} \mathbf{A}_{\mu_1 \nu_1}^{eff}(t_{\kappa_1}^{(1)}, \omega) E_{\nu_1} = \omega E_{\mu_1} \\
\text{CIS(D)} & : \quad \omega^{\text{CIS(D)}} = \sum_{\mu_1 \nu_1} E_{\mu_1}^{\text{CIS}} \mathbf{A}_{\mu_1 \nu_1}^{eff}(t_{\kappa_1}^{(1)}, \omega) E_{\nu_1}^{\text{CIS}}
\end{aligned}$$

The attentive reader has probably observed that the partitioned, doubles amplitude-direct formulation for second order methods – although it removes the need to store complete sets of any doubles amplitudes – does alone not reduce much the storage requirements of these methods: the calculation of the doubles amplitudes requires the electron repulsion integrals (ERIs) in the (modified) MO basis, which are obtained through four-index transformations from the AO integrals, as e.g. in Eqs. (26) and (32). Efficient implementations of such transformations require the storage of an array with half-transformed integrals of the size of $\frac{1}{2} O^2 N^2$, where O is the number of occupied and N the number of atomic orbitals, which is even slightly more than needed for the doubles amplitudes. For CC2 and also for the other second order methods for excited states and in the calculation of gradients for the MP2 energies, the doubles amplitudes need to be contracted in addition with two-electron integrals with three occupied or virtual indices, $(ai|jk)$ and $(ai|bc)$, which within the schemes sketched above would give rise to even larger storage requirements. The problem can be solved with the resolution-of-the-identity approximation for electron repulsion integrals.

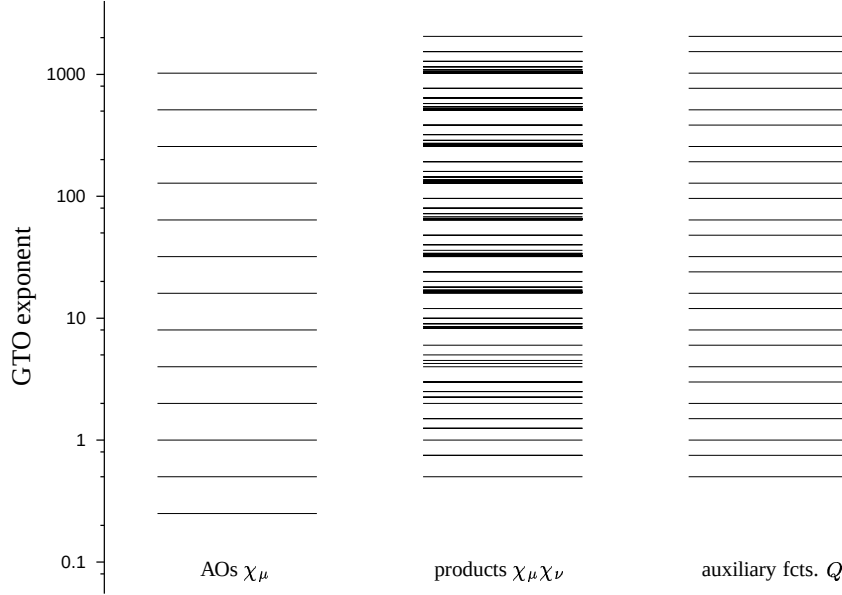


Figure 1. The left column shows exponents α_μ of an even-tempered (13s) atomic Gaussian type orbital (GTO) basis $\chi_\mu(r) = \exp(-r^2\alpha_\mu)$ and the column in the middle the exponents of all 169 overlap Gaussian functions resulting on the same atom from the products $\chi_\mu\chi_\nu$. The right column shows the exponents of an even-tempered (25s) auxiliary basis $Q(r) = \exp(-r^2\alpha_Q)$ set which could be used to expand these products.

3 The Resolution-of-the-Identity Approximation for ERIs

The main idea behind the resolution-of-the-identity approximation^{49–56} for electron repulsion integrals can be sketched as follows: With increasing atomic orbital basis sets the products of AOs appearing for the electrons 1 and 2 in the expression for the four-index two-electron integrals,

$$(\alpha\beta|\gamma\delta) = \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \chi_\alpha(\vec{r}_1) \chi_\beta(\vec{r}_1) \frac{1}{r_{12}} \chi_\gamma(\vec{r}_2) \chi_\delta(\vec{r}_2) d\tau_1 d\tau_2, \quad (33)$$

will soon become (numerically) highly linear dependent and thus it should be possible to expand these products which good accuracy in a basis set of auxiliary functions Q ,

$$\chi_\alpha(\vec{r}_1) \chi_\beta(\vec{r}_1) \approx \sum_Q Q(\vec{r}_1) c_{Q,\alpha\beta} \quad (34)$$

with a dimension much smaller than that of the original product space, $N(N+1)/2$, as illustrated in Fig. 1 for an atom with only s -type functions. The coefficients $c_{Q,\alpha\beta}$ can be determined through a least square procedure. Defining the remaining error in the expansion of an orbital pair

$$R_{\alpha\beta}(\vec{r}_1) = \chi_\alpha(\vec{r}_1) \chi_\beta(\vec{r}_1) - \sum_Q Q(\vec{r}_1) c_{Q,\alpha\beta}, \quad (35)$$

the quadratic error in the coulomb repulsion integrals $(\alpha\beta|\gamma\delta)$ can be written as

$$(R_{\alpha\beta}|R_{\gamma\delta}) = \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} R_{\alpha\beta}(\vec{r}_1) \frac{1}{r_{12}} R_{\gamma\delta}(\vec{r}_2) d\tau_1 d\tau_2 \quad (36)$$

and fulfill the Schwartz inequality

$$(R_{\alpha\beta}|R_{\gamma\delta}) \leq \sqrt{(R_{\alpha\beta}|R_{\alpha\beta})} \sqrt{(R_{\gamma\delta}|R_{\gamma\delta})} . \quad (37)$$

Minimization of $(R_{\alpha\beta}|R_{\alpha\beta})$ with respect to the expansion coefficients c leads to the linear equation:

$$\frac{d}{dc_{Q,\alpha\beta}} (R_{\alpha\beta}|R_{\alpha\beta}) = 0 \quad \Leftrightarrow \quad (R_{\alpha\beta}|Q) = 0 \quad \Leftrightarrow \quad (\alpha\beta|Q) - \sum_P c_{P,\alpha\beta} (P|Q) = 0 \quad (38)$$

with

$$(P|Q) = \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} P(\vec{r}_1) \frac{1}{r_{12}} Q(\vec{r}_2) d\tau_1 d\tau_2 , \quad (39)$$

$$(\alpha\beta|Q) = \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \chi_\alpha(\vec{r}_1) \chi_\beta(\vec{r}_1) \frac{1}{r_{12}} Q(\vec{r}_2) d\tau_1 d\tau_2 . \quad (40)$$

Arranging the two-center integrals in a matrix $V_{PQ} = (P|Q)$ the expansion coefficients can be expressed as

$$c_{Q,\alpha\beta} = \sum_P (\alpha\beta|P) [V^{-1}]_{PQ} , \quad (41)$$

and one obtains for the four-index coulomb integrals the approximation

$$(\alpha\beta|\gamma\delta) \approx \sum_{QP} (\alpha\beta|Q) [V^{-1}]_{QP} (P|\gamma\delta) . \quad (42)$$

We have above derived Eq. (42) as result of a least square fitting procedure for the overlap densities $\chi_\alpha(\vec{r})\chi_\beta(\vec{r})$, which is why this approximation is also known as “density fitting”^{19,20}. Eq. (42) can be compared with the expression for an (approximate) resolution of the identity for square integrable functions in three-dimensional space,

$$\mathbf{1} \approx \sum_{QP} |Q\rangle [S^{-1}]_{QP} \langle P| \quad \text{with} \quad S_{PQ} = \int_{\mathbb{R}^3} Q(\vec{r}) P(\vec{r}) d\tau , \quad (43)$$

applied to four-center overlap integrals

$$\int_{\mathbb{R}^3} \chi_\alpha(\vec{r}) \chi_\beta(\vec{r}) \chi_\gamma(\vec{r}) \chi_\delta(\vec{r}) d\tau = \langle \alpha\beta|\delta\gamma \rangle \approx \sum_{QP} \langle \alpha\beta|Q \rangle [S^{-1}]_{QP} \langle P|\delta\gamma \rangle . \quad (44)$$

We see that Eq. (42) can alternatively be viewed as an (approximate) resolution of the identity in a Hilbert space where the coulomb operator $1/r_{12}$ is used to define the scalar product as in Eqs. (33) and (36). This approximation has thus all properties expected from a resolution-of-the-identity or basis set approximation as e.g. that the norm of the error in the expansion $||R_{\alpha\beta}|| = (R_{\alpha\beta}|R_{\alpha\beta})$ will always decrease with an extension of the auxiliary basis and that the approximation becomes exact in the limit of a complete auxiliary basis set $\{Q\}$.

It is important to note that the resolution-of-the-identity approximation does not – or at least not in general – reduce the computational costs for the calculation of AO four-index electron repulsion integrals, since the right hand side of Eq. (42) is more complicated to evaluate than the left hand side. A reduction of the computational costs is only achieved if the decomposition of the four-index integrals into three- and two-index intermediates, provided by this approximation, can be exploited to simplify contractions of the AO coulomb integrals with other intermediates.

A common bottleneck of all second order correlation methods (for ground and excited states) is the four-index transformation of the AO ERIs $(\alpha\beta|\gamma\delta)$ to ERIs in a molecular orbital basis (possibly modified as in Eq. (26) or (32)) with two occupied and two virtual indices:

$$(ai|bj) = \sum_{\alpha} C_{\alpha a} \sum_{\gamma} C_{\gamma b} \sum_{\beta} C_{\beta i} \sum_{\delta} C_{\delta j} (\alpha\beta|\gamma\delta). \quad (45)$$

Efficient algorithms for this transformation require a number of floating point multiplications that scales for the individual partial transformations with $\frac{1}{2}ON^4 + \frac{1}{2}O^2N^3 + \frac{1}{2}O^2VN^2 + \frac{1}{2}O^2V^2N$ (ignoring possible sparsities in the integrals or coefficients) and, as already pointed out above, disc space in the order of $\frac{1}{2}O^2N^2$.

Using the resolution-of-the-identity approximation, the four-index integrals in the MO basis can be obtained as

$$(ai|bj) \approx \sum_P B_{P,ai} B_{P,bj} \quad (46)$$

with

$$B_{P,ai} = \sum_Q [V^{-1/2}]_{PQ} \sum_{\alpha} C_{\alpha a} \sum_{\beta} C_{\beta i} (Q|\alpha\beta) \quad (47)$$

which requires only $ON^2N_x + OVN N_x + OVN_x^2 + \frac{1}{2}O^2V^2N_x$ floating point multiplications and memory or disc space in the order $ON N_x$. With auxiliary basis sets optimized^{54,57,58} for the application in second order methods N_x is typically $2-4 \times N_x$. Assuming that $O \ll V \approx N$ (usually given in correlated calculations), one finds that the number of floating point operations is by the RI approximation reduced by a factor of $\approx (N/O + 3)N/N_x$. With doubly polarized or correlation-consistent triple- ζ basis sets (e.g. TZVPP or cc-pVTZ) as often used with MP2 or CC2, the RI approximation typically reduces the CPU time for the calculation of the $(ai|bj)$ integrals by more than an order of magnitude. Some typical examples for MP2 calculations for the ground state correlation energy are given in Table 1. These also demonstrate how the reduction in CPU time obtained with the RI approximation increases with the size of the orbital basis set. An important point for calculations on weakly bonded (i.e. hydrogen-bridged or van der Waals) systems is that the efficiency of the integral prescreening, which is important for the performance of conventional implementations using 4-index AO ERIs, diminishes if diffuse functions are included in the basis set. For weakly bonded complexes such diffuse functions are, however, needed for an accurate description of the long range electrostatic, exchange-correlation, and dispersion interactions. As seen at the calculations for benzene with the QZVPP and the aug-cc-pVQZ basis, RI-MP2 calculations are much less sensitive to such effects: while the CPU time for the conventional MP2 calculation increases from

Table 1. Comparison of elapsed wall-clock timings for RI-MP2 vs. conventional integral-direct MP2 energy calculations (# fcts. is the number of basis functions and # e^- the number of correlated electrons, T_{MP2} timings obtained with the mpgrad code of the TURBOMOLE package⁵⁹).

molecule	basis	# fcts.	# e^-	T_{MP2}	T_{RI-MP2}
benzene ^a	QZVPP	522	30	28 min	24 sec
benzene ^a	aug-cc-pVTZ	756	30	3.8 h	1.2 min
Fe(CO) ₅ ^a	QZVPP	670	66	11.3 h	8.7 min
Fe(C ₅ H ₅) ₂ ^a	QZVPP	970	66	843 h	45 min
C ₆₀ ^{a,b}	cc-pVTZ	1800	240	112 h	171 min
Calix[4]arene ^{b,c}	cc-pVTZ	1528	184	39.3 h	5.6 h

^a RI-MP2 timings for `ricc2` code of the TURBOMOLE package⁵⁹; ^b from Ref. 60;

^c RI-MP2 timings for `rimp2` code of the TURBOMOLE package⁵⁹;

QZVPP to aug-cc-pVQZ by more than a factor of 8, the CPU time needed for the RI-MP2 calculation increases only by a factor of 3.

However, for large scale applications at least as important is that the scaling of the storage requirements in the calculation of the integrals $(ai|bj)$ with the system size is reduced to $\mathcal{O}(ONN_x)$. In combination with the doubles amplitude-direct formulation outlined in the previous subsection, the RI approximation completely removes the need to store any intermediates larger than $\mathcal{O}(ONN_x)$ on disc or in memory. For example the MP2 ground state energy can now be calculated using the following algorithm:

```

precompute  $B_{Q,ai}$ 
do  $i = 1, \text{nocc}$ 
  do  $j = 1, \text{nocc}$ 
     $I_{ab}^{ij} = \sum_Q B_{Q,ai} B_{Q,bj} \quad \forall a, b$  (matrix-matrix multiply)
    do  $a = 1, \text{nvirt}$ 
      do  $b = 1, \text{nvirt}$ 
         $t_{ab}^{ij} = I_{ab}^{ij} / (\epsilon_i - \epsilon_a + \epsilon_j - \epsilon_b)$ 
         $E_{MP2} = E_{MP2} + (2 - \delta_{ij}) \{ 2I_{ab}^{ij} - I_{ba}^{ij} \} t_{ab}^{ij}$ 
      end do
    end do
  end do
end do
end do

```

The reductions are even larger for CC2 and other second order methods for excited states and for the $\mathcal{O}(N^5)$ -scaling steps in the calculation of MP2 gradients. It turns out that all contractions which involve other four-index integrals in the MO basis than those of $(ia|jb)$ -type, needed in second order methods, can with the decomposition given by Eq. (42) reformulated such that an explicit calculation of the four-index MO integrals can be avoided.

Together with the reduction in the CPU time the elimination of the storage bottleneck

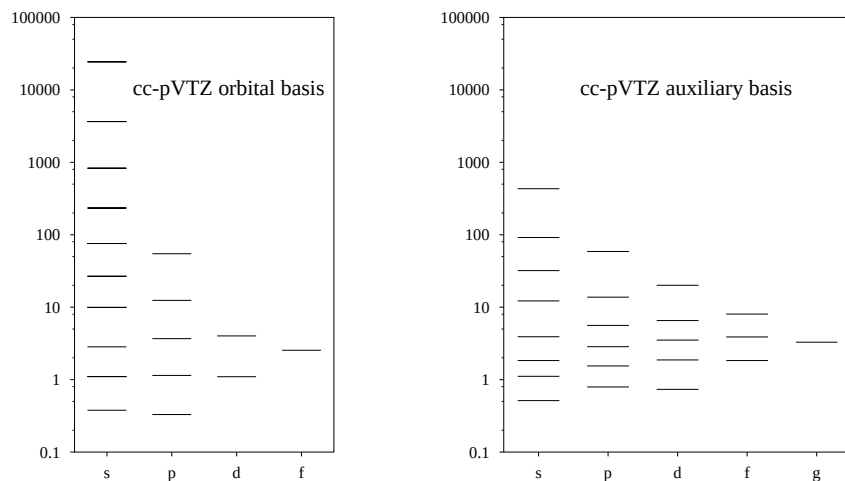


Figure 2. Exponents of the primitive GTOs in the cc-pVTZ orbital⁶¹ (on the left) and auxiliary^{57,58} (on the right) basis sets for the neon atom.

opened the possibility to apply MP2 and CC2 to much larger systems as was feasible with conventional implementations based on four-index AO ERIs. Since the steep increase of the computational costs with the basis set size is reduced by the RI approximation from $\mathcal{O}(N^4)$ to $\mathcal{O}(N^2 N_x)$ it is also easier than before to carry out such calculations with accurate basis sets, as needed to exploit fully the accuracy of MP2, CC2 or the other second order methods.

At this point it becomes necessary to ask what are the errors introduced by the RI approximation? As is obvious from the above discussion, the accuracy (but also the efficiency) of the RI approximation depends on the choice of the auxiliary basis sets. For a balanced treatment the auxiliary basis set should be optimized for the particular orbital basis used in the calculation. Firstly, because the orbital products that need to be well represented depend strongly on the orbital basis and, secondly, because the accuracy of the approximation should increase with increasing accuracy of the orbital basis to make sure that eventually a correct basis set limit will be obtained. To fully exploit the potential of the approximation it is advantageous to further “tune” the auxiliary basis set for the integrals most important in the employed electronic structure method. For second order methods these are, as shown above, $(ai|bj)$ -type integrals. The auxiliary basis functions are thus used to expand products of occupied with virtual molecular orbitals:

$$\phi_a(\vec{r})\phi_i(\vec{r}) \approx \sum_Q Q(\vec{r})c_{Q,ai} . \quad (48)$$

If we consider an atom, all products will be linear combinations of Gaussian type functions centered at the atom with angular momenta up to $l_{aux} = l_{orb} + l_{occ}$, where l_{orb} is the highest angular momentum included in the orbital basis set and l_{occ} the highest angular momentum of an occupied orbital. Also the range of exponents that should be covered by the auxiliary basis can be deduced from similar considerations, but it should be taken

into account that the importance of the orbital products $\phi_a \phi_i$ for electron correlation varies over orders of magnitudes. E.g., the contributions of core orbitals and similar those over very high lying tight virtual orbitals (sometimes referred to as “anti core” orbitals) is small because of large orbital energy denominators in the expression for the amplitudes. This limits the importance of tight functions in the auxiliary basis, in particular if a frozen core approximation is used and the core orbitals cannot at all contribute to the correlation treatment. In the other direction, the most diffuse exponent needed in the auxiliary basis set is bound by the exponent of any atomic orbital contributing significantly to an occupied orbital, irrespectively how diffuse functions are included in the basis set. A typical composition of an orbital basis and a respective auxiliary basis set of correlated calculations with a second order method is shown in Fig. 2 at the example of the cc-pVTZ basis sets for the neon atom.

It turns out that the above arguments, although strictly only valid for atoms, apply in practice usually also well to moleculesⁱ. Therefore, the auxiliary basis sets can be optimized once at the atoms for each orbital basis and then stored in a basis set library. On the TURBOMOLE web page⁵⁹ optimized auxiliary basis sets for correlated calculations with second order methods are available for several orbital basis sets including SVP⁶², TZVP⁶³, TZVPP⁵⁴, and QZVPP⁶⁴ and most of the correlation-consistent basis sets^{61, 65–70} (cc-pVXZ, aug-cc-pVXZ, cc-pwCVXZ, etc.). These have been optimized^{54, 57, 58} such that the RI error, i.e. the additional error introduced by the RI approximation, is for the ground state correlation energies (MP2 or CC2) about 2–3 orders of magnitudes smaller than the one-electron (orbital) basis set error of the respective orbital basis set. The correlation-consistent basis sets cc-pVXZ with $X = D, T, Q, \dots$ and the series SVP, TZVPP, QZVPP, $\dots$ constitute hierarchies that converge to the (valence) basis set limit and are thus a good example to demonstrate how orbital and auxiliary basis sets converge in parallel. Fig. 3 shows the results of an error analysis for the MP2 valence correlation energies for 72 molecules containing first and second row atoms (H, He, B–Ne, Al–Ar). The RI errors are somewhat larger for other properties than for ground state correlation energies, for which they have been optimized. In particular in response calculations for excited states the diffuse functions and also some other integral types become more important than they are for ground state calculations. But, still the RI error remains between one and two orders of magnitudes smaller than the orbital basis set error as is shown in Fig. 4 by an error analysis for RI-CC2 calculations on excited states with the aug-cc-pVTZ basis sets. Since the RI approximation is a basis set expansion approach the RI error is a smooth and usually extremely flat function of the coordinates. Therefore most of the error cancels out in the calculation of energy differences, as e.g. reaction enthalpies, and the errors in geometries are very small – typically a few 10^{-3} pm and, thus, usually below the convergence thresholds applied in geometry optimizations.

In summary, the major advantages of the resolution-of-the-identity approximation for the electron repulsion integrals for correlated second order methods are

- It allows efficient doubles amplitude-direct implementations and eliminates the need to store any $\mathcal{O}(\mathcal{N}^4)$ arrays in memory or on disc.

ⁱAn exception are the atoms with only s orbitals occupied in the ground state configuration, in particular H and Li, which in chemical bonds are often strongly polarized. For these atoms the auxiliary basis sets contain usually functions up to $l_{orb} + 1$ (instead of only l_{orb}) and are often optimized on small molecules.

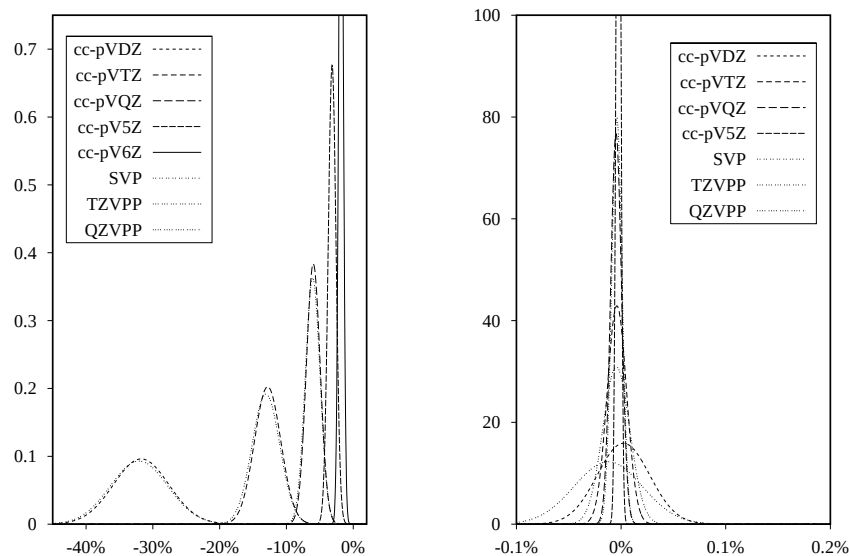


Figure 3. On the left: one-electron basis set errors in the MP2 valence correlation energy (in % of the estimated limiting value) shown as normalized Gaussian distributions determined from Δ and Δ_{std} for a test set of 72 small and medium sized molecules with row 1 (B–Ne) and row 2 (Al–Ar) atoms^{57,58}. On the right: error in the MP2 valence correlation energies due to the resolution-of-the-identity approximation for ERIs for the same test set^{57,58}. Note that the scales on the abscissa differ by about three orders of magnitude!

- The CPU time for the correlation treatment is reduced by about an order of magnitude and more.
- It is applicable in response calculations for excited states since it does not depend on the locality of any intermediates.

Another important point related to the elimination of the huge storage demands for $\mathcal{O}(\mathcal{N}^4)$ scaling intermediates (i.e. two-electron integrals or amplitudes) is that the parallelizability of these methods is improved since less data needs to be communicated between computer nodes participating in a parallel calculation. We will come back to this point in the next section.

4 Parallel Implementation of RI-MP2 and RI-CC2 for Distributed Memory Architectures

As discussed above, the time-determining steps in RI-MP2 and other second order methods implemented with the RI approximation are the computation of the electron repulsion integrals in the MO basis ($ia|jb$) and/or the double excitation amplitudes t_{ab}^{ij} and their contraction with integrals or other amplitudes to new intermediates, as for example

$$Y_{Q,ai} = \sum_{bj} t_{ab}^{ij} B_{Q,bj} . \quad (49)$$

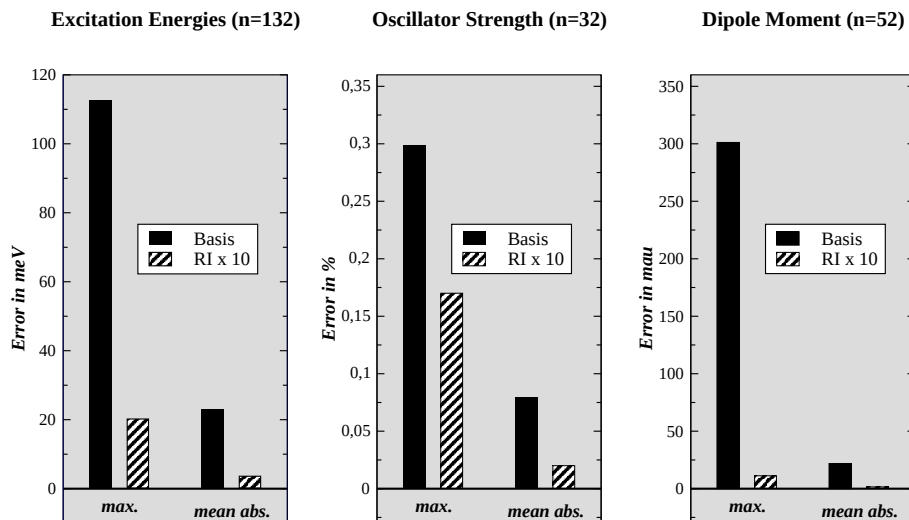


Figure 4. Mean and maximum of the one-electron orbital and the RI errors in RI-CC2 calculations for excited states with the aug-cc-pVTZ basis sets^{61,67,57}. On the left: errors in excitation energies for 132 states. In the middle: errors in the oscillator strengths for 32 states. On the right: errors in the dipole moments of 52 excited states. For the test sets used and the technical details see Ref. 71, from where the data has been taken.

Also for this step the computational costs increase as $\mathcal{O}(O^2 V^2 N_x)$. As described in Refs. 53, 71–74, $Y_{Q,ai}$ and all other intermediates calculated from t_{ab}^{ij} can efficiently be calculated in a loop over two indices for occupied orbitals with $\mathcal{O}(N^2)$ memory demands. The time-determining steps of RI-MP2 can thus efficiently parallelized over pairs of indices for occupied orbitals since these are common to all steps scaling with $\mathcal{O}(O^2 V^2 N_x)$ or $\mathcal{O}(O^2 V^3)$. An alternative could be pairs of virtual orbitals, but this would result in short loop lengths and diminished efficiency for medium sized molecules. A parallelization over auxiliary basis functions would require the communication of 4-index MO integrals between computer nodes, which would require high-performance networks. Such a solution would restrict the applicability of the program to high-end supercomputer architectures. TURBOMOLE, however, has been designed for low-cost PC clusters with standard networks (e.g. Fast Ethernet or Gigabit). Therefore we choose for the `ricc2` code a parallelization over pairs of occupied orbitals and accepted that this results in an implementation which will not be suited for massively parallel systems, since a good load balance between the participating CPUs will only be achieved for $O \gg n_{CPU}$ (vide infra).

A key problem for the parallelization of RI-MP2 and RI-CC2 is with this strategy the distribution of pairs of occupied orbitals (ij) over distributed memory nodes such that

- the symmetry of $(ia|jb)$ with respect to permutation of $ia \leftrightarrow jb$ can still be exploited
- the demands on the individual computer nodes for accessing and/or storing the three-index intermediates $B_{Q,ai}$ and $Y_{Q,ai}$ are as low as possible.

To achieve this, we partition the occupied orbitals into n_{CPU} batches $\mathcal{I}_m$ of (as much as possible) equal size, where n_{CPU} is the number of computer nodes. The pairs of

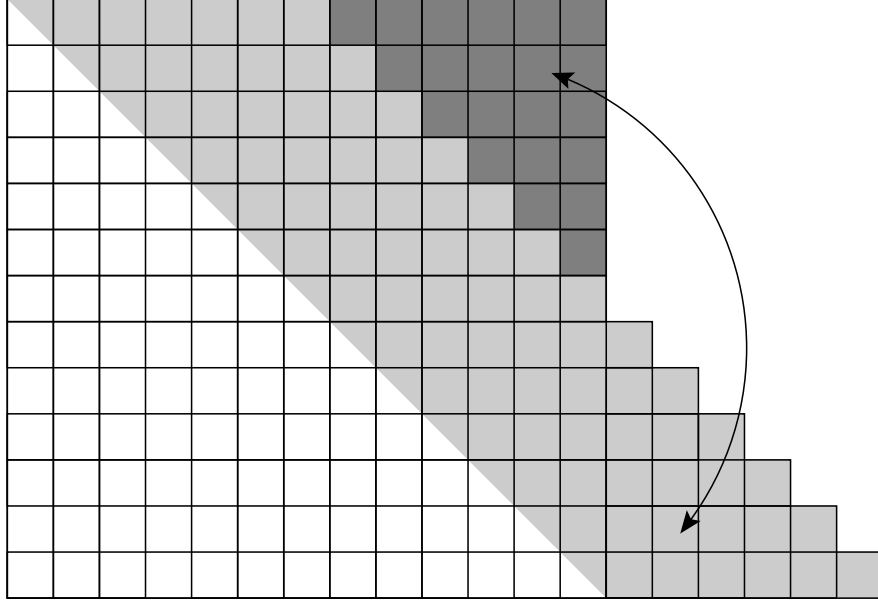


Figure 5. Arrangement of the pairs of batches $m \leq m'$ with active occupied orbitals on the upper triangle of a symmetric matrix or on block diagonal stripes.

batches $(\mathcal{I}_m, \mathcal{I}_{m'})$ with $m \leq m'$ can be ordered either on the upper triangle of a symmetric matrix or on block diagonal stripes as shown in Fig. 5. Now, each computer node gets assigned in a suitable way one block from of each diagonal, such that each node needs only access a minimal number of batches $\mathcal{I}_m$ of $B_{Q,ai}$ and $Y_{Q,ai}$. The minimal number of batches a node needs to access – in the following denoted as n_{blk} – increases approximately with $\sqrt{n_{CPU}}$. The calculation of these three-index ERIs $B_{Q,ai}$ would require about $\mathcal{O}(N^2 N_X) + \mathcal{O}(O N^2 N_X) \times n_{blk}/n_{CPU}$ floating point multiplications. Similar computational costs arise for some steps that involve $Y_{Q,ai}$ and other intermediates that follow the $\mathcal{O}(O^2 N^2 N_x)$ -scaling construction of this intermediate. Thus, a conflict between minimization of the operation count and communication arises:

- If the three-index intermediates $B_{Q,ai}$ and $Y_{Q,ai}$ are communicated between the nodes to avoid multiple integral evaluations, the communication demands per node become relatively large, $\sim N N_x \times O/\sqrt{n_{CPU}}$.
- If the communication of three-index intermediates is avoided by evaluating on each node all integrals needed, the operation count for the steps which are in RI-MP2 and RI-CC2 the next expensive ones after the $\mathcal{O}(O^2 V^2 N_X)$ steps decreases only with $1/\sqrt{n_{CPU}}$.

The first option requires a high bandwidth for communication while the second option can also realized with a low bandwidth, but on the expense of a less efficient parallelization. For both ways a prerequisite for a satisfactory efficiency is that the total computational costs are dominated by those for the $\mathcal{O}(N^5)$ steps such that the time needed for multiple

calculations ($\mathcal{O}(\mathcal{N}^4)$) or communication ($\mathcal{O}(\mathcal{N}^3)$) of three-index intermediates is a negligible fraction of the total time for the calculation. Both options have been realized in our parallel implementation of the `ricc2` code and shall in the following be denoted as modes for “slow communication” and “fast communication”.

To implement the blocked distribution of occupied orbital indices and index pairs sketched above we define at the beginning of the calculation the following index sets:

- $\mathcal{I}_m$: a block of occupied orbitals i assigned to node m
- $\mathcal{J}_m$: merged set of the n_{blk} blocks $\mathcal{I}_n$ for which node m needs the three-index ERIs $B_{Q,ai}$ or calculates a contribution to $Y_{Q,ai}$
- $\mathcal{S}_m$: the set of all columns in the blocked distribution to which node m calculates contributions.
- $\mathcal{R}_m(n)$: the indices of the rows in column n assigned in this distribution to node m

With this concept one obtains an efficient parallelization of most program parts that involve at least one occupied index. These parts use only three- and two-index AO integrals and include all steps that scale with $\mathcal{O}(\mathcal{N}^4)$ or $\mathcal{O}(\mathcal{N}^5)$ in RI-MP2 single point calculations for energies or RI-CC2 calculations for excitation energies and spectra. For a discussion of additional demanding steps in the computation of analytic derivatives (gradients) the interested reader is referred to Refs. 53, 73–75. Here, we only sketch how the computation of the intermediate $Y_{Q,ai}$ can be implemented without any MPI communication once each computer node has calculated or received all integral intermediates $B_{Q,ai}$ needed there:

```

loop  $n \in \mathcal{S}_m$ , loop  $I$  (where  $I \subseteq \mathcal{I}_n$ )
  read  $B_{Q,ai}$  for all  $i \in I$ 
  loop  $n' \in \mathcal{R}_m(n)$ , loop  $j \in \mathcal{I}_{n'}$  with  $j \leq i$ 
    * read  $B_{Q,bj}$ 
    *  $t_{ab}^{ij} \leftarrow B_{Q,ai} B_{Q,bj} / \{\epsilon_i - \epsilon_a + \epsilon_j - \epsilon_b\}$ 
    *  $Y_{P,ai} \leftarrow (2t_{ab}^{ij} - t_{ba}^{ij}) B_{P,bj}$  and for  $j \neq i$  also  $Y_{P,bj} \leftarrow (2t_{ab}^{ij} - t_{ba}^{ij}) B_{P,ai}$ 
  end loop  $j$ , loop  $n'$ 
  store  $Y_{P,ai}$  and  $Y_{P,bj}$  on disk (distributed)
end loop  $I$ , loop  $n$ 

```

If only the RI-MP2 energy is needed, it can be evaluated directly after the calculation of the integrals $\langle ia|jb \rangle$ and amplitudes t_{ab}^{ij} as described in Section 2.1 and the calculation of $Y_{Q,ai}$ can be skipped. If the latter intermediates are needed, the contributions to the $Y_{Q,ai}$ intermediate can be added and redistributed (after the loop over n has been closed) such that each node has the complete results for $Y_{P,ai}$ for all $i \in \mathcal{J}_m$ (requiring the communication of $\approx 2OV N_x / \sqrt{n_{CPU}}$ floating point numbers per node).

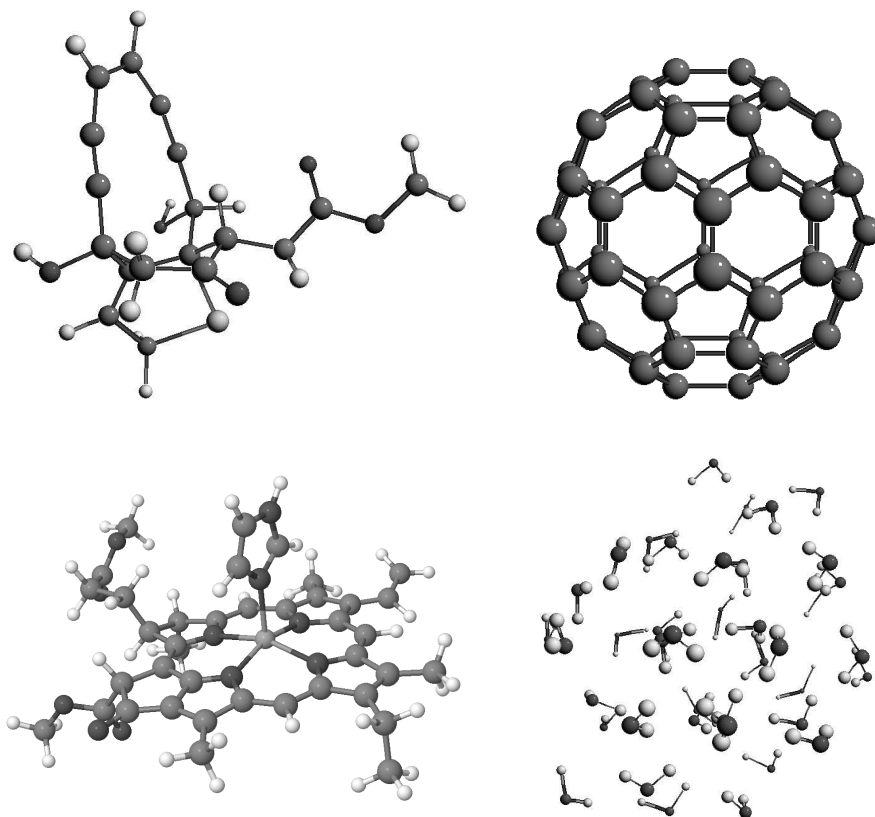


Figure 6. Structures of the four test examples used to benchmark the performance of parallel RI-MP2 calculations. For the details of the basis sets and the number of correlated electrons see text.

4.1 Performance for Parallel RI-MP2 Energy Calculations

To benchmark the calculation of MP2 energies we used four typical test systems with structures as shown in Fig. 6:

- A calicheamicine model taken from Ref. 76, which has also no point group symmetry. These calculations have been done in the cc-pVTZ basis sets^{61, 65, 66} with 934 orbital and 2429 auxiliary functions and 124 electrons have been correlated.
- The fullerene C_{60} , which has I_h symmetry, but the calculations reported here exploited only the Abelian subgroup D_{2h} . The cc-pVTZ basis set has been used, which in this case comprises 1800 orbital and 4860 auxiliary basis functions and the 240 valence electrons were correlated.
- A chlorophyll derivative which has also no point group symmetry. The cc-pVDZ basis with in total 918 orbital and 3436 auxiliary functions have been used and 264 electrons have been correlated.

- A cluster of 40 water molecules as an example for a system where integral pre-screening leads to large reductions in the costs in conventional MP2 calculations. The basis sets are 6-31G* for the orbital⁷⁷ and cc-pVDZ for the auxiliary⁵⁷ basis with, respectively, 760 and 3840 functions; the point group is C_1 and the 320 valence electrons have been correlated.

The maximum amount of core memory used by the program was in all calculations limited to 750 Mb. The calculations were run on two different Linux cluster: one cluster with ca. 100 Xeon Dual 2.8 GHz nodes connected through a cascaded Gigabit network and a second cluster with ca. 64 Athlon 1800MP MHz nodes connected through a 100 MBit fast Ethernet network. Due to a much larger load on the first cluster and its network the transfer rates reached in the benchmark calculations varied between ca. 80–200 MBit/sec per node. On the Athlon Cluster with the 100 MBit network we reached transfer rates of ca. 20–50 MBit/sec per node.

Fig. 7 shows timings for the calculation of MP2 energies for the C_{60} fullerene. On both architectures in sequential runs about 55% of the time are spend in the matrix multiplication for the $\mathcal{N}^5$ step. With increasing number of nodes this ratio slowly decreases. In case of the “slow communication” mode because the costs for the integral evaluation take an increasing fraction of the total wall time; in the “fast communication” mode (and here in particular on the cluster with the slower network) because of the increasing fraction of time spent in the communication of the 3-index MO integral intermediate $B_{Q,ai}$. Not parallelized steps – as e.g. the evaluation of the matrix V_{PQ} of 2-index ERIs, its Cholesky decomposition and formation of the inverse – take only a marginal fraction of the total wall time and the fraction of the time spend in the I/O stays approximately constant with the number of nodes used for the calculation. Another important message from Fig. 7 is, that even with a relatively slow network it is advantageous to communicate the 3-index intermediates, although on the cluster with the slower network the difference in performance between the two modes is not large. We note, however, that this depends also on the size of the system and the basis sets.

Because of the symmetry of the molecule, an RI-MP2 energy calculation for C_{60} is today not really a large scale application. The same holds for the other three test examples. Nevertheless, already for these (for parallel calculations) small examples the speed ups obtained with the present implementation are reasonable as Fig. 8 shows. The speed up obtained increases with the system size as the computational costs become dominated by the $\mathcal{N}^5$ -scaling matrix multiplication in the construction of the MO 4-index ERIs and the less good parallelizing calculation and/or communication of the 3-index MO integrals becomes unimportant for the total wall time.

5 RI-MP2 Calculations for the Fullerenes C_{60} and C_{240}

An important aspect of the parallel implementation of RI-MP2 is that it allows to combine the fast RI-MP2 approach with *parallel* Hartree-Fock self-consistent field (HF-SCF) calculations, available today in many program packages for electronic structure calculations, to optimize geometries for relatively large molecules at the MP2 level. An example for such a calculation is the determination of the MP2 basis set limit for the ground state equilibrium structure of C_{60} . The structure of C_{60} has been studied before at the MP2

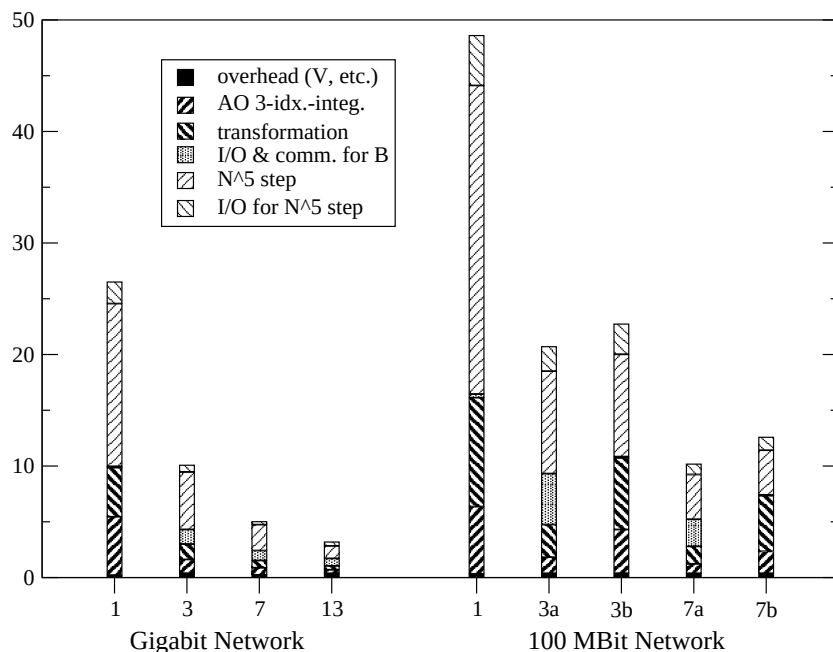


Figure 7. Timings for the most important steps in parallel RI-MP2 energy calculations for C_{60} in the cc-pVTZ basis (240 electrons correlated). For technical details of the machines used see text. At the abscissa we indicated the number of CPUs used for the calculations. For the cluster with a 100 MBit Network letters "a" and "b" are added, respectively, for calculations in the "fast" and "slow" communication modes. On the other cluster only the former program mode has been tested. The fraction denoted "overhead" includes most non-parallel steps, as the calculation of the Coulomb metric V and the inverse of its Cholesky decomposition, I/O and communication of MO coefficients, etc. With "AO 3-idx.-integ" we denoted the time spend for the calculation of the AO 3-index integrals $(P|\mu\nu)$ and with "transformation" and "I/O & comm. for B" the fractions spend in the three-index transformations for the intermediates B_{Qa}^i and for saving these intermediates on disk and/or distributing them to other computer nodes. "N⁵ step" and "I/O for N⁵ step" are the fractions spend, respectively, in the N^5 -scaling matrix multiplication and the I/O of B intermediates during the calculation of two-electron MO integrals. For parallel calculations idle times caused by non-perfect load-balance are included under the point "I/O for N⁵ step".

level by Häser and Almlöf⁷⁹ in 1991, but due to the large computational costs of MP2 the calculations had to be limited to a singly polarized TZP basis set ([5s3p1d], 1140 basis functions), which is known to cover only about 75% of the correlation energy. With the parallel implementation of RI-MP2 it was now possible repeat this calculation using cc-pVTZ basis ([4s3p2d1f], 1800 basis functions), which gives typically correlation energies almost within 90% of the basis set limit, and the cc-pVQZ basis ([5s4p3d2f1g], 3300 basis functions), which usually cuts the remaining basis set errors again into half. The results for the bond lengths and the total energies are summarized in Table 2 together with the results from Ref. 79 and the available experimental data. As anticipated from the quality of the basis sets, the result for the correlation energy increases by about 15% from the MP2/TZP to the MP2/cc-pVTZ calculation and again by about 6% from the cc-pVTZ to the cc-pVQZ basis. Also the changes in the bond lengths from the MP2/TZP to the MP2/cc-pVQZ level

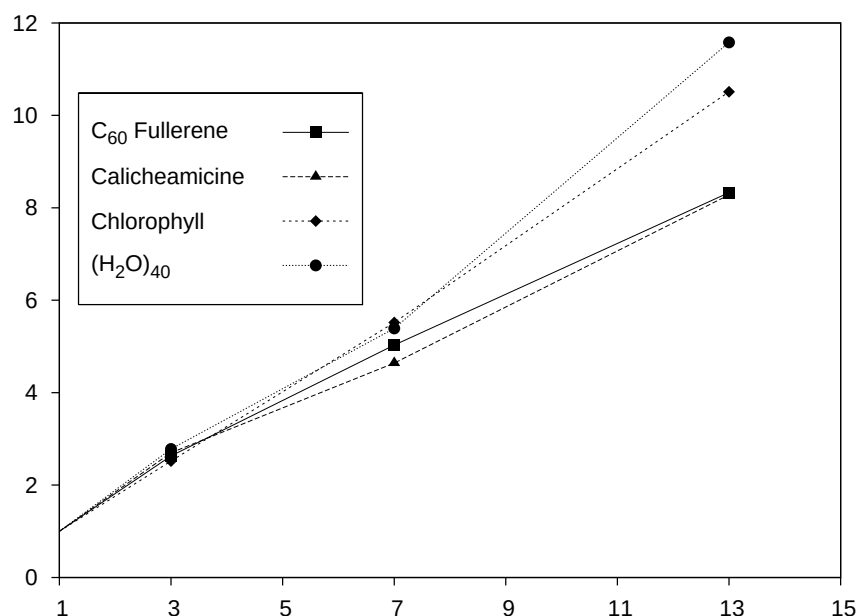


Figure 8. Speed up obtained for parallel RI-MP2 energy calculations on the Linux cluster with Gigabit network with four test examples. The number of nodes is given on the abscissa and the speed up (defined as wall time of parallel calculation divided by the wall time of the sequential run) is indicated on the ordinate.

are with 0.004–0.005 Å of the same magnitudes as between the MP2/DZP and MP2/TZP calculations. But the difference between the two C–C distances remains almost unchanged, and also the comparison with the experimental data is not effected, since the error bars of the latter are with about ± 1 pm of the same order of magnitude as the basis set effects. The inclusion of core correlation effects would lead to a further slight contraction of the bond lengths, but the largest uncertainty comes from higher-order correlation effects which would probably increase the bond lengths in this system, but likely not more than 0.005 Å. Therefore, it is estimated that the MP2/cc-pVQZ results for the equilibrium bond distances (r_e) of the buckminster fullerene C₆₀ are accurate within ± 0.005 Å. This is slightly less than the uncertainty of the presently available experimental data. Within their uncertainties the ab initio calculations and the experiments are thus in good agreement.

Another example demonstrating which system sizes can be handled with the parallel implementation of RI-MP2 is the next larger icosahedral homologue of the Buckminster fullerene C₆₀: the C₂₄₀ molecule. The correlation consistent triple- ζ basis cc-pVTZ comprises for this molecules 7200 basis functions and, if the 1s core orbitals are kept frozen, 960 electrons have to be correlated. This calculation has been run on a Linux cluster with Dual Xeon 2.8 GHz nodes connected by a Gigabit network. Because the memory demands of implementation increase for non-Abelian point groups with the square of the dimension of the irreducible representations the calculation was carried out in the D_{2h} subgroup of the molecular point group I_h. On 19 CPUs the RI-MP2 calculation was completed after 16 hours and 6 minutes. About 12.5% of the time was spend in the evaluation and distribution

Table 2. Equilibrium bond distances of C_{60} ; d_{C-C} denotes the distance between adjacent C atoms in a five-ring and $d_{C=C}$ the distance between the C-C bond shared between two six-rings. The bond distances are given in Ångström (Å) and the total energies in Hartrees (H).

Method	$d_{C-C}/\text{\AA}$	$d_{C=C}/\text{\AA}$	Energy/hartree
SCF/DZP ^a	1.450	1.375	-2272.10290
SCF/TZP ^a	1.448	1.370	-2272.33262
MP2/DZP ^b	1.451	1.412	-2279.73496
MP2/TZP ^b	1.446	1.406	-2280.41073
MP2/cc-pVTZ ^c	1.443	1.404	-2281.65632
MP2/cc-pVQZ ^c	1.441	1.402	-2282.34442
exp. ^d	1.458(6)	1.401(10)	
exp. ^e	1.45	1.40	
exp. ^f	1.432(9)	1.388(5)	

^a from Ref. 78; ^b from Ref. 79; ^c from Ref. 74, at the MP2/cc-pVTZ optimized structure the SCF energy is -2272.40406 hartree; ^d gas phase electron diffraction, Ref. 80; ^e solid state NMR, Ref. 81; ^f X-ray of $C_{60}(\text{OsO}_4)(4\text{-tert-butylpyridine})_2$, Ref. 82;

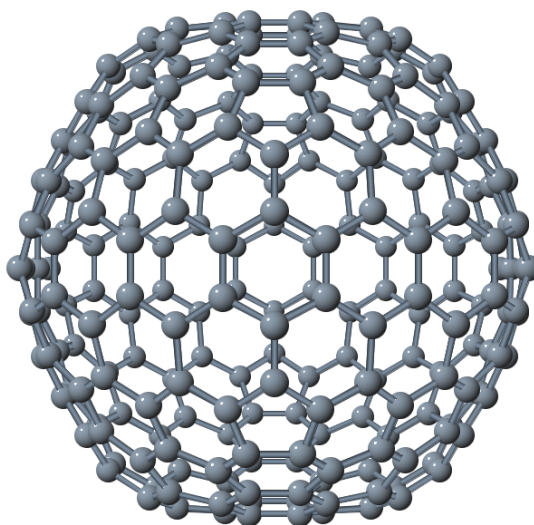


Figure 9. Structure of the icosahedral fullerene C_{240} .

of the two- and three-index integrals and 85% in the $\mathcal{O}(O^2V^2N_x)$ scaling construction of the four-index integrals in the MO basis ($ia|jb$). In D_{2h} symmetry about 6×10^{11} four-index MO integrals (≈ 4.8 TByte) had to be evaluated to calculate the MP2 energy. This

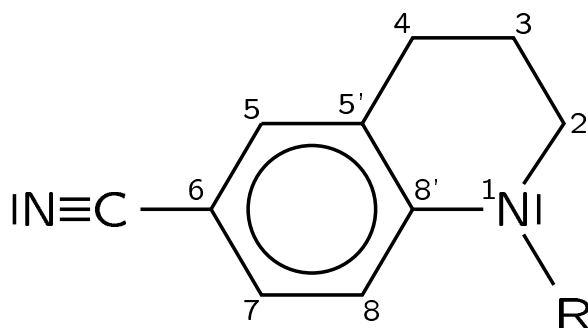


Figure 10. Enumeration of the atoms in NMC6 (R = methyl) and NTC6 (R = *tert*butyl). For DMABN R = methyl and aliphatic six-ring is replaced by a (second) methyl group at the N-atom.

shows that such a calculation would with a conventional (non-RI) MP2 require either an enormous amount of disc space or many costly re-evaluations of the four-index AO two-electron integrals and would thus even on a massively parallel architecture difficult to carry out. To the best of our knowledge this is the largest canonical MP2 calculation done until today. With the parallel implementation of the RI-MP2 approach calculations of this size can now be carried out on PC clusters build with standard (and thus low cost) hardware and are expected to become soon routine applications.

The total energy of C_{240} obtained with MP2/cc-pVTZ at the BP86^{83–85}/SVP^{62,56} optimized structure⁸⁶ is -9128.832558 H. For the buckminster fullerene C_{60} a single point MP2/cc-pVTZ calculation at the BP86/SVP optimized geometry gives a total energy of -2281.645107 H. Neglecting differential zero-point energy effects, which in this case are expected to be small, we obtain from our calculations an estimate for the reaction enthalpy of $4 \times C_{60} \rightarrow C_{240}$ of -2.25 H, i.e. a change in the enthalpy of formation per carbon atom of -9.4 mH or -25 kJ/mol. This can be compared with the experimental result⁸⁷ for $\Delta_f H^0$ of C_{60} relative to graphite of 39.25 ± 0.25 kJ/mol. Thus, the present calculations predict that the strain energy per carbon atom in C_{240} is with ≈ 15 kJ/mol only about 35% of the respective value in C_{60} .

6 Geometry Optimizations for Excited States with RI-CC2: The Intramolecular Charge Transfer States in Aminobenzonitrile Derivatives

An example for the optimization of excited state equilibrium structures with RI-CC2 are recent investigations^{88,89} on N-alkyl-substituted aminobenzonitriles (see Fig. 10). A problem discussed for this class of molecules in the literature since several decades in many publications has been the structure of a so-called intramolecular charge-transfer (ICT) state which is observed in fluorescence and femtosecond spectroscopic experiments close to a so-called locally excited (LE) state.^{90–94} The two states belong to the two lowest singlet hypersurfaces S1 and S2, which are connected through a conical intersection seam. Experimental and theoretical results^{95–99} indicate that the reaction coordinate which connects

Table 3. Calculated bond lengths (pm) and angles ($^{\circ}$) of the ground states of DMABN, NMC6, and NTC6 in comparison (from Ref. 89, for the enumeration of the atoms see Fig. 10).

	DMABN	NMC6	NTC6
$d(C_{Ph}-N_1)^a$	137.7	138.1	139.0
$d(C_8C_{8'})$	141.4	141.2	141.2
$d(C_{8'}C_{5'})$	141.4	141.9	141.1
$d(C_7C_8)$	138.7	138.7	138.9
$d(C_5C_{5'})$	138.7	138.9	138.6
$d(C_6C_7)$	140.2	140.0	139.9
$d(C_5C_6)$	140.2	140.2	140.3
$d(C_6C_{CN})$	142.7	142.6	142.6
$d(CN)$	118.2	118.1	118.1
τ^b	0	0.1	27.9
ϕ_1^c	23	24.8	18.9
ϕ_2^d	< 1	1	1.5

^a bond distance between phenyl ring and amino group. ^b torsion angle, defined as dihedral angle of the normals defined by the planes $C_8-C_{8'}-C_{5'}$ and $C_2-N_1-C_R$ and the bond $C_{8'}-N_1$. ^c out-of-plane angle of the bond $C_{8'}-N_1$ with respect to the plane $C_2-N_1-C_R$ (“wagging” angle). ^d out-of-plane angle of the bond $C_{8'}-N_1$ with respect to the plane $C_8-C_{8'}-C_5$.

the minima on the two surfaces through the conical intersection involves a Kekulé-like distortion of the phenyl ring and a twist of the amino group, which for the N,N-dimethylaminobenzonitrile (DMABN) is known to be in the ground state almost coplanar with the phenyl ring. That the twisting coordinate is involved probably explains distinct effects of different aliphatic substituents at the amino group on the fluorescence properties (vide infra) which are intensively discussed in the literature. In 1-*tert*-butyl-6-cyano-1,2,3,4-tetrahydroquinoline (NTC6) and 1-methyl-6-cyano-1,2,3,4-tetrahydroquinoline (NMC6) a twist of the amino group is restricted by the aliphatic ring to a certain range of torsion angles, but on the other side the sterically demanding bulky *tert*-butyl substituent in NTC6 disfavors a coplanar orientation. CC2/TZVPP calculations⁸⁹ predict for the ground state of NMC6 an almost coplanar orientation of the phenyl and amino moieties, but for NTC6 a tilted geometry with a twist angle of about 28° (cmp. Table 3).

Table 4 gives an overview on the CC2/TZVPP results for some spectroscopic properties of DMABN, NMC6 and NTC6, e.g. the absorption and emission energies and the dipole moments in comparison with the available experimental data. For the ICT states we found for NMC6 and NTC6 three conformations. Table 5 summarizes the results for the energetically lowest-lying structures and the ones with the highest dipole moments denoted as, respectively, ICT-1 and ICT-2, in comparison with the structure of the single conformer in the ICT state of DMABN. In all three molecules the ICT equilibrium geometries display marked quinoid distortions of the aromatic ring system. An important finding, which was not anticipated from the experimental data that has been available in the literature, is that

Table 4. Calculated absorption and emission energies and dipole moments for DMABN, NMC6 and NTC6 in comparison with experimental data. The CC2 results for absorption and emission energies are vertical electronic transition energies; the dipole moments were calculated as analytic derivatives of the CC2 total energies.

	DMABN		NMC6		NTC6	
	CC2 ^a	exp.	CC2 ^a	exp.	CC2 ^a	exp.
absorption (S ₁) [eV]	4.41 ^b	4.25 ^c	4.31 ^d		4.33 ^d	
absorption (S ₂) [eV]	4.77 ^b	4.56 ^c	4.58 ^d	4.32 ^e	4.43 ^d	4.14 ^e
osc. strengths (S ₁)	0.03 ^{b,f}		0.03 ^f		0.03 ^f	
osc. strengths (S ₂)	0.62 ^{b,f}		0.49 ^f		0.51 ^f	
T _e (LE) [eV]	4.14		4.07		3.91	
emission (LE) [eV]	3.78 ^g	3.76 ^h	3.67 ^g	3.67 ^e	3.34 ^g	3.50 ^e
T _e (ICT) [eV]	4.06–4.16 ⁱ		4.18		3.71	
emission (ICT) [eV]	2.49–3.27 ^{i,g}	2.8–3.2 ^j	2.53 ^{g,k}		2.51 ^{g,k}	2.8 ^l –3.3 ^e
dipole (GS) [D]	7.4	6.6 ^j	7.5	6.8 ^m	7.7	6.8 ^m
dipole (LE) [D]	10.1	9.7 ^j	10.4	10.6 ^m	12.6	
dipole (ICT) [D]	13.3–15.1 ⁱ	17±1 ^j	12.7 ^k		13.5 ^k	17–19 ^m

^a Unless otherwise indicated the CC2 results for DMABN are taken from Ref. 88 and those for NMC6 and NTC6 from Ref. 89. ^b CC2/TZVPP (Ref. 100). ^c EELS band maximum (Ref. 101). ^d Vertical excitation energy to the L_a (or S₂) state which has a significantly larger oscillator strength. ^e Experimental band maximum in *n*-hexane (Ref. 92). ^f Oscillator strength for vertical electronic transition calculated at the CC2/TZVPP level in length gauge. ^g Vertical energy separation from ground state at the excited state equilibrium structure. ^h Maximum of dispersed emission from jet-cooled DMABN (Ref. 102). ⁱ The first value is the result for the gas phase equilibrium structure and the second value is obtained at the C_{2v} symmetric saddle point (Ref. 88). ^j Emission energy from ICT state from maxima of fluorescence bands; ground state dipole moment derived from the dielectric constant and refractive index in dioxane and the excited state dipole moments from time-resolved microwave conductivity measurements in dioxane (Ref. 103). ^k Value refers to the ICT-2 conformer. ^l Experimental band maximum in methanol (Ref. 92). ^m Derived from solvatochromic shift of fluorescence maximum (Ref. 92).

the aromatic ring is no longer confined to planarity in the excited state. Rather, the carbon atom labeled 8' in Fig. 10 is pyramidalized. Therefore the aliphatic six-ring can accommodate twist angles of the amino group of up to 60–70°, as illustrated in Fig. 11, and in this way energetically low-lying twisted ICT states can be realized even in NTC6 and NMC6. In the literature it was before assumed that the aliphatic six-ring, which connects the amino group with the phenyl ring restricts these molecules to “planarized” structures and makes such a twist impossible.

The transition to the ICT state is at the ground state geometry dominated by the one-electron HOMO→LUMO excitation in these molecules. Both orbitals are of Ph-N anti-binding character, but the orbital energy of the LUMO decreases slightly faster with increasing twisting angle than the energy of the HOMO and already such a simple model

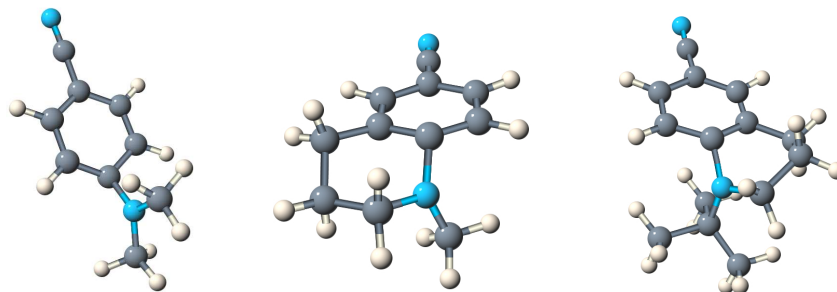


Figure 11. Equilibrium structures of the ICT states in DMABN, NMC6, and NTC6.

predicts for the ICT state close to the ground state geometry a gradient directed to a twisted structure. With increasing twisting angle the transition assumes an increasing contribution from the HOMO-2 $\rightarrow$ LUMO excitation. The HOMO-2 is the Ph-N binding counterpart of the HOMO and increases in energy with the twisting angle and mixes with the HOMO. As the angle approaches 90° one of the two orbitals becomes the lone-pair at the amino N-atom while the other is localized in the aromatic system and the transition to the ICT state is dominated by the $n \rightarrow \pi^*$ excitation. In a many electron picture this change in the character of the excitation corresponds to an avoided crossing of S2 with another, at the ground state structure energetically higher lying, charge-transfer state – in DMABN according to DFT/SCI calculation in Ref. 104 the S5 state. The avoided crossing with this state is the main driving force for the formation of the TICT structures (twisting and pyramidilization at the C_{8'} atom) in DMABN, NTC6, NMC6 and other alkyl-substituted amino-benzonitrils. It leads to a pronounced stabilization of the ICT state at large twisting angles and enhances the charge-transfer character, as it is apparent from the expectation values for the dipole moment (see Table 4). For all three molecules, DMABN, NMC6, and NTC6, one finds a similar change in the electronic character from the vertical excitation in the Franck-Condon region to the equilibrium geometries of the ICT states. This is in line with the interpretation of recent measurements of the short-time dynamics in DMABN derivatives after excitation to S₂.^{95,105–108}

For NTC6 the increase in the twist angle from the ground to the excited ICT states reduces the steric strain of the *tert*-butyl group and thus compensates for the hindrance of the twist by the aliphatic bridge. We obtain at the CC2/TZVPP level that for NTC6 and DMABN the ICT states are energetically slightly below the LE state, which is reached by an one-electron transition from the PH-N antibinding HOMO to a Ph-N non-binding orbital. For NMC6, however, the inhibition of a 90° twist is not compensated by the release of a similar strain since the methyl substituent is sterically much less demanding. Thus, in difference to DMABN and NTC6 the LE $\rightarrow$ ICT reaction for NMC6 is predicted by the RI-CC2 calculations to be slightly endotherm. This explains why NMC6 is not dual fluorescent,^j in contrast to DMABN and NTC6.

Table 5. Calculated bond lengths (pm) and angles ($^{\circ}$) and weights of the two most important one-electron excitations (%) for the intramolecular charge-transfer states of DMABN, NMC6, and NTC6 in comparison (from Ref. 89, for the enumeration of the atoms see Figure 10).

	DMABN	NMC6		NTC6	
	ICT	ICT-1	ICT-2	ICT-1	ICT-2
$d(C_{Ph}-N_1)^a$	144.3	146.8	145.0	146.8	145.7
$d(C_8C_{8'})$	144.6	143.5	144.8	142.9	144.9
$d(C_8C_{5'})$	144.6	146.2	144.5	146.0	143.6
$d(C_7C_8)$	137.2	137.2	137.7	137.3	137.8
$d(C_5C_{5'})$	137.2	138.0	136.9	137.9	137.1
$d(C_6C_7)$	142.9	143.4	142.4	143.8	142.4
$d(C_5C_6)$	142.9	141.8	143.7	141.9	144.0
$d(C_6C_{CN})$	140.9	141.2	140.9	141.1	140.8
$d(CN)$	118.9	118.8	118.9	118.8	118.9
τ^b	90	54.3	66.6	58.5	65.0
ϕ_1^b	0	24.1	14.7	20.7	5.2
ϕ_2^b	41	43.9	44.6	36.4	43.4
HOMO→LUMO		65	62	69	64
HOMO-2→LUMO		15	17	25	16

^a bond distance between phenyl ring and amino group. ^b for the definition of the torsion and the out-of-plane angles see Table 3.

7 Summary

The computational costs of wavefunction based correlated ab initio methods that treat the electron–electron interaction correctly through second order (so-called second order or approximate doubles methods) have in conventional implementations been dominated by the huge operation counts for the calculation of the four-index electron repulsion integrals in the AO basis and their transformation to the MO basis. The costs for these steps increase rapidly with the size of the system studied and the basis sets used. In addition, also the huge storage demands for the four-index transformation hindered applications on large systems.

With the resolution-of-the-identity approximation for the electron repulsion integrals the CPU time for the calculation of the MO integrals needed in second order methods is reduced by about an order of magnitude (and sometimes even much more) and the scaling of the storage demands is reduced from $\mathcal{O}(O^2N^2)$ to $\mathcal{O}(OVN_x)$. If optimized auxiliary basis sets are used, as they today are available for many orbital basis sets, the errors due to RI approximation are insignificant compared to the errors due to the incompleteness of the orbital basis sets.

In combination with a new parallel implementation in TURBOMOLE for distributed memory architectures (e.g. PC clusters) it became now possible to carry out RI-MP2 calculations for energies and structures with several thousands of basis functions and several hundreds of correlated electrons. This extends the applicability of MP2 to systems which

else can only be treated with SCF or DFT methods. Calculations on excited states using e.g. the approximate coupled-cluster singles and doubles method CC2 or the perturbative doubles correction to configuration interaction singles, CIS(D), are somewhat more involved and structure optimizations for excited states are (because of weakly avoided crossings or conical intersections) much less straightforward than for ground states. With the parallel implementation of RI-CC2 they become still feasible for molecules with more than 30 atoms and many hundred basis functions even if the molecular structure has no point group symmetry.

Acknowledgments

The author is indebted to A. Köhn and A. Hellweg for their contributions to the RICC2 program and to the calculations reviewed in this manuscript. Financial support by the Deutsche Forschungsgemeinschaft (DFG) for the reported work is gratefully acknowledged.

References

1. P. Pulay. *Chem. Phys. Lett.*, 100:151–154, 1983.
2. M. Schütz, G. Hetzer, and H.-J. Werner. *J. Chem. Phys.*, 111:5691–5705, 1999.
3. G. Hetzer, M. Schütz, H. Stoll, and H. J. Werner. *J. Chem. Phys.*, 113:9443–9455, 2000.
4. G. E. Scuseria and P. Y. Ayala. *J. Chem. Phys.*, 111:8330–8343, 1999.
5. P. E. Maslen and M. Head-Gordon. *J. Chem. Phys.*, 109:7093–7099, 1998.
6. S. H. Li, J. Ma, and Y. S. Jiang. *J. Comp. Chem.*, 23:237–244, 2002.
7. K. Morokuma. *Philosophical Transactions of the Royal Society of London Series A – Mathematical Physical and Engineering Sciences*, 360:1149–1164, 2002.
8. M. Schütz and H. J. Werner. *J. Chem. Phys.*, 114:661–681, 2001.
9. M. Schütz. *Phys. Chem. Chem. Phys.*, 4:3941–3947, 1993.
10. S. Sæbo and P. Pulay. *J. Chem. Phys.*, 87:3975–3983, 2001.
11. D. S. Lambrecht and C. Ochsenfeld. *J. Chem. Phys.*, 123:184102, 2005.
12. S. Grimme. *J. Chem. Phys.*, 118:9095–9102, 2003.
13. S. Grimme. *J. Comput. Chem.*, 24:1529–1537, 2003.
14. Y. S. Jung, R. C. Lochan, A. D. Dutoi, and M. Head-Gordon. *J. Chem. Phys.*, 121:9793–9802, 2004.
15. R. C. Lochan, Y. S. Jung, and M. Head-Gordon. *J. Phys. Chem. A*, 109:7598–7605, 2005.
16. T. D. Crawford and R. A. King. *Chem. Phys. Lett.*, 366:611–622, 2002.
17. T. Korona and H. J. Werner. *J. Chem. Phys.*, 114:3006–3019, 2003.
18. T. Korona, K. Pflüger, and H. J. Werner. *Phys. Chem. Chem. Phys.*, 6:2059–2065, 2004.
19. M. Schütz and F. R. Manby. *Phys. Chem. Chem. Phys.*, 5:3349–3358, 2003.
20. H.-J. Werner, F. R. Manby, and P. J. Knowles. *J. Chem. Phys.*, 118:8149, 2003.
21. T. Helgaker, P. Jørgensen, and J. Olsen. *Molecular Electronic-Structure Theory*. John Wiley & Sons, New York, 2000.

22. S. Grimme and M. Waletzke. *Phys. Chem. Chem. Phys.*, 2:2075–2081, 2000.
23. B. O. Roos, K. Andersson, M. P. Fülscher, P. A. Malmqvist, L. Serrano-Andres, K. Pierloot, and M. Merchán. *Advances in Quantum Chemistry*, 93:219–331, 1996.
24. J. Olsen and P. Jørgensen. *J. Chem. Phys.*, 82:3235, 1985.
25. J. Olsen and P. Jørgensen. *Time-Dependent Response Theory with Applications to Self-Consistent Field and Multiconfigurational Self-Consistent Field Wave Functions*, In D. R. Yarkony, editor, *Modern Electronic Structure Theory*, volume 2, chapter 13, pages 857–990. World Scientific, Singapore, 1995.
26. O. Christiansen, P. Jørgensen, and C. Hättig. *Int. J. Quantum Chem.*, 68:1–52, 1998.
27. C. Hättig. *Accurate Coupled Cluster Calculation of Nonlinear Optical Properties of Molecules*, In M. G. Papadopoulos, editor, *On the non-linear optical response of molecules, solids and liquids: Methods and applications*. Research Signpost, 2003.
28. P. Salek, O. Vahtras, T. Helgaker, and H. Agren. *J. Chem. Phys.*, 117:9630–9645, 2002.
29. O. Christiansen, S. Coriani, J. Gauss, C. Hättig, P. Jørgensen, F. Pawłowski, and A. Rizzo. *Accurate NLO Properties for small molecules. Methods and results*, In M. G. Papadopoulos, J. Leszczynski, and A. J. Sadlej, editors, *Nonlinear optical properties: From molecules to condensed phases*. Kluwer Academic Publishers, London, 2006.
30. O. Christiansen, H. Koch, and P. Jørgensen. *Chem. Phys. Lett.*, 243:409–418, 1995.
31. O. Christiansen, H. Koch, P. Jørgensen, and T. Helgaker. *Chem. Phys. Lett.*, 263:530, 1996.
32. J. Paldus and X. Z. Li. *Advances in Quantum Chemistry*, 110:1–175, 1999.
33. R. J. Bartlett. *Recent Advances in Coupled-Cluster Methods*, volume 3 of *Recent Advances in Computational Chemistry*. World Scientific, Singapore, 1997.
34. R. J. Bartlett. Coupled-cluster theory: An overview of recent developments. In D. R. Yarkony, editor, *Modern Electronic Structure Theory*, pages 1047–1131, Singapore, 1995. World Scientific.
35. M. Head-Gordon, M. Oumi, and D. Maurice. *Mol. Phys.*, 96:593–574, 1999.
36. M. Head-Gordon, R. J. Rico, M. Oumi, and T. J. Lee. *Chem. Phys. Lett.*, 219:21–29, 1994.
37. J. Schirmer. *Phys. Rev. A*, 26:2395–2416, 1981.
38. A. B. Trofimov and J. Schirmer. *J. Phys. B*, 28:2299–2324, 1995.
39. E. S. Nielsen, P. Jørgensen, and J. Oddershede. *J. Chem. Phys.*, 73:6238–6246, 1980.
40. M. J. Packer, E. K. Dalskov, T. Enevoldsen, H. J. A. Jensen, and J. Oddershede. *J. Chem. Phys.*, 105:5886–5900, 1980.
41. O. Christiansen, K. L. Bak, H. Koch, and S. P. A. Sauer. *Chem. Phys. Lett.*, 284:47–55, 1998.
42. A. I. Krylov, C. D. Sherrill, E. F. C. Byrd, and M. Head-Gordon. *J. Chem. Phys.*, 109:10669–10678, 1998.
43. T. B. Pedersen, H. Koch, and C. Hättig. *J. Chem. Phys.*, 110:8318–8327, 1999.
44. K. Raghavachari, G. W. Trucks, J. A. Pople, and M. Head-Gordon. *Chem. Phys. Lett.*, 157:479, 1989.
45. H. Koch, O. Christiansen, P. Jørgensen, A. Sánchez de Merás, and T. Helgaker. *J. Chem. Phys.*, 106:1808, 1997.
46. O. Christiansen, H. Koch, and P. Jørgensen. *J. Chem. Phys.*, 105:1451–1459, 1996.
47. O. Christiansen, H. Koch, and P. Jørgensen. *J. Chem. Phys.*, 103:7429–7441, 1995.

48. C. Hättig and A. Köhn. *J. Chem. Phys.*, 117:6939–6951, 2002.
49. J. L. Whitten. *J. Chem. Phys.*, 58:4496–4501, 1973.
50. B. I. Dunlap, J. W. D. Connolly, and J. R. Sabin. *J. Chem. Phys.*, 71:3396–3402, 1979.
51. O. Vahtras, J. E. Almlöf, and M. W. Feyereisen. *Chem. Phys. Lett.*, 213:514–518, 1993.
52. M. W. Feyereisen, G. Fitzgerald, and A. Komornicki. *Chem. Phys. Lett.*, 208:359–363, 1993.
53. F. Weigend and M. Häser. *Theor. Chem. Acc.*, 97:331, 1997.
54. F. Weigend, M. Häser, H. Patzelt, and R. Ahlrichs. *Chem. Phys. Letters*, 294:143, 1998.
55. A. P. Rendell and T. J. Lee. *J. Chem. Phys.*, 101:400–408, 1994.
56. K. Eichkorn, O. Treutler, H. Öhm, M. Häser, and R. Ahlrichs. *Chem. Phys. Lett.*, 240:283–289, 1995. Erratum *ibid.* 242 (1995) 652.
57. F. Weigend, A. Köhn, and C. Hättig. *J. Chem. Phys.*, 116:3175, 2001.
58. C. Hättig. *Phys. Chem. Chem. Phys.*, 7:59–66, 2005.
59. TURBOMOLE, Program Package for ab initio Electronic Structure Calculations , <http://www.cosmologic.de/turbomole.html>.
60. A. Köhn and C. Hättig. *Chem. Phys. Lett.*, 358:350–353, 2002.
61. T. H. Dunning. *J. Chem. Phys.*, 90:1007–1023, 1989.
62. A. Schäfer, H. Horn, and R. Ahlrichs. *J. Chem. Phys.*, 97(4):2571–2577, 1992.
63. A. Schäfer, C. Huber, and R. Ahlrichs. *J. Chem. Phys.*, 100(8):5829–5835, 1994.
64. F. Weigend, F. Furche, and R. Ahlrichs. *J. Chem. Phys.*, 119:12753–12762, 2003.
65. R. A. Kendall, T. H. Dunning, and R. J. Harrison. *J. Chem. Phys.*, 96:6796–6806, 1992.
66. D. E. Woon and T. H. Dunning. *J. Chem. Phys.*, 98:1358–1371, 1993.
67. D. E. Woon and T. H. Dunning. *J. Chem. Phys.*, 100:2975–2988, 1994.
68. A. K. Wilson, D. E. Woon, K. A. Peterson, and T. H. Dunning. *J. Chem. Phys.*, 110:7667–7676, 1999.
69. T. H. Dunning, K. A. Peterson, and A. K. Wilson. *J. Chem. Phys.*, 114:9244–9253, 2001.
70. K. A. Peterson and T. H. Dunning. *J. Chem. Phys.*, 117:10548–10560, 2002.
71. C. Hättig and A. Köhn. *J. Chem. Phys.*, 117:6939–6951, 2002.
72. C. Hättig and F. Weigend. *J. Chem. Phys.*, 113:5154–5161, 2000.
73. C. Hättig. *J. Chem. Phys.*, 118:7751–7761, 2003.
74. C. Hättig, A. Hellweg, and A. Köhn. *Phys. Chem. Chem. Phys.*, 2005. submitted for publication.
75. A. Köhn and C. Hättig. *J. Chem. Phys.*, 119:5021–5036, 2003.
76. P. Pulay, S. Saebø, and K. Wolinski. *Chem. Phys. Lett.*, 344:543, 2001.
77. W. J. Hehre, R. Ditchfield, and J. A. Pople. *J. Chem. Phys.*, 56:2257, 1972.
78. G. E. Scuseria. *Chem. Phys. Lett.*, 176:423, 1991.
79. M. Häser, J. Almlöf, and G. E. Scuseria. *Chem. Phys. Lett.*, 181:497–500, 1991.
80. K. Hedberg, L. Hedberg, D. S. Bethune, C. A. Brown, H. C. Dorn, R. D. Johnson, and M. De Vries. *Science*, 254:410–412, 1991.
81. C. S. Yannoni, P. P. Bernier, D. S. Bethune, G. Meijer, and J. R. Salem. *J. Am. Chem. Soc.*, 113:3190, 1991.

82. J. M. Hawkins, A. Meyer, T. A. Lewis, S. Lorin, and F. J. Hollander. *Science*, 252:312, 1991.
83. J. P. Perdew. *Phys. Rev. B*, 33(12):8822–8824, 1986.
84. J. P. Perdew. *Phys. Rev. B*, 34:7046, 1986.
85. A. D. Becke. *Phys. Rev. A*, 38(6):3098–3100, 1988.
86. F. Furche, 2005. Private communication.
87. B. V. Lebedev, L. Y. Tsvetkova, and K. B. Zhogova. *Thermochimica Acta*, 299:127–131, 1997.
88. A. Köhn and C. Hättig. *J. Am. Chem. Soc.*, 126:7399–7410, 2004.
89. A. Hellweg, C. Hättig, and A. Köhn. *J. Am. Chem. Soc.*, 2005. submitted for publication.
90. W. Rettig, B. Bliss, and K. Dirnberger. *Chem. Phys. Lett.*, 305:8–14, 1999.
91. K. A. Zachariasse. *Chem. Phys. Lett.*, 320:8–13, 2000.
92. K. A. Zachariasse, S. I. Druzhinin, W. Bosch, and R. Machinek. *J. Am. Chem. Soc.*, 126:1705–1715, 2004.
93. S. Techert and K. Zachariasse. *J. Am. Chem. Soc.*, 126:5593–5600, 2004.
94. T. Yoshihara, S. Druzhinin, and K. Zachariasse. *J. Am. Chem. Soc.*, 126:8535–3539, 2004.
95. W. Fuß, K. K. Pushpa, W. Rettig, W. E. Schmid, and S. A. Trushin. *Photochem. Photobiol. Sci.*, 1:255–262, 2002.
96. S. Zilberg and Y. Haas. *J. Phys. Chem. A*, 106:1–11, 2002.
97. W. M. Kwok, C. Ma, P. Matousek, A. W. Parker, D. Phillips, W. T. Toner, M. Towrie, and S. Umaphathy. *J. Phys. Chem. A*, 105:984–990, 2001.
98. J. Dreyer and A. Kummrow. *J. Am. Chem. Soc.*, 122:2577–2585, 2000.
99. D. Rappoport and F. Furche. *J. Am. Chem. Soc.*, 126:1277–1284, 2004.
100. A. Köhn. PhD thesis, Universität Karlsruhe, 2003.
101. C. Bulliard, M. Allan, G. Wirtz, E. Haselbach, K. A. Zachariasse, N. Detzer, and S. Grimme. *J. Phys. Chem. A*, 103:7766–7772, 1999.
102. U. Lommantzsich, A. Gerlach, C. Lahmann, and B. Brutschy. *J. Phys. Chem. A*, 102:6421–6435, 1998.
103. W. Schuddeboom, S. A. Jonker, J. M. Warman, U. Leinhos, W. Kühnle, and K. A. Zachariasse. *J. Phys. Chem.*, 96:10809–10819, 1992.
104. A. B. J. Parusel, G. Köhler, and S. Grimme. *J. Phys. Chem. A*, 102:6297–6306, 1998.
105. S. A. Trushin, T. Yatsushashi, W. Fuß, and W. E. Schmid. *Chem. Phys. Lett.*, 376:282–291, 2003.
106. T. Yatsushashi, S. A. Trushin, W. Fuß, , W. Rettig, W. E. Schmid, and S. Zilberg. *Chem. Phys.*, 296:1, 2004.
107. W. Fuß, W. Rettig, W. E. Schmid, S. A. Trushin, and T. Yatsushashi. *Farad. Disc.*, 127:23, 2004.
108. W. Fuß, 2005. Private communication.

The Korringa-Kohn-Rostoker (KKR)

Green Function Method

II. Impurities and Clusters in the Bulk and on Surfaces

Peter H. Dederichs, Samir Lounis, and Rudolf Zeller

Institute for Solid State Research
Forschungszentrum Jülich
52425 Jülich, Germany
E-mail: {p.h.dederichs, s.lounis, ru.zeller} @fz-juelich.de

1 Green Function Method

In density functional calculations the solution of the Kohn-Sham equations for the single-particle wave functions $\varphi_\alpha(\mathbf{r})$ and the corresponding eigenvalues ε_α , the single-particle energies, represents the central problem. Thus most of electronic structure calculations follow this route, i.e. calculating eigenfunctions φ_α and eigenvalues ε_α . However, the calculation of φ_α and ε_α can be avoided, if instead the single-particle Green function $G(\mathbf{r}, \mathbf{r}'; E)$ of the Kohn-Sham equation is determined, since this quantity contains all the information about the ground state. In particular the charge density and the local density of states can be directly calculated from the Green function, which is the solution of the Schrödinger equation for an energy E with a source at position $\mathbf{r}'$:

$$(-\partial_r^2 + V(\mathbf{r}) - E) G(\mathbf{r}, \mathbf{r}'; E) = -\delta(\mathbf{r} - \mathbf{r}') \quad , \quad (1)$$

with atomic units $\hbar^2/2m = 1$ used. Using the spectral representation for the (retarded) Green function

$$G(\mathbf{r}, \mathbf{r}'; E + i\epsilon) = \sum_{\alpha} \frac{\psi_{\alpha}(\mathbf{r})\psi_{\alpha}^*(\mathbf{r}')}{E + i\epsilon - E_{\alpha}} \quad . \quad (2)$$

it is easy to show that the charge density $n(\mathbf{r})$ can be directly expressed by an energy integral over the imaginary part of the Green function:

$$n(\mathbf{r}) = 2 \sum_{\substack{\alpha \\ E_{\alpha} < E_F}} |\psi_{\alpha}(\mathbf{r})|^2 = -\frac{2}{\pi} \int^{E_F} dE \operatorname{Im} G(\mathbf{r}, \mathbf{r}; E) \quad (3)$$

This relation directly allows calculation of the charge density from the imaginary part of the Green function, which can be interpreted as the local density of states at the position $\mathbf{r}$. The local density of states of a particular atom in a volume V is obtained by integrating over this volume

$$n_V(E) = -\frac{2}{\pi} \int_V d\mathbf{r} \operatorname{Im} G(\mathbf{r}, \mathbf{r}; E) \quad . \quad (4)$$

In this way the evaluation of the wave-functions $\psi_{\alpha}(\mathbf{r})$ can be avoided. Due to the strong energy-dependent structure of the density of states, the evaluation of the energy integral is

usually very cumbersome and typically about 10^3 energy points are needed in an accurate evaluation of this integral.

The numerical effort can be strongly decreased, if the analytical properties of the Green function $G(z)$ for complex energies $z = E + i\Gamma$ are used. Since $G(z)$ is analytical in the whole complex energy plane, the energy integral can be transformed into a contour integral in the complex energy plane

$$n(\mathbf{r}) = -\frac{2}{\pi} \text{Im} \int_{E_B}^{E_F} dz G(\mathbf{r}, \mathbf{r}; z) \quad (5)$$

where the contour starts at an energy E_B below the bottom of the valence bands, goes into the complex plane and comes back to the real axis at the Fermi level. Since for complex energies all structures of the Green function are broadened by the imaginary part Γ , the contour integral can be accurately evaluated using rather few energy points, typically 20-30, leading to a large saving of computer time. In this way Green function methods are competitive to diagonalization methods. Additional advantages occur for systems with two- or three-dimensional symmetry, since as a result of the energy broadening the $\mathbf{k}$ -integration over the Brillouin zone for complex energies requires much less $\mathbf{k}$ -points. In the evaluation of the contour integral, special care is necessary for the piece of the path close to E_F , since here the full structure of $G(E)$ on the real axis reappears. Therefore the energy mesh should become increasingly denser when approaching E_F .

The integration over a complex energy contour can also be extended to finite temperatures by using the analytical properties of the Fermi-Dirac distribution. Here the essential point is that the contour close to E_F is replaced by a sum over Matsubara energies $z_j = E_F + i\pi(2j-1)kT$, $j = 1, 2, \dots$. Then only complex energies are needed, since the energy point closest to E_F has still an imaginary part of πkT . This is of particular advantage, when a discrete k -mesh is used, like e.g. in the special points method.

The real problem is the evaluation of the Green function for the system of interest. For this one does not go back to Eq. (2), since this would mean the evaluation of all eigenvalues ε_α and wave functions φ_α , which one wants to avoid. Rather one relates the Green function G

$$G(E) = \frac{1}{E + i\varepsilon - H} = \frac{1}{E + i\varepsilon - H_o - V} \quad (6)$$

of a system with Hamiltonian $H = H_o + V$ to the Green function $G_o = \{E + i\varepsilon - H_o\}^{-1}$ of a reference system, which is analytically known or easy to calculate. Then $G(E)$ can be obtained from the Dyson equation

$$G(E) = G_o(E) + G_o(E) V G(E) = G_o \frac{1}{1 - VG_o} \quad (7)$$

For instance, for a bulk crystal one starts with the free space Green function $G_o(H_o = -\partial_r^2)$, such that V is the sum of the potentials of all atoms (for details see the lecture of Mavropoulos et Papanikolaou on page 131). For the surface Green function G_o is identified with the bulk Green function, such that V is the difference between the potentials at the surface and in the bulk. Analogously for an impurity in a crystal one starts again with the bulk Green function G_o , such that V represents the change of the impurity potential with respect to the bulk potential as well as the perturbation of the potentials of the neighboring host atoms. Most important is, that the perturbed potential V is well localized near the

impurity, while the perturbed wave functions are not localized and accurately described by the Dyson equation.

Such impurity problems are often described by an 'Ersatzgeometry', e.g. an impurity in a relatively small cluster of bulk atoms or by a supercell geometry with a periodic array of impurities. In these cases the boundary conditions for the wave functions are changed violently, since e.g. for a cluster all wave functions are restricted to the size of the cluster. Therefore the introduction of the host Green function G_o solves the so-called "embedding problem", since it correctly describes the embedding of the impurity in the infinite bulk system. Needless to say, that the Green function method can not only be applied to a single impurity, but also to a small cluster of impurity atoms in the bulk, provided that the perturbation of the potential is localized in a restricted area. Moreover, once the Green function of the surface is known, one can calculate the electronic structure of an impurity or of small clusters at surfaces with an analogous Dyson equation.

2 KKR Green Function Method for Impurities

In the method of Korringa, Kohn and Rostoker (KKR)¹ the Schrödinger equation is solved by multiple scattering theory, describing the propagation of a wave in the solid as a repetition of single scattering events at the different atoms. Thus first the single scattering event of the wave at the potential of the different single atoms n is calculated, described by the single site "t-matrix" t_n , and then the multiple scattering at the given arrangement of the atoms in the crystal. The resulting equations show a beautiful separation between potential and structural properties, which are typical for the KKR method. In the following we summarize here the most important results; for more details, we refer to Ref. 2 and the lecture of Ph. Mavropoulos in this school.

In the KKR Green function method one divides the whole space into non-overlapping and space-filling cells centered at positions $\mathbf{R}^n$. In each cell the electrons are scattered by potentials v^n , which in this section are assumed to be spherically symmetric and centered at $\mathbf{R}^n$. By introducing cell-centered coordinates the Green function $G(\mathbf{r} + \mathbf{R}^n, \mathbf{r}' + \mathbf{R}^{n'}; E)$ can then be expanded in each cell as a function of $\mathbf{r}$ and $\mathbf{r}'$ into spherical harmonics:

$$G(\mathbf{r} + \mathbf{R}^n, \mathbf{r}' + \mathbf{R}^{n'}; E) = \delta_{nn'} \sqrt{E} \sum_L H_L^n(\mathbf{r}_>; E) R_L^n(\mathbf{r}_<; E) + \sum_{LL'} R_L^n(\mathbf{r}; E) G_{LL'}^{nn'}(E) R_{L'}^{n'}(\mathbf{r}; E) \quad . \quad (8)$$

Here $\mathbf{r}$ and $\mathbf{r}'$ are restricted to the cells n and n' and $\mathbf{r}_<$ and $\mathbf{r}_>$ denote the one of the two vectors $\mathbf{r}$ and $\mathbf{r}'$ which has the smaller or larger absolute value. The $R_L^n(\mathbf{r}; E)$ and $H_L^n(\mathbf{r}; E)$ are the product of spherical harmonics and radial eigenfunctions to the central potential $v^n(r)$:

$$R_L^n(\mathbf{r}; E) = R_l^n(r; E) Y_L(\mathbf{r}) \quad (9)$$

$$H_L^n(\mathbf{r}; E) = H_l^n(r; E) Y_L(\mathbf{r}). \quad (10)$$

Here $R_L^n(\mathbf{r}, E)$ is the regular solution which varies at the origin as r^l and which represents the solution for an incoming spherical Bessel function $j_l(\sqrt{E}r)Y_L(\mathbf{r})$, while H_l^n is the corresponding irregular solution varying as $1/r^{l+1}$ at the origin and being identical

with the spherical Hankel function $h_l(\sqrt{E}r)$ outside the range of the potential. Both radial functions are connected by the Wronskian relation, which guarantees that the first term in Eq. (8) represents the exact Green function for the single potential $v^n(r)$ in free space. Since this term satisfies already the source condition $-\delta(\mathbf{r} - \mathbf{r}')$ for the Green function of Eq. (1), the second term is source free and contains in the double angular momentum expansion only the regular solutions R_L^n and $R_{L'}^{n'}$.

By construction, the expression in Eq. (8) for the Green function satisfies in each cell n the general solution of the Schrödinger equation (1) for the Green function, while the matrix $G_{LL'}^{nn'}(E)$, the so-called *structural Green function*, describes the connection of the solutions in the different cells and thus contains all the information about the multiple scattering problem, which is in this way reduced to the solution of an algebraic problem. The clear separation between the single-site properties, described by the radial solutions $R_L^n(\mathbf{r})$ and $H_L^n(\mathbf{r})$ and the multiple scattering properties as described by the matrix $G_{LL'}^{nn'}$, is the main advantage of the KKR method.

In principle, the structural Green function matrix $G_{LL'}^{nn'}(E)$ can be determined by matching the solutions of the neighboring cells at the cell boundaries. However at the cell boundaries the angular momentum expansion converges rather slowly, so that presumably a large $l_{\max}$ cut-off would be needed. The more elegant and at the same time more efficient way consists in using the power of multiple scattering theory, where the Green function is basically only needed in the inner region of the cell, where the potential is strong, so that the l -convergence represents no problem. As shown by Beeby and others¹, the structural Green function matrix can be determined from the corresponding matrix g in free space by the Dyson equation

$$G_{LL'}^{nn'}(E) = g_{LL'}^{nn'}(E) + \sum_{n''L''} g_{LL''}^{nn''}(E) t_{l''}^{n''}(E) G_{L''L'}^{n''n'}(E) \quad (11)$$

where the t -matrix t_l^n for the potential $v^n(r)$ is given by

$$t_l^n(E) = \int_0^R r^2 dr j_l(\sqrt{E}r) v^n(r) R_l^n(r; E) \quad (12)$$

The derivation of this equation is lengthy and straightforward, so that we refer for this to the literature cited above. An elementary derivation, valid also for the full-potential case, has been given by Zeller³.

Once the structural Green function $\overset{\circ}{G}_{LL'}^{nn'}(E)$ of the ideal crystal is known, the Green function $G_{LL'}^{nn'}(E)$ for the crystal with impurity can be evaluated by a modified Dyson equation

$$G_{LL'}^{nn'} = \overset{\circ}{G}_{LL'}^{nn'} + \sum_{n''L''} \overset{\circ}{G}_{LL''}^{nn''} \Delta t_{l''}^{n''} G_{L''L'}^{n''n'} \quad , \quad \Delta t_l^n = t_l^n - \overset{\circ}{t}_l^n \quad (13)$$

where Δt_l^n is the difference $t_l^n - \overset{\circ}{t}_l^n$ between the t -matrices in the perturbed and in the ideal lattice. Since this difference, determined by the perturbation of the potential, is restricted to the vicinity of the impurity, the Green function in this subspace can be easily determined in real space by matrix inversion. The rank of the matrices to be inverted is given by $(l_{\max} + 1)^2 \cdot n_d$, i.e. the number n_d of perturbed potentials times the number $(l_{\max} + 1)^2$ of angular momenta used. Here $l_{\max}$ is the maximum angular momentum used in the calculations, e.g. $l_{\max} = 3$.

For a single impurity it is often sufficient to neglect the perturbation of the neighboring host atoms and to take into account in Eq. (13) only the perturbation due to the impurity potential into account. This so-called single site approximation gives a quite reasonable description of the electronic structure of the impurity and is the essential ingredient of the coherent potential approximation for random alloys. For a more accurate description the perturbations of the neighbors have to be included. The size of the perturbation naturally increases, if impurity pairs, trimers or larger clusters of impurities are included. As a rule one should not only take the strong scattering centers into account, but also the perturbations of all first neighbor atoms.

If we consider an impurity or an adatom on a surface, the structure of the Dyson equation (13) is the same. One has only to replace the host Green function $\overset{\circ}{G}_{LL'}^{nn'}$ by the Green function of the ideal (unperturbed) surface and has to identify Δt_l^n by the change of the t -matrix on site n with respect to the value of the unperturbed surface. Thus it is the structural Green function $\overset{\circ}{G}_{LL'}^{nn'}$, which describes the correct embedding in the local environment. Therefore the calculation of $\overset{\circ}{G}_{LL'}^{nn'}$ represents the high entrance fee one has to pay in Green function calculations.

3 Full-Potentials, Forces and Lattice Relaxations

All-electron methods based on a spherical potential of muffin-tin type or on the atomic sphere approximation (ASA) have in general proven to be very successful and efficient for the description of the electronic structure of solids. However systems with lower symmetry and/or open structures require a more accurate treatment going beyond the spherical approximation. In particular this is necessary, if forces and lattice relaxations are calculated, since for these problems the spherical approximation fails completely.

The fundamental equation (8) for the KKR Green function is also valid in the full-potential case, so that the important separation between the single-potential problem and the multiple-scattering problem fully survives. However the single-site eigenfunctions $R_L^n(\mathbf{r}; E)$ and $H_L^n(\mathbf{r}; E)$ are now the solutions for the general potential $v^n(\mathbf{r})$ being no longer spherical³. For instance, $R_{L'}^n(\mathbf{r}; E)$ is the solution of the Schrödinger equation for a spherical wave $j_{l'}(\sqrt{E}r)Y_{L'}(\mathbf{r})$ incident on the potential $v(\mathbf{r})$

$$R_{L'}(\mathbf{r}; E) = j_{l'}(\sqrt{E}r)Y_{L'}(\mathbf{r}) + \int d^3r' g(\mathbf{r}, \mathbf{r}'; E) v(\mathbf{r}') R_{L'}(\mathbf{r}'; E) \quad (14)$$

where $g(\mathbf{r}, \mathbf{r}'; E)$ is the Green function for free space. Clearly the index L' refers to the angular momentum of the incoming partial wave. Solving Eq. (14) in this form would require a three dimensional integration. By expanding both the potential as well as the wave function $R_{L'}(\mathbf{r}; E)$ into spherical harmonics:

$$v(\mathbf{r}) = \sum_L v_L(r) Y_L(\mathbf{r}) \quad (15)$$

$$R_{L'}(\mathbf{r}; E) = \sum_L R_{LL'}(r) Y_L(\mathbf{r}) \quad (16)$$

we obtain coupled radial equations for the double indexed radial functions $R_{LL'}$

$$R_{LL'}(r; E) = \delta_{LL'} j_{l'}(\sqrt{E}r) + \int_0^S dr' r'^2 g_l(r, r'; E) \sum_{L''} v_{LL''}(r) R_{L''L'}(r'; E). \quad (17)$$

Here the first index L refers to the $\mathbf{r}$ -coordinate of the outgoing partial wave and the second one L' to the angular momentum of the incoming wave. The radial integral extends up to the range S of the potential. Moreover

$$v_{LL'}(r) = \sum_{L''} C_{LL'L''} v_{L''}(r) \quad (18)$$

where $C_{LL'L''} = \int d\Omega Y_L Y_{L'} Y_{L''}$ are the Gaunt coefficients.

The solution of the integral equation (17) or of the equivalent differential equation is rather complicated⁴. In order to avoid numerical problems one transforms Eq. (17) into a modified integral equation⁴, where the effect of the spherical part of the potential is already included in the incident radial wave function $\overset{\circ}{R}_l(r; E)$

$$R_{LL'}(r; E) = \delta_{LL'} \overset{\circ}{R}_l(r; E) + \int_0^S dr' r'^2 G_l(r, r'; E) \sum_{L''} \Delta v_{LL''}(r) R_{L''L'}(r'; E) \quad (19)$$

and where $G_l(r, r'; E)$ is the l -dependent radial Green function for the spherical component of the potential

$$G_l(r, r'; E) = \sqrt{E} \overset{\circ}{H}_l(r_>; E) \overset{\circ}{R}_l(r_<; E) \quad (20)$$

and $\Delta v_{LL'}$ the non-spherical component of the potential

$$\Delta v_{LL'}(r) = \sum_{L'' \neq 0} C_{LL'L''} v_{L''}(r) \quad (21)$$

which provides the coupling between the different angular momenta. Since the non-spherical potential Δv is always rather small, we solve Eq. (19) by iteration, being equivalent to a Born series expansion in powers of Δv . Usually 2-4 iterations are sufficient for convergence.

While for non-spherical potentials the general Eq. (8) for the Green function remains valid and only $R_L(\mathbf{r})$ and $H_L(\mathbf{r})$ have to be replaced by the single-site solutions for the anisotropic potential, the same is also true for the Dyson equations (11) and (13) describing the multiple scattering. Only the t -matrix $t_l^n(E)$ has to be replaced by the t -matrix $t_{LL'}^n(E)$ for a general potential $v^n(\mathbf{r})$ being given by

$$t_{LL'}^n(E) = \int_0^S dr' r'^2 j_l(\sqrt{E}r') \sum_{L''} v_{LL''}^n(r') R_{L''L'}^n(r'; E) \quad (22)$$

Since the Green functions occurring in the Dyson equation are anyhow L, L' -matrices, the numerical effort in solving the multiple scattering problem is the same for both spherical and non-spherical potentials. Therefore the additional numerical effort for full potentials scales in the KKR method only linearly with the number N of non-equivalent atoms,

meaning that in typical calculations the full-potential method does not require a significant increase in computing time.

In case of a full-potential treatment the Wigner-Seitz (WS) spheres used in the atomic sphere approximation (ASA) have to be replaced by the exact WS cells, which are non-overlapping and fill up the whole space completely. This is done by the use of a step function $\Theta(\mathbf{r})$ which is one inside the WS polyhedron and zero outside, and is used to truncate the potential outside the cell. All integrals are convoluted with $\Theta(\mathbf{r})$ which is expanded in spherical harmonics:

$$\Theta(\mathbf{r}) = \sum_L \Theta_L(r) Y_L(\mathbf{r}) \quad . \quad (23)$$

The expansion coefficients can be calculated for polyhedra of arbitrary shape⁵. Note that the expansion (Eq. (23)) is converging very slowly. However this is not a real problem for the calculations, since expanding the wave functions in spherical harmonics and restricting the angular momentum expansion to a cut-off l_{max} imposes a natural cut-off of $2l_{max}$ for the charge density and the potential. Therefore in the evaluation of the Coulomb integrals naturally only Θ_L -coefficients up to $4l_{max}$ are required, and this cut-off is also highly accurate for the exchange-correlation terms.

Force Calculations

An accurate treatment of the full-potential is crucial for the calculation of forces, since in deriving an expression for the force, the extremal properties of the total energy are used, so that the force formula is no longer variationally invariant. By taking the derivative of the total energy with respect to the coordinate $\mathbf{R}^m$ of atom m , the force is given by

$$\mathbf{F}^m = -\frac{\partial E}{\partial \mathbf{R}^m} \Big|_{n(\mathbf{r}; \mathbf{R}^m)} - \int d\mathbf{r} \frac{\delta E}{\delta n(\mathbf{r})} \frac{\partial n(\mathbf{r}; \mathbf{R}^m)}{\partial \mathbf{R}^m} = Z^m \frac{\partial V}{\partial \mathbf{r}} \Big|_{\mathbf{r}=\mathbf{R}^m} \quad (24)$$

The first term, to be evaluated for constant density $n(\mathbf{r}; \mathbf{R}^n)$, is the Hellmann-Feynman (HF) force, being given by the electric field $\frac{\partial V}{\partial \mathbf{r}}$ of the electrons on the nuclear charge Z^m of the nucleus. The second term gives corrections due to approximations made in the solution of the Kohn-Sham equations. It vanishes in an exact treatment, since then $\frac{\delta E}{\delta n(\mathbf{r})} = E_F$ is a constant. Within the full-potential KKR formalism, the Kohn-Sham equations for the valence electrons are solved practically exactly, with the only approximation being the l_{max} cut-off. However, the use of the HF-formula, i.e. the first term in Eq. (24), requires also a full-potential treatment of the core electrons. If an atom is shifted, the charge density of the core electrons experiences in a solid a small anisotropic distortion induced by the crystal field, which leads to an important contribution to the force on the nucleus and which unfortunately cannot be described in a spherical-core treatment. This problem can be overcome by making a spherical ansatz for the core density entering in the total energy expression. The force is then calculated as the derivative of the total energy with respect to the nuclear position assuming that the Kohn-Sham equations are solved exactly for the valence electrons only. The resulting expression for the force $\mathbf{F}^m$ on the atom $\mathbf{R}^m$ is given by⁶

$$\mathbf{F}^m = Z^m \frac{\partial V_M(\mathbf{r})}{\partial \mathbf{r}} \Big|_{\mathbf{r}=\mathbf{R}^m} - \int d\mathbf{r} n_c(\mathbf{r} - \mathbf{R}^m) \frac{\partial V(\mathbf{r})}{\partial \mathbf{r}} \quad (25)$$

where Z^m is the nuclear charge and $n_c(\mathbf{r} - \mathbf{R}^m)$ the core charge density of atom m . Furthermore $V_M(\mathbf{r})$ is the Madelung potential and $V(\mathbf{r})$ the Kohn-Sham potential. While the first term is the force on the nucleus as given by the Hellmann-Feynman theorem (but without the contribution from the core electrons at atom $\mathbf{R}^m$), the second term represents the force on these core electrons and also includes an exchange-correlation contribution, arising from the exchange between valence and core electrons. Thus basically Eq. (25) gives the force on the ion consisting of the nuclear charge and the core charge of atom m . Due to the vector character of the potential derivatives in Eq. (25), only the $l = 1$ components of the potentials $V_M(\mathbf{r})$ and $V(\mathbf{r})$ are needed for the force. Since in the present full-potential treatment the coefficients $V_L(r)$ are anyhow calculated during the selfconsistency cycles, the calculation of the force does not require additional efforts. Moreover the $l = 1$ components of the potentials are essentially determined by the $l = 1$ components of the valence charge density $n(\mathbf{r})$. Therefore one obtains only contributions from the interference of wave functions differing by $\Delta l = \pm 1$, i.e. sp, pd, fg, ... interference terms.

Lattice Relaxations

In contrast to the simplicity of the force calculation, the description of lattice relaxation effects is rather complicated within the KKR method. The main reason is the site-centered angular momentum expansions used in the Green function. In the case of lattice relaxations one needs an angular momentum expansion around the shifted position, i. e. around a non-lattice site. While in principle the host Green function can be calculated for any interstitial site by introducing a supercell with additional empty positions, this is a cumbersome procedure for the small lattice relaxations occurring for substitutional defects. In this case a transformation formalism, used e.g. in Ref. 6, is more convenient. The structural host Green function is transformed from the unshifted coordinates to the new ones being shifted by $\mathbf{s}^n$

$$\tilde{G}_{LL'}^{nn'}(E) = \sum_{L''L'''} U_{LL''}(\mathbf{s}^n; E) G_{L''L'''}^{nn'}(E) U_{L'L'''}(\mathbf{s}^{n'}; E) \quad (26)$$

where

$$U_{L'L}(\mathbf{s}; E) = 4\pi \sum_{L''} i^{l'+l''-l} C_{LL'L''} j_{l''}(\sqrt{E}s) Y_{l''}(\mathbf{s}) \quad (27)$$

The $\tilde{G}$ is the host Green function but expanded in the shifted coordinate system. An analogous U transformation has to be done for the t-matrix. Finally we must solve the following Dyson equation for the structural Green function to obtain the new Green function for potentials or t-matrices on the shifted sites.

$$\tilde{G} = \tilde{G}^0 + \tilde{G}^0 [t - \tilde{t}^0] \tilde{G} \quad (28)$$

where $\tilde{G}^0$ and $\tilde{t}^0$ are the host Green function and host t-matrix in the angular momentum expansion around the shifted sites (Eq. (26)). Thus, apart from the U -transformation, the structure of the Dyson equation is unchanged. While the U -transformation Eq. (26) is exact, if the sums over L'' and L''' are extended over infinite angular momenta, in practical calculations an l_{max} -cut-off is used. As can be seen from Eq. (27), for small s the $U_{LL'}$ -matrix couples states with $l - l' = \pm 1$. Thus a relatively high l_{max} has to be chosen and

the error increases with increasing displacements. Typically calculations for substitutional defects with $l_{\max} = 4$ are sufficiently accurate up to displacements of 10 % of the nearest neighbor distance. For larger displacements or interstitial defects the Green function $\tilde{G}_{LL'}^{nn'}$ has to be determined by Brillouin zone integration.

4 Impurities in Metals and Semiconductors

4.1 Lattice Relaxations around Impurities

Substitutional impurities have in general a different size than the host atoms, which leads to lattice displacements of the neighboring host atoms away from their positions in the ideal lattice. Due to the high coordination number, in metallic systems these relaxations are usually very small, not more than a few percent of the nearest neighbor distance, so that calculations without lattice relaxations give usually quite reasonable results. As an example Figure 1 shows the displacements of the nearest neighbors (NN) around 3d impurities in Cu⁶. The displacements are given in terms of the NN distance. In most cases the Cu lattice is dilated due to the impurities, except for the case of Fe, Co and Ni. The triangles with error bars are the results of EXAFS measurements, which allow a direct determination of the relaxations. The bump in the curve with the maximum at Cr is a magneto-elastic effect. The transition metal impurities Cr, Mn and Fe are magnetic with rather large local moments of $2.9 \mu_B$ (Cr), $3.4 \mu_B$ (Mn) and $2.5 \mu_B$ (Fe). Calculations without spin polarization lead the expected parabolic behavior of the displacements across the 3d-series.

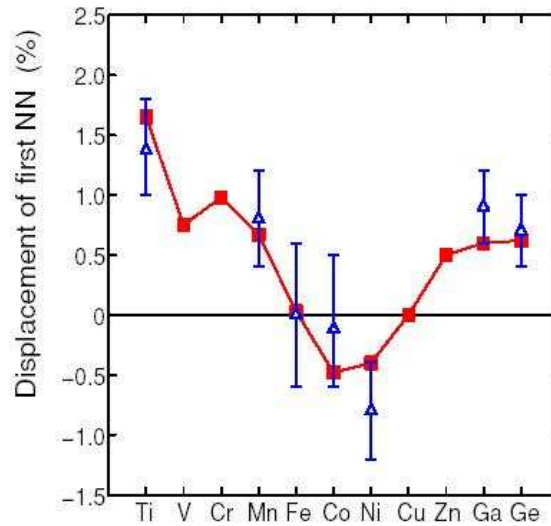


Figure 1. Calculated lattice relaxations around 3d and 4sp impurities in Cu. The displacements of the nearest neighbor Cu atoms (in percentage of the nearest-neighbor distance) are given. The triangles with error bars refer to EXAFS results⁶.

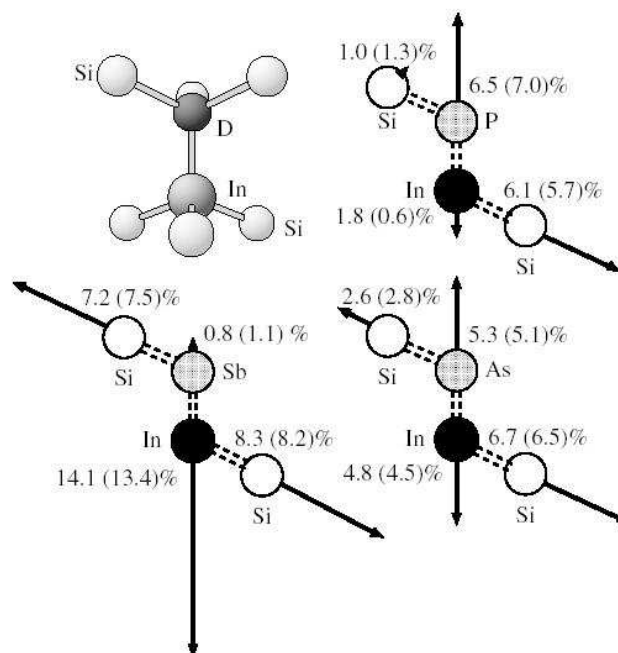


Figure 2. Ground-state configuration of In-donor complexes in Si (InSb, InP, InAs). The numbers denote the relaxations in percentage of the nearest-neighbor distance. The first number refers to KKR calculations, the second one (in brackets) to pseudopotential results⁷.

In contrast to metallic alloys lattice relaxations around defects in semiconductors can be much larger, being a result of the more open lattice structure and the lower coordination number of the diamond or zinc blende lattice. Here we discuss the relaxations around acceptor-donor complexes on nearest neighbor sites in Si, such as InSb, InAs and InP pairs in Si. These defects are electrically and magnetically inactive and experimental information about the structure is difficult to obtain. One of the few methods to investigate such defects are perturbed angular correlation (PAC) experiments which measure the electric field gradients. In Figure 2 we present the calculated atomic configurations for In-P, In-Sb, In-As pairs in Si. The full-potential KKR results are compared with the results obtained from pseudopotential calculations. The atomic configurations obtained using both *ab initio* methods are essentially the same, but the KKR can give direct access to properties that are determined by the core electrons, like hyperfine fields or electric field gradients. The electric field gradients (EFG) of the Cd-donor pairs in Si and Ge have been measured. While calculations without lattice relaxations give the wrong trend with respect to the atomic numbers of the donor atoms, the agreement greatly improves, if the relaxed configurations, as e.g. given in Figure 2 for the corresponding In-pairs in Si, are considered⁷. Thus a reliable calculation of the relaxations is decisive for understanding the EFG. (The EFG is basically determined by the second derivative of the electrostatic potential at the nucleus. Thus it has some similarity to the force, which is determined by the first derivative at the nucleus.)

4.2 Isomer Shifts and Charge Transfer in Fe Alloys

Isomer shifts of Mössbauer nuclei are directly related to the local charge density $\rho(0)$ at the nuclear position. They therefore provide unique information about charge-density changes upon alloy formation. Unfortunately the changes $\Delta\rho(0)$ are very much smaller than the total charge density $\rho(0)$ at the nucleus, $\Delta\rho(0) \sim 10^{-6}\rho(0)$, so that great numerical precision is needed to calculate isomer shifts. The standard Mössbauer nucleus is Fe. Therefore a large number of Mössbauer data exist for Fe alloys, in particular dilute Fe alloys. Here in addition to the bulk Mössbauer line a satellite is measured arising from the charge perturbation of the Fe nuclei adjacent to the impurities and being shifted from the main line. It is generally believed that these lines give direct information about the charge transfer in these alloys.

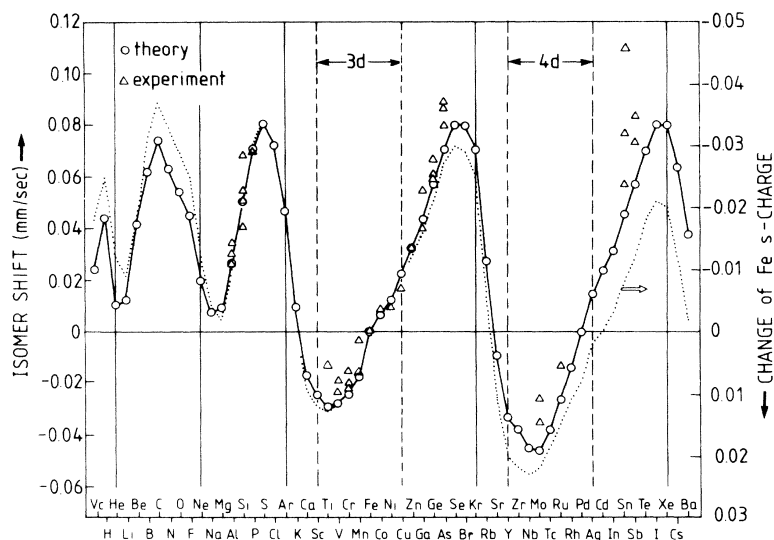


Figure 3. Calculated isomer shift of a nearest neighbor Fe atom as a function of the nuclear charge Z of the impurity (left hand scale). The triangles are experimental values for dilute Fe alloys. The dotted line gives the change Δn_s of the number of Fe s electrons and refers to the right-hand scale⁸.

Figure 3 shows the calculated isomer-shift values for the NN Fe atom around the impurities and the comparison with the experimental data for these alloys as a function of the nuclear charge Z of the impurity⁸. The triangles refer to experimental data. The general agreement is fairly good and the experimental trends are well reproduced. In general the shifts show a minimum at the beginning of each row of the periodic system and slowly increase to a maximum at the end of the row, dropping sharply to the minimum at the beginning of the next row, etc. However, the periodicity is far from being perfect. For instance, with increasing row number the maxima systematically shift to larger Z values and the minima get more pronounced. We also see systematic deviations between the calculated and experimental values, in particular for the 4d and 5p impurities. For these oversized impurities lattice relaxations are expected to be significant, which are not included in the

calculations.

Since a negative (positive) isomer shift means that the charge $\rho(0)$ at the NN Fe nucleus has increased (decreased), the behavior seen in Figure 3 follows simple electronegativity arguments that the charge flows from the element with the lower electronegativity to the one with the higher electronegativity. For instance, in the 3d series the charge should flow from the impurity to the NN Fe atoms if the impurity is on the left-hand side of Fe in the periodic table, whereas the charge should flow to the impurity if it is on the right-hand side of Fe. However, this is certainly a too-simplified description, since the isomer shift only measures the change of the charge density $\Delta\rho(0)$ at the Fe nucleus, which moreover is only of s character. No direct information is obtained about p - or d -electrons. The question therefore arises as to how the changes $\Delta\rho(0)$ are related to the changes of the valence charges, i.e., the changes Δn_s , Δn_p , and Δn_d of the numbers of s , p , and d valence electrons of the NN Fe atom.

Indeed the calculations show that the trend of the isomer shifts of the NN Fe atoms is determined by the change of the Fe s -valence charge, which is shown by the dotted line in Figure 3. Compared to this the change of the s -core charge density at the origin is relatively small and shows no strong systematic trend. Moreover the changes of the p - and d -charges, which might screen the s -wave function and in this way might change the s -density at the nucleus, are not significant. In fact, if this would be the case, the trends for the isomer shifts would be very different, since in the sp -impurity series the Fe d -charge strongly increases, will the s -charge moderately decreases. Thus we see, that for the dilute Fe-alloys, the isomer shift is determined by the change Δn_s of the valence s -charge⁸, which seems to be generally valid in metallic alloys.

5 Non-Collinear Configurations of 3d Impurities on Ferromagnetic Surfaces

In this section we consider some recent calculations⁹ for non-collinear configurations of 3d-dimers and multimers on the surfaces of ferromagnets. Here the unperturbed surfaces are ferromagnetic with a collinear moment configuration. Thus the Green function $\overset{\circ}{G}$ and t -matrix $\overset{\circ}{t}$ of the ideal surface are diagonal in spin-space

$$\overset{\circ}{G}(E) = \begin{Bmatrix} \overset{\circ}{G}_{\uparrow\uparrow}(E) & 0 \\ 0 & \overset{\circ}{G}_{\downarrow\downarrow}(E) \end{Bmatrix} \quad \overset{\circ}{t}(E) = \begin{Bmatrix} \overset{\circ}{t}_{\uparrow\uparrow} & 0 \\ 0 & \overset{\circ}{t}_{\downarrow\downarrow} \end{Bmatrix} \quad (29)$$

while non-collinear states lead to non-diagonal t -matrices for the impurity atoms and the surrounding substrate neighbors.

$$t(E) = \begin{Bmatrix} t_{\uparrow\uparrow} & t_{\uparrow\downarrow} \\ t_{\downarrow\uparrow} & t_{\downarrow\downarrow} \end{Bmatrix} \quad (30)$$

The basic approximation with respect to non-collinearity consists of the assumption, that the exchange-correlation potential of each atom $\mathbf{R}^n$ has a unique quantization axis $\mathbf{e}^n$, being common to the whole cell n and determined by the direction of the local moment $\mathbf{M}^n$ in cell n . In this local reference frame, the t -matrix t^n is diagonal

$$t_{loc}^n = \begin{Bmatrix} t_{\uparrow\uparrow}^n & 0 \\ 0 & t_{\downarrow\downarrow}^n \end{Bmatrix} \quad (31)$$

and the local radial functions R_ℓ^n and H_ℓ^n are spin dependent as in a collinear calculation. However the Dyson equation describing the multiple scattering events has to be evaluated in a common global frame of reference, as e.g. determined by the magnetization direction of the substrate. The corresponding transformed t^n -matrices are given by

$$t_{glob}^n(E) = U_n t_{loc}^n(E) U_n^t \quad (32)$$

where the rotation matrix U_n of spin space is given by

$$U_n \begin{pmatrix} \cos(\frac{\theta_n}{2}) e^{-i\phi_n/2} & -\sin(\frac{\theta_n}{2}) e^{-i\phi_n/2} \\ \sin(\frac{\theta_n}{2}) e^{i\phi_n/2} & \cos(\frac{\theta_n}{2}) e^{i\phi_n/2} \end{pmatrix} \quad (33)$$

Here θ_n and ϕ_n are the polar angles defining the direction of the local moment $\mathbf{M}^n$ with respect to the substrate moments.

The basic reason for non-collinear states is "frustration", arising from the competition between ferromagnetic and antiferromagnetic coupling. In addition, also spin-orbit coupling can lead to a non-alignment of the local moments. However this is a very weak effect for transition metals, for which frustration is much more important. We will illustrate this in the following for transition metal dimers on the Ni(001) surface.

Let us start with single 3d adsorbate atoms on Ni(001). The calculations show, that the 3d adatoms have large and stable local moments. The moments of the Co, Fe and Mn adatoms couple ferromagnetically to the substrate moments, while the V and Cr moments prefer an antiferromagnetic coupling to the substrate. The situation of two 3d-adatoms forming a dimer is illustrated in Figure 4. Three kind of dimers are shown: Dimer 1 with the adatoms on nearest neighbor sites, Dimer 2 with the adatoms on second neighbor sites and Dimer 3 on fourth neighbor site. For the Dimer 2 and dimers with larger separation the interaction of the dimer atoms is very small and the configuration is dominated by the interaction with the substrate, meaning that these dimers show the same behavior as the single adatoms, coupling antiferromagnetically to the substrate in the case of V and Cr and ferromagnetically in the case of Mn and Fe, such that both adatoms are parallel aligned to each other. The same is also correct for the NN dimers of Fe or V, where the dimer atom interaction is strongly ferromagnetic (for Fe) or weakly antiferromagnetic (for V). In the case of the Cr and Mn dimers the situation is more complicated, since the interaction of the dimer atoms is strongly antiferromagnetic, favoring an antiferromagnetic pairing of the two moments. However, this is in contradiction to the interaction with the substrate moments, which as explained above, favors a parallel alignment of the impurity moments. Therefore frustration occurs, which can lead to a non-collinear ground state.

The situation is most easily explained, if a model operator in form of the classical Heisenberg model

$$H = -\frac{1}{2} \sum_{i \neq j} J_{ij} \mathbf{e}_i \cdot \mathbf{e}_j \quad (34)$$

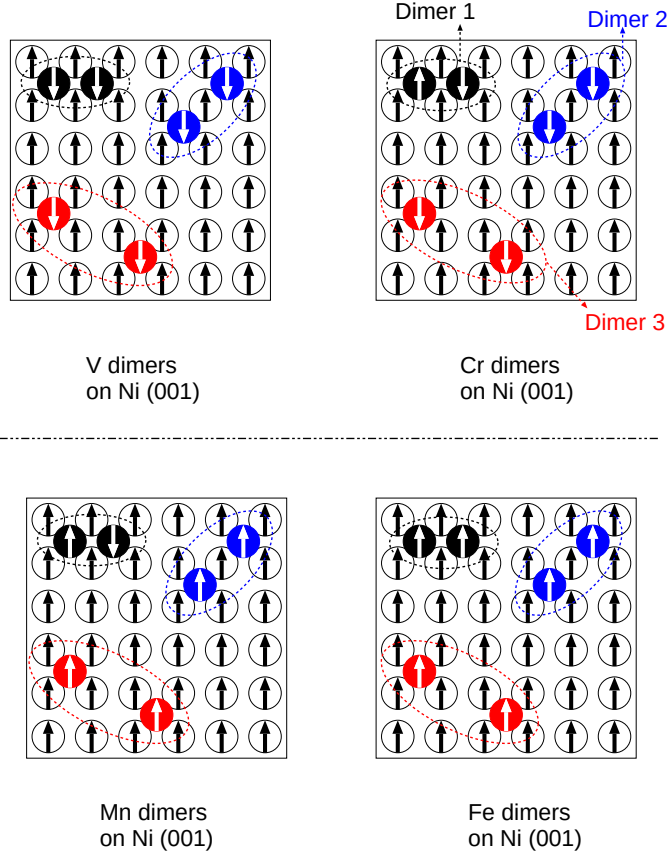


Figure 4. Different geometrical configurations considered for dimers at the surface of Ni(001). Dimer-1-type corresponds to the case where the atoms are first neighboring atoms, dimer-2-type where the atoms are second NN and finally dimer-3-type to fourth NN. The collinear magnetic ground state are also shown for V, Cr, Mn and Fe dimers.

is used. Here J_{ij} is the exchange integral between the atoms i and j and $\mathbf{e}_i$ defines the direction of the local moment M_i . By applying this to the interaction of the two adatoms $A = 1, 2$ and their interaction with the Ni moments, which for simplicity are assumed to be fixed, the Hamiltonian is

$$H = -J_{A-A} \cos(\theta_1 - \theta_2) - 4J_{A-Ni}(\cos \theta_1 + \cos \theta_2) \quad (35)$$

where θ_1 and θ_2 are the angles with the respect to the substrate magnetization.

Let us now consider two typical spin configurations, shown in Figure 5(a) and Figure 5(b). Figure 5(a) refers to a collinear configuration, which we call ferrimagnetic, since the two moments, being antiferromagnetically aligned, are not equivalent anymore, resulting in a small, but finite total moment. This configuration is also a selfconsistent solution of the Kohn-Sham equations, if the collinear constraint is removed. This can be

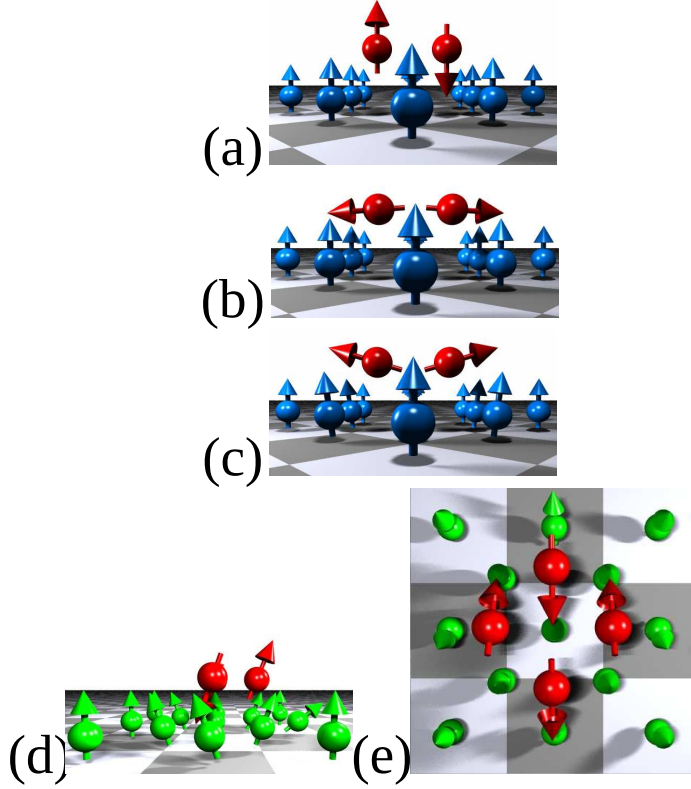


Figure 5. Most stable configurations of Cr and Mn dimer on Ni (in blue) obtained with (a) the collinear KKR method and (b)-(c) the non-collinear KKR method⁹. For Cr(Mn) the rotation angle with respect to the z axis is equal to 94.2° (72.6°) and the collinear(non-collinear) state is the ground state (see text). (d) shows a side view of the stable Cr trimer on the fcc Fe(001) surface (in green), with two Cr atoms pointing down (the second one cannot be seen), one Cr atom pointing up. (e) shows the stable Cr tetramer on the same fcc Fe surface.

understood e.g. from the Heisenberg model Eq. (35), since $\theta_1 = 0$, $\theta_2 = 180^\circ$ and small variations around these angles change the cos-values in Eq. (35) only in second order, so that the total energy is an extremum. The configuration in Figure 5(b) is non-collinear, but has the same energy as the collinear configuration (a), since in configuration (a) the interaction of the two adatoms with the substrate atoms cancel each other, while in configuration in Figure 5(b) they vanish for both atoms since $\cos \theta_1 = 0 = \cos \theta_2$. However this configuration is not a selfconsistent solution of the non-collinear Kohn-Sham equations, since a small variation $\Delta\theta_1$ and $\Delta\theta_2$ around the values of 90° , respectively 270° , changes the energy linearly in $\Delta\theta_1$ and $\Delta\theta_2$. Thus there exists a force which tilts the moments slightly towards or away from the surface, depending on the sign of J_{A-Ni} . In fact the configuration (b) is the non-collinear solution for a Cr dimer. With a rotation angle of 94° , deviating only slightly from 90° (which can hardly be seen in the figure), a small energy is gained due to the antiferromagnetic coupling with the substrate ($J_{Cr-Ni} < 0$). In contrast to this the configuration in Figure 5(c) is the selfconsistent solution for a Mn-dimer, which prefers a ferromagnetic coupling with the substrate atoms ($J_{Mn-Ni} > 0$). Here the angle

with respect to the z-axis is 73° , the deviation from 90° is much larger. The *ab initio* calculation shows, that this is the ground state for the Mn-dimer. However for the Cr dimer the collinear solution of Figure 5(a) is the ground state, which is in contradiction to the Heisenberg model and arises from small changes of the local moments upon rotation, an effect which cannot be described by this model.

In Figure 5(d) and (e) we show two other non-collinear configurations obtained in the *ab initio* calculations, the configurations for compact Cr trimers and Cr tetramers on fcc $\text{Fe}_{3ML}/\text{Cu}(001)$. In this case the exchange interactions are very similar, except that the antiferromagnetic coupling of the Cr adatoms to the Fe substrate atoms is considerably stronger. In both cases the Cr-Cr interaction is strongly antiferromagnetic. For the trimer, it is most important, that the effective interaction with the substrate moments is non-zero in the collinear configuration, but zero in the planar configuration. Thus the (basically) collinear configuration with the outer Cr atoms antiferromagnetically aligned to the surface moments and the central Cr atom ferromagnetically aligned is favored. However an additional small tilting occurs, in particular for the wrongly aligned central Cr atom, which further lowers the energy, so that also this configuration becomes non-collinear.

For the tetramer, the neighboring Cr atoms couple again antiferromagnetically. For the in-plane configuration, similar to the dimer, the effective interaction with the substrate moments vanishes, however slight tiltings of the moments towards the surface lead to an additional energy gain stabilizing the in-plane configuration.

6 Quantum Corrals on the Cu(111) Surface

Over the last decades a great deal of experimental and theoretical efforts has been devoted to study electrons in two-dimensional (2D) surface states. Here the (111) surface of the noble metals has served as a model system, exhibiting a surface state in the gap around the L -point of the bulk Brillouin zone. For Cu(111) this state shows a parabolic dispersion with a minimum at 0.39 eV below the Fermi level. The corresponding band structure projected on the $\bar{\Gamma} - \bar{M}$ line of the 2D-Brillouin zone is shown in Figure 6. The shaded regions give the regions in $E - \mathbf{k}_{\parallel}$ space, for which bulk eigenstates (Bloch waves) exist. Surface states can only exist in the white "gap"-regions. Two such states are indicated. Of special interest is the parabolic band with the minimum close to E_F , since this state is only partially occupied and gives rise to a two-dimensional metallic behavior, which is of great interest for the following.

In this lecture we are interested in defects and small clusters in the bulk and on surfaces. As is well known, in metallic systems point defects cause long ranged charge oscillations (Friedel oscillations), which are governed by the Fermi surface properties. For defects in the bulk, these Friedel oscillations of the charge perturbation vary for large distances r as $1/r^3$ times an oscillatory function and are in the jellium model proportional to:

$$\delta n(r) \sim \frac{\cos(2k_F r + \delta)}{r^3} . \quad (36)$$

However in the case of adatoms on surfaces, the charge response decays for long in-plane distances ρ slower than in the bulk and is determined by the surface states. In a free electron model, being well suited for the above surface state for Cu(111), for large distances the charge density is ρ proportional to

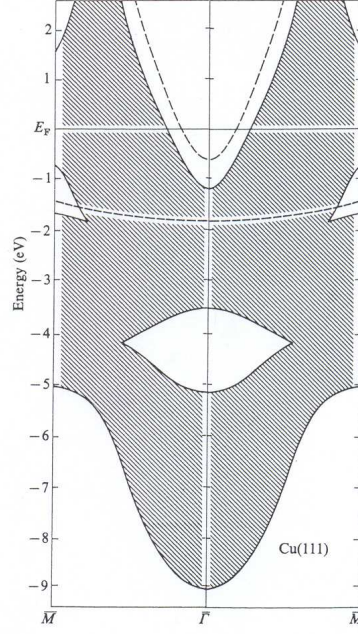


Figure 6. Surface states (dashed curves) and bulk projected bands at a Cu(111) surface according to a six-layer surface band structure calculation¹⁰.

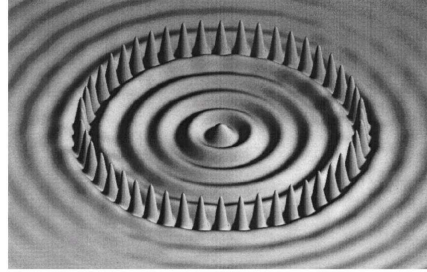


Figure 7. LDOS at the Fermi energy in and around a 142.8 Å diameter circular corral of 48 Fe atoms on Cu(111)¹².

$$\delta n(\rho) \sim \frac{\sin(2k_F \rho + \delta)}{\rho^2} . \quad (37)$$

However, since also bulk states exist, which span most of the phase space (see Figure 6), the short range screening of the defect is dominated by these states, while only the long ranged behavior is determined by the surface state, which has a small wave vector k_F leading to long wave length oscillations.

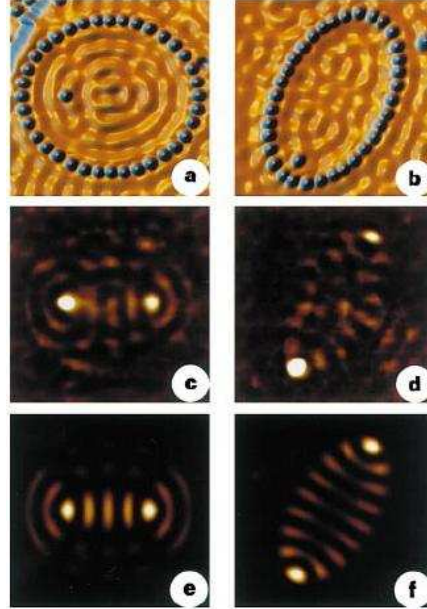


Figure 8. Visualization of the quantum mirage. **a, b**, Topographs showing the $e = \frac{1}{2}$ (**a**) and $e = 0.786$ (**b**) ellipse each with a Co atom at the left focus. **c, d**, associated $\frac{dI}{dV}$ difference map showing the Kondo effect projected to the empty right focus, resulting in a Co atom mirage. **e** and **f**, Calculated eigenmodes at E_F (magnitude of the wave function is plotted)¹³.

Many authors have observed such long ranged oscillations around adatoms, small clusters and steps on the Cu(111) surface in STM experiments. Most prominent among these is the work of Eigler *et al.*¹². By atomic manipulations they were able to construct a corral of Fe atoms on the (111) Cu surface, and have shown that the surface states in the corral are more or less localized and form a discrete spectrum of resonant states. As an illustration of these we show in Figure 7 the result of KKR calculations of Crampin *et al.*¹¹ for a circular corral of 48 Fe atoms on the Cu(111) surface. Shown are the local density of states at the Fermi energy and 5 Ångström above the surface. Within the corral one sees a quantum well state with five maxima, corresponding to a localized state being more or less completely confined to the corral. Outside one sees oscillations arising from scattered surface state electrons at the corral, which decay with distance.

Let us shortly discuss the reason for the strong scattering of the surface state electrons at the Fe atoms. Basically in the vacuum region the full-potential of Fe acts as a scattering center for the surface wave, being much stronger than the scattering at an Fe impurity in the bulk, where only the change of the Fe potential with respect to the host potential is effective. Moreover the wave vector k_F is relatively small, such that the wave length is considerably larger than the spacing between the Fe atoms. Therefore the surface wave does not “see” the corrugation of the Fe ring and is strongly reflected as in cylindrical well. In fact the sequence of resonances can be well described by such a quantum well model, as has been shown recently¹⁴. The most fascinating corral experiments are the observation

of atomic mirages in an elliptical quantum well¹³. An ellipse has the well known property that all classical waves emanating from one of the two focus points in every direction are reflected from the ellipse wall and focused in the second point, where these waves add up coherently since each such partial wave has the same path length and therefore the same phase shift. This is illustrated in Figure 8 taken from the letter of Manoharan *et al.*¹³. Figure 8(a) and (b) show the STM topography for two ellipses with different eccentricities, each including one Co atom at the left focus point. Figure 8(c) and (d) show the $\frac{dI}{dV}$ difference maps, *i.e.* the change of the STM intensity map with respect to a small bias voltage V , which corresponds in the calculations to the local density of states in the vacuum region at the height of the STM tip. We see clearly two intensity spots, the real Co atom at the left focus and its image at the right focus. Thus in the empty focus we see the same accumulation of charge in the surface state as around the Co atom; therefore the image is called a quantum mirage. In fact the Co atom is a Kondo impurity and a strong and sharp Kondo peak appears only in a very small energy region of about 10 meV around the Fermi level. Moreover the large mirage only appears, if one of the quantum well states falls into this energy region. Figure 8(e) and (f) show the calculated localized eigenstate observed in the experiment. The calculated local density of states compares very well with the $\frac{dI}{dV}$ curves shown in Figure 8(c) and (d). Thus several conditions have to be satisfied for the Co mirage to appear: (i) the Co-atom has to sit in a focus point; if it sits at another position away from the focus point, no image appears, (ii) the bias voltage has to be such, that it coincides with an eigenstate of the ellipsoidal corral having maxima at the focus points, (iii) finally the image is particularly intense, if the eigenvalue coincides with the Kondo resonance.

7 Conclusions

In the lecture of Mavropoulos and in this one you have seen that the KKR Green function method is an excellent tool for ground state calculations, which is particularly well suited for nanostructures like surfaces, multilayers, impurities and clusters on surfaces etc. However, compared to other methods it is somewhat complicated and demanding, in particular when lattice relaxations are needed. Its basic advantage is the availability of the Green function, which allows to do more than just ground state calculations. One can do linear response with respect to an electric or magnetic field and can treat quantum transport based on Kubo-Greenwood or Landauer formalism. Another example are calculations for disordered alloys based on the coherent potential approximation (KKR-CPA). Also for the problem of excited states the availability of the Green function is an important plus.

References

1. T. H. Dupree, Ann. Phys. (N. Y.), **15**, 63 (1961); J. L. Beeby, Proc. Roy. Soc. London Ser. A **302**, 113 (1967); G. J. Morgan, Proc. Phys. Soc. **89**, 365 (1966).
2. W. Kohn and N. Rostoker, Phys. Rev. **94**, 1111 (1954).
3. R. Zeller, J. Phys. C **20**, 2347 (1987).
4. B. Drittler, Dissertation, Rheinisch-Westfälische Technische Hochschule Aachen (1991).

5. N. Stefanou and R. Zeller, J. Phys.: Condens. Matter **3**, 7599 (1991).
6. N. Papanikolaou, R. Zeller, P. H. Dederichs, and N. Stefanou, Phys. Rev. B **55**, 4157 (1997).
7. A. Settels, T. Korhonen, N. Papanikolaou, R. Zeller, and P. H. Dederichs, Phys. Rev. Lett. **83**, 4369 (1999).
8. H. Akai, S. Blügel, R. Zeller, and P. H. Dederichs, Phys. Rev. Lett. **56**, 2407 (1986).
9. S. Lounis, Ph. Mavropoulos, P. H. Dederichs, and S. Blügel, Phys. Rev. B **72**, 224437 (2205).
10. A. Euceda, D. M. Bylander, and L. Kleinman, Phys. Rev. B **28**, 528 (1983).
11. S. Crampin and O. R. Bryant, Phys. Rev. B **54**, 17367 (1996).
12. M. F. Crommie, C. P. Lutz, and D. M. Eigler, Science **262**, 218 (1993).
13. H. C. Manoharan, C. P. Lutz, and D. M. Eigler, Nature **403**, 512 (2000).
14. B. Lazarovits, B. Újfalussy, L. Szunyogh, B. L. Györffy and P. Weinberger, J. Phys.: Condens. Matter **17** S1037 (2005).

The Optimized Effective Potential Method and LDA + U

Stefan Kurth and Stefano Pittalis

Institute of Theoretical Physics
Freie Universität Berlin
Arnimallee 14, 14195 Berlin, Germany
E-mail: Stefan.Kurth@physik.fu-berlin.de

The fundamental quantity of density functional theory is the exchange-correlation energy functional. The exact form of this functional is unknown and it has to be approximated in practice. The exchange-correlation potentials corresponding to approximate functionals which depend explicitly on the Kohn-Sham orbitals have to be computed with the optimized effective potential (OEP) method. Here, the OEP method is reviewed and commonly used orbital functionals are introduced. One particular approach, the so-called LDA+U method which is often applied to describe strongly correlated systems is discussed from the OEP perspective. Finally, a few selected numerical results for atoms, molecules, and solids are presented.

1 Introduction

Density functional theory (DFT) has become the most widely used tool to study the electronic structure of atoms, molecules and solids. As it is used today, it essentially is an ingenious reformulation of the many-body problem. Instead of trying to solve the Schrödinger equation of interacting electrons directly, the problem is cast in a way such as to make it tractable in an approximate, but in many cases surprisingly accurate way. The success of DFT is largely due to the availability of increasingly accurate approximations to the central quantity of DFT, the so-called exchange-correlation energy functional. While the simple local density approximation (LDA) proved to be surprisingly accurate especially in solid state physics, only the advent of the so-called generalized gradient approximations (GGA's) with their increased accuracy led to an explosion of applications of DFT in quantum chemistry.

The development of new, improved functionals is an ongoing effort. In this contribution we are dealing with a particular class of approximations which are explicit functionals of the Kohn-Sham orbitals rather than explicit functionals of the density (such as LDA or GGA). Treatment of orbital functionals in the DFT framework requires the use of the so-called optimized effective potential (OEP) method to compute the corresponding effective single-particle potentials. Here, we will review this method, and derive its central equation. Several approximate and exact schemes to solve this equation will be discussed as well as some properties of the resulting potentials. An extension of the OEP method for the description of non-collinear magnetism will briefly be described.

We discuss possible ways to construct orbital functionals and present some of them which have been used in the literature. Furthermore, we discuss a connection between orbital functionals and the so-called LDA+U method. This method tries to combine insights gained from studies of model Hamiltonians into the DFT framework in order to describe strongly correlated systems. Finally, we discuss some numerical results obtained with orbital functionals for atoms, molecules, and solids.

2 Kohn-Sham Density Functional Theory

We are interested in systems of N interacting electrons moving in some external electrostatic potential $v(\mathbf{r})$. The Hamiltonian to describe such systems is given by (atomic units are used throughout)

$$\hat{H} = - \sum_{i=1}^N \frac{\nabla_i^2}{2} + \sum_{i=1}^N v(\mathbf{r}_i) + \frac{1}{2} \sum_{\substack{i,j=1 \\ i \neq j}}^N \frac{1}{|\mathbf{r}_i - \mathbf{r}_j|} := \hat{T} + \hat{V} + \hat{V}_{ee} \quad (1)$$

with the kinetic energy operator $\hat{T}$, the potential energy $\hat{V}$ and the operator $\hat{V}_{ee}$ of the electron-electron interaction.

The first key insight is given by the Hohenberg-Kohn theorem¹ which states that the ground-state energy of a system of interacting electrons can be written as a functional of the electron density $n(\mathbf{r})$ alone and takes its minimum at the true ground state density. This functional can be written as

$$E_v[n] = F[n] + \int d^3r v(\mathbf{r})n(\mathbf{r}) \quad (2)$$

where $F[n]$ is a *universal* functional of the density, i.e., it is independent of the external potential $v(\mathbf{r})$ and therefore it is the same functional of the density for all Coulombic systems such as atoms, molecules and solids. Using the “constrained search” formulation of Levy^{2,3}, this functional may formally be defined as

$$F[n] = \min_{\Psi \rightarrow n} \langle \Psi | \hat{T} + \hat{V}_{ee} | \Psi \rangle = \langle \Psi_n^{\min} | \hat{T} + \hat{V}_{ee} | \Psi_n^{\min} \rangle \quad (3)$$

where the search runs over all possible, antisymmetric N -electron wave functions which yield the density $n(\mathbf{r})$ and $\Psi_n^{\min}$ is the minimizing wave function.

Eq. (3) is a purely formal definition and cannot be used in practice to find the minimizing density of Eq. (2). To make the Hohenberg-Kohn density variational principle practical, Kohn and Sham⁴ suggested to construct a system of non-interacting electrons in such a way that the ground state density of the auxiliary non-interacting system is equal to the ground state density of the interacting system. To this end, $F[n]$ is partitioned as

$$F[n] = T_s[n] + U[n] + E_{xc}[n] \quad (4)$$

where $T_s[n]$ is the kinetic energy functional of *non-interacting* electrons with density n and

$$U[n] = \frac{1}{2} \int d^3r \int d^3r' \frac{n(\mathbf{r})n(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} \quad (5)$$

is the classical electrostatic or Hartree energy of the charge distribution $n(\mathbf{r})$. $E_{xc}[n]$ is the so-called exchange-correlation (xc) energy functional which will be the main quantity of interest later in this work. Using this decomposition the total energy functional (Eq. (2)) may be written as

$$E_v[n] = T_s[n] + \int d^3r v(\mathbf{r})n(\mathbf{r}) + U[n] + E_{xc}[n] \quad (6)$$

The Hohenberg-Kohn theorem can also be applied to non-interacting electrons moving in some potential $v_s(\mathbf{r})$. The corresponding ground-state energy functional then reads

$$E_{v_s}^{non-int}[n] = T_s[n] + \int d^3r v_s(\mathbf{r})n(\mathbf{r}) . \quad (7)$$

Due to the Hohenberg-Kohn variational principle one can determine the ground state density of this non-interacting system by minimizing the functional (Eq. (7)), i.e.,

$$\frac{\delta E_{v_s}^{non-int}[n]}{\delta n(\mathbf{r})} = 0 . \quad (8)$$

Alternatively, however, we can also determine the ground state density of a system of N non-interacting electrons by solving the single-particle Schrödinger equation

$$\left(-\frac{\nabla^2}{2} + v_s(\mathbf{r})\right) \varphi_{i\sigma}(\mathbf{r}) = \varepsilon_{i\sigma} \varphi_{i\sigma}(\mathbf{r}) \quad (9)$$

and compute the density from the $N = N_\uparrow + N_\downarrow$ orbitals with lowest energy eigenvalues $\varepsilon_{i\sigma}$ by

$$n(\mathbf{r}) = \sum_{\sigma=\uparrow,\downarrow} \sum_{i=1}^{N_\sigma} |\varphi_{i\sigma}(\mathbf{r})|^2 . \quad (10)$$

Here, N_σ is the total number of electrons with spin σ . Applying the variational principle to the energy functional (Eq. (6)) one realizes that the resulting variational equation has the same structure as the one for non-interacting electrons if one defines an effective potential

$$v_s(\mathbf{r}) = v(\mathbf{r}) + v_H(\mathbf{r}) + v_{xc}(\mathbf{r}) \quad (11)$$

with the Hartree potential

$$v_H(\mathbf{r}) = \frac{\delta U[n]}{\delta n(\mathbf{r})} = \int d^3r' \frac{n(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} \quad (12)$$

and the exchange-correlation potential

$$v_{xc}(\mathbf{r}) = \frac{\delta E_{xc}[n]}{\delta n(\mathbf{r})} . \quad (13)$$

As a consequence, the ground state density of the *interacting* system may also be calculated by solving the single-particle equation (9) with the effective potential given by Eq. (11). Since v_H and v_{xc} depend on the density (10), the whole procedure has to be solved self-consistently. This constitutes the Kohn-Sham scheme of density functional theory⁴.

It is worth noting that this scheme in principle yields the exact ground state density and, via Eq. (6), also the exact ground state energy, provided that the exact exchange-correlation energy functional $E_{xc}[n]$ is used. This exchange-correlation functional is something like the holy grail of density functional theory: although its formal definition has already been given in the original work of Hohenberg and Kohn¹ and further been clarified in the constrained-search formulation of Levy^{2,3}, these definitions cannot be used in practice and one needs to resort to approximations.

Fortunately, it is possible to construct simple approximations which perform, sometimes surprisingly, well. The first and probably most widely known approximation is the

so-called Local Density Approximation (LDA), which is based on the model of the uniform electron gas. Formally, the LDA is defined as

$$E_{xc}^{LDA}[n] = \int d^3r n(\mathbf{r}) \epsilon_{xc}^{unif}(n(\mathbf{r})) \quad (14)$$

where $\epsilon_{xc}^{unif}(n)$ is the exchange-correlation energy per electron of the uniform electron gas with (constant) density n . This function is known from Quantum Monte Carlo calculations⁵ and several simple parametrizations have been suggested⁶⁻⁸. Despite its simplicity, LDA has proven to be surprisingly accurate and hard to improve.

A major breakthrough in the development of more accurate exchange-correlation functionals came with the advent of so-called “generalized gradient approximations” (GGA’s) which have the general form

$$E_{xc}^{GGA}[n] = \int d^3r f(n, \nabla n) . \quad (15)$$

While the input $\epsilon_{xc}^{unif}(n)$ in LDA is unique, the function $f(n, \nabla n)$ in Eq. (15) is not and many forms have been suggested⁹⁻¹⁷. Although LDA had already widely been used in solid state physics, only with the increased accuracy of the GGA’s the Kohn-Sham scheme became a standard tool for electronic structure calculations in theoretical chemistry as well.

3 Orbital Functionals and the Optimized Effective Potential Method

The explicit dependence on the density is known both for LDA and GGA functionals. More recently, it has been recognized that functionals depending explicitly on the Kohn-Sham orbitals may also be viewed as density functionals since the orbitals are, through Eq. (9), functionals of the density-dependent effective potential $v_s(\mathbf{r})$. Since the explicit dependence of the orbitals on the density remains unknown, these orbital functionals are sometimes called *implicit* density functionals.

The idea of implicit density functionals might appear surprising at first sight. However, this idea is already used in the original Kohn-Sham formalism because the kinetic energy $T_s[n]$ of the non-interacting system may be expressed without approximation in terms of the Kohn-Sham orbitals, i.e.,

$$T_s[n] = \sum_{\sigma=\uparrow,\downarrow} \sum_{i=1}^{N_\sigma} \int d^3r \varphi_{i\sigma}^*(\mathbf{r}) \left(-\frac{\nabla^2}{2} \right) \varphi_{i\sigma}(\mathbf{r}) . \quad (16)$$

Turning to the exchange-correlation energy one also realizes that orbital-dependent expressions provide a natural framework for approximations to E_{xc} : if one performs a power series expansion of the functional $E_{xc}[n]$ in terms of the interaction strength e^2 (where e is the elementary charge) one obtains as leading term in this series the exact exchange energy. This term is a known explicit functional of the orbitals and reads as

$$E_x^{EXX}[n] = E_x^{EXX}[\{\varphi_i\}] = -\frac{1}{2} \sum_{\sigma=\uparrow,\downarrow} \sum_{j,k=1}^{N_\sigma} \int d^3r \int d^3r' \frac{\varphi_{j\sigma}^*(\mathbf{r}) \varphi_{k\sigma}^*(\mathbf{r}') \varphi_{j\sigma}(\mathbf{r}') \varphi_{k\sigma}(\mathbf{r})}{|\mathbf{r} - \mathbf{r}'|} \quad (17)$$

which is nothing but the Fock exchange energy but evaluated with Kohn-Sham orbitals. The derivation of Eq. (17) as well as possible routes to go beyond exchange will be discussed in more detail in Section 4.

3.1 Optimized Effective Potential Equation: Two Derivations

Once we accept to use an expression for E_{xc} which explicitly depends on the Kohn-Sham orbitals, the main question is how to compute the corresponding exchange-correlation potential of Eq. (13).

The way to accomplish this task is indicated by the Hohenberg-Kohn theorem. Application of this theorem to the Kohn-Sham system shows that there is a one to one correspondence between $n(\mathbf{r})$ and $v_s(\mathbf{r})$. The latter, in turn, determines the Kohn-Sham orbitals from which the density can be obtained. Thus, we can formally write any of these quantities as functional of one of the other quantities and an explicit dependence on one of them introduces an implicit dependence on the others. We use this observation in order to rewrite Eq. (13) using the chain rule of functional differentiation as follows

$$v_{xc}(\mathbf{r}) = \sum_{\sigma=\uparrow,\downarrow} \sum_{j=1}^{N_\sigma} \int d^3 r' \left(\frac{\delta E_{xc}}{\delta \varphi_{j\sigma}(\mathbf{r}')} \frac{\delta \varphi_{j\sigma}(\mathbf{r}')}{\delta n(\mathbf{r})} + c.c. \right). \quad (18)$$

For simplicity, we assumed here that the exchange-correlation functional is given as a functional of the occupied Kohn-Sham orbitals only. The functional derivative of E_{xc} with respect to the orbitals can easily be calculated from its explicit functional form. In order to calculate the functional derivative of the orbitals with respect to the density, we now view the orbitals as functional of $v_s(\mathbf{r})$ and use, for a second time, the chain rule of functional differentiation to obtain

$$v_{xc}(\mathbf{r}) = \sum_{\sigma=\uparrow,\downarrow} \sum_{j=1}^{N_\sigma} \int d^3 r' \int d^3 r'' \left(\frac{\delta E_{xc}}{\delta \varphi_{j\sigma}(\mathbf{r}')} \frac{\delta \varphi_{j\sigma}(\mathbf{r}')}{\delta v_s(\mathbf{r}'')} \frac{\delta v_s(\mathbf{r}'')}{\delta n(\mathbf{r})} + c.c. \right). \quad (19)$$

The third functional derivative on the right hand side of this equation may now be identified as the inverse of the static density response function of the Kohn-Sham system defined by

$$\chi_s(\mathbf{r}, \mathbf{r}') = \frac{\delta n(\mathbf{r})}{\delta v_s(\mathbf{r}')} . \quad (20)$$

Operating with χ_s on Eq. (19) from the right one obtains

$$\int d^3 r' v_{xc}(\mathbf{r}') \chi_s(\mathbf{r}', \mathbf{r}) = \sum_{\sigma=\uparrow,\downarrow} \sum_{j=1}^{N_\sigma} \int d^3 r' \left(\frac{\delta E_{xc}}{\delta \varphi_{j\sigma}(\mathbf{r}')} \frac{\delta \varphi_{j\sigma}(\mathbf{r}')}{\delta v_s(\mathbf{r})} + c.c. \right). \quad (21)$$

Now all the terms in this equation can be expressed in terms of the Kohn-Sham orbitals and eigenvalues. The functional derivative of the orbitals with respect to the potential can be obtained exactly from first order perturbation theory and reads

$$\frac{\delta \varphi_{j\sigma}(\mathbf{r}')}{\delta v_s(\mathbf{r})} = G_{s,j\sigma}(\mathbf{r}', \mathbf{r}) \varphi_{j\sigma}(\mathbf{r}) \quad (22)$$

where we have defined

$$G_{s,j\sigma}(\mathbf{r}, \mathbf{r}') = \sum_{\substack{k \\ k \neq j}} \frac{\varphi_{k\sigma}(\mathbf{r}) \varphi_{k\sigma}^*(\mathbf{r}')}{\varepsilon_{j\sigma} - \varepsilon_{k\sigma}} . \quad (23)$$

For simplicity, we have assumed here that the single-particle levels are non-degenerate. The static linear density response function of the Kohn-Sham system may be written as

$$\chi_s(\mathbf{r}, \mathbf{r}') = \sum_{\sigma=\uparrow, \downarrow} \sum_{j=1}^{N_\sigma} (\varphi_{j\sigma}^*(\mathbf{r}) G_{s,j\sigma}(\mathbf{r}, \mathbf{r}') \varphi_{j\sigma}(\mathbf{r}') + c.c.) . \quad (24)$$

Substituting (22) and (24) into Eq. (21) yields

$$\sum_{\sigma=\uparrow, \downarrow} \sum_{j=1}^{N_\sigma} \int d^3 r' (\varphi_{j\sigma}^*(\mathbf{r}') (v_{xc}(\mathbf{r}') - u_{xc,j\sigma}(\mathbf{r}')) G_{s,j\sigma}(\mathbf{r}', \mathbf{r}) \varphi_{j\sigma}(\mathbf{r}) + c.c.) = 0 \quad (25)$$

where we have defined

$$u_{xc,j\sigma}(\mathbf{r}') := \frac{1}{\varphi_{j\sigma}^*(\mathbf{r}') \delta \varphi_{j\sigma}(\mathbf{r}')} \frac{\delta E_{xc}}{\delta \varphi_{j\sigma}(\mathbf{r}')} . \quad (26)$$

Introducing the so-called orbital shifts^{18–20}

$$\psi_{j\sigma}^*(\mathbf{r}) := \int d^3 r' \varphi_{j\sigma}^*(\mathbf{r}') (v_{xc}(\mathbf{r}') - u_{xc,j\sigma}(\mathbf{r}')) G_{s,j\sigma}(\mathbf{r}', \mathbf{r}) \quad (27)$$

allows to rewrite Eq. (25) in a very compact way

$$\sum_{\sigma=\uparrow, \downarrow} \sum_{j=1}^{N_\sigma} (\psi_{j\sigma}^*(\mathbf{r}) \varphi_{j\sigma}(\mathbf{r}) + c.c.) = 0 \quad (28)$$

which is the standard form of the so-called Optimized Effective Potential (OEP) equation.

The name of this equation suggests that the resulting potential is optimal, but in which sense is it so? In DFT, the value of the ground-state total energy corresponds to the minimum value of the total energy functional. This minimum is obtained only if the exact ground-state particle density is inserted. Now the idea of the Kohn-Sham method is that the density can be obtained from single-particle orbitals solving a single-particle Schrödinger equation. Hence, in order to produce those single-particle orbitals we need to use a proper single-particle potential which is nothing but the optimized effective potential. Formally, this potential follows from minimizing the total energy functional (Eq. (6)). Using again the chain rule one obtains

$$\frac{\delta E_v}{\delta v_s(\mathbf{r})} = \int d^3 r' \frac{\delta E_v}{\delta n(\mathbf{r}')} \frac{\delta n(\mathbf{r}')}{\delta v_s(\mathbf{r})} = 0 \quad (29)$$

where the last equality follows from the Hohenberg-Kohn variational principle, Eq. (8).

Interpreting $E_v[n]$ now as a functional of the orbitals, Eq. (29) may be written as

$$0 = \frac{\delta E_v}{\delta v_s(\mathbf{r})} = \sum_{\sigma=\uparrow, \downarrow} \sum_{j=1}^{N_\sigma} \int d^3 r' \left(\frac{\delta E_v}{\delta \varphi_{j\sigma}^*(\mathbf{r}')} \frac{\delta \varphi_{j\sigma}^*(\mathbf{r}')}{\delta v_s(\mathbf{r})} + c.c. \right) . \quad (30)$$

The functional derivative of E_v with respect the Kohn-Sham orbitals is easily expressed as

$$\frac{\delta E_v}{\delta \varphi_{j\sigma}^*(\mathbf{r}')} = \left(-\frac{\nabla^2}{2} + v(\mathbf{r}) + v_H(\mathbf{r}) \right) \varphi_j(\mathbf{r}) + \frac{\delta E_{xc}}{\delta \varphi_{j\sigma}^*(\mathbf{r}')} \quad (31)$$

which, by using the Kohn-Sham equation, may be expressed as

$$\frac{\delta E_v}{\delta \varphi_{j\sigma}^*(\mathbf{r})} = (\varepsilon_{j\sigma} - v_{xc}(\mathbf{r})) \varphi_j(\mathbf{r}) + \frac{\delta E_{xc}}{\delta \varphi_{j\sigma}^*(\mathbf{r}')} . \quad (32)$$

Inserting this expression into Eq. (30) and using the orthonormality of the Kohn-Sham orbitals, one finally arrives at the integral equation (25) for $v_{xc}(\mathbf{r})$. In other words, the optimized effective potential is that *local* potential which yields single-particle orbitals which minimize the total energy^{21,22}. This potential is nothing but the Kohn-Sham potential^{23,24}.

3.2 Solution of the OEP Equation: KLI and CEDA Approximation

The OEP equation (25) is an integral equation to be solved for the exchange-correlation potential $v_{xc}(\mathbf{r})$. Historically, this solution first has been achieved for systems with very high symmetry²². In order to reduce the computational effort, simplifying yet accurate approximations to the full OEP equations have been suggested which will be discussed in this section.

We see that an important ingredient of the OEP equation is the Green function of Eq. (23) which involves a summation over occupied and unoccupied Kohn-Sham orbitals. Sharp and Horton²¹ and later Krieger, Li, and Iafrate (KLI)^{25,26} proposed to approximate the Green function (Eq. (23)) by replacing the energy denominators by a constant value, independent of the particle indices j and k . Using the completeness of the Kohn-Sham orbitals, this approximation leads to

$$\begin{aligned} G_{s,j\sigma}^{KLI}(\mathbf{r}, \mathbf{r}') &= \frac{1}{\Delta} \left(\sum_{k=1}^{\infty} \varphi_{k\sigma}^*(\mathbf{r}) \varphi_{j\sigma}(\mathbf{r}') - \varphi_{j\sigma}^*(\mathbf{r}) \varphi_{j\sigma}(\mathbf{r}') \right) \\ &= \frac{1}{\Delta} \left(\delta(\mathbf{r} - \mathbf{r}') - \varphi_{j\sigma}^*(\mathbf{r}) \varphi_{j\sigma}(\mathbf{r}') \right) . \end{aligned} \quad (33)$$

Substitution into Eq. (27) gives

$$\sum_{\sigma=\uparrow,\downarrow} \sum_{j=1}^{N_\sigma} n_{j\sigma}(\mathbf{r}) \left(v_{xc}^{KLI}(\mathbf{r}) - u_{xc,j\sigma}(\mathbf{r}) - (\bar{v}_{xc,j\sigma}^{KLI} - \bar{u}_{xc,j\sigma}) \right) + c.c. = 0 \quad (34)$$

where we have used the definitions

$$n_{j\sigma}(\mathbf{r}) := \varphi_{j\sigma}^*(\mathbf{r}) \varphi_{j\sigma}(\mathbf{r}), \quad (35)$$

and the constants

$$\bar{v}_{xc,j\sigma}^{KLI} := \int d^3r \varphi_{j\sigma}^*(\mathbf{r}) v_{xc}^{KLI}(\mathbf{r}) \varphi_{j\sigma}(\mathbf{r}) . \quad (36)$$

The constants $\bar{u}_{xc,j\sigma}$ are defined in an analogous way as orbital averages of $u_{xc,j\sigma}(\mathbf{r})$ with respect to the orbital $\varphi_{j\sigma}$. Solving Eq. (34) for $v_{xc}^{KLI}(\mathbf{r})$ yields

$$v_{xc}^{KLI}(\mathbf{r}) = \frac{1}{2n(\mathbf{r})} \sum_{\sigma=\uparrow,\downarrow} \sum_{j=1}^{N_\sigma} n_{j\sigma}(\mathbf{r}) \left(u_{xc,j\sigma}(\mathbf{r}) + \bar{v}_{xc,j\sigma}^{KLI} - \bar{u}_{xc,j\sigma} + c.c. \right) . \quad (37)$$

This so-called KLI equation has transformed the OEP integral equation into an algebraic equation which is easier to solve in practical applications. Although this transformation is approximate, in many cases the deviations from full OEP results are small.

As we have seen, in the KLI approximation the Kohn-Sham orbital energy differences $\varepsilon_{j\sigma} - \varepsilon_{k\sigma}$ are approximated by one and the same constant, irrespective of the sign of different terms. In a similar spirit, a different approximation known as Common Energy Denominator Approximation (CEDA)²⁷ or Localized Hartree-Fock (LHF)²⁸ approximation has been proposed which only replaces the energy differences for occupied-unoccupied orbital pairs by a constant while it retains the energy differences for the occupied-occupied pairs. For the Green function this gives

$$G_{s,j\sigma}^{CEDA}(\mathbf{r}, \mathbf{r}') = \sum_{\substack{k=1 \\ k \neq j}}^{N_\sigma} \frac{\varphi_{k\sigma}^*(\mathbf{r})\varphi_{k\sigma}(\mathbf{r}')}{\varepsilon_j - \varepsilon_k} + \frac{1}{\Delta} \sum_{k > N_\sigma} \varphi_{k\sigma}^*(\mathbf{r})\varphi_{k\sigma}(\mathbf{r}') . \quad (38)$$

Adding and subtracting the contribution of the occupied orbitals with the common energy denominator and using the completeness of the Kohn-Sham orbitals yields

$$G_{s,j\sigma}^{CEDA}(\mathbf{r}, \mathbf{r}') = \frac{\delta(\mathbf{r} - \mathbf{r}')}{\Delta} + \sum_{\substack{k=1 \\ k \neq j}}^{N_\sigma} \frac{\varphi_{k\sigma}^*(\mathbf{r})\varphi_{k\sigma}(\mathbf{r}')}{\varepsilon_j - \varepsilon_k} - \frac{1}{\Delta} \sum_{k=1}^{N_\sigma} \varphi_{k\sigma}^*(\mathbf{r})\varphi_{k\sigma}(\mathbf{r}') , \quad (39)$$

which, when inserted into Eq. (25) leads to the following equation for the exchange-correlation potential in the CEDA approximation

$$\begin{aligned} v_{xc}^{CEDA}(\mathbf{r}) &= \frac{1}{2n(\mathbf{r})} \sum_{\sigma=\uparrow,\downarrow} \sum_{j=1}^{N_\sigma} \left(n_{j\sigma}(\mathbf{r}) u_{xc,j\sigma}(\mathbf{r}) \right. \\ &\quad \left. + \sum_{i=1}^{N_\sigma} \varphi_{i\sigma}^*(\mathbf{r}) \varphi_{j\sigma}(\mathbf{r}) \left(\overline{v}_{xc,ji\sigma}^{CEDA} - \overline{u}_{xc,ji\sigma} \right) + c.c. \right) \end{aligned} \quad (40)$$

with

$$\overline{v}_{xc,ji\sigma}^{CEDA} := \int d^3r \varphi_{j\sigma}^*(\mathbf{r}) v_{xc}^{CEDA}(\mathbf{r}) \varphi_{i\sigma}(\mathbf{r}) \quad (41)$$

and similarly for $\overline{u}_{xc,ji\sigma}$. One immediately sees that the CEDA reduces to the KLI approximation if all off-diagonal terms $i \neq j$ in the second sum are neglected.

Both the KLI as well as the CEDA approximation (Eqs. (37) and (40), respectively) can easily be implemented within a self-consistent scheme with essentially the same effort. From a theoretical point of view, CEDA has the advantage of being invariant under unitary transformations of the occupied orbitals while KLI is not. From a practical point of view, CEDA and KLI results are often very similar.

3.3 Exact Transformation of the OEP Equation

In this section we discuss a way to transform the OEP equations exactly which both emphasizes the role of the orbital shifts (Eq. (27)) and allows to motivate the KLI approximation from a different perspective. We start by noting that the non-interacting Green function satisfies the following differential equations

$$\left(\hat{h}_s(\mathbf{r}) - \varepsilon_{j\sigma} \right) G_{s,j\sigma}(\mathbf{r}', \mathbf{r}) = - \left(\delta(\mathbf{r} - \mathbf{r}') - \varphi_{j\sigma}^*(\mathbf{r}) \varphi_{j\sigma}(\mathbf{r}') \right) \quad (42)$$

where $\hat{h}_s(\mathbf{r})$ is the Kohn-Sham Hamiltonian. Acting with the operator $(\hat{h}_s(\mathbf{r}) - \varepsilon_{j\sigma})$ on Eq. (27) yields a differential equation which uniquely determines¹⁸ the orbital shifts

$$\left(\hat{h}_s(\mathbf{r}) - \varepsilon_{j\sigma}\right) \psi_{j\sigma}^*(\mathbf{r}) = -\left(v_{xc}(\mathbf{r}) - u_{xc,j\sigma}(\mathbf{r}) - (\bar{v}_{xc,j\sigma} - \bar{u}_{xc,j\sigma})\right) \varphi_{j\sigma}^*(\mathbf{r}). \quad (43)$$

Solving for $v_s(\mathbf{r})\psi_{j\sigma}^*(\mathbf{r})$ gives

$$\begin{aligned} v_s(\mathbf{r})\psi_{j\sigma}^*(\mathbf{r}) = & -\left(v_{xc}(\mathbf{r}) - u_{xc,j\sigma}(\mathbf{r}) - (\bar{v}_{xc,j\sigma} - \bar{u}_{xc,j\sigma})\right) \varphi_{j\sigma}^*(\mathbf{r}) \\ & + \left(\frac{\nabla^2}{2} + \varepsilon_{j\sigma}\right) \psi_{j\sigma}^*(\mathbf{r}). \end{aligned} \quad (44)$$

Multiplication of this equation with $\varphi_{j\sigma}$, subsequent summation over all occupied orbitals and using the OEP equation in the form (28) eventually leads to

$$\begin{aligned} v_{xc}(\mathbf{r}) = \frac{1}{2n(\mathbf{r})} \sum_{\sigma=\uparrow,\downarrow} \sum_{j=1}^{N_\sigma} \left\{ n_{j\sigma}(\mathbf{r}) (u_{xc,j\sigma}(\mathbf{r}) + \bar{v}_{xc,j\sigma} - \bar{u}_{xc,j\sigma}) \right. \\ \left. + \left(\frac{\nabla^2}{2} \psi_{j\sigma}^*(\mathbf{r}) + \varepsilon_{j\sigma} \psi_{j\sigma}^*(\mathbf{r})\right) \varphi_{j\sigma}(\mathbf{r}) + c.c. \right\}. \end{aligned} \quad (45)$$

The second term in the curled brackets can be rewritten by using the Kohn-Sham equations and the OEP equation again which finally leads to an exact reformulation of the OEP equation as

$$\begin{aligned} v_{xc}(\mathbf{r}) = \frac{1}{2n(\mathbf{r})} \sum_{\sigma=\uparrow,\downarrow} \sum_{j=1}^{N_\sigma} \left\{ n_{j\sigma}(\mathbf{r}) (u_{xc,j\sigma}(\mathbf{r}) + \bar{v}_{xc,j\sigma} - \bar{u}_{xc,j\sigma}) \right. \\ \left. - \nabla \cdot (\psi_{j\sigma}^*(\mathbf{r}) \nabla \varphi_{j\sigma}(\mathbf{r})) + c.c. \right\}. \end{aligned} \quad (46)$$

Clearly, if the term involving the orbital shifts $\psi_{j\sigma}^*(\mathbf{r})$ is neglected in this expression, one again obtains the KLI approximation.

The orbital shifts $\psi_{j\sigma}(\mathbf{r})$ also play a central role in an iterative scheme to the solution of the full OEP equation recently suggested by Kümmel and Perdew^{19,20}. The idea of this scheme is to solve Eq. (43) for the orbital shifts directly in the following way: for a given approximate solution $v_{xc}(\mathbf{r})$ to the OEP equation, compute the right hand side of Eq. (43) and then solve this equation for the orbital shift $\psi_{j\sigma}(\mathbf{r})$ subject to the orthogonality constraint

$$\int d^3r \psi_{j\sigma}^*(\mathbf{r}) \varphi_{j\sigma}(\mathbf{r}) = 0 \quad (47)$$

which follows from the definition in Eq. (27) by the orthonormality of the Kohn-Sham orbitals. With the resulting orbital shifts compute the quantity

$$S(\mathbf{r}) = \sum_{\sigma=\uparrow,\downarrow} \sum_{j=1}^{N_\sigma} \psi_{j\sigma}^*(\mathbf{r}) \varphi_{j\sigma}(\mathbf{r}) + c.c. \quad (48)$$

and then compute a new potential by

$$v_{xc}^{\text{new}}(\mathbf{r}) = v_{xc}^{\text{old}}(\mathbf{r}) + cS(\mathbf{r}) \quad (49)$$

with some positive constant c . With this new v_{xc} , recompute the right hand side of Eq. (43) and then solve again for a new orbital shift. This is iterated a few times for fixed $\varphi_{j\sigma}$, $\hat{h}_s$ and $\varepsilon_{j\sigma}$ before eventually these quantities are also recomputed during the regular Kohn-Sham self-consistency cycle. This scheme has been applied successfully^{19,20} to compute the OEP potential (in exact exchange approximation) not only for highly symmetric systems such as atoms but also to small sodium clusters where a direct solution of the OEP integral equation (28) is a much more difficult task.

3.4 Derivative Discontinuity of the Exchange-Correlation Energy

Orbital dependent functionals have an important advantage over standard explicit density functionals like LDA or GGA: they may reproduce the derivative discontinuity of the exchange-correlation functional as a function of particle number which occurs at integer particle number N ²⁹⁻³¹. This is a very important property of the exact functional which was overlooked for a long time.

In order to discuss this property one has to generalize the definition (3) of the Hohenberg-Kohn functional to non-integer particle numbers, i.e., to densities with

$$\int d^3r n(\mathbf{r}) = N + \omega \quad (50)$$

where N is an integer and $0 \leq \omega \leq 1$. This generalization of the Hohenberg-Kohn functional can be given as

$$F_{fac}[n] = \min_{\hat{D} \rightarrow n} \text{Tr} \left\{ \hat{D}(\hat{T} + \hat{V}_{ee}) \right\} \quad (51)$$

where the search runs over all statistical mixtures

$$\hat{D} = (1 - \omega)|\Psi_N\rangle\langle\Psi_N| + \omega|\Psi_{N+1}\rangle\langle\Psi_{N+1}| \quad (52)$$

of an N -particle state $|\Psi_N\rangle$ and an $(N + 1)$ -particle state $|\Psi_{N+1}\rangle$ which yield the given density $n(\mathbf{r})$. The density and energy of the ensemble described by $\hat{D}$ are then given as

$$n(\mathbf{r}) = \text{Tr} \left\{ \hat{D} \hat{n}(\mathbf{r}) \right\} = (1 - \omega)n_N(\mathbf{r}) + \omega n_{N+1}(\mathbf{r}) \quad (53)$$

and

$$E_v[n] = F_{fac}[n] + \int d^3r v(\mathbf{r})n(\mathbf{r}) = (1 - \omega)E_N + \omega E_{N+1} \quad (54)$$

where $n_N(\mathbf{r})$ and E_N are the density and energy corresponding to the state $|\Psi_N\rangle$ and similarly for n_{N+1} and E_{N+1} . Since the functional (Eq. (54)) is defined on the domain of densities with non-integer particle numbers, its minimum has to be found under the subsidiary condition of integer particle number which is incorporated by introducing the chemical potential μ as a the Lagrange multiplier. At the minimizing density, the variational equation then is

$$\mu(N) = \left. \frac{\delta E_v[n]}{\delta n(\mathbf{r})} \right|_{n_N} = \frac{\partial E_N}{\partial N} \quad (55)$$

Eqs. (54) and (55) show that the chemical potential $\mu(N)$ jumps discontinuously as N passes through an integer if E_N and E_{N+1} are separated by a finite energy gap. We can define this discontinuity as

$$\Delta(N) = \lim_{\omega \rightarrow 0} \left(\left. \frac{\delta E_v[n]}{\delta n(\mathbf{r})} \right|_{n_{N+\omega}} - \left. \frac{\delta E_v[n]}{\delta n(\mathbf{r})} \right|_{n_{N-\omega}} \right). \quad (56)$$

This expression can be written in terms of the ionization potential $I(N)$ and electronic affinity $A(N)$ of the N -particle system as

$$\Delta(N) = I(N) - A(N) = E(N+1) - 2E(N) + E(N-1) \quad (57)$$

where $E(N)$ is the ground state energy of the N -particle system. For infinite systems $\Delta(N)$ represents the so-called fundamental energy gap while for finite systems it is twice the chemical hardness. Insertion of Eq. (6) into Eq. (56) yields

$$\Delta(N) = \Delta_{KS}(N) + \Delta_{xc}(N) \quad (58)$$

where

$$\Delta_{KS} = \lim_{\omega \rightarrow 0} \left(\left. \frac{\delta T_s[n]}{\delta n(\mathbf{r})} \right|_{n_{N+\omega}} - \left. \frac{\delta T_s[n]}{\delta n(\mathbf{r})} \right|_{n_{N-\omega}} \right) = \varepsilon_{N+1}(N) - \varepsilon_N(N) \quad (59)$$

is the Kohn-Sham gap. $\varepsilon_j(N)$ is the j -th Kohn-Sham energy eigenvalue of the N -particle system and

$$\Delta_{xc}(N) = \lim_{\omega \rightarrow 0} \left(\left. \frac{\delta E_{xc}[n]}{\delta n(\mathbf{r})} \right|_{n_{N+\omega}} - \left. \frac{\delta E_{xc}[n]}{\delta n(\mathbf{r})} \right|_{n_{N-\omega}} \right) \quad (60)$$

is the derivative discontinuity of the exchange-correlation functional.

In Section 6.3 we will discuss some numerical results for the discontinuity and its contribution to the energy gap of semiconductors and insulators.

3.5 Orbital Functionals in Non-Collinear Spin Density Functional Theory

Until now we have discussed the OEP method in the framework of pure density functional theory, where only the particle density is used as fundamental variable. Spin-density functional theory (SDFT)³² is an extension of the original DFT suited both for the description of systems with unpaired electrons and for systems in an external magnetic field. However, in SDFT only the coupling of the magnetic field to the spin degrees of freedom is taken into account. SDFT is mostly (but not always) applied under the assumption of collinear magnetism, i.e., both the external and Kohn-Sham magnetic field are assumed to have a constant direction in space and the magnetization vector is only allowed to be parallel or antiparallel to this given direction. In this situation, SDFT leads to two decoupled Kohn-Sham equations for spin-up and spin-down electrons. The OEP method is also well established for this situation. Here, however, we will present the OEP equations of SDFT without this restriction, i.e., for the non-collinear situation³³.

Instead of using Eq. (6), one usually starts by writing the ground-state energy $E_{v,\mathbf{B}}[n, \mathbf{m}]$ of a system of interacting electrons with ground state $|\Psi_0\rangle$ moving in the electrostatic potential $v(\mathbf{r})$ and the external magnetic field $\mathbf{B}(\mathbf{r})$ as

$$E_{v,\mathbf{B}}[n, \mathbf{m}] = T_s[n, \mathbf{m}] + \int d^3r v(\mathbf{r})n(\mathbf{r}) - \int d^3r \mathbf{B}(\mathbf{r})\mathbf{m}(\mathbf{r}) + U[n] + E_{xc}[n, \mathbf{m}]. \quad (61)$$

Here, $E_{v,\mathbf{B}}$, T_s , and E_{xc} are expressed as functionals of the particle density

$$n(\mathbf{r}) = \langle \Psi_0 | \hat{\Psi}^\dagger(\mathbf{r})\hat{\Psi}(\mathbf{r}) | \Psi_0 \rangle \quad (62)$$

and the magnetization density

$$\mathbf{m}(\mathbf{r}) = -\mu_B \langle \Psi_0 | \hat{\Psi}^\dagger(\mathbf{r})\boldsymbol{\sigma}\hat{\Psi}(\mathbf{r}) | \Psi_0 \rangle \quad (63)$$

where $\hat{\Psi}(\mathbf{r})$ is the field operator for Pauli spinors, μ_B is the Bohr magneton and $\boldsymbol{\sigma}$ is the vector of Pauli matrices. Under the usual assumption of non-interacting v -representability, i.e., assuming that the densities $n(\mathbf{r})$ and $\mathbf{m}(\mathbf{r})$ can also be obtained as densities of a fictitious system of non-interacting electrons, one can then derive the Kohn-Sham equation

$$\left(-\frac{\nabla^2}{2} + v_s(\mathbf{r}) + \mu_B \boldsymbol{\sigma} \mathbf{B}_s(\mathbf{r}) \right) \Phi_j(\mathbf{r}) = \varepsilon_j \Phi_j(\mathbf{r}) \quad (64)$$

where the $\Phi_i(\mathbf{r})$ are two-component, single-particle Pauli spinors. The effective potentials are given by

$$v_s(\mathbf{r}) = v(\mathbf{r}) + \int d^3r' \frac{n(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} + v_{xc}(\mathbf{r}) \quad (65)$$

with the exchange-correlation potential

$$v_{xc}(\mathbf{r}) = \frac{\delta E_{xc}[n, \mathbf{m}]}{\delta n(\mathbf{r})} \quad (66)$$

and

$$\mathbf{B}_s(\mathbf{r}) = \mathbf{B}(\mathbf{r}) + \mathbf{B}_{xc}(\mathbf{r}) \quad (67)$$

with the exchange-correlation magnetic field

$$\mathbf{B}_{xc}(\mathbf{r}) = -\frac{\delta E_{xc}[n, \mathbf{m}]}{\delta \mathbf{m}(\mathbf{r})} \quad (68)$$

and the (universal) exchange-correlation energy functional $E_{xc}[n, \mathbf{m}]$ which has to be approximated in practice.

If E_{xc} is approximated as an *explicit* functional of n and $\mathbf{m}$, the functional derivatives (Eq. (66)) and (Eq. (68)) are easily evaluated. However, here we are more interested in the situation when E_{xc} is given as an explicit functional of the (spinor) orbitals (i.e., only as an *implicit* functional of the densities). Again, the exchange-correlation potentials need to be calculated with the OEP method described below. For simplicity, we assume that the approximate E_{xc} is given as a functional of the occupied spinors only, $E_{xc} = E_{xc}[\{\Phi_j\}]$. At this point, it should be noted that the spinor orbitals Φ_j (and therefore also $E_{xc}[\{\Phi_j\}]$) are, through Eq. (64), unique functionals of the potentials v_s and $\mathbf{B}_s$.

In order to obtain the OEP equations for the exchange-correlation potentials, we use the chain rule of functional derivatives to calculate

$$\begin{aligned}\frac{\delta E_{xc}}{\delta v_s(\mathbf{r})} &= \int d^3 r' \left(v_{xc}(\mathbf{r}') \frac{\delta n(\mathbf{r}')}{\delta v_s(\mathbf{r})} - \mathbf{B}_{xc}(\mathbf{r}') \frac{\delta \mathbf{m}(\mathbf{r}')}{\delta v_s(\mathbf{r})} \right) \\ &= \sum_{k=1}^N \int d^3 r' \left(\frac{\delta E_{xc}}{\delta \Phi_k(\mathbf{r}')} \frac{\delta \Phi_k(\mathbf{r}')}{\delta v_s(\mathbf{r})} + h.c. \right)\end{aligned}\quad (69)$$

and

$$\begin{aligned}\frac{\delta E_{xc}}{\delta \mathbf{B}_s(\mathbf{r})} &= \int d^3 r' \left(v_{xc}(\mathbf{r}') \frac{\delta n(\mathbf{r}')}{\delta \mathbf{B}_s(\mathbf{r})} - \mathbf{B}_{xc}(\mathbf{r}') \frac{\delta \mathbf{m}(\mathbf{r}')}{\delta \mathbf{B}_s(\mathbf{r})} \right) \\ &= \sum_{k=1}^N \int d^3 r' \left(\frac{\delta E_{xc}}{\delta \Phi_k(\mathbf{r}')} \frac{\delta \Phi_k(\mathbf{r}')}{\delta \mathbf{B}_s(\mathbf{r})} + h.c. \right).\end{aligned}\quad (70)$$

Using Eq. (64), the functional derivatives of the orbitals (and therefore also of the densities) with respect to the potentials can be computed from first-order perturbation theory. Following steps similar to the ones given in Section 3.1, one can write the OEP equations in the simple form

$$\sum_{k=1}^N \left(\Phi_k^\dagger(\mathbf{r}) \Psi_k(\mathbf{r}) + h.c. \right) = 0 \quad (71)$$

$$-\mu_B \sum_{k=1}^N \left(\Phi_k^\dagger(\mathbf{r}) \boldsymbol{\sigma} \Psi_k(\mathbf{r}) + h.c. \right) = 0 \quad (72)$$

where we have defined the orbital shifts^{18–20}

$$\Psi_k(\mathbf{r}) = \sum_{\substack{j \\ j \neq k}} \frac{D_{kj} \Phi_j(\mathbf{r})}{\varepsilon_k - \varepsilon_j} \quad (73)$$

and

$$D_{kj} = \int d^3 r' \Phi_j^\dagger(\mathbf{r}') \left((v_{xc}(\mathbf{r}') + \mu_B \boldsymbol{\sigma} \mathbf{B}_{xc}(\mathbf{r}')) \Phi_k(\mathbf{r}') - \frac{\delta E_{xc}}{\delta \Phi_k^\dagger(\mathbf{r}')} \right). \quad (74)$$

The OEP equations (71) and (72) can be interpreted in a simple manner: the optimized local potentials v_s and $\mathbf{B}_s$, i.e., the potentials giving the lowest total energy for a given energy functional, are the ones which yield a vanishing change in the density $n(\mathbf{r})$ and magnetization density $\mathbf{m}(\mathbf{r})$ in first-order perturbation theory when applying a perturbation $\hat{H}'$ to the Kohn-Sham equations where the perturbation is chosen such that the matrix elements with respect to the Kohn-Sham orbitals are given by $\langle \Phi_j | \hat{H}' | \Phi_k \rangle = -D_{kj}$.

Similar to the discussion in the previous section about the transformation of the OEP equation (28) to the form (46) in the DFT framework, one can also exactly transform the OEP equations (71) and (72) in non-collinear spin-DFT. Again, one starts with the differential equation for the orbital shifts

$$(\hat{h}_s(\mathbf{r}) - \varepsilon_k) \Psi_k(\mathbf{r}) = - (v_{xc}(\mathbf{r}) + \mu_B \boldsymbol{\sigma} \mathbf{B}_s(\mathbf{r})) \Phi_k(\mathbf{r}) + \frac{\delta E_{xc}}{\delta \Phi_k^\dagger(\mathbf{r})} + D_{kk} \Phi_k(\mathbf{r}). \quad (75)$$

This equation, together with the OEP equations (71) and (72) eventually leads to the transformed OEP equations for non-collinear spin-DFT which may most conveniently be written in matrix notation as

$$\begin{pmatrix} n(\mathbf{r}) & -m_1(\mathbf{r}) & -m_2(\mathbf{r}) & -m_3(\mathbf{r}) \\ -m_1(\mathbf{r}) & \mu_B^2 n(\mathbf{r}) & 0 & 0 \\ -m_2(\mathbf{r}) & 0 & \mu_B^2 n(\mathbf{r}) & 0 \\ -m_3(\mathbf{r}) & 0 & 0 & \mu_B^2 n(\mathbf{r}) \end{pmatrix} \begin{pmatrix} v_{xc}(\mathbf{r}) \\ B_{xc,1}(\mathbf{r}) \\ B_{xc,2}(\mathbf{r}) \\ B_{xc,3}(\mathbf{r}) \end{pmatrix} = \begin{pmatrix} g_1(\mathbf{r}) \\ g_2(\mathbf{r}) \\ g_3(\mathbf{r}) \\ g_4(\mathbf{r}) \end{pmatrix}. \quad (76)$$

Here we have defined

$$g_1(\mathbf{r}) = \frac{1}{2} \sum_{k=1}^N \left(u_{xc,k}(\mathbf{r}) - \nabla \left((\nabla \Phi_k^\dagger(\mathbf{r})) \Psi_k(\mathbf{r}) \right) + n_k(\mathbf{r}) D_{kk} + h.c. \right) \quad (77)$$

with the orbital densities

$$n_k(\mathbf{r}) = \Phi_k^\dagger(\mathbf{r}) \Phi_k(\mathbf{r}) \quad (78)$$

and

$$u_{xc,k}(\mathbf{r}) = \frac{\delta E_{xc}}{\delta \Phi_k(\mathbf{r})} \Phi_k(\mathbf{r}). \quad (79)$$

The other terms on the right-hand side of Eq. (76) are given by

$$g_{\alpha+1}(\mathbf{r}) = \frac{\mu_B}{2} \sum_{k=1}^N \left(\tilde{u}_{xc,k,\alpha}(\mathbf{r}) - \nabla \left((\nabla \Phi_k^\dagger(\mathbf{r})) \sigma_\alpha \Psi_k(\mathbf{r}) \right) + m_{k,\alpha}(\mathbf{r}) D_{kk} + h.c. \right) \quad (80)$$

with $\alpha = 1, 2, 3$ and the Pauli matrices $\sigma_1 = \sigma_x$ etc. as well as the orbital magnetization

$$m_{k,\alpha}(\mathbf{r}) = -\mu_B \Phi_k^\dagger(\mathbf{r}) \sigma_\alpha \Phi_k(\mathbf{r}) \quad (81)$$

and

$$\tilde{u}_{xc,k,\alpha}(\mathbf{r}) = \frac{\delta E_{xc}}{\delta \Phi_k(\mathbf{r})} \sigma_\alpha \Phi_k(\mathbf{r}). \quad (82)$$

Of course, under the common assumption that the Kohn-Sham exchange-correlation magnetic field has only a non-vanishing z -component, $\mathbf{B}_{xc}(\mathbf{r}) = (0, 0, B_{xc,3}(\mathbf{r}))^T$, out of the four equations of Eq. (76) only two are non-trivial and these can easily be transformed to the OEP equations of collinear spin-DFT¹⁸.

4 Orbital Functionals for Exchange and Correlation

In this section we will discuss several ways towards the construction of orbital-dependent approximations to the exchange-correlation energy. We present some important concepts which offer some insight into the exchange-correlation energy and which serve as starting points for various approximate exchange-correlation functionals. We then give a brief overview over existing orbital functional approximations.

4.1 Adiabatic Connection

The adiabatic connection method^{34–38} derives its name from the idea that one can connect the Hamiltonian of an interacting system to that of a non-interacting system via introduction of a coupling constant:

$$\hat{H}_\lambda = \hat{T} + \hat{V}_\lambda + \lambda \hat{V}_{ee} . \quad (83)$$

In the context of density functional theory, the external potential $\hat{V}_\lambda$ for the coupling constant λ is chosen such that the ground-state density of $\hat{H}_\lambda$ is independent of λ . For $\lambda = 1$, $\hat{V}_{\lambda=1}$ is just the external potential of the interacting system under consideration. Since for $\lambda = 0$ the ground state density of the corresponding non-interacting system is required to be equal to the density of the interacting system, $\hat{V}_{\lambda=0}$ is the corresponding Kohn-Sham potential.

In order to make the dependence of the results of this section on the strength of the electron-electron interaction explicit, we use

$$\hat{V}_{ee} = \sum_{\substack{i,j=1 \\ i \neq j}}^N \frac{e^2}{|\mathbf{r}_i - \mathbf{r}_j|} \quad (84)$$

with e being the elementary charge.

The Hellmann-Feynman theorem for $\hat{H}_\lambda$ reads

$$\frac{dE_\lambda}{d\lambda} = \left\langle \Psi_n^{\min,\lambda} \left| \frac{\partial \hat{H}_\lambda}{\partial \lambda} \right| \Psi_n^{\min,\lambda} \right\rangle \quad (85)$$

where E_λ is the ground-state energy at coupling constant λ and $\Psi_n^{\min,\lambda}$ is the ground-state wave function of $\hat{H}_\lambda$, i.e., that wave function which yields the ground-state density $n(\mathbf{r})$ and minimizes the expectation value $\langle \hat{T} + \lambda \hat{V}_{ee} \rangle$. From Eq. (85), the following expression for the ground-state energy of the interacting system (at $\lambda = 1$) immediately follows:

$$E(\lambda = 1) = E_v[n] = T_s[n] + \int d^3r v(\mathbf{r})n(\mathbf{r}) + \int_0^1 d\lambda \langle \Psi_n^{\min,\lambda} | \hat{V}_{ee} | \Psi_n^{\min,\lambda} \rangle . \quad (86)$$

Comparison with Eq. (6) then yields an expression for the exchange-correlation energy,

$$E_{xc}[n] = \int_0^1 d\lambda \langle \Psi_n^{\min,\lambda} | \hat{V}_{ee} | \Psi_n^{\min,\lambda} \rangle - U[n] =: \int_0^1 d\lambda W_\lambda[n] , \quad (87)$$

where

$$W_\lambda[n] = \langle \Psi_n^{\min,\lambda} | \hat{V}_{ee} | \Psi_n^{\min,\lambda} \rangle - U[n] . \quad (88)$$

It is often convenient to partition $E_{xc}[n]$ into its exchange and correlation part,

$$E_{xc}[n] = E_x[n] + E_c[n] . \quad (89)$$

One may define the exchange energy as the value of the coupling-constant integrand (Eq. (88)) at $\lambda = 0$:

$$E_x[n] = W_{\lambda=0}[n] = \langle \Phi_n^{\min} | \hat{V}_{ee} | \Phi_n^{\min} \rangle - U[n] \quad (90)$$

where $\Phi_n^{\min}$ is the noninteracting ground-state wave function which yields n , i.e., the Kohn-Sham ground-state Slater determinant. Expressed in terms of single-particle orbitals, one obtains the usual Fock exchange energy expression of Eq. (17), i.e,

$$E_x[n] = E_x^{EXX}[n] = -\frac{e^2}{2} \sum_{\sigma=\uparrow,\downarrow} \sum_{j,k=1}^{N_\sigma} \int d^3r \int d^3r' \frac{\varphi_{j\sigma}(\mathbf{r})\varphi_{j\sigma}^*(\mathbf{r}')\varphi_{k\sigma}^*(\mathbf{r})\varphi_{k\sigma}(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}. \quad (91)$$

An alternative, but equivalent definition of the exchange energy makes use of uniform coordinate scaling³⁹. Define a density

$$n_\gamma(\mathbf{r}) = \gamma^3 n(\gamma\mathbf{r}), \quad (92)$$

which is scaled uniformly with a scaling parameter γ . Then, $E_x[n]$ is defined as the following limit

$$E_x[n] = \lim_{\gamma \rightarrow \infty} \frac{1}{\gamma} E_{xc}[n_\gamma] \quad (93)$$

which implies that E_x scales linearly under uniform coordinate scaling, i.e.

$$E_x[n_\gamma] = \gamma E_x[n]. \quad (94)$$

The relationship between the adiabatic connection and coordinate scaling extends much further than only for the definition of the exchange energy. Using coordinate scaling arguments, Görling and Levy⁴⁰ have suggested a perturbation expansion of the correlation energy in terms of the coupling constant λ . They were able to relate the λ -dependent external potential $v_\lambda(\mathbf{r})$ to functional derivatives at full coupling ($\lambda = 1$):

$$v_\lambda(\mathbf{r}) = v_{\lambda=0}(\mathbf{r}) - \lambda \left(v_H(\mathbf{r}) + v_x(\mathbf{r}) + \lambda \frac{\delta E_c[n_{1/\lambda}]}{\delta n(\mathbf{r})} \right) \quad (95)$$

where the Hartree potential is given by Eq. (12) and the exchange potentials is

$$v_x(\mathbf{r}) = \frac{\delta E_x[n]}{\delta n(\mathbf{r})}. \quad (96)$$

Görling and Levy used Eq. (95) to calculate the ground state energy of $\hat{H}_\lambda$ with perturbation theory in the expansion parameter λ . To first order, the result is (at full coupling, $\lambda = 1$):

$$E^{(0)}[n] + E^{(1)}[n] = T_s[n] + \int d^3r v(\mathbf{r})n(\mathbf{r}) + U[n] + E_x[n] \quad (97)$$

which is formally identical to the Hartree-Fock energy expression evaluated with Kohn-Sham orbitals. The second-order term is given by

$$\begin{aligned} E^{(2)}[n] &= E_c^{GL2}[n] \\ &= \frac{1}{2} \sum_{\sigma, \sigma'} \sum_{\varepsilon_{i\sigma}, \varepsilon_{j\sigma'} \leq \varepsilon_F < \varepsilon_{k\sigma}, \varepsilon_{l\sigma'}} \frac{(ik\sigma||jl\sigma')[(jl\sigma'||ik\sigma) - \delta_{\sigma,\sigma'}(jl\sigma||ki\sigma)]}{\varepsilon_{i\sigma} + \varepsilon_{j\sigma'} - \varepsilon_{k\sigma} - \varepsilon_{l\sigma'}} \\ &\quad + \sum_{\sigma} \sum_{\varepsilon_{i\sigma} \leq \varepsilon_F < \varepsilon_{l\sigma}} \frac{1}{\varepsilon_{i\sigma} - \varepsilon_{l\sigma}} \left| \langle i\sigma|v_x|l\sigma \rangle + \sum_{\varepsilon_{j\sigma} \leq \varepsilon_F} (ij\sigma||j\sigma) \right|^2 \end{aligned} \quad (98)$$

where we have defined the two-electron integral

$$(ik\sigma||jl\sigma') = \int d^3r \int d^3r' \varphi_{i\sigma}^*(\mathbf{r}) \varphi_{k\sigma}(\mathbf{r}) \frac{e^2}{|\mathbf{r} - \mathbf{r}'|} \varphi_{j\sigma'}^*(\mathbf{r}') \varphi_{l\sigma'}(\mathbf{r}') \quad (99)$$

and

$$\langle i\sigma|v_x|j\sigma' \rangle = \int d^3r \varphi_{i\sigma}^*(\mathbf{r}) v_x(\mathbf{r}) \varphi_{j\sigma'}(\mathbf{r}) . \quad (100)$$

The similarity to traditional quantum-chemical methods is again apparent. The first term in Eq. (98) is the familiar energy expression of second-order Møller-Plesset perturbation theory (see, e.g., Ref. 41) while the second term originates from the first-order correction of the Kohn-Sham Slater determinant towards the Hartree-Fock wave function. As the exchange energy could be defined through coordinate scaling (Eq. (93)), the second-order correlation energy may also be obtained as the $\gamma \rightarrow \infty$ limit of the correlation energy under uniform coordinate scaling,

$$E_c^{GL2}[n] = \lim_{\gamma \rightarrow \infty} E_c[n_\gamma] > -\infty . \quad (101)$$

The inequality sign only applies to densities which integrate to a finite number of electrons⁴².

Both E_x of Eq. (91) and E_c^{GL2} of Eq. (98) are given explicitly in terms of the Kohn-Sham orbitals (and orbital energies) and the corresponding exchange and correlation potentials have to be calculated using the OEP method. While the exchange energy $E_x[n]$ is only a functional of the occupied orbitals, E_c^{GL2} also is a functional of all the unoccupied orbitals and the orbital energies.

4.2 Fluctuation-Dissipation Theorem

A different representation of E_{xc} , which we want to introduce now, relates the exchange-correlation energy to the (dynamic) density-density response function of the interacting system of interest. Obviously, the description of response functions requires a formal extension of density functional theory to the time-dependent domain⁴³. The central ideas of time-dependent density functional theory (TDDFT) are completely analogous to the static case (for a review of TDDFT, see Ref. 44). One first proves a one-to-one mapping between time-dependent external potentials and time-dependent densities, then one constructs an auxiliary non-interacting system with the same time-dependent density as the interacting system. The equations to describe this auxiliary system have the structure of a Schrödinger equation for non-interacting electrons moving in a time-dependent, local single-particle potential $v_s(\mathbf{r}, t)$.

As the name implies, the linear density response function χ describes the change of the density $\delta n(\mathbf{r}, t)$ in response to some external time-dependent perturbation $\delta v(\mathbf{r}, t)$:

$$\delta n(\mathbf{r}, t) = \int_{t_0}^t dt' \int d^3r' \chi(\mathbf{r}, \mathbf{r}', t - t') \delta v(\mathbf{r}', t') . \quad (102)$$

We assume that the system is in its ground state before the perturbation is switched on at time t_0 . Using the fluctuation-dissipation theorem at zero temperature, one can express

the exchange-correlation energy in terms of the Fourier transform of the linear density response function:

$$E_{xc}[n] = \frac{1}{2} \int_0^1 d\lambda \int d^3r \int d^3r' \frac{e^2}{|\mathbf{r} - \mathbf{r}'|} \left(-\frac{1}{\pi} \int_0^\infty d\omega \operatorname{Im} \chi^\lambda(\mathbf{r}, \mathbf{r}', \omega) - n(\mathbf{r})\delta(\mathbf{r} - \mathbf{r}') \right). \quad (103)$$

Again, we have used the idea of the adiabatic connection, i.e., we consider the interacting response function χ^λ for a scaled electron-electron interaction $\lambda e^2/|\mathbf{r} - \mathbf{r}'|$ while the density is kept fixed for all values of λ .

The time-dependent Kohn-Sham scheme allows to calculate the interacting density response function as solution of the following integral equation which relates χ^λ to the response function χ_s of the non-interacting Kohn-Sham system^{45,46}:

$$\chi^\lambda(\mathbf{r}, \mathbf{r}', \omega) = \chi_s(\mathbf{r}, \mathbf{r}', \omega) + \int d^3x \int d^3y \chi_s(\mathbf{r}, \mathbf{x}, \omega) \left(\frac{\lambda e^2}{|\mathbf{x} - \mathbf{y}|} + f_{xc}^\lambda(\mathbf{x}, \mathbf{y}, \omega) \right) \chi^\lambda(\mathbf{y}, \mathbf{r}', \omega). \quad (104)$$

Here, the so-called exchange-correlation kernel is defined as functional derivative of the time-dependent exchange-correlation potential, evaluated at the ground-state density of the static (unperturbed) system:

$$f_{xc}^\lambda(\mathbf{r}, \mathbf{r}', \omega) = \int d(t - t') \exp(i\omega(t - t')) \left. \frac{\delta v_{xc}^\lambda[n](\mathbf{r}, t)}{\delta n(\mathbf{r}', t')} \right|_{n(\mathbf{r}, t) = n(\mathbf{r})}. \quad (105)$$

The Kohn-Sham linear density response function χ_s may be expressed in terms of the Kohn-Sham orbitals and orbital energies through

$$\chi_s(\mathbf{r}, \mathbf{r}', \omega) = \sum_\sigma \sum_{i,j} (f_{i\sigma} - f_{j\sigma}) \frac{\varphi_{i\sigma}^*(\mathbf{r}) \varphi_{j\sigma}(\mathbf{r}) \varphi_{j\sigma}^*(\mathbf{r}') \varphi_{i\sigma}(\mathbf{r}')}{\omega - (\varepsilon_{j\sigma} - \varepsilon_{i\sigma}) + i\delta} \quad (106)$$

where $f_{i\sigma}$ is the occupation number (0 or 1) of the orbital $\varphi_{i\sigma}(\mathbf{r})$ in the Kohn-Sham ground-state Slater determinant. Insertion of Eqs. (106) and (104) into Eq. (103) represents the exact exchange-correlation energy as an explicitly orbital-dependent functional.

It should be noted that, similar to Görling-Levy perturbation theory of the previous section, also the representation (Eq. (103)) allows for E_{xc} to be expanded in a power series in e^2 , i.e.,

$$E_{xc} = \sum_{j=1}^{\infty} e^{2j} E_{xc}^{(j)} = E_x + E_c^{(2)} + \dots \quad (107)$$

where the Fock exchange energy constitutes the first-order and E_c^{GL2} of Eq. (98) the second-order term. The series (Eq. (107)) has to be understood as a series expansion of the *functional* E_{xc} which can be evaluated with any set of single-particle orbitals. In particular, it usually is evaluated on the Kohn-Sham orbitals which themselves (through the Kohn-Sham potential v_s) also depend on e^2 .

Since the exchange-correlation potential is computed from E_{xc} as the functional derivative (Eq. (13)), Eq. (107) implies that also v_{xc} can be written as a power series in

e^2 ,

$$v_{xc}(\mathbf{r}) = \sum_{j=1}^{\infty} e^{2j} v_{xc}^{(j)}(\mathbf{r}) = v_x(\mathbf{r}) + v_c^{(2)}(\mathbf{r}) + \dots \quad (108)$$

where the first order term is again the exact exchange contribution.

In a sense, the representation (Eq. (103)) of E_{xc} is much more powerful than the Görling-Levy expansion. While the latter provides an order-by-order expansion of E_{xc} , the former provides a prescription for resummation of an infinite power series if for the response function χ^λ one inserts a solution of the integral equation (104). The simplest of these resummed functionals is the so-called random phase approximation (RPA) which results if the exchange-correlation kernel is set to zero, $f_{xc}^{RPA} \equiv 0$.

4.3 Exact Exchange Energy

As we have seen in the previous section, the exact exchange energy functional (Eq. (91)) is nothing but the Fock term evaluated with Kohn-Sham orbitals and may be viewed as the leading term in a power series expansion of E_{xc} in powers of e^2 . Moreover, if one does not include any correlation, the corresponding total energy functional (97) is exactly of the Hartree-Fock form. Therefore a comparison between exact-exchange only DFT and Hartree-Fock theory is instructive.

In Hartree-Fock theory, the energy functional (Eq. (97)) is minimized with respect to the orbitals without restriction leading to the non-local Hartree-Fock potential. Consequently, the Hartree-Fock ground state Slater determinant gives the lowest energy when inserted into the Hartree-Fock energy functional. In contrast, in exchange-only OEP (EXX) one minimizes the same functional under the constraint that the orbitals are eigenfunctions to one and the same *local* potential. Since this minimization is not unconstrained, the resulting total energy from the OEP orbitals must be higher than the Hartree-Fock energy. However, OEP gives the optimal local potential and therefore any approximate local potential such as exchange-only KLI or CEDA yields orbitals which, when inserted into the Hartree-Fock energy expression, lead to total energies higher than the OEP ones. In practice, it turns out that total energies in Hartree-Fock are typically very close to exchange-only OEP, KLI and CEDA results.

Both Hartree-Fock and EXX theories are free of self-interaction for the occupied orbitals. This means that the orbital $\varphi_{i\sigma}$ does not “feel” the electrostatic potential formally created by itself as part of the Hartree potential because this term is exactly cancelled by a corresponding term in the exchange potential. It should be noted that standard functionals like LDA or GGA do not have this property. Moreover, while the Hartree-Fock potential is *not* self-interaction free for unoccupied states, the EXX potential is. As a consequence, Hartree-Fock only leads to few unoccupied bound states. On the other hand, the EXX potential decays asymptotically as $-1/r$ for finite systems for *all* orbitals and therefore supports a whole Rydberg series of unoccupied bound states.

The latter statement needs some clarification. It is generally said that the EXX potential for finite systems behaves as^{22, 25, 47}

$$\lim_{r \rightarrow \infty} v_x^{EXX}(\mathbf{r}) \rightarrow -\frac{1}{r}. \quad (109)$$

Although this statement is true for $\mathbf{r}$ far away from the system in most directions, it has been found^{48,20} that if one approaches the asymptotic region on nodal surfaces of the highest occupied orbital the Kohn-Sham exchange potential may actually approach a non-vanishing constant value.

For metals, the Hartree-Fock single-particle density of states at the Fermi energy vanishes. This unphysical feature does not appear for the EXX density of states.

4.4 Correlation Energy Functionals from GL2 to RPA and Beyond

Although both the adiabatic connection method of Section 4.1 as well as the fluctuation-dissipation theorem of Section 4.2 provide a way to construct orbital functionals for the correlation energy, much less work has been done using these functionals than with the EXX functional. The reason is mainly a practical one: while EXX depends on the occupied orbitals only, these functionals for correlation depend on *all* Kohn-Sham orbitals, occupied and unoccupied, as well as on the Kohn-Sham eigenenergies. Although the OEP equations for the corresponding correlation potentials can easily be written down, their solution is much more difficult in practice. However, some features of these orbital functionals for correlation are known and further investigation of their properties appears both interesting and promising.

One of the interesting properties of orbital-dependent correlation energy functionals is their ability to give the proper C_6/R^6 long-range van-der-Waals contribution to the total energy for two closed-shell atoms separated by distance R . Zaremba and Kohn⁴⁹ have derived an expression for the C_6 van-der-Waals coefficient in terms of the dynamic polarizability of the separated subsystems. The latter can easily be expressed in terms of the linear density response function of Eq. (104). C_6 coefficients have been evaluated⁵⁰ for some atoms using an approximate linear response function where χ on the right hand side of Eq. (104) is replaced by the Kohn-Sham linear response function χ_s and the PGG approximation⁴⁶ for the exchange-correlation kernel f_{xc} was used. The results were found to be accurate to within 10-20 % for light atoms but considerably worse for heavier ones.

Along similar lines, Engel et. al.⁵¹ have mapped out the binding energy curve of rare gas dimers as a function of atomic separation using the Görling-Levy second order functional of Eq. (98). They found a qualitatively correct description, however, a full quantitative description apparently requires higher-order correlations to be taken into account.

The results of Refs. 50 and 51 were obtained from non-selfconsistent calculations. In more recent work⁵² a problem with the self-consistent calculation of the correlation potential from the second-order functional (Eq. (98)) has been pointed out. It was found that for finite systems the corresponding correlation potential diverges exponentially for large distances r . This is in contrast to an exact result⁵³ which says that the exact correlation potential decays asymptotically as $\sim 1/r^4$. The work of Facco Bonetti et. al.⁵² has sparked a discussion^{54,55} where theoretical arguments have been put forward which indicate that the correct asymptotic behavior can only be obtained when subtle cancellation effects (related to the exact closure relation and discretization of the continuum) are taken into account⁵⁶. Self-consistent calculations for atoms using E_c^{GL2} have also been performed in a basis set representation in Ref. 57.

Recently, there has been some effort in going beyond the second-order approximation and use Eq. (103) with a density response function which solves the Dyson-like equation

(104) for some approximation to the exchange-correlation kernel f_{xc} . Non-selfconsistent calculations have been performed in RPA and beyond, both for simple model systems such as jellium slabs⁵⁸ but also for molecules^{59,60}. For molecules, also van-der-Waals binding energy curves for rare-gas dimers have been calculated⁶⁰.

It is known that in order to construct a correlation energy functional which is compatible with exact exchange, the correlation part has to cancel the long-range part of the exact exchange hole. Local functionals like LDA or GGA have a local correlation hole and therefore combination of these functionals with exact exchange is less accurate than if also an LDA or GGA approximation for exchange is used. The correlation energy functionals discussed in this section have a chance to be compatible with exact exchange due to their nonlocal nature. In fact, a functional has been constructed which uses E_c^{GL2} as input and interpolates the coupling-constant integrand W_λ of Eq. (88). This so-called interaction strength interpolation (ISI) functional has been demonstrated to give accurate results for molecular atomization energies when combined with exact exchange⁶¹.

4.5 Self-Interaction Correction to LDA

Unlike the exact exchange functional discussed in Section 4.3, explicit density functionals like LDA or GGA typically are not free of self-interaction, i.e., the exchange-correlation energy does not cancel exactly the self-interaction contained in the Hartree energy (Eq. (5)). One of the consequences is the incorrect exponential asymptotic decay of the corresponding exchange-correlation potentials for finite systems.

Some time ago, Perdew and Zunger⁷ suggested to make any approximate E_{xc}^{approx} self-interaction free by removing the self-interaction explicitly for each orbital. Their self-interaction corrected (SIC) exchange-correlation energy is constructed by (in the formulation of collinear spin-DFT)

$$E_{xc}^{SIC}[n_\uparrow, n_\downarrow] = E_{xc}^{approx}[n_\uparrow, n_\downarrow] - \sum_{\sigma=\uparrow,\downarrow} \sum_{j=1}^{N_\sigma} \left(U[n_{j\sigma}] + E_{xc}^{approx}[n_{j\sigma}, 0] \right) \quad (110)$$

where the $n_{j\sigma}(\mathbf{r})$ are the orbital densities defined by Eq. (35). In the original work⁷, the corresponding total energy functional was minimized with respect to the orbitals, leading to orbital-dependent effective single-particle potentials. However, Eq. (110) certainly is an orbital-dependent functional and may therefore be treated with the OEP formalism⁶².

The self-interaction correction should be particularly important for localized orbitals such as *d*- or *f*-states in transition metals or rare earths. The SIC functional has successfully been applied to such materials in the solid state^{63,64}. However, here it is important that the functional is *not* evaluated with the Kohn-Sham eigenstates, since in a solid these states are Bloch orbitals and often extend throughout the whole system. For such extended states, the self-interaction correction vanishes. Therefore, in order to apply SIC, one has to construct localized orbitals which are then used to evaluate the SIC correction.

4.6 Other Orbital Functionals: Hybrids and Meta-GGA's

Finally we want to discuss two other classes of functionals which fall into the domain of orbital functionals.

One of them are the so-called hybrid functionals which are constructed by approximating the exchange energy by a *fraction* of exact exchange plus some GGA part for the remainder, i.e.,

$$E_x^{hyb}[n] = aE_x^{EXX}[n] + (1 - a)E_x^{GGA}[n] \quad (111)$$

where a is a constant parameter. Usually, this approximate exchange functional is combined with some GGA for correlation. Hybrids have been introduced in quantum chemistry^{65–68} and have been found to yield accurate results for many energetic properties. Probably the most widely known hybrid functional is the three-parameter functional with the acronym B3LYP which is defined by

$$E_{xc}^{B3LYP}[n] = E_{xc}^{LDA}[n] + a_0(E_x^{EXX}[n] - E_x^{LDA}[n]) + a_x E_x^{B88}[n] + a_c E_c^{LYP}[n] \quad (112)$$

Here E_x^{B88} is the GGA exchange energy of Becke¹⁴. E_c^{LYP} is the GGA correlation functional suggested by Lee, Yang, and Parr¹⁵ which is derived from the explicitly orbital-dependent functional of Colle and Salvetti⁶⁹.

Hybrid functionals are implemented in a wide range of quantum chemistry program packages. Interpretation of the results of these packages, however, requires some care. The reason is that the self-consistency cycle in many cases does not use the OEP method or any approximation to it to evaluate the corresponding effective single-particle potentials. Instead, the results are obtained by minimizing the energy expression with respect to the single-particle orbitals. Due to the appearance of the Fock term $E_x^{EXX}[n]$ in the energy expression, this leads to *non-local* effective single-particle potentials and thus the results are outside the realm of density functional theory.

The second class of functionals we would like to mention here are the so-called meta-GGA's^{70–72} which are of the general form

$$E_{xc}^{MGA}[n] = \int d^3r g(n, \nabla n, \tau) \quad (113)$$

where

$$\tau(\mathbf{r}) = \frac{1}{2} \sum_{\sigma=\uparrow,\downarrow} \sum_{i=1}^{N_\sigma} |\nabla \varphi_{i\sigma}(\mathbf{r})|^2 \quad (114)$$

is the kinetic energy density of the Kohn-Sham orbitals. Through their dependence on τ , meta-GGA's also become orbital functionals and their exchange-correlation potentials have to be calculated with the OEP method. As in the case of the hybrids, this is not always done. However, meta-GGA's have been implemented self-consistently in the LHF approximation to OEP to calculate magnetic response properties⁷³.

5 The LDA+U Method

In this section we will discuss a method which has been used to describe strongly correlated systems in a modified DFT framework. This method uses insights gained from studies of model Hamiltonians for strong correlations like the Hubbard model⁷⁴ and tries to incorporate them into the density functional framework. The resulting so-called LDA+U method⁷⁵ has been found, e.g., to correctly predict the antiferromagnetic insulating ground state of La_2CuO_4 ⁷⁶. Although usually the LDA+U procedure is not discussed in the context of the

OEP method, structural similarities to DFT with an orbital-dependent energy functional may also be found here.

The rationale behind the LDA+U method is the observation that the LDA does not properly describe strongly correlated systems such as transition metal oxides or rare-earth compounds where localized d - and f -electrons play an important role. One tries to correct the LDA treatment of those orbitals and to identify those regions of space (usually “atomic spheres”) where the atomic character of the localized orbitals has survived. Within these spheres one expands the electronic states in a localized basis $|inlm\sigma\rangle$ of atomic eigenstates where i is a site index, n is the principal quantum number, l the angular momentum, m the magnetic quantum number and σ the spin index. For simplicity, we restrict ourselves to situations where only one particular nl shell is contributing. The matrix elements of the density matrix for this shell are then given by

$$f_{mm'}^\sigma = -\frac{1}{\pi} \int^{\varepsilon_F} dE \operatorname{Im} (G_{inlm;inlm'}^{s,\sigma}(E)) \quad (115)$$

where $G_{inlm;inlm'}^{s,\sigma}(E) = \langle inlm\sigma | (E - \hat{H}_s)^{-1} | inlm'\sigma \rangle$ is the matrix element of the single-particle Green’s function corresponding to the non-interacting Hamiltonian $\hat{H}_s$. One then writes the total energy functional in the LDA+U method in terms of the matrix elements (Eq. (115)) as⁷⁷

$$E^{LDA+U}[n, \{f^\sigma\}] = E^{LDA}[n] + E^U[\{f^\sigma\}] - E^{dc}[\{f^\sigma\}]. \quad (116)$$

Here, $E^{LDA}[n]$ is the usual LDA total energy functional and

$$\begin{aligned} E^U[\{f^\sigma\}] = & \frac{1}{2} \sum_{\{m\}, \sigma}^{loc} \left(\langle mm'' | \hat{V}_{ee}^{scr} | m'm''' \rangle f_{mm'}^\sigma f_{m''m'''}^{-\sigma} \right. \\ & \left. + \left(\langle mm'' | \hat{V}_{ee}^{scr} | m'm''' \rangle - \langle mm'' | \hat{V}_{ee}^{scr} | m'''m' \rangle \right) f_{mm'}^\sigma f_{m''m'''}^\sigma \right) \end{aligned} \quad (117)$$

where $\hat{V}_{ee}^{scr}$ is the screened Coulomb interaction between the nl electrons and the sum runs only over these localized states. Finally, $E_{dc}[\{f^\sigma\}]$ corrects for double counting and is given by

$$E^{dc}[\{f^\sigma\}] = \frac{1}{2} U N(N-1) - \frac{1}{2} J (N_\uparrow(N_\uparrow-1) + N_\downarrow(N_\downarrow-1)) \quad (118)$$

where N_σ is the number of σ electrons in the nl shell and $N = N_\uparrow + N_\downarrow$. U is the Hubbard parameter which is an average measure for the on-site (screened) Coulomb interaction while, similarly, J is an average (screened) exchange parameter.

Eqs. (116), (117), and (118) define the total energy in terms of the localized basis functions $|inlm'\sigma\rangle$ which is somewhat unusual. In order to get some insight into the nature of this definition we consider the very simple case of an isolated atom. We further take $\hat{H}_s$ as the Kohn-Sham Hamiltonian corresponding to the LDA+U energy functional. It should be noted that this is a deviation from the LDA+U procedure as it is typically applied in the literature. There the Hamiltonian is orbital dependent and results from a minimization of Eq. (116) with respect to the orbitals.

One now can choose the eigenfunctions for such an atomic, orbital-independent $\hat{H}_s$ as basis $|inlm\sigma\rangle$. Then the density matrix (115) becomes diagonal, $f_{mm'}^\sigma = f_{mm}^\sigma \delta_{m,m'}$ and

the energy E^U reads

$$E^U[\{f^\sigma\}] = \frac{1}{2} \sum_{\{m\}, \sigma}^{loc} \left(\langle mm' | \hat{V}_{ee}^{scr} | mm' \rangle f_{mm}^\sigma f_{m'm'}^{-\sigma} + \left(\langle mm' | \hat{V}_{ee}^{scr} | mm' \rangle - \langle mm' | \hat{V}_{ee}^{scr} | m'm \rangle \right) f_{mm}^\sigma f_{m'm'}^\sigma \right). \quad (119)$$

This can be rewritten in terms of the density

$$n^{loc}(\mathbf{r}) = \sum_{\{m\}}^{loc} f_{mm}^\sigma |\varphi_{nlm\sigma}(\mathbf{r})|^2 \quad (120)$$

and single-particle density matrix

$$\gamma_\sigma^{loc}(\mathbf{r}, \mathbf{r}') = \sum_{\{m\}, \sigma}^{loc} f_{mm}^\sigma \varphi_{nlm\sigma}^*(\mathbf{r}) \varphi_{nlm\sigma}(\mathbf{r}') \quad (121)$$

of the localized states as an Hartree-Fock-like energy with respect to the screened interaction, i.e.,

$$E^U[\{\varphi_{nlm\sigma}\}] = \frac{1}{2} \int d^3r \int d^3r' n^{loc}(\mathbf{r}) V_{ee}^{scr}(\mathbf{r}, \mathbf{r}') n^{loc}(\mathbf{r}') - \frac{1}{2} \sum_{\sigma} \int d^3r \int d^3r' |\gamma_\sigma^{loc}(\mathbf{r}, \mathbf{r}')|^2 V_{ee}^{scr}(\mathbf{r}, \mathbf{r}'). \quad (122)$$

This shows that the LDA+U functional tries to correct LDA by treating localized orbitals on a Hartree-Fock-like footing with respect to the screened interaction. The connection between the screened interaction in the LDA+U method and the screened Coulomb interaction of Hedin's GW approach⁷⁸ has been discussed in Ref. 79. In practice one expresses the matrix elements of the screened interaction V_{ee}^{scr} in terms of known Slater integrals and the parameters U and J which are determined from LDA supercell calculations^{75, 77, 79}.

For simplicity, we discussed only the case of an isolated atom but it is clear that one can view the LDA+U method in the framework of orbital functionals for any system if the functional E^U of Eq.(122) is interpreted as a functional of the *localized* Kohn-Sham orbitals only. Of course, this requires a selection of a set of localized orbitals by physical insight into the system under consideration.

In a sense the LDA+U method is similar to the philosophy both of the self-interaction corrected (SIC) functional of Section 4.5 and the hybrids of Section 4.6. It is similar to SIC because Eq. (122) is free of self-interaction of the localized orbitals by construction. These are the orbitals for which the self-interaction error of LDA is largest. On the other hand it is similar to the hybrids because it uses a fraction of “screened” exchange, although not in a global sense but by correcting only a particular set of orbitals, i.e., the localized ones.

6 Numerical Results

In this section we report some results obtained with the OEP method for various systems ranging from atoms and molecules to solids. The selection of results presented is, of

course, both subjective and incomplete. We concentrate on results obtained with the exact exchange functional and, if data are available, correlation energy functionals either from perturbation theory or from the fluctuation-dissipation formula.

6.1 Atoms

We start by presenting some results of exchange-only calculations for atoms using the exact exchange functional in spin-unrestricted OEP and KLI, as well as pure exchange-only spin density functionals such as LDA, Becke's GGA for exchange (B88)¹⁴ and the Perdew-Wang GGA (PW91)¹⁶. Spin-unrestricted Hartree-Fock⁸⁰ (SUHF) energies are also given as a reference. In Table 1 we present total ground state energies in these different approximations for atoms with nuclear charge $Z \leq 10$. Results for heavier atoms can be found, e.g., in Ref. 18. The calculations have been done solving the radial part of the Schrodinger equation by the Numerov method. For open shell configurations the Kohn-Sham potential has been calculated with fractional occupation of all orbitals within this shell. This kind of angular averaging leads to spherically symmetric Kohn-Sham potentials and the angular part can be treated analytically.

We first compare the SUHF and exchange-only OEP results. In this case, the two schemes differ only in the constraints used when minimizing the same total energy functional. In the OEP method, the minimization is subject to the constraint that the single-particle wave functions are eigenfunctions of one *local* effective potential. In the Hartree-Fock method, one computes those orbitals which minimize the Hartree-Fock total energy without any constraint (except orthonormality) and a non local effective potential is obtained. As a consequence of the additional constraint, the OEP total energies have to be higher (or equal at best) than the corresponding SUHF ones. This is confirmed by the data in Table (1). It can also be seen that the OEP total energies are very close to the SUHF results.

Comparing the KLI with the OEP results one sees that the KLI approximation gives higher total energies for almost all cases except for one- and two-electron systems for

Atom	SUHF	OEP	KLI	B88	PW91	LDA
H	0.5000	0.5000	0.5000	0.4979	0.4953	0.4571
He	2.8617	2.8617	2.8617	2.8634	2.8552	2.7236
Li	7.4328	7.4325	7.4324	7.4288	7.4172	7.1934
Be	14.5730	14.5724	14.5723	14.5664	14.5543	14.2233
B	24.5293	24.5283	24.5281	24.5173	24.5035	24.0636
C	37.6900	37.6889	37.6887	37.6819	37.6658	37.1119
N	54.4046	54.4034	54.4030	54.4009	54.3824	53.7093
O	74.8136	74.8121	74.8117	74.8148	74.7964	73.9919
F	99.4108	99.4092	99.4087	99.4326	99.4130	98.4740
Ne	128.5471	128.5454	128.5448	128.5901	128.5689	127.4907

Table 1. Absolute ground state energies (in atomic units) for different exchange-only approximations. Values are taken from¹⁸ and references therein.

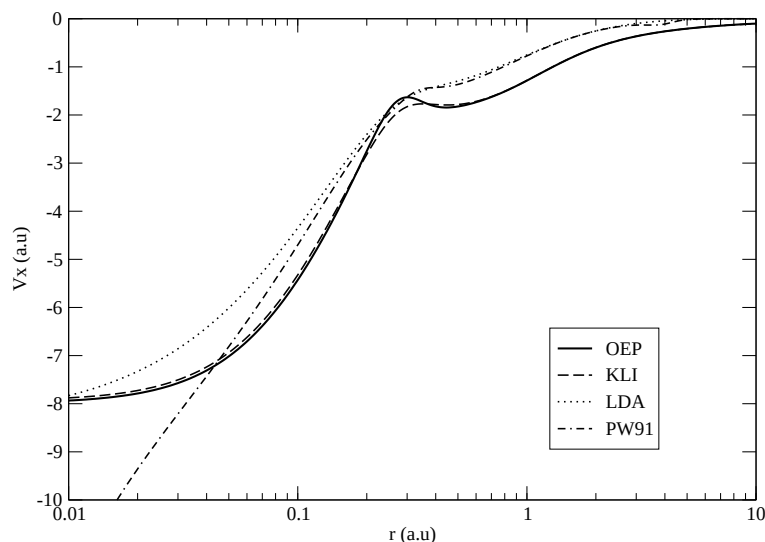


Figure 1. Exchange potentials for Ne from various self-consistent exchange-only calculations.

which the KLI approximation is exact and coincides with both OEP and HF. The difference between KLI and OEP total energies is in general very small. Moreover it tends to decrease almost monotonically with increasing atomic numbers. At least from the point of view of the total energies, the KLI scheme provides a very good approximation to the full OEP results. This statement is often true also for other quantities such as single-particle eigenvalues and potentials. However, as we will see later, there are properties for which the KLI approximation is not so accurate.

Comparing OEP results with those of standard density functionals, one sees that LDA and OEP total energies show large differences. Something can be gained by considering GGA functionals such as B88 and PW91. It should be noted that for some atoms the GGA functionals give total energies lower than OEP.

In Figure 1 we show the self-consistent exchange potentials for the Ne atom in various approximations. Unlike the exact exchange functional both LDA and GGA's suffer from the self-interaction error. The self-interaction energy contained in the Hartree energy is not exactly cancelled by LDA or GGA exchange which leads to the incorrect asymptotic behavior of the corresponding potentials: both LDA and GGA potentials decay exponentially fast far away from the nucleus. On the other hand, both OEP and KLI potentials are free of self-interaction and therefore their asymptotics is dominated by the $-\frac{1}{r}$ term. The KLI potential follows the OEP potential rather closely in most regions of space. The difference is largest in the atomic inter-shell region where the OEP potentials shows a pronounced hump while in the KLI potential the hump is less prominent.

The self-interaction problem has another consequence as well. Since both LDA and GGA potentials decay exponentially fast, a neutral atom does not exhibit a Rydberg series of excited states in these approximations. Also, negative ions are not stable. On the other hand, OEP and KLI both show a Rydberg series and are also able to support negative

Atom/Ion	KLI		BLYP	PW91	exact
	EXX	EXX+CS			
He	0.9180	0.9446	0.5849	0.5833	0.9037
Be ²⁺	5.6671	5.6992	4.8760	4.8701	5.6556
Be	0.3089	0.3294	0.2009	0.2072	0.3426
C ²⁺	1.6933	1.7226	1.4804	1.4856	1.7594
Ne	0.8494	0.8841	0.4914	0.4942	0.7945
Na ⁺	1.7959	1.8340	1.3377	1.3416	1.7410

Table 2. Ionization potentials (in a.u.) from highest occupied Kohn-Sham orbital energies for different exchange-correlation functionals. EXX+CS are results for the exact exchange functional combined with the orbital functional for correlation of Ref. 69 in KLI approximation. Values are taken from Ref. 18.

ions. Moreover, if one calculates the ionization potential as the negative eigenvalue of the highest occupied Kohn-Sham orbital⁵³ one finds a much better agreement with experiment in OEP than in GGA. This is illustrated for some atoms and ions in Table 2.

While in OEP the self-interaction is absent for both occupied and unoccupied orbitals, in Hartree-Fock only the occupied orbitals are self-interaction free. Therefore the unoccupied orbitals in Hartree-Fock are usually too high in energy leading to HOMO-LUMO gaps which are too large.

In Section 3.2, in addition to the KLI approximation we also discussed a slightly modified approximate scheme for the solution of the OEP equation, the CEDA approximation. How do CEDA results compare to OEP or KLI ones? As a matter of fact, total energies, single-particle energies and potentials of atoms are very close to each other in the three different schemes⁸¹. Table 3 shows atomic ground state energy differences between CEDA and SUHF as well as KLI and SUHF results. One can see that the KLI and CEDA ground state energies are nearly identical. It should be mentioned here, however, that CEDA and KLI results not for all properties agree so well among each other. An example where CEDA and KLI results are not so close will be discussed in Section 6.2.

Finally, in Table 4 we report some results for atomic correlation energies from evaluation of Eq. (103) with different approximations for the linear response function. The results were obtained for two different exchange-correlation kernels, the RPA kernel $f_{xc}^{RPA} \equiv 0$ and the kernel f_{xc}^{PGG} of Petersilka et. al.⁴⁶. The second order correlation energies were obtained by inserting the response function obtained from Eq. (104) with χ^λ on the right

Atom	$E^{KLI} - E^{SUHF}$	$E^{CEDA} - E^{SUHF}$
Be	0.00060	0.00060
Ne	0.0022	0.0022
Ar	0.0048	0.0052

Table 3. Atomic energy differences between KLI, CEDA and HF total energies (in hartrees). Values taken from Ref. 81.

Atom	second order		infinite order		exact
	$f_{xc} = 0$	f_{xc}^{PGG}	$f_{xc} = 0$	f_{xc}^{PGG}	
He	0.0961	0.0481	0.0830	0.0444	0.0420
Be	0.2419	0.1254	0.1753	0.1011	0.0950
Ne	0.7157	0.4013	0.5799	0.3293	0.3929

Table 4. Atomic correlation energies ($-E_c$ in hartrees) in second order and to infinite order for different exchange-correlation kernels. Exact results are from Ref. 82.

hand side replaced by χ_s . The results to infinite order used the response function solving the integral equation (104) for the given approximate f_{xc} . The results are evaluated with exchange-only KLI orbitals represented with the Slater-type basis set used in Ref. 83.

6.2 Molecules

DFT with orbital functionals has been applied to molecular systems as well. The first results were obtained with the KLI approximation^{84–86}. Later, both CEDA⁸¹ and full OEP results were reported^{87,88,20}. For total energies, binding energies and vibrational frequencies the exchange-only KLI, CEDA, and OEP results are rather close to each other and also close to Hartree-Fock values. In Table 5 we show few deviations of total ground state energies from Hartree-Fock results both in KLI and CEDA approximations which are almost identical.

For molecular binding energies, exchange-only KLI, CEDA, and OEP results are of rather poor quality. The errors, which are close to Hartree-Fock errors, are on average more than twice as large as in LDA and almost an order of magnitude worse than GGA results⁸⁸. This is not unexpected since the exchange hole has a long-range component in the dissociation limit which has to be compensated by a corresponding long-range component of the correlation hole. Correlation functionals with this property are notoriously difficult to construct⁶¹. Combination of the exact exchange energy with LDA or GGA correlation leads to some improvement, leading to errors for the atomization energies of the same order of magnitude as LDA results (but typically with the opposite sign). This confirms that the success of explicit functionals relies on an error cancellation between the approximate exchange and correlation parts of the total energy.

Not for all properties KLI, CEDA and OEP results are so similar as for total energies and related quantities. Quite large differences emerge⁸¹ for linear and non-linear response

Molecule	$E^{KLI} - E^{HF}$	$E^{CEDA} - E^{HF}$
N ₂	0.0078	0.0077
CO	0.0076	0.0076
H ₂ O	0.00392	0.00364

Table 5. Molecular energies differences between KLI, CEDA and HF total energies (in hartrees). Values from Ref. 81.

α	H ₄	H ₆	H ₈	H ₁₂	H ₁₈
HF	32.0	56.4	82.3	137.6	222.3
OEP	32.2	56.6	84.2	138.1	
CEDA		59.3		149.4	244.2
KLI	33.1	60.2	90.6	156.3	260.7

Table 6. Linear polarizabilities α (in a.u.) for different hydrogen chains in various exchange-only treatments. Values from Ref. 89 and Ref. 81.

properties of molecular chains where the high mobility of the valence electrons along the backbone leads to a large directional electronic response.

The response properties of these molecular chains can be characterized by the linear polarizabilities α and the hyperpolarizabilities γ , defined as the first and third derivatives, $\partial\mu_z/\partial E_z$ and $\partial^3\mu_z/\partial^3 E_z$, of the dipole moment μ_z with respect to the electric field E_z along the direction of the chain, respectively.

In LDA and GGA α and γ are usually overestimated by orders of magnitude in comparison to Hartree-Fock results which are in reasonable agreement with results obtained from correlated quantum chemical methods^{81,89}. KLI and CEDA values give some improvement over standard functionals. This is due to the orbital structure of the CEDA and KLI exchange potentials which produces an exchange field counteracting the external field. Moreover, there are pronounced differences between KLI and CEDA results (see Tables 6 and 7). The latter gives considerable improvement as compared to KLI, but in order to reach Hartree-Fock quality a full solution of the OEP equations is required⁸⁹.

Some molecular calculations using RPA-type correlation energy functionals have been reported recently^{59,60,90}. RPA atomization energies are much improved over EXX ones^{59,60} with an error close to but slightly worse than standard GGA functionals. Similar results were obtained with kernels derived from standard GGA's⁶⁰.

The binding energy curve of the H₂ molecule has been studied in detail in Ref. 90 using the RPA functional (in a non-selfconsistent way). There the total density obtained from unrestricted exchange-only calculations is used to obtain an accurate *spin-restricted* Kohn-Sham potential. The orbitals resulting from this potential are then used as input for the RPA calculation. The resulting ground state energies are accurate not only around equilibrium bond length but also at larger interatomic separation beyond the Coulson-Fisher point.

$\gamma/10^3$	H ₄	H ₆	H ₈	H ₁₂	H ₁₈
HF		29.8		147	301.3
OEP	9.3	30	68	144	
CEDA		34.7		209.2	468.4
KLI	10.7	36	90	300	778.1

Table 7. Hyperpolarizabilities γ (in a.u.) for different hydrogen chains in various exchange-only treatments. Values from Ref. 89 and Ref. 81.

However at intermediate separations an unphysical bump appears, which may be due to the absence of double excitations in adiabatic linear response theory. Similar results are obtained for RPA-like functionals if the exact exchange kernel is included. In another work⁹¹ a different orbital-dependent exchange-correlation functional has been employed in the OEP framework within the CEDA approximation reproducing very well the entire H_2 potential curve.

6.3 Solids

A number of successful applications of the exact exchange functional have been reported for solids. The first application is due to Kotani⁹², who treated the exact exchange potential within the linear muffin-tin-orbital method in the atomic-sphere approximation. Later, Görling⁹³ proposed a procedure to solve the OEP equation for a solid exactly in a plane wave basis. This technique was then applied^{94,95} to several semiconductors. In the same work, the exchange discontinuity Δ_x has been computed as well.

Lattice constants predicted by exact exchange plus LDA correlation generally agree as well with experiment as the full LDA ones⁹⁵. On the other hand, bulk moduli are overestimated and a treatment of correlations beyond LDA is needed⁹⁵.

Probably the most interesting result of these calculations are the ones for the band gaps of semiconductors. Standard functionals such as LDA or GGA typically give Kohn-Sham band gaps which are too small, often by a factor of two. This is a manifestation of two main shortcomings of these functionals: on the one hand there is the self-interaction error and on the other hand they fail to reproduce a finite derivative discontinuity.

The exact exchange functional eliminates the self-interaction problem and the corresponding Kohn-Sham gaps often provide excellent estimates for the experimental gaps as can be seen in Table 8. On the other hand, EXX also leads to a derivative discontinuity Δ_x which is actually quite large. If this is included in the calculation of the gap, the agreement is ruined and the gaps are too large. This finding can also be understood from a different perspective: in Eq. (57) we have expressed the fundamental gap in terms of total energies of the $N - 1$, N , and $N + 1$ -particle system. Experience tells us that EXX and Hartree-Fock total energies are very close. Therefore the fundamental gap in EXX is expected to be close to the Hartree-Fock gap which is usually far too large.

Not always the EXX Kohn-Sham gaps are in such a good agreement with experimental gaps as in semiconductors. In Table 9 we show some results for noble-gas solids⁹⁶. We see that the Kohn-Sham gaps differ from the fundamental band gaps by several eV and

Solid	xcLDA	EXX+cLDA	Δ_x	Exp.
Si	0.52	1.43	5.84	1.17
C	4.16	5.06	8.70	5.47
GaN (Γ)	1.90	3.46	7.63	3.30
InN (Γ)	-0.18	1.40	6.14	1.95

Table 8. Calculated Kohn-Sham and experimental energy gaps in in semiconductors (in eV). EXX is exact exchange, cLDA denotes correlation in LDA. Values taken from⁹⁴ and references therein.

Solid	Δ_{KS}	Δ_{exp}	Δ_{opt}
Ne	14.15	21.4	17.4
Ar	9.61	14.2	12.2
Kr	7.87	11.6	10.2
Xe	6.69	9.8	8.4

Table 9. Calculated Kohn-Sham gaps, Δ_{KS} , and experimental fundamental energy gaps, Δ_{exp} , in noble-gas solids (in eV). Δ_{opt} is the experimental optical gap. Values taken from Ref. 96.

reproduce about 80% of the experimental optical gaps.

Very recently, self-consistent RPA calculations have been performed to study the band gap problem⁹⁷. To simplify the calculation, a plasmon pole approximation for the RPA (or GW) self-energy was used. The resulting Kohn-Sham gaps turned out to be rather close to LDA results, i.e., they were considerably smaller than Kohn-Sham gaps from EXX calculations. However, if the exchange-correlation discontinuity was taken into account, the resulting gaps were close to GW and experimental gaps.

As a final application of the OEP formalism to extended systems we mention a very recent calculation within the non-collinear spin-DFT framework³³ for a magnetically frustrated monolayer of chromium which shows how intra-atomic non-collinearity may be underestimated by local functionals.

Acknowledgements

We gratefully acknowledge financial support through the Deutsche Forschungsgemeinschaft and through the EU Network of Excellence NANOQUANTA.

References

1. P. Hohenberg and W. Kohn. Inhomogeneous electron gas. *Phys. Rev.*, 136:B864, 1964.
2. M. Levy. Universal variational functionals of electron densities, first-order density matrices, and natural spin-orbitals and solution of the v -representability problem. *Proc. Natl. Acad. Sci. USA*, 76:6062, 1979.
3. M. Levy. Electron densities in search of Hamiltonians. *Phys. Rev. A*, 26:1200, 1982.
4. W. Kohn and L.J. Sham. Self-consistent equations including exchange and correlation. *Phys. Rev.*, 140:A1133, 1965.
5. D.M. Ceperley and B.J. Alder. Ground State of the Electron Gas by a Stochastic Method. *Phys. Rev. Lett.*, 45:566, 1980.
6. S.H. Vosko, L. Wilk, and M. Nusair. Accurate spin-dependent electron liquid correlation energies for local spin-density calculations - a critical analysis. *Can. J. Phys.*, 58:1200, 1980.
7. J.P. Perdew and A. Zunger. Self-interaction correction to density-functional approximations for many-electron systems. *Phys. Rev. B*, 23:5048, 1981.

8. J.P. Perdew and Y. Wang. Accurate and simple analytic representation of the electron gas correlation energy. *Phys. Rev. B*, 45:13244, 1992.
9. S.-K. Ma and K.A. Brueckner. Correlation Energy of an Electron Gas with a Slowly Varying High Density. *Phys. Rev.*, 165:18, 1968.
10. M. Rasolt and D.J.W. Geldart. Exchange and correlation energy in a nonuniform fermion fluid. *Phys. Rev. B*, 34:1325, 1986.
11. J.P. Perdew. Accurate Density Functional for the Energy: Real-Space Cutoff of the Gradient Expansion for the Exchange Hole. *Phys. Rev. Lett.*, 55:1665, 1985.
12. J.P. Perdew and Y. Wang. Accurate and simple exchange functional for the electronic exchange energy: Generalized gradient approximation. *Phys. Rev. B*, 33:8800, 1986. *ibid.* **40**, 3399 (1989) (E).
13. A.D. Becke. Density functional calculations of molecular bond energies. *J. Chem. Phys.*, 84:4524, 1986.
14. A.D. Becke. Density functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A*, 38:3098, 1988.
15. C. Lee, W. Yang, and R.G. Parr. Development of the Colle-Salvetti correlation energy formula into a functional of the electron density. *Phys. Rev. B*, 37:785, 1988.
16. J.P. Perdew. Unified theory of exchange and correlation beyond the local density approximation. In P. Ziesche and H. Eschrig, editors, *Electronic Structure of Solids '91*, page 11, Berlin, 1991. Akademie Verlag.
17. J.P. Perdew, K. Burke, and M. Ernzerhof. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.*, 77:3865, 1996. *ibid.* **78**, 1396 (1997)(E).
18. T. Grabo, T. Kreibich, S. Kurth, and E.K.U. Gross. Orbital functionals in density functional theory: the optimized effective potential method. In V. Anisimov, editor, *Strong Coulomb Correlations in Electronic Structure Calculations: Beyond Local Density Approximations*, page 203, Amsterdam, 2000. Gordon and Breach.
19. S. Kümmel and J.P. Perdew. Simple iterative construction of the optimized effective potential for orbital functionals, including exact exchange. *Phys. Rev. Lett.*, 90:043004, 2003.
20. S. Kümmel and J.P. Perdew. Optimized effective potential made simple: Orbital functionals, orbital shifts, and the exact Kohn-Sham exchange potential. *Phys. Rev. B*, 68:035103, 2003.
21. R.T. Sharp and G.K. Horton. A Variational Approach to the Unipotential Many-Electron Problem. *Phys. Rev.*, 90:317, 1953.
22. J.D. Talman and W.F. Shadwick. Optimized effective atomic central potential. *Phys. Rev. A*, 14:36, 1976.
23. V. Sahni, J. Gruenebaum, and J.P. Perdew. Study of the density-gradient expansion for the exchange energy. *Phys. Rev. B*, 26:4371, 1982.
24. J.P. Perdew and M.R. Norman. Electron removal energies in Kohn-Sham density-functional theory. *Phys. Rev. B*, 26:5445, 1982.
25. J.B. Krieger, Y. Li, and G.J. Iafrate. Construction and application of an accurate local spin-polarized Kohn-Sham potential with integer discontinuity: Exchange-only theory. *Phys. Rev. A*, 45:101, 1992.
26. J.B. Krieger, Y. Li, and G.J. Iafrate. Systematic approximations to the optimized effective potential: Application to orbital-density-functional theory. *Phys. Rev. A*, 46:5453, 1992.

27. O.V. Gritsenko and E.J. Baerends. Orbital structure of the Kohn-Sham exchange potential and exchange kernel and the field-counteracting potential for molecules in an electric field. *Phys. Rev. A*, 64:42506, 2001.
28. F. Della Sala and A. Görling. Efficient localized Hartree-Fock methods as effective exact-exchange Kohn-Sham methods for molecules. *J. Chem. Phys.*, 115:5718, 2001.
29. J.P. Perdew, R.G. Parr, M. Levy, and J.L. Balduz. Density functional theory for fractional particle number: Derivative discontinuity of the energy. *Phys. Rev. Lett.*, 49:1691, 1982.
30. L.J. Sham and M. Schlüter. Density-functional theory of the energy gap. *Phys. Rev. Lett.*, 51:1888, 1983.
31. L.J. Sham and M. Schlüter. Density-functional theory of the band gap. *Phys. Rev. B*, 32:3883, 1985.
32. U. von Barth and L. Hedin. A local exchange-correlation potential for the spin polarized case. *J. Phys. C*, 5:1629, 1972.
33. S. Sharma, J.K. Dewhurst, C. Ambrosch-Draxl, N. Helbig, S. Kurth, E.K.U. Gross, S. Shallcross, and L. Nordström. Optimized effective potential method for non-collinear magnetism. cond-mat/0510800 (2005).
34. J. Harris and R.O. Jones. The surface energy of a bounded electron gas. *J. Phys. F*, 4:1170, 1974.
35. D.C. Langreth and J.P. Perdew. The exchange-correlation energy of a metallic surface. *Solid State Commun.*, 17:1425, 1975.
36. D.C. Langreth and J.P. Perdew. Exchange-correlation energy of a metallic surface: Wavevector analysis. *Phys. Rev. B*, 15:2884, 1977.
37. O. Gunnarsson and B.I. Lundqvist. Exchange and correlation in atoms, molecules, and solids by the spin-density-functional formalism. *Phys. Rev. B*, 13:4274, 1976.
38. A. Görling, M. Levy, and J.P. Perdew. Expectation values in density-functional theory, and kinetic contribution to the exchange-correlation energy. *Phys. Rev. B*, 47:1167, 1993.
39. M. Levy and J.P. Perdew. Hellmann-Feynman, virial, and scaling requisites for the exact universal density functionals. shape of the correlation potential and diamagnetic susceptibility for atoms. *Phys. Rev. A*, 32:2010, 1985.
40. A. Görling and M. Levy. Correlation-energy functional and its high-density limit obtained from a coupling-constant perturbation expansion. *Phys. Rev. B*, 47:13105, 1993.
41. A. Szabo and N.S. Ostlund. *Modern Quantum Chemistry*. MacMillan, New York, 1982.
42. M. Levy. Density-functional exchange correlation through coordinate scaling in adiabatic connection and correlation hole. *Phys. Rev. A*, 43:4637, 1991.
43. E. Runge and E.K.U. Gross. Density-Functional Theory for Time-Dependent Systems. *Phys. Rev. Lett.*, 52:997, 1984.
44. E.K.U. Gross, J.F. Dobson, and M. Petersilka. Density functional theory of time-dependent phenomena. In R.F. Nalewajski, editor, *Density Functional Theory II*, volume 181 of *Topics in Current Chemistry*, page 81. Springer, Berlin, 1996.
45. E.K.U. Gross and W. Kohn. Time-dependent density-functional theory. *Adv. Quantum Chem.*, 21:255, 1990.
46. M. Petersilka, U.J. Gossmann, and E.K.U. Gross. Excitation energies from time-

- dependent density-functional theory. *Phys. Rev. Lett.*, 76:1212, 1996.
47. T. Kreibich, S. Kurth, T. Grabo, and E.K.U. Gross. Asymptotic Properties of the Optimized Effective Potential. *Adv. Quantum Chem.*, 33:31, 1999.
 48. F. Della Sala and A. Görling. The asymptotic region of the Kohn-Sham exchange potential in molecules. *J. Chem. Phys.*, 116:5374, 2002.
 49. E. Zaremba and W. Kohn. Van der Waals interaction between an atom and a solid surface. *Phys. Rev. B*, 13:2270, 1976.
 50. M. Lein, J.F. Dobson, and E.K.U. Gross. Towards the description of van der Waals interactions within density functional theory. *J. Comput. Chem.*, 20:12, 1999.
 51. E. Engel, A. Höck, and R.M. Dreizler. Van der Waals bonds in density functional theory. *Phys. Rev. A*, 61:032502, 2000.
 52. A. Facco Bonetti, E. Engel, R.N. Schmid, and R.M. Dreizler. Investigation of the correlation potential from Kohn-Sham perturbation theory. *Phys. Rev. Lett.*, 86:2241, 2001.
 53. C.O. Almbladh and U. von Barth. Exact results for the charge and spin densities, exchange-correlation potentials, and density-functional eigenvalues. *Phys. Rev. B*, 31:3231, 1985.
 54. Y.M. Niquet, M. Fuchs, and X. Gonze. Comment on “Investigation of the correlation potential from Kohn-Sham perturbation theory”. *Phys. Rev. Lett.*, 90:219301, 2003.
 55. A. Facco Bonetti, E. Engel, R.N. Schmid, and R.M. Dreizler. Reply to Y.M. Niquet, M. Fuchs and X. Gonze. *Phys. Rev. Lett.*, 90:219302, 2003.
 56. Y.M. Niquet, M. Fuchs, and X. Gonze. Asymptotic behavior of the exchange--correlation potentials from the linear-response Sham-Schlüter equation. *J. Chem. Phys.*, 118:9504, 2003.
 57. I. Grabowski, S. Hirata, S. Ivanov, and R.J. Bartlett. Ab initio density functional theory: OEP-MBPT(2). A new orbital-dependent correlation functional. *J. Phys. Chem.*, 116:4415, 2002.
 58. J.F. Dobson, J. Wang, and T. Gould. Correlation energies of inhomogeneous many-electron systems. *Phys. Rev. B*, 66:081108(R), 2002.
 59. F. Furche. Molecular tests of the random phase approximation to the exchange-correlation energy functional. *Phys. Rev. B*, 64:195120, 2001.
 60. F. Furche and T. Van Voorhis. Fluctuation-dissipation theorem density-functional theory. *J. Chem. Phys.*, 122:164106, 2005.
 61. M. Seidl, J.P. Perdew, and S. Kurth. Simulation of all-order density functional perturbation theory, using the second-order and the strong-correlation limit. *Phys. Rev. Lett.*, 84:5070, 2000.
 62. J. Chen, J.B. Krieger, Y. Li, and G.J. Iafrate. Kohn-Sham calculations with self-interaction-corrected local-spin-density exchange-correlation energy functional for atomic systems. *Phys. Rev. A*, 54:3939, 1996.
 63. A. Svane and O. Gunnarsson. Transition-metal oxides in the self-interaction-corrected density-functional formalism. *Phys. Rev. Lett.*, 65:1148, 1990.
 64. Z. Szotek, W.M. Temmerman, and H. Winter. Self-interaction corrected, local spin density description of the γ - α transition in Ce. *Phys. Rev. Lett.*, 72:1244, 1994.
 65. A.D. Becke. A new mixing of Hartree-Fock and local density-functional theories. *J. Chem. Phys.*, 98:1372, 1993.
 66. A.D. Becke. Density-functional thermochemistry: III. The role of exact exchange.

J. Chem. Phys., 98:5648, 1993.

67. A.D. Becke. Density-functional thermochemistry: IV. A new dynamical correlation functional and implications for exact-exchange mixing. *J. Chem. Phys.*, 104:1040, 1996.
68. C. Adamo and V. Barone. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *J. Chem. Phys.*, 110:6158, 1999.
69. R. Colle and O. Salvetti. Approximate calculation of the correlation energy for closed shells. *Theoret. Chim. Acta*, 37:329, 1975.
70. J.P. Perdew, S. Kurth, A. Zupan, and P. Blaha. Accurate density functional with correct formal properties: A step beyond the generalized gradient approximation. *Phys. Rev. Lett.*, 82:2544, 1999. *ibid.* **82**, 5179 (1999)(E).
71. S. Kurth, J.P. Perdew, and P. Blaha. Molecular and solid state tests of density functional approximations: LSD, GGA's, and Meta-GGA's. *Int. J. Quantum Chem.*, 75:889, 1999.
72. J. Tao, J.P. Perdew, V.N. Staroverov, and G.E. Scuseria. Climbing the density functional ladder: Nonempirical meta-generalized gradient approximation designed for molecules and solids. *Phys. Rev. Lett.*, 91:146401, 2003.
73. A.V. Arbuznikov and M. Kaupp. The self-consistent implementation of exchange--correlation functionals depending on the local kinetic energy density. *Chem. Phys. Lett.*, 381:495, 2003.
74. J. Hubbard. Electron correlations in narrow energy bands. *Proc. R. Soc. London*, A 276:238, 1963.
75. V.I. Anisimov, J. Zaanen, and O.K. Andersen. Band theory and Mott insulators: Hubbard U and Stoner J. *Phys. Rev. B*, 44:943, 1991.
76. M.T. Czyżyk and G.A. Sawatzky. Local-density functional and on-site correlations: The electronic structure of La_2CuO_4 and LaCuO_3 . *Phys. Rev. B*, 49:14211, 1994.
77. A.I. Liechtenstein, V.I. Anisimov, and J. Zaanen. Density-functional theory and strong interactions: Orbital ordering in Mott-Hubbard insulators. *Phys. Rev. B*, 52:R5467, 1995.
78. L. Hedin. New method for calculating the one-particle Green's function with application to the electron-gas problem. *Phys. Rev.*, 139:A796, 1965.
79. V.I. Anisimov, F. Aryasetiawan, and A.I. Liechtenstein. First-principles calculations of the electronic structure and spectra of strongly correlated systems: the LDA+U method. *J. Phys. Condens. Matter*, 9:767, 1997.
80. Y. Li, J.B. Krieger, and G.J. Iafrate. Self-consistent calculations of atomic properties using self-interaction-free exchange-only Kohn-Sham potentials. *Phys. Rev. A*, 47:165, 1993.
81. M. Grüning, O.V. Gritsenko, and E.J. Baerends. Exchange potential from the common energy denominator approximation for the Kohn-Sham Green's function: Application to (hyper)polarizabilities of molecular chains. *J. Chem. Phys.*, 116:6435, 2002.
82. J.B. Krieger, J. Chen, G.J. Iafrate, and A. Savin. Construction of an accurate self-interaction-corrected correlation energy functional based on an electron gas with a gap. In A. Gonis and N. Kioussis, editors, *Electron Correlations and Materials Properties*, page 463, New York, 1999. Plenum.
83. N.E. Dahlen and U. von Barth. Variational second-order Møller-Plesset theory based on the Luttinger-Ward functional. *J. Chem. Phys.*, 120:6826, 2004.

84. T. Grabo and E.K.U. Gross. The optimized effective potential method of density functional theory: Applications to atomic and molecular systems. *Int. J. Quantum Chem.*, 64:95, 1997.
85. Y.-H. Kim, M. Städele, and R.M. Martin. Density-functional study of small molecules within the Krieger-Li-Iafrate approximation. *Phys. Rev. A*, 60:3633, 1999.
86. E. Engel, A. Höck, and R.M. Dreizler. Accuracy of the Krieger-Li-Iafrate approximation for molecules. *Phys. Rev. A*, 62:042502, 2000.
87. S. Ivanov, S. Hirata, and R.J. Bartlett. Exact exchange treatment for molecules in finite-basis-set Kohn-Sham theory. *Phys. Rev. Lett.*, 83:5455, 1999.
88. A. Görling. New KS method for molecules based on an exchange charge density generating the exact local KS exchange potential. *Phys. Rev. Lett.*, 83:5459, 1999.
89. S. Kümmel, L. Kronik, and J.P. Perdew. Electrical response of molecular chains from density functional theory. *Phys. Rev. Lett.*, 93:213002, 2004.
90. M. Fuchs, Y.-M. Niquet, X. Gonze, and K. Burke. Describing static correlation in bond dissociation by Kohn-Sham density functional theory. *J. Chem. Phys.*, 122:094116, 2005.
91. M. Grüning, O.V. Gritsenko, and E.J. Baerends. Exact exchange-correlation treatment of dissociating H_2 in density functional theory. *J. Chem. Phys.*, 118:7183, 2003.
92. T. Kotani. Exact exchange-potential band-structure calculations by the LMTO-ASA method: MgO and CaO. *Phys. Rev. B*, 50:14816, 1994.
93. A. Görling. Exact treatment of exchange in Kohn-Sham band-structure schemes. *Phys. Rev. B*, 53:7024, 1996.
94. M. Städele, J.A. Majewski, P. Vogl, and A. Görling. Exact Kohn-Sham exchange potential in semiconductors. *Phys. Rev. Lett.*, 79:2089, 1997.
95. M. Städele, M. Moukara, J.A. Majewski, P. Vogl, and A. Görling. Exact exchange Kohn-Sham formalism applied to semiconductors. *Phys. Rev. B*, 59:10031, 1999.
96. R. Magyar, A. Fleszar, and E.K.U. Gross. Exact-exchange density-functional calculations for noble-gas solids. *Phys. Rev. B*, 69:045111, 2004.
97. M. Grüning, A. Marini, and A. Rubio. Unpublished.

Many-Body Perturbation Theory: The GW Approximation

Christoph Friedrich and Arno Schindlmayr

Institute of Solid State Research
Forschungszentrum Jülich
52425 Jülich, Germany
E-mail: {c.friedrich, a.schindlmayr}@fz-juelich.de

In this lecture we present many-body perturbation theory as a method to determine quasiparticle excitations in solids, especially electronic band structures, accurately from first principles. The main ingredient is the electronic self-energy that, in principle, contains all many-body exchange and correlation effects beyond the Hartree potential. As its exact mathematical expression is unknown, approximations must be used in practical calculations. The GW approximation is obtained using a systematic algebraic approach on the basis of Green function techniques. It constitutes an expansion of the self-energy up to linear order in the screened Coulomb potential, which describes the interaction between the quasiparticles and includes dynamic screening through the creation of exchange-correlation holes around the bare particles. The implementation of the GW approximation relies on a perturbative treatment starting from density functional theory. Besides a detailed mathematical discussion we focus on the underlying physical concepts and show some illustrative applications.

1 Introduction

In the previous lectures we have seen that density functional theory (DFT) is the method of choice when we are interested in the ground-state properties of a many-electron system. DFT is based on the Hohenberg-Kohn theorem,¹ which states that there is (a) a one-to-one correspondence between the ground-state density $n_0(\mathbf{r})$ and the external potential as well as (b) a variational principle for the energy functional $E[n_0] \leq E[n]$. The second statement allows to obtain the ground state of a many-electron system by variation of its density, a quantity that is much less complicated than the many-electron wave function $\Psi_0(\mathbf{r}_1, \dots, \mathbf{r}_N)$, where N is the particle number. The first statement implies that the many-particle Hamiltonian is a functional of the ground-state density. Since the diagonalization of the Hamiltonian yields the complete excitation spectrum, the excited states can ultimately be regarded as functionals of the ground-state density as well. However, the Hohenberg-Kohn theorem does not provide us with an explicit mathematical form. In this lecture we show that excited-state properties can be accessed more directly with a purpose-built method, the so-called many-body perturbation theory.^{2,3} Incidentally, in practice its implementation within the GW approximation⁴ for the electronic self-energy is based on a perturbative evaluation with Kohn-Sham orbitals and can, therefore, finally be interpreted as the desired density functional.

The solution of the Kohn-Sham equation⁵ of DFT yields a whole spectrum of single-particle states, and one is tempted to identify the corresponding eigenvalues with excitation energies. Strictly speaking, such an interpretation is wrong: the Kohn-Sham wave functions and eigenvalues must be considered as mathematical tools and cannot be endowed with a physical meaning. The only exception is the energy of the highest occupied state,

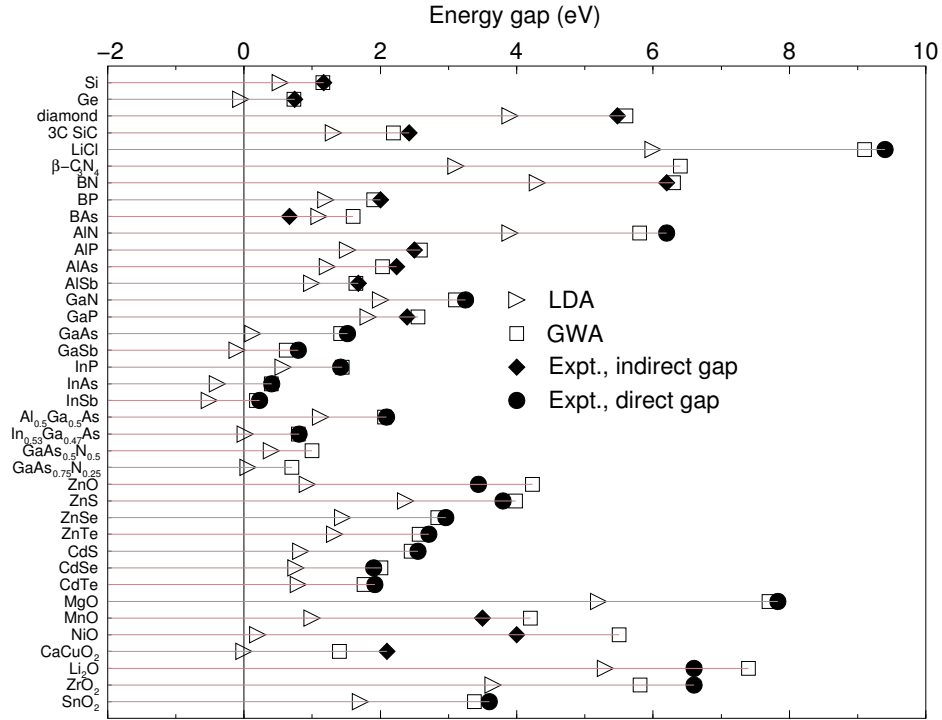


Figure 1. Comparison of LDA, GW and experimental band gaps for a variety of materials. Taken from Ref. 8.

which equals the exact ionization potential (or chemical potential for metals).^{6,7} Consequently, while often qualitatively correct, the DFT band structure fails to give reliable quantitative values for the band gaps of insulators and semiconductors, which are often underestimated by as much as 1.0 eV or more. In the case of Ge the local-density approximation (LDA) of DFT even predicts a semi-metal with a negative band gap rather than a semiconductor. In this lecture we demonstrate that the Kohn-Sham eigenvalues can be corrected using Green function techniques and the GW approximation for the electronic self-energy. Figure 1 shows a comparison of LDA and self-energy corrected band gaps with respective experimental values for a variety of materials. The underestimation within the LDA as well as the improvement by the GW approximation are evident.

Band gaps are experimentally measured by photoelectron spectroscopy. Figure 2 gives a schematic illustration. In direct photoelectron spectroscopy a photon with energy $\hbar\omega$ impinges on the sample and ejects an electron, whose kinetic energy E_{kin} is subsequently measured. The binding energy ϵ_i of this electron is given by the difference $\epsilon_i = E_{\text{kin}} - \hbar\omega$. Actually, we already simplified the argumentation here, as the formulation “binding energy of an electron” suggests that the electrons are independent. In reality they are correlated through the Coulomb interaction, and the ejection of an electron is always a many-body process. In this general sense ϵ_i equals the difference $\epsilon_i = E_0^N - E_i^{N-1}$ between the total energy E_0^N of the N -particle ground state Ψ_0^N and the energy E_i^{N-1} of the $(N-1)$ -particle state Ψ_i^{N-1} that remains after the emission. Inverse photoelectron spectroscopy is

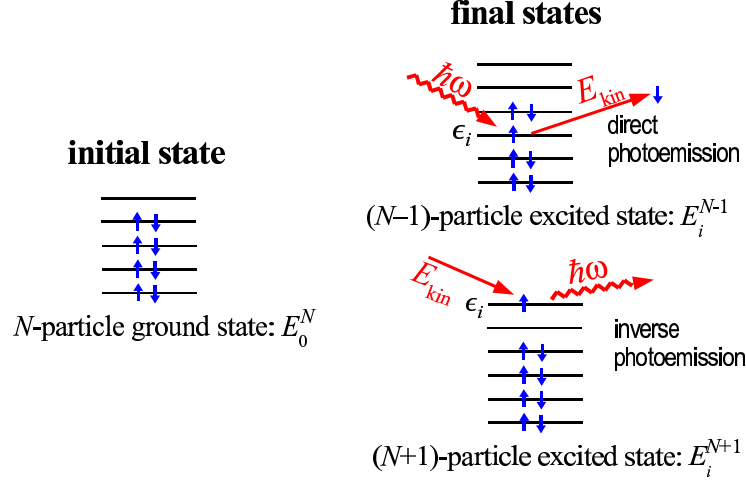


Figure 2. Schematic illustration of direct and inverse photoelectron spectroscopy. In both processes the particle number changes. The measured energy difference $E_{\text{kin}} - \hbar\omega$ corresponds to $\epsilon_i = E_0^N - E_i^{N-1}$ in direct and $\epsilon_i = E_i^{N+1} - E_0^N$ in inverse photoelectron spectroscopy.

the complementary process: electrons are injected into the sample, and the energy of the emitted photon is measured. The number of electrons in the system thus increases from N to $N + 1$, and we can identify $E_{\text{kin}} - \hbar\omega$ with the energy difference $\epsilon_i = E_i^{N+1} - E_0^N$ of the many-electron systems.

The fact that the independent-electron picture breaks down due to the strong Coulomb interaction questions single-electron concepts like band structure or Fermi surface. Still, in practice these work surprisingly well. In fact, we can at least retain a nearly-independent-particle picture if we consider quasiparticles instead of electrons (or holes). In the case of electron injection into a sample the repulsive Coulomb interaction creates a Coulomb hole around the additional electron (see Figure 3). Analogously, if an electron leaves the system, its Coulomb hole also disappears. Relative to the ground-state N -electron system, the addition (removal) of an electron in indirect (direct) photoelectron spectroscopy hence creates (annihilates) an ensemble consisting of the bare electron and its oppositely charged Coulomb hole. This ensemble behaves in many ways like a single-particle and is thus called “quasiparticle”. Since the Coulomb hole reduces the total charge of the quasiparticle, the effective interaction between quasiparticles is screened and considerably weaker than the bare Coulomb interaction between electrons. In fact, the screened interaction is sufficiently small so that the quasiparticles can be regarded as approximately independent, which finally justifies the independent-particle approximation and explains the success of mean-field theories.

A theoretical description of processes involving the ejection or injection of electrons requires a framework that links the N -particle with the $(N \pm 1)$ -particle systems. For this purpose we employ many-body perturbation theory. The central variable is the time-ordered Green function $G(\mathbf{r}t, \mathbf{r}'t')$. As we will see, it contains the excitation energies ϵ_i and even the excitation lifetimes. Besides, we can directly obtain the ground-state electron density, the expectation values of one-particle operators and the ground-state total energy

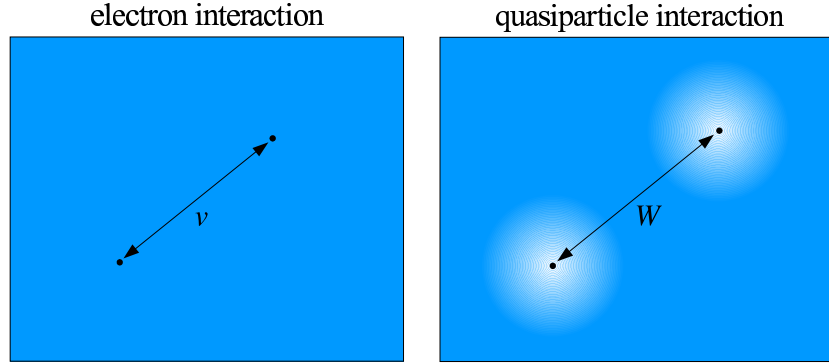


Figure 3. The electrons in a many-electron system are correlated by the strong Coulomb interaction v . The motion of one electron depends on the motion of all other electrons. A nearly-independent-particle picture can be recovered within the quasiparticle concept. Due to exchange and correlation effects a Coulomb hole forms around an electron and behaves together with it like a single entity, which is called quasiparticle. Quasiparticles interact via a weak screened interaction W instead of the strong Coulomb interaction.

from it. The Green function is hence capable of giving access to the same observables as the ground-state electron density. In contrast to the DFT expression $E[n]$, the functional $E[G]$ is even known exactly.² While the Green function contains much more information than the electron density, it is also a more complicated function and thus rarely applied to ground-state properties. In the present lecture we will, therefore, concentrate on the calculation of excited states.

Section 2 lays the theoretical foundations of the method. More complicated derivations are deferred to the appendix. The GW approximation is discussed in Section 3.1, and some aspects of its numerical implementation are given in Section 3.2. As an illustration, Section 3.3 presents a number of selected applications. Section 4 contains the summary.

2 Theory

2.1 Green Function

In this section we introduce the time-ordered Green function and examine its properties. We use the second-quantization formulation of quantum mechanics.^{2,3} For the present purpose it is sufficient to know that this formalism involves field operators $\hat{\psi}(\mathbf{r})$ and $\hat{\psi}^\dagger(\mathbf{r})$ that describe the annihilation and the creation of an electron at the position $\mathbf{r}$, respectively. We will not take spin dependence explicitly into account. If necessary, the spin quantum number can simply be added to the formulas by considering it to be part of the spatial coordinate $\mathbf{r}$.

The Green function $G^e(\mathbf{r}t, \mathbf{r}'t')$ is defined such that $i\hbar G^e(\mathbf{r}t, \mathbf{r}'t')$ is the probability amplitude for the propagation of an additional electron from $(\mathbf{r}', t')$ to $(\mathbf{r}, t)$ in a many-electron system with the Hamiltonian (Eq. 41)). This process brings the system from the N -electron ground state $|\Psi_0^N(t')\rangle$ to a final state $\hat{\psi}(\mathbf{r})U(t, t')\hat{\psi}^\dagger(\mathbf{r}')|\Psi_0^N(t')\rangle$. The final state is constructed by the successive action of the electron creation operator $\hat{\psi}^\dagger(\mathbf{r}')$, the evolution operator $\hat{U}(t, t') = \exp[-i\hat{H}(t - t')/\hbar]$, which takes the system from t' to a later

time $t > t'$, and the electron annihilation operator $\hat{\psi}(\mathbf{r})$ on the N -electron ground state. As the probability amplitude is given by the overlap of the final state with $|\Psi_0^N(t)\rangle$, the Green function becomes

$$\begin{aligned} G^e(\mathbf{r}t, \mathbf{r}'t') &= -\frac{i}{\hbar} \left\langle \Psi_0^N(t) \left| \hat{\psi}(\mathbf{r}) \hat{U}(t, t') \hat{\psi}^\dagger(\mathbf{r}') \right| \Psi_0^N(t') \right\rangle \theta(t - t') \\ &= -\frac{i}{\hbar} \left\langle \Psi_0^N \left| \hat{\psi}(\mathbf{r}) \hat{\psi}^\dagger(\mathbf{r}') \right| \Psi_0^N \right\rangle \theta(t - t'), \end{aligned} \quad (1)$$

where $\theta(t - t')$ is the Heaviside step function defined by

$$\theta(t - t') = \begin{cases} 1 & \text{if } t > t', \\ 0 & \text{if } t < t'. \end{cases} \quad (2)$$

For the last equality in Eq. (1) we changed from the Schrödinger to the Heisenberg picture, where the expression is particularly simple. States and operators in the two pictures are related by

$$|\Psi_H\rangle = \hat{U}(0, t) |\Psi_S(t)\rangle \quad \text{and} \quad \hat{A}_H(t) = \hat{U}(0, t) \hat{A}_S \hat{U}(t, 0). \quad (3)$$

In the following we always omit the indices S and H. Similarly, we have the Green function

$$G^h(\mathbf{r}'t', \mathbf{r}t) = -\frac{i}{\hbar} \left\langle \Psi_0^N \left| \hat{\psi}^\dagger(\mathbf{r}'t') \hat{\psi}(\mathbf{r}t) \right| \Psi_0^N \right\rangle \theta(t' - t) \quad (4)$$

for the propagation of an additional hole from $(\mathbf{r}, t)$ to $(\mathbf{r}', t')$. As a matter of convenience, we combine the two expressions in one time-ordered Green function

$$G(\mathbf{r}t, \mathbf{r}'t') = G^e(\mathbf{r}t, \mathbf{r}'t') - G^h(\mathbf{r}'t', \mathbf{r}t) = -\frac{i}{\hbar} \left\langle \Psi_0^N \left| \hat{T} \left[\hat{\psi}(\mathbf{r}t) \hat{\psi}^\dagger(\mathbf{r}'t') \right] \right| \Psi_0^N \right\rangle, \quad (5)$$

where we used the time-ordering operator $\hat{T}$, which rearranges a series of field operators in order of ascending time arguments from right to left with a factor (-1) for each pair permutation. Depending on the time order, Eq. (5) describes either electron ($t > t'$) or hole ($t < t'$) propagation. The electron density $n(\mathbf{r})$ can be expressed in terms of the Green function as

$$n(\mathbf{r}t) = \left\langle \Psi_0^N \left| \hat{\psi}^\dagger(\mathbf{r}t) \hat{\psi}(\mathbf{r}t) \right| \Psi_0^N \right\rangle = -i\hbar G(\mathbf{r}t, \mathbf{r}t + \eta). \quad (6)$$

Here and in the following η is an infinitesimal positive number. It serves only to enforce the correct order of the field operators. Its unit should always be clear from the context; presently it is an infinitesimal time.

Let us consider the time-ordered Green function $G(\mathbf{r}, \mathbf{r}'; \tau)$ of a stationary system with $\tau = t - t'$. If we insert the closure relation $\sum_i |\Psi_i^{N\pm 1}\rangle \langle \Psi_i^{N\pm 1}| = 1$ between the two field operators in Eq. (5), where $\{|\Psi_i^{N\pm 1}\rangle\}$ is the complete set of state vectors of the $(N \pm 1)$ -particle system, transform to the Schrödinger picture and use the definitions

$$\psi_i^{N-1}(\mathbf{r}) = \left\langle \Psi_i^{N-1} \left| \hat{\psi}(\mathbf{r}) \right| \Psi_0^N \right\rangle \quad \text{and} \quad \psi_i^{N+1}(\mathbf{r}) = \left\langle \Psi_0^N \left| \hat{\psi}(\mathbf{r}) \right| \Psi_i^{N+1} \right\rangle \quad (7)$$

together with the excitation energies

$$\epsilon_i^{N-1} = E_0^N - E_i^{N-1} \quad \text{and} \quad \epsilon_i^{N+1} = E_i^{N+1} - E_0^N, \quad (8)$$

then we obtain

$$G(\mathbf{r}, \mathbf{r}'; \tau) = -\frac{i}{\hbar} \sum_i \psi_i^{N+1}(\mathbf{r}) \psi_i^{N+1*}(\mathbf{r}') e^{-i\epsilon_i^{N+1}\tau/\hbar} \theta(\tau) + \frac{i}{\hbar} \sum_i \psi_i^{N-1}(\mathbf{r}) \psi_i^{N-1*}(\mathbf{r}') e^{-i\epsilon_i^{N-1}\tau/\hbar} \theta(-\tau). \quad (9)$$

The sums run over the ground state and all excited states of the $(N-1)$ - and $(N+1)$ -particle system, respectively. Expression (Eq. (9)) can be interpreted as follows: The state after the addition of an electron ($\tau > 0$) is represented by a linear combination of excited states

$$\hat{\psi}^\dagger(\mathbf{r}') |\Psi_0^N\rangle = \sum_i \psi_i^{N+1*}(\mathbf{r}') |\Psi_i^{N+1}\rangle \quad (10)$$

that subsequently evolve according to their respective phase factors $\exp(-i\epsilon_i^{N+1}\tau/\hbar)$. The resulting state is then probed at the point $\mathbf{r}$ by the projections $\psi_i^{N+1}(\mathbf{r})$. The case $\tau < 0$ (hole propagation) is analogous. Consequently, the Green function indeed contains the complete excitation spectrum of the $(N \pm 1)$ -particle system. Fourier transformation of Eq. (9) to the frequency axis using the Fourier transform of the Heaviside step function

$$\theta(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \theta(\tau) e^{i\omega\tau - \eta|\tau|} d\tau = \frac{i}{2\pi(\omega + i\eta)} \quad (11)$$

finally yields the Lehmann representation of the Green function

$$G(\mathbf{r}', \mathbf{r}; \omega) = \sum_i \frac{\psi_i^{N+1}(\mathbf{r}) \psi_i^{N+1*}(\mathbf{r}')}{\hbar\omega - \epsilon_i^{N+1} + i\eta} + \sum_i \frac{\psi_i^{N-1}(\mathbf{r}) \psi_i^{N-1*}(\mathbf{r}')}{\hbar\omega - \epsilon_i^{N-1} - i\eta}. \quad (12)$$

We observe that the Green function has poles at the true many-particle excitation energies $\epsilon_i^{N\pm 1}$. These energies correspond to excitations of an $(N-1)$ -particle and an $(N+1)$ -particle system and hence to those processes measured in direct and inverse photoelectron spectroscopy. In the case of a noninteracting (or mean-field) system the $\psi_i^{N+1}(\mathbf{r})$ are simply the unoccupied and the $\psi_i^{N-1}(\mathbf{r})$ the occupied single-particle wave functions, the $\epsilon_i^{N\pm 1}$ are the corresponding single-particle energies. In order not to overload the notation, we will drop the $(N \pm 1)$ superscripts from now on.

2.2 Spectral Function

In connection with Eq. (9) we can define the spectral function $A(\mathbf{r}, \mathbf{r}'; \omega)$, i.e., the density of the excited (or quasiparticle) states that contribute to the electron or hole propagation. In a finite system this density is a series of delta functions at the excitation energies

$$A(\mathbf{r}, \mathbf{r}'; \omega) = \sum_i \psi_i(\mathbf{r}) \psi_i^*(\mathbf{r}') \delta(\hbar\omega - \epsilon_i) \quad (13)$$

weighted by the products of the corresponding projections (Eq. (7)). This allows us to rewrite Eq. (12) as an integral over frequencies

$$G(\mathbf{r}, \mathbf{r}'; \omega) = \hbar \int_{-\infty}^{\infty} \frac{A(\mathbf{r}, \mathbf{r}'; \omega')}{\hbar\omega - \hbar\omega' + \text{sgn}(\hbar\omega' - \mu) i\eta} d\omega' \quad (14)$$

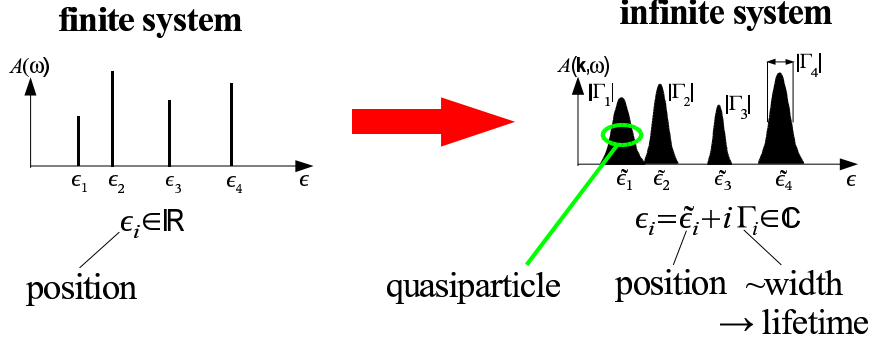


Figure 4. The excitation peaks of a finite system in the spectral function $A(\omega)$ merge into quasiparticle peaks of finite width in the case of an infinite system. This gives rise to finite excitation lifetimes determined by the inverse of the peak widths.

with $\max(\epsilon_i^{N-1}) \leq \mu \leq \min(\epsilon_i^{N+1})$. In an infinite system μ corresponds to the chemical potential. The inequality $\max(\epsilon_i^{N-1}) \leq \min(\epsilon_i^{N+1})$ follows from the convexity of the total energy as a function of the particle number, i.e., $E_0^{N-1} - E_0^N \geq E_0^N - E_0^{N+1}$: we lose more energy when removing an electron than we gain by adding one. With the identity

$$\frac{1}{x \mp i\eta} = \mathcal{P}\left(\frac{1}{x}\right) \pm i\pi\delta(x) \quad (15)$$

in the limit $\eta \rightarrow 0^+$, where $\mathcal{P}(1/x)$ is the principal value of $1/x$, we find that

$$A(\mathbf{r}, \mathbf{r}'; \omega) = -\text{sgn}(\hbar\omega - \mu) \frac{1}{\pi} \text{Im} G(\mathbf{r}, \mathbf{r}'; \omega). \quad (16)$$

The closure relation of the functions in Eq. (7) yields another important property

$$\int_{-\infty}^{\infty} A(\mathbf{r}, \mathbf{r}'; \omega) d\omega = \sum_i \psi_i(\mathbf{r}) \psi_i^*(\mathbf{r}') = \delta(\mathbf{r} - \mathbf{r}'). \quad (17)$$

When we change from a finite to an infinite system, the delta functions in $A(\mathbf{r}, \mathbf{r}'; \omega)$ merge and form a series of smooth peaks with finite line widths instead of sharp resonances (see Figure 4). However, if the resulting spectral features are of Lorentzian form, i.e.,

$$A(\mathbf{r}, \mathbf{r}'; \omega) = \sum_i \psi_i(\mathbf{r}) \psi_i^*(\mathbf{r}') \frac{\Gamma_i}{(\hbar\omega - \tilde{\epsilon}_i)^2 + \Gamma_i^2}, \quad (18)$$

where the $\tilde{\epsilon}_i$ are the peak positions and $|\Gamma_i|$ the corresponding peak widths, then we can perform the integration in Eq. (14) analytically and again obtain a discrete sum over i as in Eq. (12), provided that the energies are defined as complex numbers $\epsilon_i = \tilde{\epsilon}_i + i\Gamma_i$. Consequently, the form of the Fourier transform (Eq. (9)) remains unchanged, too. The imaginary component of ϵ_i leads to a damping term $\exp(-|\Gamma_i|\tau/\hbar)$, revealing that the excitation has a finite lifetime of $\hbar/|\Gamma_i|$. Physically, the de-excitation proceeds via Auger transitions that create electron-hole pairs on the way. The damping of the particle propagator G may seem surprising, as it suggests that the particle gradually disappears. However, one must keep in mind that we deal with an infinite system, i.e., $N \rightarrow \infty$, and an additional particle

(electron or hole) can “dissipate” into the Fermi sea. In this sense, one often speaks of finite quasiparticle lifetimes and calls $\psi_i(\mathbf{r})$ and ϵ_i the quasiparticle wave functions and energies, respectively. The quasiparticle equation (22) introduced in the next section holds for infinite systems if one uses an analytic continuation of the self-energy into the complex frequency plane.

2.3 Dyson Equation

Appendix A shows that the time-ordered Green function $G(\mathbf{r}, \mathbf{r}'; \omega)$ of the interacting system obeys an integral equation, the Dyson equation

$$G(\mathbf{r}, \mathbf{r}'; \omega) = G_0(\mathbf{r}, \mathbf{r}'; \omega) + \iint G_0(\mathbf{r}, \mathbf{r}''; \omega) \Sigma(\mathbf{r}'', \mathbf{r}'''; \omega) G(\mathbf{r}''', \mathbf{r}'; \omega) d^3 r'' d^3 r''', \quad (19)$$

where $G_0(\mathbf{r}, \mathbf{r}'; \omega)$ is the Green function of a mean-field system defined by

$$\hat{h}_0 \varphi_i^0(\mathbf{r}) = \epsilon_i^0 \varphi_i^0(\mathbf{r}) \quad (20)$$

with the single-particle Hamiltonian

$$\hat{h}_0(\mathbf{r}) = -\frac{\hbar^2}{2m} \nabla^2 + V_{\text{ext}}(\mathbf{r}) + \frac{e^2}{4\pi\epsilon_0} \int \frac{n(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d^3 r'. \quad (21)$$

The quantities $V_{\text{ext}}(\mathbf{r})$, m , e and ϵ_0 are defined as in Eq. (41). The Green function $G_0(\mathbf{r}, \mathbf{r}'; \omega)$ is obtained from Eq. (12) with the wave functions $\varphi_i^0(\mathbf{r})$ and energies ϵ_i^0 . The nonlocal and frequency-dependent function $\Sigma(\mathbf{r}, \mathbf{r}', \omega)$ is the non-Hermitian self-energy operator, which contains all many-body exchange and correlation effects beyond the electrostatic Hartree potential. This can be more easily seen in a reformulation of the Dyson equation. By inserting the Lehmann representation (Eq. (12)) into Eq. (19), we find that the wave functions $\psi_i(\mathbf{r})$ and energies ϵ_i obey the quasiparticle equation

$$\hat{h}_0(\mathbf{r}) \psi_i(\mathbf{r}) + \int \Sigma(\mathbf{r}, \mathbf{r}'; \epsilon_i/\hbar) \psi_i(\mathbf{r}') d^3 r' = \epsilon_i \psi_i(\mathbf{r}) \quad (22)$$

(see Appendix B), which is nonlinear in ϵ_i . Although it looks very similar to the one-particle equations of mean-field approaches like Hartree, Hartree-Fock or DFT, it does *not* constitute a mean-field formulation, since the self-energy takes all dynamic many-electron processes into account. Consequently, the functions $\psi_i(\mathbf{r})$ and energies ϵ_i must not be understood as single-particle quantities. In fact, they are defined in Eqs. (7) and (8) as properties of the many-electron system. From the nonlinearity of the quasiparticle equation it follows that the wave functions $\psi_i(\mathbf{r})$ are not orthonormal, in contrast to single-particle wave functions. However, they do fulfill the closure relation (Eq. (17)).

The Dyson equation (19) can be rewritten in the form of a geometric series by subsequently replacing G on the right-hand side by $G_0 + G_0 \Sigma G$, which leads to, symbolically written,

$$G = G_0 + G_0 \Sigma G_0 + G_0 \Sigma G_0 \Sigma G_0 + G_0 \Sigma G_0 \Sigma G_0 \Sigma G_0 + \dots \quad (23)$$

This is a typical equation of scattering theory, where the different terms of the geometric series describe single, double, triple, etc., scattering processes, and Σ is the scattering potential. Such a succession of scattering processes can be illustrated by Feynman diagrams,

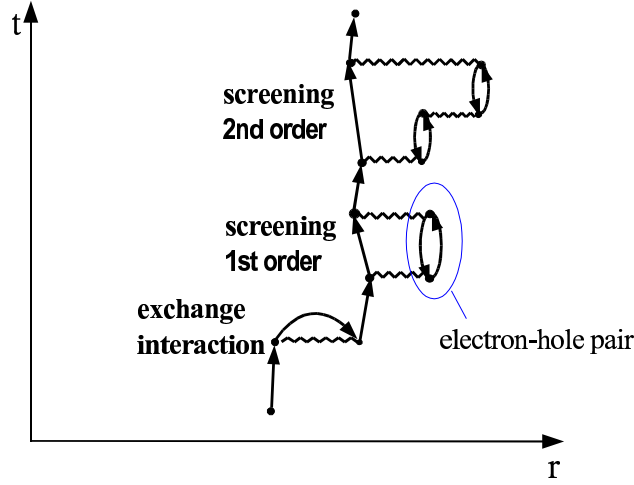


Figure 5. Illustration of a series of scattering processes using Feynman diagrams. All zigzag lines representing the instantaneous Coulomb interaction must be drawn horizontally. Arrows going forward in time represent electron and those going backward in time hole propagators. The self-energy is the sum of all possible single scattering processes.

where G_0 is drawn as a straight arrow and the Coulomb interaction as a zigzag line. According to Eq. (23), a diagrammatic representation of a multiple scattering process should involve a series of arrows (G_0) divided by single scattering processes (Σ). In the example of Figure 5 these are the exchange interaction, the creation of an electron-hole pair (the “bubble” diagram) and finally the creation of a pair that itself creates another pair. In order to obtain the complete Green function, we have to sum all multiple scattering processes, of which the one shown in Figure 5 is merely one example. The self-energy is given by the sum of all single scattering processes. The interpretation in terms of scattering processes allows to construct approximations for Σ by the summation of diagrams considered essential for the physical behavior of a given electron system. In general, however, such approximations are rarely convergent, and too many processes turn out to be quantitatively important. Therefore, we apply a systematic algebraic method instead.

3 Implementation and Applications

3.1 GW Approximation

In practice we must use an approximation for the self-energy, such as the GW approximation, which contains the electron exchange and a large part of the electron correlation. It is formally derived in Appendix A and has a very simple mathematical form in the time domain

$$\Sigma^{GW}(\mathbf{r}, \mathbf{r}'; \tau) = i\hbar G_0(\mathbf{r}, \mathbf{r}'; \tau) W(\mathbf{r}, \mathbf{r}'; \tau + \eta). \quad (24)$$

In order to calculate the self-energy contribution to the quasiparticle energies, we need the Fourier transform on the frequency axis

$$\Sigma^{GW}(\mathbf{r}, \mathbf{r}'; \omega) = \frac{i\hbar}{2\pi} \int_{-\infty}^{\infty} G_0(\mathbf{r}, \mathbf{r}'; \omega + \omega') W(\mathbf{r}, \mathbf{r}'; \omega') e^{i\omega' \eta} d\omega'. \quad (25)$$

The first function on the right-hand side is the Green function of the noninteracting system defined by Eq. (20) and the second function the dynamically screened interaction $W(\mathbf{r}, \mathbf{r}'; \omega)$, which is related to the bare Coulomb potential $v(\mathbf{r}, \mathbf{r}') = e^2 / (4\pi\epsilon_0 |\mathbf{r} - \mathbf{r}'|)$ through the inverse of the dielectric function

$$W(\mathbf{r}, \mathbf{r}'; \omega) = \int \epsilon^{-1}(\mathbf{r}, \mathbf{r}''; \omega) v(\mathbf{r}'', \mathbf{r}') d^3 r'' = v(\mathbf{r}, \mathbf{r}') + \int n_{\text{ind}}(\mathbf{r}, \mathbf{r}''; \omega) v(\mathbf{r}'', \mathbf{r}') d^3 r''. \quad (26)$$

The screened interaction $W(\mathbf{r}, \mathbf{r}'; \omega)$ is the effective potential at $\mathbf{r}'$ induced by a quasiparticle at $\mathbf{r}$: the Coulomb potential of the electron repels other electrons in its neighborhood and thus gives rise to the formation of an exchange and correlation hole, whose effective positive charge $n_{\text{ind}}(\mathbf{r}, \mathbf{r}''; \omega)$ screens the bare Coulomb potential $v(\mathbf{r}, \mathbf{r}')$ (see Figure 6). Analogously, an effective negative charge screens the Coulomb potential of a hole. The screened interaction is considerably weaker than the bare Coulomb interaction. The GW approximation uses the random-phase approximation (RPA)

$$\epsilon(\mathbf{r}, \mathbf{r}'; \omega) = \delta(\mathbf{r} - \mathbf{r}') - \int v(\mathbf{r}, \mathbf{r}'') P(\mathbf{r}'', \mathbf{r}'; \omega) d^3 r'', \quad (27)$$

$$P(\mathbf{r}, \mathbf{r}'; \tau) = -i\hbar G_0(\mathbf{r}, \mathbf{r}'; \tau) G_0(\mathbf{r}', \mathbf{r}; -\tau). \quad (28)$$

It corresponds to a subset of scattering processes in the many-electron system. Some of the respective diagrams are just the ones shown in Figure 5. Using expression in Eq. (12) for the Green function G_0 of the noninteracting system we observe that the Fourier transform of the polarization function in Eq. (28) is given by

$$P(\mathbf{r}, \mathbf{r}'; \omega) = \sum_i^{\text{occ.}} \sum_j^{\text{unocc.}} \varphi_i^0(\mathbf{r}) \varphi_j^{0*}(\mathbf{r}) \varphi_i^{0*}(\mathbf{r}') \varphi_j^0(\mathbf{r}') \times \left(\frac{1}{\omega + \epsilon_i^0 - \epsilon_j^0 + i\eta} - \frac{1}{\omega - \epsilon_i^0 + \epsilon_j^0 - i\eta} \right) \quad (29)$$

in terms of the wave functions $\varphi_i^0(\mathbf{r})$ and energies ϵ_i^0 .

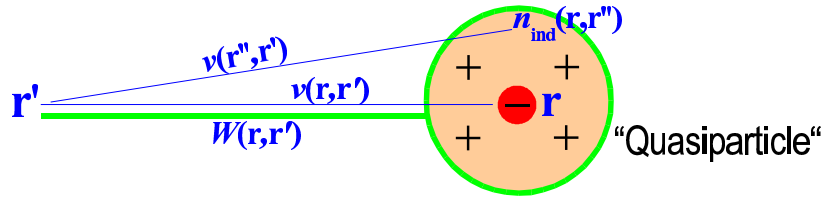


Figure 6. The formation of the Coulomb hole around an electron at $\mathbf{r}$ screens its Coulomb potential $v(\mathbf{r}, \mathbf{r}')$. This leads to the definition of the screened interaction $W(\mathbf{r}, \mathbf{r}')$ that takes into account the combined potentials of the bare electron and its screening cloud n_{ind} . The ensemble consisting of the electron and its polarization cloud is called "quasiparticle".

The well-known Hartree-Fock equations can be recovered from Eq. (22) if we use the energy-independent self-energy

$$\Sigma^{\text{HF}}(\mathbf{r}, \mathbf{r}') = i\hbar G_0(\mathbf{r}, \mathbf{r}'; -\eta)v(\mathbf{r}, \mathbf{r}') \quad (30)$$

(given in the time domain) instead. By comparison with Eq. (24), we see that the GW approximation constitutes an expansion of the self-energy up to first order in the screened interaction as opposed to the bare Coulomb interaction in Eq. (30). This approximates the exact self-energy considerably better, because W is much smaller than v . Due to the similarity of the two self-energy expressions, the GW approximation can formally be regarded as a Hartree-Fock approach with a dynamically screened interaction W instead of the static Coulomb interaction v .

3.2 Numerical Implementation

For band structure calculations it is more efficient to obtain the ϵ_i directly from the quasiparticle equation (22) instead of solving the Dyson integral equation (19) and searching for the poles of the Green function. Furthermore, it is then possible to exploit the formal similarity to the Kohn-Sham equation

$$\hat{h}_0\varphi_i^{\text{KS}}(\mathbf{r}) + V_{\text{xc}}(\mathbf{r})\varphi_i^{\text{KS}}(\mathbf{r}) = \epsilon_i^{\text{KS}}\varphi_i^{\text{KS}}(\mathbf{r}), \quad (31)$$

where $V_{\text{xc}}(\mathbf{r})$ is the local exchange-correlation potential. In many cases the Kohn-Sham eigenvalues ϵ_i^{KS} already provide a reasonable estimate of the band structure and are in qualitative agreement with experiment. For systems where the quasiparticle wave functions are known, one also finds $\varphi_i^{\text{KS}}(\mathbf{r}) \approx \psi_i(\mathbf{r})$.⁹ This observation indicates that the self-energy correction $\Sigma(\mathbf{r}, \mathbf{r}'; \epsilon_i/\hbar) - V_{\text{xc}}(\mathbf{r})\delta(\mathbf{r} - \mathbf{r}')$ is small and justifies the use of first-order perturbation theory to obtain approximate energies

$$\epsilon_i \approx \epsilon_i^{\text{KS}} + \langle \varphi_i^{\text{KS}} | \Sigma(\epsilon_i/\hbar) - V_{\text{xc}} | \varphi_i^{\text{KS}} \rangle. \quad (32)$$

A solution of this nonlinear equation still requires the knowledge of the frequency dependence of the self-energy, which is not known in general. Therefore, we use the linear expansion

$$\Sigma(\mathbf{r}, \mathbf{r}'; \epsilon_i/\hbar) \approx \Sigma(\mathbf{r}, \mathbf{r}'; \epsilon_i^{\text{KS}}/\hbar) + \frac{\epsilon_i - \epsilon_i^{\text{KS}}}{\hbar} \frac{\partial \Sigma(\mathbf{r}, \mathbf{r}'; \epsilon_i^{\text{KS}}/\hbar)}{\partial \omega}, \quad (33)$$

which leads to

$$\epsilon_i \approx \epsilon_i^{\text{KS}} + Z_i \langle \varphi_i^{\text{KS}} | \Sigma(\epsilon_i^{\text{KS}}/\hbar) - V_{\text{xc}} | \varphi_i^{\text{KS}} \rangle. \quad (34)$$

The quasiparticle renormalization factor is given by

$$Z_i = \left(1 - \left\langle \varphi_i^{\text{KS}} \left| \frac{1}{\hbar} \frac{\partial \Sigma(\epsilon_i^{\text{KS}}/\hbar)}{\partial \omega} \right| \varphi_i^{\text{KS}} \right\rangle \right)^{-1} \quad (35)$$

and equals the quasiparticle weight

$$Z_i = \int |\psi_i(\mathbf{r})|^2 d^3r < 1. \quad (36)$$

With the decomposition of W into the bare Coulomb interaction v and the remainder $W - v$, the GW self-energy (Eq. (24)) splits into exchange and correlation parts, symbolically written as

$$\Sigma^{GW} = i\hbar G_0^{\text{KS}} W = i\hbar G_0^{\text{KS}} v + i\hbar G_0^{\text{KS}} (W - v) = \Sigma_x^{GW} + \Sigma_c^{GW}. \quad (37)$$

Instead of G_0 we use the Kohn-Sham Green function G_0^{KS} . After inserting this decomposition into Eq. (25), we must evaluate the convolutions

$$\Sigma_x^{GW}(\mathbf{r}, \mathbf{r}'; \omega) = \frac{i\hbar}{2\pi} \int_{-\infty}^{\infty} G_0^{\text{KS}}(\mathbf{r}, \mathbf{r}'; \omega + \omega') v(\mathbf{r}, \mathbf{r}') e^{i\omega' \eta} d\omega', \quad (38a)$$

$$\Sigma_c^{GW}(\mathbf{r}, \mathbf{r}'; \omega) = \frac{i\hbar}{2\pi} \int_{-\infty}^{\infty} G_0^{\text{KS}}(\mathbf{r}, \mathbf{r}'; \omega + \omega') [W(\mathbf{r}, \mathbf{r}'; \omega') - v(\mathbf{r}, \mathbf{r}')] d\omega'. \quad (38b)$$

The integral (Eq. (38a)) can be evaluated analytically and leads to the well-known expression for the Hartree-Fock exchange term

$$\langle \varphi_i^{\text{KS}} | \Sigma_x^{GW} | \varphi_i^{\text{KS}} \rangle = -\frac{e^2}{4\pi\epsilon_0} \sum_j^{\text{occ.}} \int \frac{\varphi_i^{\text{KS}*}(\mathbf{r}) \varphi_j^{\text{KS}}(\mathbf{r}) \varphi_j^{\text{KS}*}(\mathbf{r}') \varphi_i^{\text{KS}}(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d^3r d^3r'. \quad (39)$$

In general, the second convolution (Eq. (38b)) must be computed numerically. For this purpose the integration contour is usually deformed to the complex plane, where the analytical continuations of G_0 and W are smoother.

Let the Kohn-Sham wave functions be represented in a basis $\{\zeta_\alpha(\mathbf{r})\}$. According to Eq. (29) we can then write the polarization function and all related quantities in terms of products $\chi_\mu(\mathbf{r}) = \zeta_\alpha^*(\mathbf{r})\zeta_\beta(\mathbf{r})$ with the composite index $\mu = (\alpha, \beta)$ as

$$P(\mathbf{r}, \mathbf{r}'; \omega) = \sum_{\mu, \nu} P_{\mu\nu}(\omega) \chi_\mu^*(\mathbf{r}) \chi_\nu(\mathbf{r}'). \quad (40)$$

The Eqs. (24) to (28) are solved by matrix operations:

1. A self-consistent DFT loop produces the Kohn-Sham wave functions $\varphi_i^{\text{KS}}(\mathbf{r})$ and energies ϵ_i^{KS} . At this point we can already evaluate the exchange term (Eq. (39)).
2. The polarization matrix $P_{\mu\nu}(\omega)$ is calculated according to Eq. (29).
3. The dielectric matrix is obtained from $\epsilon_{\mu\nu}(\omega) = \delta_{\mu\nu} - \sum_\gamma v_{\mu\gamma} P_{\gamma\nu}(\omega)$ and inverted.
4. Next the screened interaction $W_{\mu\nu}(\omega) = \sum_\gamma \epsilon_{\mu\gamma}^{-1}(\omega) v_{\gamma\nu}$ is calculated from a matrix multiplication of the inverse dielectric function with the Coulomb matrix.
5. The correlation term $\langle \varphi_i^{\text{KS}} | \Sigma_c^{GW} | \varphi_i^{\text{KS}} \rangle$ is evaluated according to Eq. (38b) with a numerical contour integration on the complex frequency plane.
6. Finally, approximate quasiparticle energies are obtained from Eqs. (34) and (35).

The computation of the dielectric function, its inversion and the convolution (Eq. (38b)) are very time-consuming. Therefore, some (especially older) codes approximate the inverse dielectric function by a so-called plasmon-pole model.^{10,11} These models replace the imaginary component of $\epsilon^{-1}(\omega)$, which has a peaked structure, by a sum of delta functions

at the corresponding frequencies. This simplification reduces the third step to a single matrix inversion of the static dielectric function at $\omega = 0$ and makes an analytic evaluation of the frequency integral (Eq. (38b)) possible.

3.3 Examples

Although Hedin's seminal article⁴ was already published in 1965, it was not before the middle of the 1980s that the first ab initio calculations for real materials were reported in the literature. In spite of several approximations in the numerical treatment, which were necessary because of the lack of computer power back then, initial results were already very promising. Hybertsen and Louie¹² as well as Godby *et al.*¹³ showed that the calculated band gap of Si fell within a margin of about 0.1 eV from the experimental value. Shortly afterwards the same authors reported band gaps for several other semiconducting materials that turned out to be equally accurate.^{14,15} After these pioneering studies the *GW* approximation was applied to a variety of semiconductors with great success (see, e.g., Figure 1). The principal effect of the *GW* self-energy correction on the band structure of a semiconductor is to rigidly shift the valence bands up and the conduction bands down, thus opening the band gap. Figure 7 shows this effect for Si as an example.

Not only the band gaps of semiconductors and insulators are improved by the *GW* self-energy correction, but the correlation-induced band narrowing of metals is also correctly described. The band narrowing reflects the higher effective mass of quasiparticles (the polarization cloud adds to the electron mass) compared to bare electrons. For this reason, the self-energy is sometimes also referred to as “mass operator”. Figure 8 shows the energy

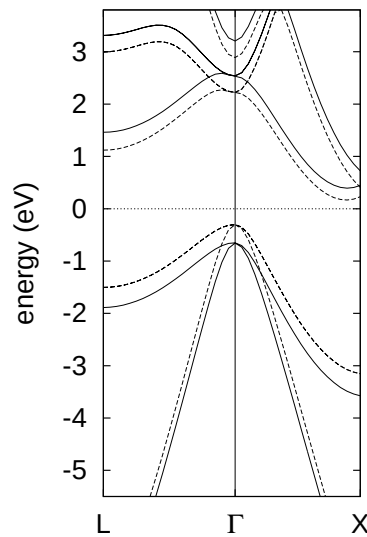


Figure 7. LDA band structure (dashed lines) of silicon with *GW* self-energy corrected valence and conduction bands (solid lines). The *GW* approximation shifts the corresponding bands up and down, respectively, but leaves the dispersion essentially unaffected.

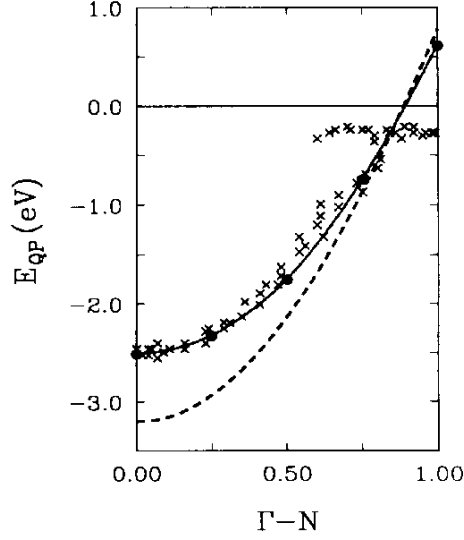


Figure 8. Comparison of LDA (dashed), quasiparticle (solid line) and experimental (crosses) bands for Na. Taken from Ref. 16.

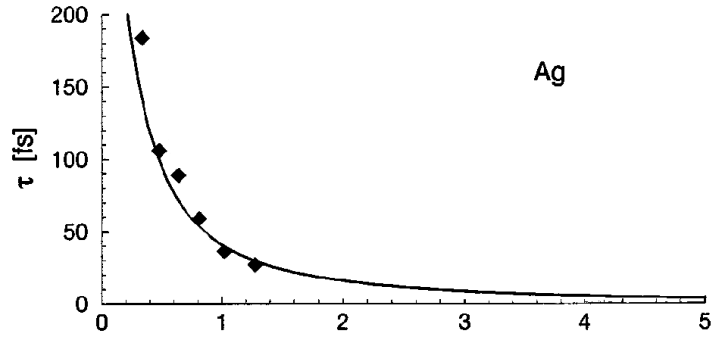


Figure 9. De-excitation dynamics measured in time-resolved two-photon photoemission spectroscopy (diamonds) and calculated with the GW approximation (solid line). Taken from Ref. 17.

dispersion of Na as an example.¹⁶ The band narrowing brought about by the GW self-energy correction leads to nearly perfect agreement with experiment.

The calculated excitation or quasiparticle lifetimes can be directly compared with two-photon photoemission spectroscopy. This experimental method allows to measure dynamical de-excitation processes in electronic systems. After a first photon has excited the electron system (creating a “hot” electron), a second photon probes the quasiparticle density of states like in ordinary direct photoelectron spectroscopy. The time delay between the two photons can be tuned such that the system can be observed in different stages of the electronic de-excitation process. From a series of measurements one can thus deduce the

lifetime, which depends on the excitation energy, i.e., the energy of the first photon. In the example of Ag in Figure 9 the theoretical curve $\tau = \hbar |\Gamma|^{-1}$ obtained from the imaginary parts of the quasiparticle energies (see Section 2.2) closely follows the experimental data points.¹⁷

4 Summary

In this lecture we presented the *GW* approximation for the electronic self-energy, which allows to calculate excited-state properties like excitation energies and lifetimes. The self-energy describes scattering processes between electrons and, in principle, contains all exchange and correlation effects beyond the electrostatic Hartree potential. The *GW* approximation includes a subset of these scattering processes. Apart from exchange it describes the creation of electron-hole pairs within the random-phase approximation (RPA) that leads to the formation of polarization clouds around the bare particles. The ensemble of an electron or a hole together with its polarization cloud behaves essentially like a single entity and is called a quasiparticle. The quasiparticles interact via a screened potential that is considerably weaker than the bare Coulomb interaction. This makes a perturbative treatment possible. In this respect, the *GW* approximation constitutes an expansion of the self-energy up to linear order in the screened interaction. It works well in a large class of systems where the polarization effects covered by the RPA play the dominant role in electron correlation, such as simple metals and semiconductors.

The *GW* approximation is by nature a perturbative approach. Actual *GW* calculations are usually based on the self-consistent Kohn-Sham wave functions and energies as a starting point. This method has its limitations in materials where DFT already gives unphysical results. It breaks down for systems with very strong electronic correlation, which is insufficiently described by the available exchange-correlation functionals. The large error in the band gap of NiO in Figure 1 is an example. In reality, NiO is a strongly correlated Mott-Hubbard insulator, whereas it comes out as a semiconductor with a very small band gap (nearly a semi-metal) in DFT calculations.

The *GW* method is designed for the analysis of excited states of the $(N \pm 1)$ -electron systems. The treatment of optical absorption processes, where the particle number does not change due to the promotion of valence electrons into unoccupied conduction states rather than emission, requires the simultaneous description of two particles, an electron and a hole, i.e., an exciton. Consequently, one must describe such a process with a two-particle Green function. In this case many-body perturbation theory leads to the so-called Bethe-Salpeter equation. Absorption spectra obtained from this equation are indeed very accurate.¹⁸ An alternative is time-dependent density functional theory,¹⁹ which also gives access to the excited states of an N -electron system.

Appendix A

Hedin equations

With the field-operators introduced in Section 2.1 we can rewrite the many-particle Hamiltonian

$$\hat{H} = \sum_i \left[-\frac{\hbar^2}{2m} \nabla_i^2 + V_{\text{ext}}(\mathbf{r}_i) \right] + \frac{1}{2} \sum_{ij} v(\mathbf{r}_i, \mathbf{r}_j), \quad (41)$$

where $V_{\text{ext}}(\mathbf{r})$ is the potential created by the atomic nuclei, $v(\mathbf{r}, \mathbf{r}') = e^2 / (4\pi\epsilon_0 |\mathbf{r} - \mathbf{r}'|)$ the Coulomb interaction, m the electron mass, e the electron charge and ϵ_0 the vacuum dielectric constant, as

$$\hat{H} = \int \hat{\psi}^\dagger(\mathbf{r}) \hat{h}(\mathbf{r}) \hat{\psi}(\mathbf{r}) d^3r + \frac{1}{2} \iint \hat{\psi}^\dagger(\mathbf{r}) \hat{\psi}^\dagger(\mathbf{r}') v(\mathbf{r}, \mathbf{r}') \hat{\psi}(\mathbf{r}') \hat{\psi}(\mathbf{r}) d^3r d^3r' \quad (42)$$

with the one-particle operator

$$\hat{h}(\mathbf{r}) = -\frac{\hbar^2}{2m} \nabla^2 + V_{\text{ext}}(\mathbf{r}). \quad (43)$$

The expression in Eq. (42) is just a mathematical reformulation of Eq. (41) and should not be mistaken for the energy expectation value in Hartree theory, although it looks similar.

From the equation of motion for the annihilation operator

$$i\hbar \frac{\partial}{\partial t} \hat{\psi}(\mathbf{r}, t) = [\hat{\psi}(\mathbf{r}, t), \hat{H}]_- = \hat{h}(\mathbf{r}) \hat{\psi}(\mathbf{r}, t) + \int v(\mathbf{r}, \mathbf{r}') \hat{\psi}^\dagger(\mathbf{r}', t) \hat{\psi}(\mathbf{r}', t) \hat{\psi}(\mathbf{r}, t) d^3r', \quad (44)$$

which describes the time evolution of a Heisenberg operator in the same way as the Schrödinger equation describes that of a wave function, we can directly deduce the equation of motion for the Green function

$$i\hbar \frac{\partial}{\partial t} G(\mathbf{r}t, \mathbf{r}'t') = \delta(\mathbf{r} - \mathbf{r}') \delta(t - t') + \hat{h}(\mathbf{r}) G(\mathbf{r}t, \mathbf{r}'t') \\ - \frac{i}{\hbar} \int v(\mathbf{r}, \mathbf{r}'') \left\langle \Psi_0^N \left| \hat{T} \left[\hat{\psi}^\dagger(\mathbf{r}'', t) \hat{\psi}(\mathbf{r}'', t) \hat{\psi}(\mathbf{r}, t) \hat{\psi}^\dagger(\mathbf{r}', t') \right] \right| \Psi_0^N \right\rangle d^3r''. \quad (45)$$

This is not a closed equation, because it involves the two-particle Green function

$$G(1234) = -\frac{1}{\hbar^2} \left\langle \Psi_0^N \left| \hat{T} \left[\hat{\psi}(1) \hat{\psi}(2) \hat{\psi}^\dagger(4) \hat{\psi}^\dagger(3) \right] \right| \Psi_0^N \right\rangle. \quad (46)$$

Here and in the following we denote the set of space-time coordinates $(\mathbf{r}_1, t_1)$ with a natural number 1, etc., and further define

$$\delta(12) = \delta(\mathbf{r}_1 - \mathbf{r}_2) \delta(t_1 - t_2), \quad (47)$$

$$v(12) = v(\mathbf{r}_1, \mathbf{r}_2) \delta(t_1 - t_2), \quad (48)$$

$$\int d1 = \int d^3r_1 \int_{-\infty}^{\infty} dt_1, \quad (49)$$

$$1^+ = (\mathbf{r}_1, t_1 + \eta), \quad (50)$$

where η is an infinitesimal positive time. With the two-particle Green function (Eq. (46)) we can rewrite Eq. (45) as

$$i\hbar \frac{\partial}{\partial t_1} G(12) = \delta(12) + \hat{h}(1)G(12) - i\hbar \int v(1^+3)G(1323^+)d3. \quad (51)$$

The additional infinitesimals in 1^+ and 3^+ make sure that the time order is the same as in Eq. (45).

In order to employ the functional-derivative method, we introduce an external potential $U(1)$ that is again set to zero at the end. Of course, all quantities from now on depend on $U(1)$, while the equations remain invariant provided that we replace $\hat{h}(1) \rightarrow \hat{h}(1) + U(1)$. We can use functional differentiation to define a number of useful quantities. The reaction of the density to changes in the external potential is governed by the linear-response function

$$R(12) = \left. \frac{\delta n(1)}{\delta U(2)} \right|_{U=0}. \quad (52)$$

The test potential and the Coulomb potential created by the induced charge can be combined into an effective potential

$$U_{\text{eff}}(1) = U(1) + \iint v(13)R(32)U(2)d2d3, \quad (53)$$

which is related to $U(1)$ via the inverse dielectric function

$$\varepsilon^{-1}(12) = \left. \frac{\delta U_{\text{eff}}(1)}{\delta U(2)} \right|_{U=0} = \delta(12) + \int v(13)R(32)d3. \quad (54)$$

With the definition of the polarization function

$$P(12) = \left. \frac{\delta n(1)}{\delta U_{\text{eff}}(2)} \right|_{U=0} \quad (55)$$

and the chain rule for functional derivatives one obtains the geometric series

$$\varepsilon^{-1}(12) = \delta(12) + \int v(13)P(32)d3 + \iiint v(13)P(34)v(45)P(52)d3d4d5 + \dots, \quad (56)$$

which can easily be inverted to yield

$$\varepsilon(12) = \delta(12) - \int v(13)P(32)d3. \quad (57)$$

If we take the Coulomb potential of an electron at 2 as the test potential, we get the screened potential

$$W(12) = \int \varepsilon^{-1}(13)v(32)d3 = v(12) + \iint v(13)P(34)W(42)d3d4 \quad (58)$$

as the effective potential at position 1.

After this interlude we can go on with the derivation. For the functional derivative of the Green function with respect to the test potential we find^{4,20}

$$\left. \frac{\delta G(12)}{\delta U(3)} \right|_{U=0} = G(12)G(33^+) - G(1323^+). \quad (59)$$

This allows us to eliminate the two-particle Green function, and the integral in Eq. (51) becomes

$$\begin{aligned}
& -i\hbar \int v(1^+3)G(1323^+)d3 \\
& = \underbrace{-i\hbar \left(\int v(13)G(33^+)d3 \right)}_{V^H(1)} G(12) + i\hbar \int v(1^+3) \frac{\delta G(12)}{\delta U(3)} d3 \\
& = V^H(1)G(12) + \int \Sigma(13)G(32)d3,
\end{aligned} \tag{60}$$

where $V^H(1)$ is the Hartree potential [cf. Eq. (21)] and

$$\begin{aligned}
\Sigma(12) &= i\hbar \iint v(1^+3) \frac{\delta G(14)}{\delta U(3)} G^{-1}(42)d3 d4 \\
&= -i\hbar \iint v(1^+3)G(14) \frac{\delta G^{-1}(42)}{\delta U(3)} d3 d4 \\
&= i\hbar \iint W(1^+3)G(14)\Gamma(42;3)d3 d4
\end{aligned} \tag{61}$$

the self-energy. For the second identity we used partial integration and for the third the chain rule for functional derivatives, the definition of the screened interaction (Eq. (58)) and the vertex function

$$\Gamma(12;3) = - \left. \frac{\delta G^{-1}(12)}{\delta U_{\text{eff}}(3)} \right|_{U=0}. \tag{62}$$

With the self-energy (Eq. (61)) the equation of motion for the Green function (Eq. (51)) now becomes

$$\left[i\hbar \frac{\partial}{\partial t_1} - \hat{h}_0(1) \right] G(12) - \int \Sigma(13)G(32)d3 = \delta(12), \tag{63}$$

where we incorporated the Hartree potential into the one-particle operator

$$\hat{h}_0(1) = \hat{h}(1) + V^H(1). \tag{64}$$

The delta function on the right-hand side of Eq. (63) demonstrates that $G(12)$ is indeed a Green function in the mathematical sense. In a noninteracting system the self-energy vanishes, and Eq. (63) becomes

$$\left[i\hbar \frac{\partial}{\partial t_1} - \hat{h}_0(1) \right] G_0(12) = \delta(12). \tag{65}$$

Multiplication of Eq. (63) with G_0 and Eq. (65) with G from the left followed by integration yields the Dyson equation

$$G(12) = G_0(12) + \iint G_0(13)\Sigma(34)G(42)d3 d4. \tag{66}$$

Finally, Eqs. (65) and (66) allow us to rewrite the vertex function (Eq. (62)) as

$$\Gamma(12;3) = \delta(12)\delta(13) + \frac{\delta \Sigma(12)}{\delta U_{\text{eff}}(3)}, \tag{67}$$

and with the identity

$$\begin{aligned}\frac{\delta G(12)}{\delta U_{\text{eff}}(3)} &= \frac{\delta}{\delta U_{\text{eff}}(3)} \iint G(14)G^{-1}(45)G(52)d4 d5 \\ &= 2\frac{\delta G(12)}{\delta U_{\text{eff}}(3)} + \iint G(14)\frac{\delta G^{-1}(45)}{\delta U_{\text{eff}}(3)}G(52)d4 d5\end{aligned}\quad (68)$$

we obtain

$$\Gamma(12; 3) = \delta(12)\delta(13) - \iiint \frac{\delta \Sigma(12)}{\delta G(45)}G(56)\Gamma(67; 3)G(74)d4 d5 d6 d7 \quad (69)$$

and analogously

$$P(12) = i\hbar \frac{\partial G(11^+)}{\partial U_{\text{eff}}(2)} = -i\hbar \iint G(13)\Gamma(34; 2)G(41)d3 d4. \quad (70)$$

The Eqs. (58), (66), (61), (69) and (70) constitute Hedin's set of integro-differential equations, whose self-consistent solution, in principle, solves the many-electron problem exactly. Unfortunately, they are not just numerical relations but contain a functional derivative in Eq. (69). Therefore, the Hedin equations cannot be solved self-consistently by computer codes, but they may be iterated analytically in order to derive useful approximations. In practice we can only perform one iteration. We start with the Green function G_0 of the noninteracting system, which corresponds to the single-particle Hamiltonian (Eq. (64)). As the corresponding self-energy vanishes in this case, the set of equations simplifies to

$$\Gamma(12; 3) = \delta(12)\delta(13), \quad (71)$$

$$P(12) = -i\hbar G_0(12)G_0(21), \quad (72)$$

$$W(12) = v(12) + \iint v(13)P(34)W(42)d3 d4, \quad (73)$$

$$\Sigma(12) = i\hbar G_0(12)W(1^+2), \quad (74)$$

$$G(12) = G_0(12) + \iint G_0(13)\Sigma(34)G(42)d3 d4. \quad (75)$$

The polarization function here corresponds to the bubble diagram of Feynman's diagrammatic approach to many-body perturbation theory and leads to the random-phase approximation for the screened interaction (cf. Figure 5). The expression for the self-energy in this first iteration coined the name *GW* approximation.

Appendix B

Quasiparticle equation

Inserting Eq. (12) into the equation of motion for the Green function of a stationary system in the frequency domain

$$\left[\hbar\omega - \hat{h}_0(\mathbf{r}) \right] G(\mathbf{r}, \mathbf{r}', \omega) - \int \Sigma(\mathbf{r}, \mathbf{r}''; \omega)G(\mathbf{r}'', \mathbf{r}'; \omega)d^3r'' = \delta(\mathbf{r} - \mathbf{r}'), \quad (76)$$

which is equivalent to the Dyson equation, yields

$$\sum_i \frac{\psi_i^*(\mathbf{r}')}{\hbar\omega - \epsilon_i \mp i\eta} \left\{ [\hbar\omega - \hat{h}_0(\mathbf{r})] \psi_i(\mathbf{r}) - \int \Sigma(\mathbf{r}, \mathbf{r}''; \omega) \psi_i(\mathbf{r}'') d^3r'' \right\} = \delta(\mathbf{r} - \mathbf{r}') . \quad (77)$$

Now we multiply with $(\hbar\omega - \epsilon_j)$ and take the limit $\omega \rightarrow \epsilon_j/\hbar$ on both sides. If we assume that the system is nondegenerate (i.e., all ϵ_i are different), the left-hand side becomes

$$\begin{aligned} \lim_{\omega \rightarrow \epsilon_j/\hbar} (\hbar\omega - \epsilon_j) \sum_i \frac{\psi_i^*(\mathbf{r}')}{\hbar\omega - \epsilon_i \mp i\eta} \left\{ [\hbar\omega - \hat{h}_0(\mathbf{r})] \psi_i(\mathbf{r}) - \int \Sigma(\mathbf{r}, \mathbf{r}''; \omega) \psi_i(\mathbf{r}'') d^3r'' \right\} \\ = \psi_j^*(\mathbf{r}') \left\{ [\epsilon_j - \hat{h}_0(\mathbf{r})] \psi_j(\mathbf{r}) - \int \Sigma(\mathbf{r}, \mathbf{r}''; \epsilon_j/\hbar) \psi_j(\mathbf{r}'') d^3r'' \right\} , \end{aligned} \quad (78)$$

and the right-hand side becomes

$$\lim_{\omega \rightarrow \epsilon_j/\hbar} (\hbar\omega - \epsilon_j) \delta(\mathbf{r} - \mathbf{r}') = 0 . \quad (79)$$

Since $\psi_j^*(\mathbf{r}')$ does not vanish for all $\mathbf{r}'$, the expression in the curly brackets must be zero. This leads directly to the quasiparticle equation

$$\hat{h}_0(\mathbf{r}) \psi_j(\mathbf{r}) + \int \Sigma(\mathbf{r}, \mathbf{r}''; \epsilon_j/\hbar) \psi_j(\mathbf{r}'') d^3r'' = \epsilon_j \psi_j(\mathbf{r}) . \quad (80)$$

It remains valid in the degenerate case, which is seen as follows. We assume that the solution of Eq. (80) leads to degenerate amplitudes $\psi_j(\mathbf{r})$ and energies ϵ_j . Then we introduce an arbitrary perturbation $\hat{\phi}$, e.g., an additional external potential in $\hat{h}_0(\mathbf{r})$, that breaks the symmetry in such a way that the degeneracy is lifted. For this (nondegenerate) case the above proof applies. The validity of Eq. (80) for the degenerate case is then established by taking the limit $\hat{\phi} \rightarrow 0$.

References

1. P. Hohenberg and W. Kohn, *Inhomogeneous Electron Gas*, Phys. Rev. **136**, B864 (1964).
2. A. L. Fetter and J. D. Walecka, *Quantum Theory of Many-Particle Systems* (Dover, New York, 2003).
3. G. D. Mahan, *Many-Particle Physics* (Plenum, New York, 1990).
4. L. Hedin, *New Method for Calculating the One-Particle Green's Function with Application to the Electron-Gas Problem*, Phys. Rev. **139**, A796 (1965).
5. W. Kohn and L. J. Sham, *Self-Consistent Equations Including Exchange and Correlation Effects*, Phys. Rev. **140**, A1133 (1965).
6. L. J. Sham and W. Kohn, *One-Particle Properties of an Inhomogeneous Interacting Electron Gas*, Phys. Rev. **145**, 561 (1966).
7. J. P. Perdew, R. G. Parr, M. Levy, and J. L. Balduz, *Density-Functional Theory for Fractional Particle Number: Derivative Discontinuities of the Energy*, Phys. Rev. Lett. **49**, 1691 (1982).

8. W. G. Aulbur, L. Jönsson, and J. W. Wilkins, *Quasiparticle calculations in solids* in Solid State Physics, edited by H. Ehrenreich and F. Spaepen (Academic, New York, 2000), Vol. 54, p. 1.
9. P. Duffy, D. P. Chong, M. E. Casida, and D. R. Salahub, *Assessment of Kohn-Sham density-functional orbitals as approximate Dyson orbitals for the calculation of electron-momentum-spectroscopy scattering cross sections*, Phys. Rev. A **50**, 4707 (1994).
10. W. von der Linden and P. Horsch, *Precise quasiparticle energies and Hartree-Fock bands of semiconductors and insulators*, Phys. Rev. B **37**, 8351 (1988).
11. G. E. Engel and B. Farid, *Generalized plasmon-pole model and plasmon band structures of crystals*, Phys. Rev. B **47**, 15931 (1993).
12. M. S. Hybertsen and S. G. Louie, *First-Principles Theory of Quasiparticles: Calculation of Band Gaps in Semiconductors and Insulators*, Phys. Rev. Lett. **55**, 1418 (1985).
13. R. W. Godby, M. Schlüter, and L. J. Sham, *Accurate Exchange-Correlation Potential for Silicon and Its Discontinuity on Addition of an Electron*, Phys. Rev. Lett. **56**, 2415 (1986).
14. M. S. Hybertsen and S. G. Louie, *Electron correlation in semiconductors and insulators: Band gaps and quasiparticle energies*, Phys. Rev. B **34**, 5390 (1986).
15. R. W. Godby, M. Schlüter, and L. J. Sham, *Quasiparticle energies in GaAs and AlAs*, Phys. Rev. B **35**, 4170 (1987).
16. J. E. Northrup, M. S. Hybertsen, and S. G. Louie, *Theory of quasiparticle energies in alkali metals*, Phys. Rev. Lett. **59**, 819 (1987).
17. R. Keyling, W.-D. Schöne, and W. Ekardt, *Comparison of the lifetime of excited electrons in noble metals*, Phys. Rev. B **61**, 1670 (2000).
18. G. Onida, L. Reining, and A. Rubio, *Electronic excitations: density-functional versus many-body Green's-function approaches*, Rev. Mod. Phys. **74**, 601 (2002).
19. M. Petersilka, U. J. Gossmann, and E. K. U. Gross, *Excitation Energies from Time-Dependent Density-Functional Theory*, Phys. Rev. Lett. **76**, 1212 (1996).
20. L. Hedin and S. Lundqvist, *Effects of Electron-Electron and Electron-Phonon Interactions on the One-Electron States of Solids* in Solid State Physics, edited by F. Seitz, D. Turnbull and H. Ehrenreich (Academic, New York, 1969), Vol. 23, p. 1.

Time-Dependent Density Functional Theory

Nikos L. Doltsinis

Chair of Theoretical Chemistry
Ruhr-Universität Bochum
44780 Bochum, Germany
E-mail: nikos.doltsinis@theochem.rub.de

Time-dependent density functional theory (TDDFT) is the generalization of stationary DFT to time-dependent potentials and electron densities. Most practical applications are concerned with the interaction of molecules with an electric laser field. In the vast majority of cases, such as the calculation of photoabsorption spectra for fixed nuclei, the electric field constitutes a small perturbation which can be treated using linear response theory. This has to be distinguished from directly solving the time-dependent Kohn-Sham equations in the time domain, i.e. dynamically propagating orbitals and nuclei.

1 Introduction

Density functional theory (DFT) in its usual time-independent form is essentially a ground state theory and, as such, excludes the interaction of matter with time-dependent fields. There is generally no rigorous way, for example, to calculate electronic excitation energies due to photoabsorption. Standard DFT can be extended to excited states representing the lowest state of a given space-spin symmetry. Beyond this, a number of *ad hoc* solutions based on Ziegler's sum method¹ have been proposed over the years^{2–13}.

The description of time-dependent phenomena, including photoexcitation, was incorporated properly into DFT by Runge and Gross¹⁴ who generalized the Hohenberg-Kohn theorem to time-dependent densities and potentials. It makes sense to distinguish between two main types of time-dependent DFT (TDDFT) calculations. The overwhelming majority of applications deal with relatively weak electric fields, e.g. photoabsorption spectra, which can be treated as a small perturbation within linear response theory. The other branch solves the TDDFT equations in the time domain to dynamically propagate electrons and nuclei^{15–20}. In the present article we will limit the discussion to the linear response aspect, the particular focus being on electronic excitation.

The past decade has seen TDDFT linear response theory^{14,21–24} become the most widely used electronic structure method for calculating vertical electronic excitation energies^{25,26}. Except for certain well-known problem cases such as, for instance, charge transfer^{27–30} Rydberg states^{31–34} and double excitations³⁵, TDDFT excitation energies are generally remarkably accurate, typically to within a fraction of an electron Volt^{36–38,31}.

Excited state analytical nuclear forces within TDDFT have only been implemented recently^{39–42} in an attempt to extend the applicability of TDDFT beyond single point calculations. One complication has been the fact that TDDFT merely provides excitation energies, but excited state wave functions are not properly defined. The first excited state geometry optimization using analytical gradients was presented by van Caillie and Amos based on a Handy-Schaefer Z-vector method^{39,40}. An extended Lagrangian ansatz was chosen by Furche and Ahlrichs⁴¹ and Hutter⁴² for their Gaussian-type basis set and plane

wave/pseudopotential implementations, respectively. The latter variant is of particular importance for condensed phase applications since it is used in conjunction with periodic boundary conditions. In order to ensure completeness, the number of Kohn-Sham (KS) orbitals included in constructing the response matrix in a molecular orbital (MO) basis must equal the number of basis functions. Since a plane wave basis typically consists of two orders of magnitude more basis functions than a Gaussian-type basis set a complete MO formulation of TDDFT is impractical. A solution to this problem is to cast the working matrix equations directly into a plane wave basis as proposed by Hutter⁴². Earlier, Doltsinis and Sprik⁴³ have proposed an alternative, *active space* approach to TDDFT in which only a subset of (active) KS orbitals is selected to construct the response matrix. For a large variety of excited states, convergence of the corresponding excitation energies has been shown to be rapid with respect to the number of orbitals included in the active space^{43,30}. Doltsinis and Kosov⁴⁴ followed this active space ansatz and derived analytical expressions for excited state nuclear forces within an MO basis. In contrast to previous work, they do not rely on a Lagrangian formulation^{41,42,45}, but employ an implicit differentiation scheme instead.

This article is organized as follows. First a general introduction to TDDFT is given before deriving the working equations of TDDFT linear response theory for the calculation of excitation energies. Subsequently, we present two routes to computing excited state nuclear gradients within TDDFT linear response theory, the popular extended Lagrangian ansatz and the implicit differentiation method. Finally, we discuss some illustrative examples of excited state energy and gradient calculations using a plane wave basis set.

2 Theory

2.1 Time-Dependent Kohn-Sham Theory

According to the Runge-Gross theorem¹⁴ there is a one to one correspondence between the time-dependent external potential, $v_{\text{ext}}(\mathbf{r}, t)$, and the time-dependent electron density, $\rho(\mathbf{r}, t)$ for a fixed initial state. This can be seen as a generalization of the usual Hohenberg-Kohn theorem⁴⁶ for electronic ground states. Similar to the static case, one can cast the many-electron problem into the Kohn-Sham non-interacting electrons form assuming non-interacting v -representability. The latter assumption means that the density of the interacting system can be reproduced by the non-interacting potential v_s , i.e.

$$\rho(\mathbf{r}, t) = \sum_i^{\text{occ}} |\phi_i(\mathbf{r}, t)|^2 \quad (1)$$

where the orbitals $\phi_i(\mathbf{r}, t)$ satisfy the time-dependent Kohn-Sham equations

$$i \frac{\partial}{\partial t} \phi_i(\mathbf{r}, t) = \left(-\frac{\nabla^2}{2} + v_s[\rho](\mathbf{r}, t) \right) \phi_i(\mathbf{r}, t) \quad (2)$$

with

$$v_s[\rho](\mathbf{r}, t) = v_{\text{ext}}(\mathbf{r}, t) + \int d\mathbf{r}' \frac{\rho(\mathbf{r}', t)}{|\mathbf{r} - \mathbf{r}'|} + v_{\text{xc}}[\rho](\mathbf{r}, t) \quad (3)$$

defining the exchange-correlation potential $v_{\text{xc}}[\rho](\mathbf{r}, t)$. In the usual adiabatic approximation⁴⁷, the exchange-correlation potential is taken to be simply the derivative of the static ground state exchange-correlation energy, E_{xc} , with respect to the density,

$$v_{\text{xc}}[\rho](\mathbf{r}, t) \approx \frac{\delta E_{\text{xc}}[\rho]}{\delta \rho} \quad . \quad (4)$$

2.2 Linear Response Theory

Let us suppose a time-dependent perturbing potential $v_1(\mathbf{r}, t)$, for instance an oscillating electric field $v_1(\mathbf{r}, t) = Ez \cos wt$, is switched on at time $t = t_0$. The external potential is then given by

$$v_{\text{ext}}(\mathbf{r}, t) = v_0(\mathbf{r}) + v_1(\mathbf{r}, t) = \begin{cases} v_0(\mathbf{r}) & , t \leq t_0 \\ v_0(\mathbf{r}) + Ez \cos wt & , t > t_0 \end{cases} \quad (5)$$

where v_0 is usually the attractive Coulomb potential between electrons and nuclei

$$v_0(\mathbf{r}) = - \sum_K^N \frac{Z_K}{|\mathbf{R}_K - \mathbf{r}|} \quad . \quad (6)$$

The first order density response for *interacting* particles to the perturbation (the first order deviation of the time-dependent density $\rho(\mathbf{r}, t)$ from the unperturbed ground state density $\rho_0(\mathbf{r})$) may be obtained from,

$$\rho(\mathbf{r}, t) - \rho_0(\mathbf{r}) \approx \rho_1(\mathbf{r}, t) = \int dt' \int d\mathbf{r}' \chi(\mathbf{r}, t, \mathbf{r}', t') v_1(\mathbf{r}', t') \quad (7)$$

with the *interacting* response function

$$\chi(\mathbf{r}, t, \mathbf{r}', t') = \left. \frac{\delta \rho(\mathbf{r}, t)}{\delta v_{\text{ext}}(\mathbf{r}', t')} \right|_{v_0} \quad . \quad (8)$$

Expressing the rhs of Eq. (7) in terms of the Kohn-Sham response function of *non-interacting* particles,

$$\chi_s(\mathbf{r}, t, \mathbf{r}', t') = \left. \frac{\delta \rho(\mathbf{r}, t)}{\delta v_s(\mathbf{r}', t')} \right|_{v_s[\rho_0]} \quad (9)$$

one arrives at

$$\rho_1(\mathbf{r}, t) = \int dt' \int d\mathbf{r}' \chi_s(\mathbf{r}, t, \mathbf{r}', t') v_{s,1}(\mathbf{r}', t') \quad (10)$$

where

$$v_{s,1}(\mathbf{r}, t) = v_1(\mathbf{r}, t) + \int d\mathbf{r}' \frac{\rho(\mathbf{r}', t)}{|\mathbf{r} - \mathbf{r}'|} + \int d\mathbf{r}' \int dt' f_{\text{xc}}[\rho_0](\mathbf{r}, t, \mathbf{r}', t') \rho_1(\mathbf{r}', t') \quad (11)$$

with the exchange-correlation kernel

$$f_{\text{xc}}[\rho_0](\mathbf{r}, t, \mathbf{r}', t') = \frac{\delta v_{\text{xc}}[\rho](\mathbf{r}, t)}{\delta \rho(\mathbf{r}', t')} \quad . \quad (12)$$

The frequency-dependent linear density response is obtained by Fourier transform as

$$\begin{aligned} \rho_1(\mathbf{r}, \omega) = & \int d\mathbf{r}' \chi_s(\mathbf{r}, \mathbf{r}'; \omega) v_1(\mathbf{r}', \omega) \\ & + \int d\mathbf{r}' \int d\mathbf{r}'' \chi_s(\mathbf{r}, \mathbf{r}'; \omega) \left(\frac{1}{|\mathbf{r}' - \mathbf{r}''|} + f_{xc}[\rho_0](\mathbf{r}', \mathbf{r}''; \omega) \right) \rho_1(\mathbf{r}'', \omega) \end{aligned} \quad (13)$$

and

$$\chi_s(\mathbf{r}, \mathbf{r}'; \omega) = \sum_{ph} \left(\frac{\phi_p(\mathbf{r}) \phi_h^*(\mathbf{r}) \phi_p^*(\mathbf{r}') \phi_h(\mathbf{r}')}{\omega - (\epsilon_p - \epsilon_h)} - \frac{\phi_p^*(\mathbf{r}) \phi_h(\mathbf{r}) \phi_p(\mathbf{r}') \phi_h^*(\mathbf{r}')}{\omega + (\epsilon_p - \epsilon_h)} \right) \quad (14)$$

ϕ_p and ϕ_h being the Kohn-Sham particle (unoccupied) and hole (occupied) molecular orbitals corresponding to the Kohn-Sham energies ϵ_p and ϵ_h , respectively.

Equation (13) for the first-order density has to be solved self-consistently. This is most conveniently done by casting Eq. (13) into a matrix eigenvalue problem^{21,22}. For this purpose, Bauernschmitt and Ahlrichs²² parametrized the first-order density

$$\rho_1(\mathbf{r}, \omega) = \sum_{ph\sigma} [P_{ph\sigma}(\omega) \phi_{p\sigma}(\mathbf{r}) \phi_{h\sigma}^*(\mathbf{r}) + P_{hp\sigma}(\omega) \phi_{p\sigma}^*(\mathbf{r}) \phi_{h\sigma}(\mathbf{r})] \quad (15)$$

Notice that we have now introduced the spin index $\sigma (= \uparrow, \downarrow)$ and the undetermined expansion coefficients $P_{ph\sigma}(\omega)$ have the meaning of the linear response of the density matrix. Equation (13) can then be rewritten as

$$\left[\begin{pmatrix} \mathbf{L} & \mathbf{M} \\ \mathbf{M}^* & \mathbf{L}^* \end{pmatrix} - \omega \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} \right] \begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix} = -v_{1,t}(\omega) \begin{pmatrix} \mathbf{V} \\ \mathbf{V}^* \end{pmatrix} \quad (16)$$

with the vector components $X_{hp\sigma} = P_{hp\sigma}(\omega)$, $Y_{hp\sigma} = P_{ph\sigma}(\omega)$, and

$$V_{hp\sigma} = \int d\mathbf{r} \phi_{h\sigma}^*(\mathbf{r}) v_{1,\mathbf{r}}(\mathbf{r}) \phi_{p\sigma}(\mathbf{r}) \quad (17)$$

having separated the time-dependent perturbing potential into a purely space-dependent and a purely time-dependent part, i.e. $v_1(\mathbf{r}, t) = v_{1,\mathbf{r}}(\mathbf{r}) v_{1,t}(t)$. The matrices $\mathbf{M}$ and $\mathbf{L}$ are given by

$$L_{hp\sigma, h'p'\sigma'} = \delta_{\sigma\sigma'} \delta_{pp'} \delta_{hh'} (\epsilon_{p\sigma} - \epsilon_{h\sigma}) + K_{hp\sigma, h'p'\sigma'} \quad (18)$$

and

$$M_{hp\sigma, h'p'\sigma'} = K_{hp\sigma, h'p'\sigma'} \quad (19)$$

with the coupling matrix

$$K_{hp\sigma, h'p'\sigma'} = \int d\mathbf{r} \int d\mathbf{r}' \phi_{h\sigma}^*(\mathbf{r}) \phi_{p\sigma}(\mathbf{r}) f_{H,xc}^{\sigma\sigma'}(\mathbf{r}, \mathbf{r}') \phi_{h'\sigma'}^*(\mathbf{r}') \phi_{p'\sigma'}(\mathbf{r}') \quad (20)$$

Here we have summarized the Hartree term and the exchange-correlation term in the response kernel

$$f_{H,xc}^{\sigma\sigma'}(\mathbf{r}, \mathbf{r}') = \frac{1}{|\mathbf{r} - \mathbf{r}'|} + \frac{\delta^2 E_{xc}}{\delta \rho^\sigma(\mathbf{r}) \delta \rho^{\sigma'}(\mathbf{r}')} \quad (21)$$

The poles of the response function (Eq. (8)) of the *interacting* system represent electronic excitation energies^{21,48,16}. They are characterized by zero eigenvalues of the matrix on

the lhs of Eq. (16) and can therefore be obtained by solving the non-Hermitian eigenvalue problem

$$\begin{pmatrix} \mathbf{L} & \mathbf{M} \\ \mathbf{M}^* & \mathbf{L}^* \end{pmatrix} \begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix} = \omega \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix} . \quad (22)$$

In the following, we shall assume the Kohn-Sham orbitals to be real. It is then possible to halve the dimension of the non-Hermitian eigenvalue problem by means of a unitary transformation⁴⁹. The essential steps are multiplication of Eq. (22) on the left by $\begin{pmatrix} \mathbf{L} & \mathbf{M} \\ \mathbf{M} & \mathbf{L} \end{pmatrix} \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}$ and substitution of Eq. (22) into the rhs of the resulting matrix equation. Subtracting the two equations obtained this way from each other, one finds

$$(\mathbf{L} - \mathbf{M})(\mathbf{L} + \mathbf{M})(\mathbf{X} + \mathbf{Y}) = \omega^2(\mathbf{X} + \mathbf{Y}) . \quad (23)$$

If $(\mathbf{L} - \mathbf{M})$ is positive definite, Eq. (23) can be multiplied on the left with $(\mathbf{L} - \mathbf{M})^{-\frac{1}{2}}$ yielding the Hermitian eigenvalue problem

$$\mathbf{\Omega} \mathbf{F} = \omega^2 \mathbf{F} \quad (24)$$

where

$$\mathbf{\Omega} = (\mathbf{L} - \mathbf{M})^{\frac{1}{2}} (\mathbf{L} + \mathbf{M}) (\mathbf{L} - \mathbf{M})^{\frac{1}{2}} \quad (25)$$

and

$$\mathbf{F} = (\mathbf{L} - \mathbf{M})^{-\frac{1}{2}} (\mathbf{X} + \mathbf{Y}) \quad (26)$$

The Hermitian coupling matrix (Eq. (25)) elements are

$$\Omega_{hp\sigma, h'p'\sigma'} = \delta_{\sigma\sigma'} \delta_{pp'} \delta_{hh'} (\epsilon_{p\sigma} - \epsilon_{h\sigma})^2 + 2\sqrt{\epsilon_{p\sigma} - \epsilon_{h\sigma}} K_{ph\sigma, p'h'\sigma'} \sqrt{\epsilon_{p'\sigma} - \epsilon_{h'\sigma}} . \quad (27)$$

The Tamm-Dancoff approximation, which amounts to setting $\mathbf{Y}$ to zero, further simplifies the problem in particular when hybrid Hartree-Fock/DFT exchange-correlation functionals are used³⁵. However, it does not reduce the computational cost or storage space compared to the full Hermitian formulation (Eq. (24)).

2.3 Excited State Gradients

Excited state analytical nuclear forces within TDDFT have only been implemented recently^{39–42,44} in an attempt to extend the applicability of TDDFT beyond single point calculations. One complication has been the fact that TDDFT merely provides excitation energies, but excited state wave functions are not properly defined. The first excited state geometry optimization using analytical gradients was presented by van Caillie and Amos based on a Handy-Schaefer Z-vector method^{39,40}. An extended Lagrangian ansatz was chosen by Furche and Ahlrichs⁴¹ and Hutter⁴² for their Gaussian-type basis set and plane wave/pseudopotential implementations, respectively. The latter variant is of particular importance for condensed phase applications since it is used in conjunction with periodic boundary conditions.

2.3.1 Extended Lagrangian Method

The starting point for the derivation of TDDFT excited state energy gradients is the construction of the extended Lagrangian^{41,42}

$$\mathcal{L} = \mathcal{L}^{\text{LR}} + \mathcal{L}^{\text{KS}} \quad (28)$$

consisting of the TDDFT linear response Lagrangian

$$\mathcal{L}^{\text{LR}} = \langle \mathbf{X}, \mathbf{Y} | \mathbf{\Lambda} | \mathbf{X}, \mathbf{Y} \rangle - \omega [\langle \mathbf{X}, \mathbf{Y} | \mathbf{\Delta} | \mathbf{X}, \mathbf{Y} \rangle - 1] \quad (29)$$

with

$$\mathbf{\Lambda} = \begin{pmatrix} \mathbf{L} & \mathbf{M} \\ \mathbf{M} & \mathbf{L} \end{pmatrix} \quad (30)$$

and

$$\mathbf{\Delta} = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} . \quad (31)$$

The corresponding non-Hermitian linear response problem (Eq. (22)) can be recovered from

$$\frac{\delta \mathcal{L}^{\text{LR}}}{\delta \langle \mathbf{X}, \mathbf{Y} |} = 0 \quad (32)$$

subject to the constraint

$$\frac{\partial \mathcal{L}^{\text{LR}}}{\partial \omega} = 0 . \quad (33)$$

The ground state Kohn-Sham contribution is given by

$$\mathcal{L}^{\text{KS}} = \sum_{ph\sigma} Z_{ph\sigma} H_{ph\sigma} - \sum_{ij} W_{ij\sigma} [\langle \phi_{i\sigma} | \phi_{j\sigma} \rangle - \delta_{ij}] \quad (34)$$

where the $Z_{ph\sigma}$ and $W_{ij\sigma}$ are Lagrange multipliers and $H_{ph\sigma}$ are matrix elements of the static Kohn-Sham Hamiltonian. The conditions

$$\frac{\partial \mathcal{L}^{\text{KS}}}{\partial Z_{ph\sigma}} = 0 \quad (35)$$

and

$$\frac{\partial \mathcal{L}^{\text{KS}}}{\partial W_{ij\sigma}} = 0 \quad (36)$$

ensure that the orbitals $\phi_{i\sigma}$ are orthonormal and satisfy the ground state Kohn-Sham equations. The important step is now to determine the unknown Lagrange multipliers $Z_{ph\sigma}$ and $W_{ij\sigma}$ from

$$\frac{\delta \mathcal{L}}{\delta \phi_{i\sigma}} = 0 . \quad (37)$$

The derivative of the excitation energy with respect to the nuclear coordinate R_α ($\alpha = 1, \dots, 3N$) for a molecule consisting of N atoms yields

$$\omega^\alpha = \mathcal{L}^\alpha = \langle \mathbf{X}, \mathbf{Y} | \mathbf{\Lambda}^\alpha | \mathbf{X}, \mathbf{Y} \rangle + \sum_{ph\sigma} Z_{ph\sigma} H_{ph\sigma}^\alpha - \sum_{ij} W_{ij\sigma} [\langle \phi_{i\sigma}^\alpha | \phi_{j\sigma} \rangle + \langle \phi_{i\sigma} | \phi_{j\sigma}^\alpha \rangle] \quad (38)$$

where we have used the short-hand notation $\frac{df}{dR_\alpha} \equiv f^\alpha$ for a general function f .

2.3.2 Implicit Differentiation Approach

An alternative to the above extended Lagrangian method is the recently proposed implicit differentiation scheme⁴⁴. In the following detailed derivation, we confine ourselves to singlet excitations (extension to triplet excitations is straightforward) and therefore drop the spin index σ .

Multiplying Eq. (24) by $\langle \mathbf{F} |$ from the left we obtain

$$\langle \mathbf{F} | \mathbf{\Omega} | \mathbf{F} \rangle = \omega^2 \quad . \quad (39)$$

Differentiation with respect to nuclear position yields

$$\omega^\alpha = \frac{1}{2\omega} \langle \mathbf{F} | \mathbf{\Omega}^\alpha | \mathbf{F} \rangle = \frac{1}{2\omega} \sum_{ph} \sum_{p'h'} F_{ph} \Omega_{ph,p'h'}^\alpha F_{p'h'} \quad , \quad (40)$$

where the F_{ph} are the components of the linear response eigenvector $\mathbf{F}$. Carrying out the differentiation of the response matrix, Eq. (40) becomes

$$\begin{aligned} \omega^\alpha = \frac{1}{\omega} & \left[\sum_{ph} (F_{ph})^2 (\epsilon_p^\alpha - \epsilon_h^\alpha) (\epsilon_p - \epsilon_h) \right. \\ & + 2 \int d\mathbf{r} \int d\mathbf{r}' \Gamma_1(\mathbf{r}) f_{H,xc}(\mathbf{r}, \mathbf{r}') \Gamma_2(\mathbf{r}') \\ & \left. + 2 \int d\mathbf{r} \Gamma_1(\mathbf{r}) \frac{\delta^3 E_{xc}}{\delta \rho(\mathbf{r})^3} \rho^\alpha(\mathbf{r}) \Gamma_1(\mathbf{r}) \right] \quad . \end{aligned} \quad (41)$$

Here we have defined the contracted densities

$$\Gamma_1(\mathbf{r}) = \sum_{ph} F_{ph} \sqrt{\epsilon_p - \epsilon_h} \Gamma_{ph}(\mathbf{r}) \quad (42)$$

and

$$\Gamma_2(\mathbf{r}) = \sum_{ph} F_{ph} \left[\frac{\epsilon_p^\alpha - \epsilon_h^\alpha}{\sqrt{\epsilon_p - \epsilon_h}} \Gamma_{ph}(\mathbf{r}) + 2 \sqrt{\epsilon_p - \epsilon_h} \Gamma_{ph}^\alpha(\mathbf{r}) \right] \quad (43)$$

with

$$\Gamma_{ij}(\mathbf{r}) = \phi_i(\mathbf{r}) \phi_j(\mathbf{r}) \quad . \quad (44)$$

In order to compute the excitation energy gradient (Eq. (41)), we require the nuclear derivatives of KS orbital energies and wave functions, ϵ_i^α and ϕ_i^α ($i = p, h$). These can be obtained using an implicit differentiation scheme as follows. We start by writing down the KS equations in matrix form

$$F_{ij} \equiv H_{ij} - \epsilon_i \delta_{ij} = 0 \quad . \quad (45)$$

For the full differential of F_{ij} we have

$$dF_{ij} = \left(\frac{\partial H_{ij}}{\partial R_\alpha} - \epsilon_i^\alpha \delta_{ij} \right) dR_\alpha + \sum_k \int d\mathbf{r} H_{ij}^k \delta \phi_k(\mathbf{r}) = 0 \quad , \quad (46)$$

where $H_{ij}^k \equiv \frac{\delta H_{ij}}{\delta \phi_k(\mathbf{r})}$. Division by dR_α yields

$$\frac{\partial H_{ij}}{\partial R_\alpha} - \epsilon_i^\alpha \delta_{ij} = - \sum_k \int d\mathbf{r} H_{ij}^k \phi_k^\alpha(\mathbf{r}) = - \sum_k \int d\mathbf{r} \int d\mathbf{r}' H_{ij}^k \delta(\mathbf{r} - \mathbf{r}') \phi_k^\alpha(\mathbf{r}') \quad . \quad (47)$$

On the rhs of Eq. (47) we have inserted a delta function, which we now express in terms of KS orbitals

$$\delta(\mathbf{r} - \mathbf{r}') = \sum_l \phi_l(\mathbf{r}) \phi_l(\mathbf{r}') \quad . \quad (48)$$

Thus Eq. (47) becomes

$$\frac{\partial H_{ij}}{\partial R_\alpha} - \epsilon_i^\alpha \delta_{ij} = - \sum_{kl} H_{ij}^{kl} \phi_k^{\alpha l} \quad , \quad (49)$$

where

$$\begin{aligned} H_{ij}^{kl} &\equiv \int d\mathbf{r} H_{ij}^k \phi_l(\mathbf{r}) \\ &= (\delta_{ik} \delta_{lj} + \delta_{jk} \delta_{li}) \epsilon_l + 2n_k \int d\mathbf{r} \int d\mathbf{r}' \Gamma_{kl}(\mathbf{r}) f_{\text{H,xc}}(\mathbf{r}, \mathbf{r}') \Gamma_{ij}(\mathbf{r}') \quad , \quad (50) \end{aligned}$$

and

$$\phi_k^{\alpha l} \equiv \int d\mathbf{r} \phi_l(\mathbf{r}) \phi_k^\alpha(\mathbf{r}) \quad , \quad (51)$$

n_k being the number of electrons occupying orbital k .

Exploiting the symmetry of the nonadiabatic coupling matrix elements (Eq. (51)), i.e. $\phi_l^{\alpha k} = -\phi_k^{\alpha l}$ and therefore $\phi_l^{\alpha l} = 0$, Eq. (49) can be rewritten as

$$\frac{\partial H_{ij}}{\partial R_\alpha} = \sum_{l < k} D_{ij}^{lk} \phi_k^{\alpha l} \quad , (i < j) \quad (52)$$

and for the diagonal terms ($i = j$)

$$\epsilon_i^\alpha = \frac{\partial H_{ii}}{\partial R_\alpha} - \sum_{k < l} D_{ii}^{kl} \phi_k^{\alpha l} \quad , \quad (53)$$

where

$$D_{ij}^{lk} = H_{ij}^{lk} - H_{ij}^{kl} = (\delta_{il} \delta_{kj} + \delta_{ik} \delta_{lj}) (\epsilon_k - \epsilon_l) + 2(n_k - n_l) K_{ij, lk} \quad . \quad (54)$$

With the definition in Eq. (54) Eq. (52) becomes

$$\frac{\partial H_{hp}}{\partial R_\alpha} = \sum_{p' h'} ((\epsilon_{p'} - \epsilon_{h'}) \delta_{pp'} \delta_{hh'} + 4K_{p' h', ph}) \phi_{h'}^{\alpha p'} \quad (55)$$

for particle-hole states, and

$$\frac{\partial H_{ij}}{\partial R_\alpha} = 4 \sum_{ph} K_{ij, ph} \phi_h^{\alpha p} + (\epsilon_i - \epsilon_j) \phi_i^{\alpha j} \quad , (i < j, ij \ni ph) \quad (56)$$

for all remaining combinations. Eq. (56) allows us to express the nonadiabatic coupling elements between non-particle-hole states analytically as

$$\phi_i^{\alpha j} = \frac{\frac{\partial H_{ii}}{\partial R_\alpha} + 4 \sum_{ph} K_{ij,ph} \phi_h^{\alpha p}}{(\epsilon_i - \epsilon_j)}, (i < j, ij \ni ph) \quad (57)$$

The system of linear equations (55) is first solved for the particle-hole nonadiabatic coupling elements $\phi_p^{\alpha h}$, which are then inserted into Eq. (57) to obtain the remaining, non-particle-hole, elements. The second term in the numerator of Eq. (57) is most conveniently evaluated by introducing the contracted density

$$\Gamma_3(\mathbf{r}) = \sum_{ph} \phi_p(\mathbf{r}) \phi_h(\mathbf{r}) \phi_h^{\alpha p} \quad (58)$$

Then

$$\sum_{ph} K_{ij,ph} \phi_p^{\alpha h} = \int d\mathbf{r} \int d\mathbf{r}' \Gamma_3(\mathbf{r}) f_{H,xc}(\mathbf{r}, \mathbf{r}') \phi_i(\mathbf{r}') \phi_j(\mathbf{r}') \equiv K'_{ij} \quad (59)$$

Thus Eq. (57) becomes

$$\phi_j^{\alpha i} = \frac{\frac{\partial H_{ii}}{\partial R_\alpha} + 4K'_{ij}}{(\epsilon_i - \epsilon_j)}, (i < j, ij \ni ph) \quad (60)$$

Similarly the KS orbital energy gradients can now be obtained from the simplified Eq. (53)

$$\epsilon_i^\alpha = \frac{\partial H_{ii}}{\partial R_\alpha} + 4K'_{ii} \quad (61)$$

Finally, the nuclear derivative of the KS orbital wave function is recovered by unfolding the nonadiabatic couplings

$$\phi_k^\alpha(\mathbf{r}) = \sum_l \phi_l(\mathbf{r}) \phi_k^{\alpha l} \quad (62)$$

Equations (1)–(31) have been implemented with periodic boundary conditions using a plane wave expansion of the KS MOs at the Γ point of the Brillouin zone. By making use of the periodic boundary conditions, the generalized densities Γ_1 , Γ_2 , Γ_3 and Γ_{ij} can be expanded in reciprocal space via the three-dimensional Fourier transform, e.g.

$$\Gamma_1(\mathbf{r}) = \sum_{\mathbf{G}} \Gamma_1(\mathbf{G}) \exp(i\mathbf{G}\mathbf{r}) \quad (63)$$

where $\mathbf{G}$ is the vector of the reciprocal lattice. The Hartree part of the matrix element $\int d\mathbf{r} \int d\mathbf{r}' \Gamma_1(\mathbf{r}) f_{H,xc}(\mathbf{r}, \mathbf{r}') \Gamma_2(\mathbf{r}')$ and $\int d\mathbf{r} \int d\mathbf{r}' \Gamma_{kl}(\mathbf{r}) f_{H,xc}(\mathbf{r}, \mathbf{r}') \Gamma_{ij}(\mathbf{r}')$ which enter the key equations (41) and (50), respectively, can be readily computed in reciprocal space, e.g.

$$\int d\mathbf{r} \int d\mathbf{r}' \Gamma_1(\mathbf{r}) \frac{1}{|\mathbf{r} - \mathbf{r}'|} \Gamma_2(\mathbf{r}') = \Omega \sum_{\mathbf{G} \neq 0} \frac{2\pi}{G^2} \Gamma_1(\mathbf{G}) \Gamma_2(\mathbf{G}) \quad (64)$$

whereas the exchange-correlation parts of the matrix elements are calculated via direct numerical integration over grid in coordinate space.

active space	ΔE_{KS}	1o/1v	5o/1v	5o/5v	5o/50v	5o/100v
exc. energy	8.39	9.47	9.40	9.40	9.33	9.29

Table 1. Dependence of N₂ TDLDA excitation energy (${}^1\Pi_g, 3\sigma_g \rightarrow 1\pi_g$) in eV using plane waves (p.w.) with a 70 Ry cutoff in a 10 a_0 periodic box on the number of occupied (o) and virtual (v) Kohn-Sham orbitals included in the active space. ΔE_{KS} is the unperturbed Kohn-Sham energy difference.

3 Example Calculations

In this section, we present some (test) results for excitation energies and gradients from the TDDFT implementations described in Refs. 43, 44. All calculations were performed with the CPMD package⁵⁰ employing periodic boundary conditions and a plane wave (PW) basis set in conjunction with Troullier-Martins normconserving pseudopotentials⁵¹.

The central idea underlying this active space (AS) approach originates from the observation that excitation energies for a large number of electronic transitions exhibit only a minor dependence on the size of the response matrix (Eq. 27)). This is illustrated in Table 1 for the $3\sigma_g \rightarrow 1\pi_g$ transition of N₂. A simple two-state HOMO–LUMO response calculation is seen to give an excitation energy which is less than 0.2 eV away from an extended treatment including all 5 occupied and 100 virtual MOs. Generally, such behavior is to be expected for excitations which can be characterized by only a few low-lying one-electron transitions without higher-lying continuum states mixing in.

Table 2 lists the first eight excitation energies of N₂ comparing the AS-PW results (30 a.u. unit cell, 70 Ry cutoff, AS consisting of 5 h and 85 p states) to atomic basis set studies as well as experimental data. We observe satisfactory agreement, the largest difference being 0.19 eV for the $3\sigma_g \rightarrow 1\pi_g$ transition.

As a further example, our results for H₂CO in a fcc lattice (30 a.u. unit cell, 70 Ry cutoff, AS contained 6 h and 70 p states) are presented in Table 3. Again, there is reason-

Transition	TDLDA					CCSD	Exp.
	ref. ²²	ref. ³⁷	ref. ³¹	ref. ⁵²	AS-PW ⁴³	ref. ⁵³	ref. ⁵⁴
(${}^1\Delta_u$) $1\pi_u \rightarrow 1\pi_g$	10.22	10.23	10.21	10.20	10.25	10.54	10.27
(${}^1\Sigma_u^-$) $1\pi_u \rightarrow 1\pi_g$	9.65	9.68	9.64	9.63	9.73	10.09	9.92
(${}^1\Pi_g$) $3\sigma_g \rightarrow 1\pi_g$	9.05	9.07	9.04	9.04	9.23	9.27	9.31
(${}^3\Pi_u$) $2\sigma_u \rightarrow 1\pi_g$	10.36	10.37	10.36	10.36	10.45	11.19	11.19
(${}^3\Sigma_u^-$) $1\pi_u \rightarrow 1\pi_g$	9.65	9.68	9.64	9.63	9.73	9.86	9.67
(${}^3\Delta_u$) $1\pi_u \rightarrow 1\pi_g$	8.82	8.84	8.82	8.80	8.92	8.93	8.88
(${}^3\Pi_g$) $3\sigma_g \rightarrow 1\pi_g$	7.54	7.58	7.54	7.53	7.66	8.05	8.04
(${}^3\Sigma_u^+$) $1\pi_u \rightarrow 1\pi_g$	7.86	7.90	7.86	7.84	7.99	7.56	7.75
polarizability [a.u.]	-	12.27	12.17	12.11	12.6		11.74

Table 2. Comparison of AS-PW N₂ excitation energies in eV to other TDLDA results obtained with atomic sets, CCSD, and experiment. In the last row, sum-over-states values for the mean static polarizability are compared to the experimental number.

Transition	TDLDA				CASPT2	Exp.
	ref. ²²	ref. ³¹	AS-PW ⁴³	AS-PW(AC) ⁴³	ref. ⁵⁵	ref. ^{56,57}
$^1A_2 (n \rightarrow \pi^*)$	3.64	3.68	3.68	3.69	3.91	3.94
$^3A_2 (n \rightarrow \pi^*)$	3.02	3.06	3.08	3.08	3.48	3.50
$^3A_1 (\pi \rightarrow \pi^*)$	6.11	6.24	6.38	6.39	5.99	5.53
$^1B_1 (\sigma \rightarrow \pi^*)$	8.70	8.79	8.94	8.94	9.09	8.68
$^3B_2 (n \rightarrow 3s)$	5.86	5.78	5.65	6.56	-	6.83
$^1B_2 (n \rightarrow 3s)$	5.93	5.82	5.76	6.89	7.30	7.09
$^3A_1 (n \rightarrow 3p)$	-	6.48	6.73	7.59	-	7.79
$^1A_1 (n \rightarrow 3p)$	6.79	6.49	6.93	7.83	8.09	7.97

Table 3. Comparison of H₂CO AS-PW excitation energies in eV with and without asymptotic correction (AC) to literature TDLDA results obtained with a gaussian set and experiment.

able agreement between the different LDA studies. The CASPT2 excitation energies are, however, much closer to the experimental values than the TDLDA results. Moreover, inspecting the data for the transitions to the $3s$ and $3p$ Rydberg orbitals the deficiency of the LDA exchange-correlation potential at long range becomes apparent. These problems persist for most GGA's. In contrast, employing the asymptotic correction proposed by Tozer and Handy³¹ these excitation energies can be drastically improved as shown in Table 3.

In our final example, we go beyond the static, zero temperature description. Figure 1 depicts an average photoabsorption spectrum of formamide obtained from the instantaneous spectra of 11 molecular configurations sampled from a CPMD run at 300 K. Although many more configurations are needed to obtain a converged spectrum, we are able to reproduce the most important features of the experimental photoabsorption spectrum. The W band (see Table 4 for a summary of this historic nomenclature) resulting from the $n \rightarrow \pi^*$ transition has been found to have a maximum around 5.8 eV⁵⁸. Gingell et al.⁵⁸ have associated the R₁ band between 6.2 and 7.0 eV with transitions from the n and π orbitals to the $3s$ Rydberg orbital which is supposed to be mixed with a σ^* valence orbital. The main absorption peak V₁ with a maximum at roughly 7.4 eV has traditionally been linked to the $\pi_2 \rightarrow \pi^*$ transition, whereas the sharp structure on its high-energy side has been assigned to transitions to $3p$ Rydberg orbitals⁵⁸. There remains considerable doubt regarding the origin of the broad band, Q, around 9.2 eV, which, for a long time, was thought to be due to the $\pi_1 \rightarrow \pi^*$ transition⁵⁹. Gingell et al.⁵⁸ have interpreted this peak as a superposition of a large number of Rydberg states.

Our theoretical LDA absorption spectrum in Figure 1 is seen to agree overall quite well with the VUV curve, only the low-energy side of the main peak seemingly being somewhat overestimated. However, analyzing the individual excitations at equilibrium geometry listed in Table 4, it becomes clear that use of the asymptotically corrected LDA exchange-correlation potential corrects for this discrepancy. According to our calculations, the R₁ band then unambiguously arises from $n \rightarrow 3s$ and $\pi \rightarrow 3s$ excitations. We cannot confirm, however, any involvement of a σ^* valence orbital as previously assumed⁵⁸. Instead we observe a small contribution (roughly 5% in the asymptotically corrected case) from the $\pi_2 \rightarrow \pi^*$ transition. Our calculations further reveal that all the sharp structure of

transition	AS-PW ⁴³	AS-PW(AC) ⁴³	HCTH(AC) ⁶⁰	CASPT2 ⁶¹	Exp. ⁵⁸
$n \rightarrow \pi^*$	5.27	5.29	5.43	5.61	5.8 (W)
$n \rightarrow 3s$	5.56	6.27	6.30	6.59	6.35 (R ₁)
$\pi \rightarrow 3s$	6.28	7.09	7.03	6.52	
$n \rightarrow 3p$	6.44	7.22	7.09	7.31	7.7 (R ₂)
$\pi \rightarrow 3p$	7.18	7.91	7.76	7.04	
$\pi \rightarrow \pi^*$	7.80	7.67	7.58	7.41	7.4 (V ₁)

Table 4. Comparison of formamide equilibrium geometry AS-PW excitation energies in eV using LDA with and without asymptotic correction (AC) to literature data.

the main peak is due to excitations from n to $3p$ orbitals, generally being more intense by one order of magnitude compared with their $\pi_2 \rightarrow 3p$ equivalents.

The most intense line found with LDA(AC) is at 7.67 eV and its character is a mixture of 37% $\pi_2 \rightarrow \pi^*$ and 63% excitations from the n or π_2 valence to $3p$ or $3d$ Rydberg orbitals. A small $\pi_2 \rightarrow \pi^*$ contribution can also be detected in our calculated Q band, which otherwise predominantly results from $n \rightarrow 3d$ transitions.

We would like to point out that our LDA(AC) results are in accord with the HCTH(AC) values of Handy⁶⁰. On the other hand, the CASPT2 results of Serrano-Andrés and Fölscher⁶¹ show larger differences; in particular the energetic ordering of excitation from the n and π_2 orbitals seems to be reversed.

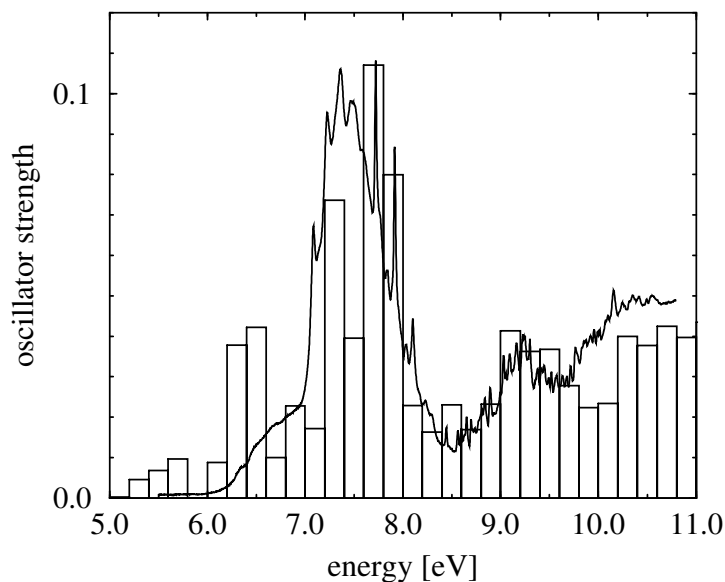


Figure 1. Theoretical room temperature photoabsorption spectrum of formamide obtained from TDLDA calculations (white bars) compared to the experimental VUV data of Gingell et al..

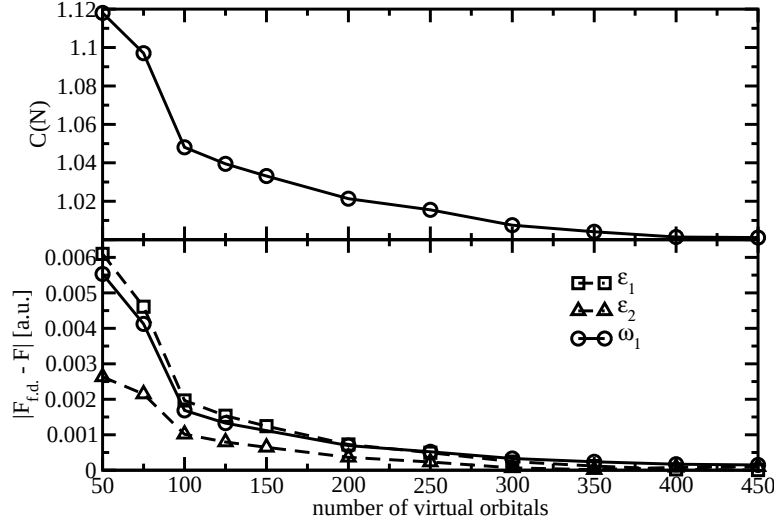


Figure 2. Upper panel: The integral in Eq. (65) as a function of the number of virtual orbitals included in the active space. Lower panel: Absolute deviation of finite difference and analytic derivatives of the KS HOMO and LUMO energies and the first singlet excitation energy of H_2 at a bond length of 1.0 a.u. as a function of the number of virtual KS orbitals included in the active space. The calculations were carried out in a cubic box of length 6 a.u. with periodic boundary conditions and a plane wave cutoff of 40 Ry.

In the following we shall illustrate the performance and the convergence behavior of the implicit differentiation method⁴⁴ (cf. Section 2.3.2) for nuclear gradients.

The upper panel of Figure 2 displays the completeness of the active space as a function of the number of virtual KS orbitals included in the space. The integral

$$C(N) = \int d\mathbf{r} \sum_{i=1}^N \phi_i(\mathbf{r}) \phi_i(0) \quad (65)$$

was used as a measure of completeness of the active space. It becomes unity when the active space of KS orbitals is complete, i.e. when the total number of the KS orbitals (virtual and occupied) equals the number of plane waves used to solve the KS equations. The total number of plane waves is 925 for the 6 a.u. cubic box and 40 Ry plane wave cutoff. With 450 virtual orbitals included the active space is almost complete and the value of the integral (Eq. (65)) deviates from unity by approximately 10^{-3} which is already comparable with the accuracy of the numerical integration. The lower panel of Figure 2 shows the absolute deviation of the analytic derivatives from the respective finite difference values for the the first singlet excitation energy, ω_1 , as well as the HOMO and LUMO KS orbital energies, ϵ_1 and ϵ_2 , of H_2 at a bond length of 1.0 a.u. as a function of the number of virtual KS orbitals included in the active space. The absolute deviation in analytical gradients vanishes rapidly as the number of virtual orbitals is increased and the errors in the analytical gradients of different states generally show the same patterns in the dependence upon the number of virtual orbitals included in the active space.

Figure 3 shows the absolute deviation of the analytical derivative from the respective finite difference value as a function of the size of the active space for the first three KS

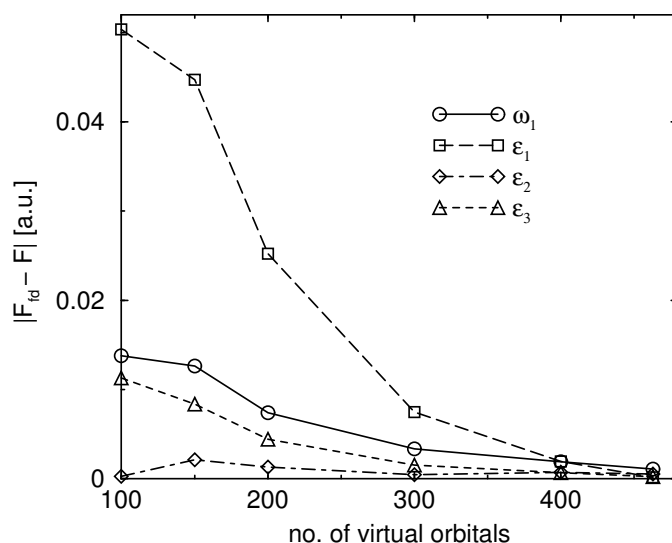


Figure 3. Absolute deviation of the analytic derivatives from the respective finite difference values for the the first singlet excitation energy, ω_1 , as well as the three lowest KS orbital energies, ϵ_i ($i = 1, 2, 3$), of N_2 at a bond length of 2.0 a.u. as a function of the number of virtual KS orbitals included in the active space. The calculations were carried out in a cubic box of length 6 a.u. with periodic boundary conditions and a plane wave cutoff of 40 Ry.

orbital energies ϵ_i ($i = 1, 2, 3$) as well as the lowest response matrix eigenvalue ω_1 of the N_2 molecule. The errors of the analytical gradients are seen to decrease rapidly as the number of orbitals included in the active space approaches the number of plane wave basis functions (in this case 925 plane waves). For the largest active space the deviations of all energies are of the order of 10^{-3} or smaller. At this point, the accuracy of the analytical derivatives is hard to assess because the finite difference reference values are also subjected to numerical errors.

To test the practical value of our derivatives, we have performed geometry optimizations of N_2 in the first excited state ($8 \times 5.6 \times 5.6$ a.u. box, 40 Ry plane-wave cutoff, i.e. 600 basis functions). When we include only 100 virtual orbitals in the active space, we obtain a bond length of 2.44 a.u., which deviates by 0.02 a.u. from the value of 2.42 a.u. determined by a series of single point energy calculations. Upon increasing the number of virtual states to 200, the optimized bond length comes out as 2.42 a.u., correct to two decimal places. Our test calculations illustrate how the size of the active space may be adjusted to achieve any desired level of accuracy. For many practical purposes it will be sufficient to work with a reduced active space which is significantly smaller than the total number of basis functions.

We would like to emphasize that the method described here is capable of providing additional information beyond excited state energy gradients. Figure 4 shows, for instance, the nonadiabatic coupling strength between the second and third KS orbitals, ϕ_2 and ϕ_3 , for the H_2 molecule as a function of its bond length. The nonadiabatic coupling values obtained from Eq. (60) exhibit a singularity at the crossing point between the two KS

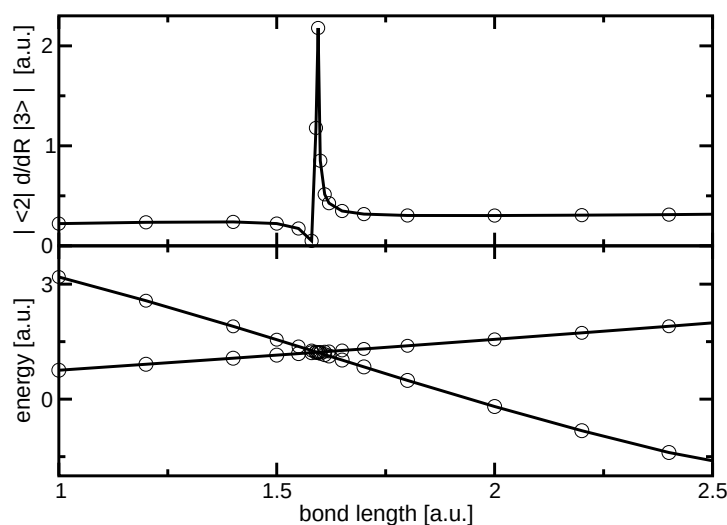


Figure 4. Upper panel: Absolute value of the nonadiabatic coupling matrix element between the KS orbitals ϕ_2 and ϕ_3 along the molecular axis of H_2 as a function of the bond length. Lower panel: KS orbital energies, ϵ_2 and ϵ_3 of H_2 as a function of bond length. The calculations were carried out in a periodic orthorhombic box of size $8 \times 5.6 \times 5.6$ a.u.³ using a plane wave cutoff of 40 Ry.

orbital energies, as one would expect due to the KS energy difference in the denominator. This feature of our formalism may be exploited in future applications of TDDFT beyond the Born-Oppenheimer approximation.

Acknowledgements

I am grateful to Daniel Kosov and Michiel Sprik for helpful discussions.

References

1. T. Ziegler, A. Rauk, and E. J. Baerends. *Theor. Chim. Acta*, 43:261, 1977.
2. C. Daul. *Int. J. Quantum Chem.*, 52:867, 1994.
3. T. Ziegler. *Chem. Rev.*, 91:651, 1991.
4. U. von Barth. *Phys. Rev. A*, 20:1693, 1979.
5. B. I. Dunlap. *Chem. Phys.*, 125:89, 1988.
6. A. C. Stückl, C. A. Daul, and H. U. Güdel. *Int. J. Quantum Chem.*, 61:579, 1997.
7. I. Frank, J. Hutter, D. Marx, and M. Parrinello. *J. Chem. Phys.*, 108:4060, 1998.
8. J. Gräfenstein, E. Kraka, and D. Cremer. *Chem. Phys. Lett.*, 288:593, 1998.
9. J. Gräfenstein and D. Cremer. *Phys. Chem. Chem. Phys.*, 2:2091, 2000.
10. M. Filatov and S. Shaik. *Chem. Phys. Lett.*, 288:689, 1998.
11. M. Filatov and S. Shaik. *J. Chem. Phys.*, 110:116, 1999.
12. M. Filatov and S. Shaik. *Chem. Phys. Lett.*, 304:429, 1999.

13. S. Grimm, C. Nonnenberg, and I. Frank. *J. Chem. Phys.*, 119:11574, 2003. *ibid.* **119**, (2003) 11585.
14. E. Runge and E. K. U. Gross. *Phys. Rev. Lett.*, 52:997, 1984.
15. J. Theilhaber. *Phys. Rev. B*, 46:12990, 1992.
16. E. K. U. Gross, J. F. Dobson, and M. Petersilka. In R. F. Nastajewski, editor, *Density Functional Theory II*, Topics in Current Chemistry. Springer, Heidelberg, 1996.
17. U. Saalman and R. Schmidt. *Z. Phys. D*, 38:153, 1996.
18. U. Saalman and R. Schmidt. *Phys. Rev. Lett.*, 80:3213, 1998.
19. T. Kunert and R. Schmidt. *Phys. Rev. Lett.*, 86:5258, 2001.
20. B. Torralva, T. A. Niehaus, M. Elstner, S. Suhai, Th. Frauenheim, and R. E. Allen. *Phys. Rev. B*, 64:153105, 2001.
21. M. E. Casida. In D. P. Chong, editor, *Recent Advances in Density-Functional Methods*, page 155. World Scientific, Singapore, 1995.
22. R. Bauernschmitt and R. Ahlrichs. *Chem. Phys. Lett.*, 256:454, 1996.
23. M. A. L. Marques and E. K. U. Gross. *Annu. Rev. Phys. Chem.*, 55:427, 2004.
24. K. Burke, J. Werschnik, and E. K. U. Gross. *J. Chem. Phys.*, 123:062206, 2005.
25. N. T. Maitra, K. Burke, H. Appel, E. K. U. Gross, and R. van Leeuwen. In K. D. Sen, editor, *Reviews in Modern Quantum Chemistry: A Celebration of the Contributions of R. G. Parr*, volume 2, page 1186. World-Scientific, Singapore, 2002.
26. D. J. Tozer and N. C. Handy. *Phys. Chem. Chem. Phys.*, 2:2117, 2000.
27. O. Gritsenko and E. J. Baerends. *J. Chem. Phys.*, 121:655, 2004.
28. L. Bernasconi, M. Sprik, and J. Hutter. *Chem. Phys. Lett.*, 394:141, 2004.
29. A. Dreuw, J. L. Weisman, and M. Head-Gordon. *J. Chem. Phys.*, 119:2943, 2003.
30. M. E. Casida, F. Gutierrez, J. Guan, F.-X. Gadea, D. Salahub, and J.-P. Daudey. *J. Chem. Phys.*, 113:7062, 2000.
31. D. J. Tozer and N. C. Handy. *J. Chem. Phys.*, 109:10180, 1998.
32. M. E. Casida, C. Jamorski, K. C. Casida, and D. R. Salahub. *J. Chem. Phys.*, 108:4439, 1998.
33. M. E. Casida and D. R. Salahub. *J. Chem. Phys.*, 113:8918, 2000.
34. S. Hirata, C.-G. Zhan, E. Aprà, T. L. Windus, and D. A. Dixon. *J. Phys. Chem. A*, 107:10154, 2003.
35. S. Hirata and M. Head-Gordon. *Chem. Phys. Lett.*, 302:375, 1999.
36. H. Appel, E. K. U. Gross, and K. Burke. *Phys. Rev. Lett.*, 90:043005, 2003.
37. P. R. T. Schipper, O. V. Gritsenko, S. J. A van Gisbergen, and E. J. Baerends. *J. Chem. Phys.*, 112:1344, 2000.
38. H. H. Neize, A. Görling, and N. Rösch. *J. Chem. Phys.*, 113:2088, 2000.
39. C. van Caillie and R. D. Amos. *Chem. Phys. Lett.*, 308:249, 1999.
40. C. van Caillie and R. D. Amos. *Chem. Phys. Lett.*, 317:159, 2000.
41. F. Furche and R. Ahlrichs. *J. Chem. Phys.*, 117:7433, 2002.
42. J. Hutter. *J. Chem. Phys.*, 118:3928, 2003.
43. N. L. Doltsinis and M. Sprik. *Chem. Phys. Lett.*, 330:563, 2000.
44. N. L. Doltsinis and D. Kosov. *J. Chem. Phys.*, 122:144101, 2005.
45. E. R. Bittner and D. S. Kosov. *J. Chem. Phys.*, 110:6645, 1999.
46. P. Hohenberg and W. Kohn. *Phys. Rev. B*, 136:864, 1964.
47. E. K. U. Gross and W. Kohn. *Adv. Quant. Chem.*, 21:255, 1990.
48. M. Petersilka, U. J. Grossmann, and E. K. U. Gross. *Phys. Rev. Lett.*, 76:1212, 1996.

49. P. Jørgensen and J. Simons. *Second Quantization Based Methods in Quantum Chemistry*. Academic, New York, 1981.
50. CPMD 3.4: J. Hutter, P. Ballone, M. Bernasconi, P. Focher, E. Fois, S. Goedecker, D. Marx, M. Parrinello, and M. Tuckerman; MPI für Festkörperforschung, Stuttgart and IBM Zurich Research Laboratory.
51. N. Troullier and J. L. Martins. *Phys. Rev. B*, 43:1993, 1991.
52. C. Jamorski, M. E. Casida, and D. R. Salahub. *J. Chem. Phys.*, 104:5134, 1996.
53. S. B. Ben-Shlomo and U. Kaldor. *J. Chem. Phys.*, 92:3680, 1990.
54. J. Oddershede, N. E. Grüner, and G. H. F. Diercksen. *Chem. Phys.*, 97:303, 1985.
55. M. Merchán and B. O. Roos. *Theor. Chim. Acta*, 92:227, 1995.
56. D. J. Clouthier and D. A. Ramsay. *Annu. Rev. Phys. Chem.*, 34:31, 1983.
57. S. Taylor, D. G. Wilden, and J. Comer. *Chem. Phys.*, 70:291, 1982.
58. J. M. Gingell, N. J. Mason, H. Zhao, I. C. Walker, and M. R. F. Siggel. *Chem. Phys.*, 220:191, 1997.
59. H. D. Hunt and W. T. Simpson. *J. Am. Chem. Soc.*, 75:4540, 1953.
60. N. C. Handy. private communication.
61. L. Serrano-Andrés and M. P. Fülscher. *J. Am. Chem. Soc.*, 118:12199, 1996.

Free Energy and Rare Events in Molecular Dynamics

Nikos L. Doltsinis

Chair of Theoretical Chemistry
Ruhr-Universität Bochum
44780 Bochum, Germany
E-mail: nikos.doltsinis@theochem.rub.de

Molecular dynamics (MD) preferentially samples regions of configuration space close to potential minima while transition state regions at the top of high reaction barriers are rarely visited. In this lecture, we will discuss a variety of improved sampling techniques that make possible the study of rare events on the finite MD time scale and the calculation of free reaction energies.

1 Introduction

Any kind of molecular dynamics (MD) simulation, be it classical (i.e. using force fields) or *ab initio*, suffers from the same fundamental limitations due to the finiteness of computational resources. Although the arguments and computational techniques presented in this article apply to both types of MD, our perspective is that of a first principles dynamicist.

Despite continuous advances in both numerical efficiency and computer technology, Car-Parrinello molecular dynamics (CP-MD) simulations^{1,2} are still limited to processes involving a few hundred atoms and occurring within a few tens of picoseconds at most. The number of chemical reactions which take place spontaneously on such a short time scale, however, is fairly limited; typically energy barriers of many kcal/mol need to be overcome. A wide variety of computational approaches have been developed over the years to force or speed up chemical reactions in *ab initio* molecular dynamics and calculate free energies.

The fundamental problem in moving from reactants to products is that the true reaction coordinate is often unknown. Choosing a realistic reaction coordinate is essential for a simulation of rare events to be useful. This task becomes exceedingly difficult for systems with large numbers of nuclear degrees of freedom. A way of systematically determining an ensemble of reaction coordinates has been suggested by Chandler and co-workers^{3,4} based on pioneering work by Pratt⁵. Although powerful and appealing, their transition path sampling method is computationally rather demanding, in particular when used in combination with CP-MD⁶. In the latter context, a variety of alternative methods has been proposed over the years. One class of approaches is based on geometric constraints, such as simple atomic distances^{7,8}, collective “target” distance⁹⁻¹² or coordination number¹³⁻¹⁶ constraints. The strategy of another class of methods is to suitably modify the potential energy surface to enhance sampling of rare events. Chemical flooding¹⁷, adiabatic free energy sampling^{18,19} and non-Markovian metadynamics²⁰⁻²⁴ all fall into this category.

The present article will cover aspects of both classes of methods, the focus being on the frequently used constraint techniques. After a general introduction to the constraint formalism in MD, we will describe in some detail two specific constraint approaches, the coordination constraint and the targeted MD (TMD) methods. In the last part of this text, the method of non-Markovian metadynamics shall be presented.

In a limited number of cases, the reaction coordinate can be approximated reasonably well by simple geometric variables, such as bond lengths or angles. Rare events of this type may be studied using standard constraint techniques^{25–28}, provided that one is able to guess the reaction mechanism prior to the simulation.

For cases where the reaction coordinate cannot be well approximated by such simple constraints, i.e. for complex reactions involving a large number of atoms, there exist a number of more advanced constraint methods. The coordination number constraint, for instance, was designed by Sprik^{13–16} to describe chemical reactions in solution. Despite being a simple scalar order parameter, the coordination constraint is capable of acting on all atoms in the system. Thus it ensures a large degree of intrinsic flexibility while retaining the numerical simplicity of a simple bond constraint.

Similarly, the TMD approach proposed by Schlitter and co-workers^{9–11} can be treated using the well-established numerical techniques developed for standard distance constraints without the need to know in advance a good low-dimensional approximation to the reaction coordinate. In the TMD method, the reaction coordinate is defined by a single mass-weighted root-mean-square “target distance” between a known initial structure and a fixed final (target) structure. By gradually reducing the constrained target distance to zero, the system is driven from the reactant to product state without explicitly defining the reaction pathway.

The method of non-Markovian metadynamics^{20–23} is based on the idea that the system can be forced to sample higher-energy regions of configuration space by gradually filling up the wells of the potential energy surface. This is achieved by adding an artificial Gaussian-shaped contribution to every region that is visited during the simulation thereby driving the system out of one local minimum to a neighboring local minimum.

2 Constraint Techniques

2.1 Basic Theory

2.1.1 Constrained Molecular Dynamics

The Lagrangian of an unconstrained system consisting of N atoms with positions $\mathbf{R} = (\mathbf{R}_1, \mathbf{R}_2, \dots, \mathbf{R}_K, \dots, \mathbf{R}_N)$ and masses M_K is

$$\mathcal{L} = \mathcal{T} - \mathcal{V} = \frac{1}{2} \sum_K^N M_K \dot{\mathbf{R}}_K^2 - \mathcal{V}(\mathbf{R}) \quad (1)$$

where $\mathcal{T}$ and $\mathcal{V}$ are the kinetic and potential energies, respectively. The equations of motion

$$\frac{\partial}{\partial t} \frac{\partial \mathcal{L}}{\partial \dot{R}_K^\alpha} = \frac{\partial \mathcal{L}}{\partial R_K^\alpha} \quad (2)$$

yield

$$M_K \ddot{R}_K^\alpha = -\frac{\partial \mathcal{V}}{\partial R_K^\alpha}, \quad K = 1, \dots, N; \alpha = x, y, z \quad (3)$$

Let us now introduce a number, n , of holonomic (only coordinate dependent) constraints

$$\sigma_i(\mathbf{R}) = \xi_i(\mathbf{R}) - \xi_i', \quad i = 1, \dots, n \quad (4)$$

where the $\xi_i(\mathbf{R})$ are geometric variables, e.g. bond lengths, bond angles, or dihedral angles, and the ξ'_i are their respective prescribed values. The Lagrangian of the constrained system is

$$\mathcal{L}' = \mathcal{L} - \sum_i^n \lambda_i \sigma_i \quad (5)$$

with a set of n undetermined Lagrange multipliers λ_i . The corresponding equations of motion are

$$\frac{\partial}{\partial t} \frac{\partial \mathcal{L}'}{\partial \dot{R}_K^\alpha} = \frac{\partial \mathcal{L}'}{\partial R_K^\alpha} \quad (6)$$

or

$$M_K \ddot{R}_K^\alpha = -\frac{\partial \mathcal{V}}{\partial R_K^\alpha} - \sum_i^n \lambda_i \frac{\partial \sigma_i}{\partial R_K^\alpha}, \quad K = 1, \dots, N; \alpha = x, y, z \quad (7)$$

A commonly used iterative scheme to solve the constrained equations of motion (Eq. (7)) in combination with the Verlet propagation algorithm has been named SHAKE^{25,29}. Alternatively, constrained dynamics may be formulated within the Hamiltonian framework³⁰ in which case one can derive the RATTLE algorithm³¹ to be used in conjunction with the velocity Verlet propagator.

2.1.2 Free Energy by Thermodynamic Integration

The free energy difference between two states, ξ^1 and ξ^2 , can be expressed in terms of the reversible work required to move from ξ^1 to ξ^2 along the reaction coordinate ξ ^{13,27,32,33,28}

$$\Delta W = W(\xi^2) - W(\xi^1) = \int_{\xi^1}^{\xi^2} d\xi' \left\langle \frac{\partial \mathcal{H}}{\partial \xi} \right\rangle_{\xi'}^{\text{cond.}} \quad (8)$$

where mean force

$$\left\langle \frac{\partial \mathcal{H}}{\partial \xi} \right\rangle_{\xi'}^{\text{cond.}} = \frac{\left\langle \frac{\partial \mathcal{H}}{\partial \xi} \delta(\xi - \xi') \right\rangle}{\langle \delta(\xi - \xi') \rangle} \quad (9)$$

is the conditional ensemble average evaluated at $\xi = \xi'$ of the generalized force, i.e. the derivative of the Hamiltonian $\mathcal{H}$ with respect to the reaction coordinate ξ .

Using the so-called blue moon ensemble method²⁶ it is possible to calculate conditional averages from time averages over constrained MD trajectories at fixed values $\xi = \xi'$,

$$\left\langle \frac{\partial \mathcal{H}}{\partial \xi} \right\rangle_{\xi'}^{\text{cond.}} = \frac{\left\langle Z^{-\frac{1}{2}} \left[\frac{\partial \mathcal{V}}{\partial \xi} - kT \frac{\partial \ln |\mathbf{J}|}{\partial \xi} \right] \right\rangle_{\xi'}}{\left\langle Z^{-\frac{1}{2}} \right\rangle_{\xi'}} \quad (10)$$

where the factor

$$Z = \sum_K^N \frac{1}{M_K} \left(\frac{\partial \xi}{\partial \mathbf{R}_K} \right)^2 \quad (11)$$

corrects for the configurational bias introduced by the constraint. The Jacobian matrix, $\mathbf{J}$, for the coordinate transformation from Cartesian coordinates, $\mathbf{R}$, to generalized coordinates, $\mathbf{u}$, including the reaction coordinate ξ , $\mathbf{R} \leftrightarrow \mathbf{u} \equiv \{\xi, \mathbf{q}\}$, is defined by

$$J_{K\kappa}^\alpha = \frac{\partial R_K^\alpha}{\partial u_\kappa} . \quad (12)$$

In practice, the rhs of Eq. (10) is difficult to evaluate because of the derivatives with respect to ξ . The problem can be simplified considerably by exploiting the fact that the Lagrange multiplier associated with the constraint (cf. Eq. (5)) is equal to the generalized force³⁴,

$$-\frac{\partial \mathcal{H}}{\partial \xi} = \lambda . \quad (13)$$

This immediately follows from the Hamiltonian equation of motion

$$\dot{p}^\xi = -\frac{\partial \mathcal{H}}{\partial \xi} - \lambda \quad (14)$$

realizing that the momentum conjugate, $\dot{p}^\xi$, to the constrained variable ξ vanishes. Applying the blue moon unbiasing procedure, the mean force (Eq. (9)) can be expressed in terms of the constrained average of the Lagrange multiplier^{27, 13, 32, 33, 28},

$$\left\langle \frac{\partial \mathcal{H}}{\partial \xi} \right\rangle_{\xi'}^{\text{cond.}} = \frac{\langle Z^{-\frac{1}{2}} (\lambda - kTG) \rangle_{\xi'}}{\langle Z^{-\frac{1}{2}} \rangle_{\xi'}} \quad (15)$$

with

$$G = \frac{1}{Z^2} \sum_{K,L}^N \frac{1}{M_K M_L} \frac{\partial \xi}{\partial \mathbf{R}_K} \frac{\partial^2 \xi}{\partial \mathbf{R}_K \partial \mathbf{R}_L} \frac{\partial \xi}{\partial \mathbf{R}_L} . \quad (16)$$

2.2 Specific Reaction Coordinates

2.2.1 Coordination Constrained Molecular Dynamics

For the study of chemical reactions in solutions, the coordination number of a specific atom $K = C$,

$$\xi = n_C(\mathbf{R}) = \sum_{K \neq C}^N [\exp[\kappa(r_K - r_c)] + 1]^{-1} \quad (17)$$

has proven particularly useful¹³. Here, r_K is the distance of atom K from atom C ; the coordination radius r_c and the width of the Fermi function κ are parameters to be set prior to the simulation.

This coordination constraint has been applied in a number of CP-MD studies of proton transfer^{14–16, 12, 35} mostly in aqueous solution. In the following, we shall discuss the example of $\text{P}(\text{OH})_5$ in liquid water^{15, 16}.

The study of pentacoordinated phosphorus compounds has played a key role in the attempt to unravel the mechanism of the hydrolysis reaction of RNA^{36–39, 15}. Experimental evidence on the first step of RNA hydrolysis — so-called cleavage transesterification — has suggested two possible reaction pathways involving a monoanionic phosphorane

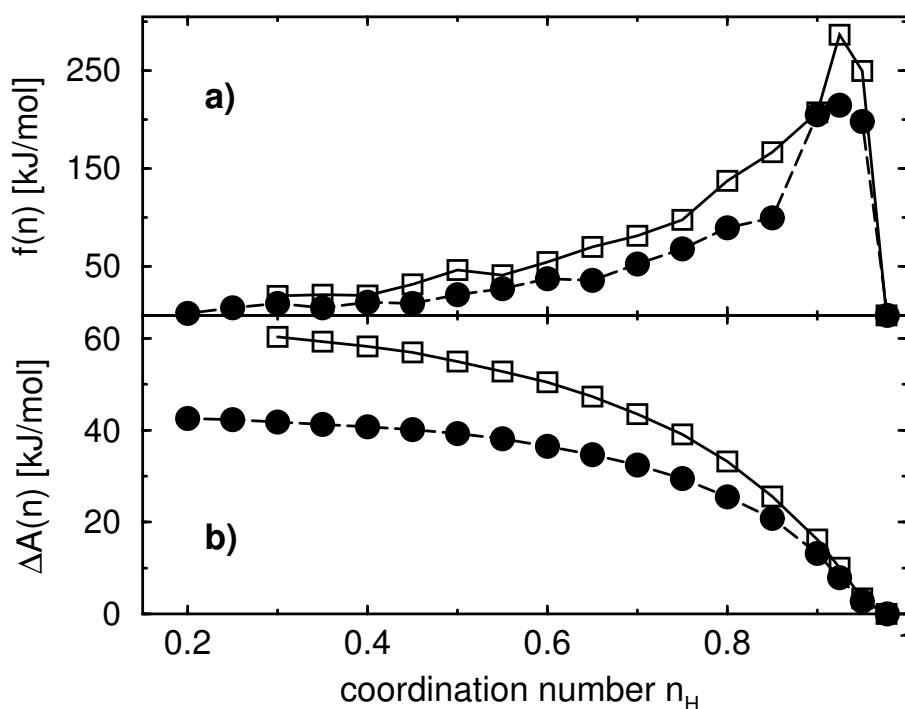


Figure 1. a) Mean force of constraint as a function of the coordination number n for the axial ($\square$) and the equatorial ($\bullet$) site b) corresponding free energy curves obtained by integration of the mean forces.

transition state and a dianionic phosphorane intermediate, respectively^{37,38}. Determination of the phosphorane's lifetime and protonation state is, therefore, of prime importance in establishing a detailed picture of the hydrolysis reaction mechanism. Knowledge of the relevant pK_a values would thus provide vital clues.

The essential quantity required in order to calculate the pK_a value of a given molecule is the Helmholtz free energy difference, ΔA , for hydrogen abstraction. This quantity can be extracted from constrained molecular dynamics by thermodynamic integration provided the control parameter is a reasonable approximation to the reaction coordinate. In high dimensional systems, a simple distance constraint often yields a rather poor approximation to the true reaction coordinate, and when used to locate transition states can lead to erroneous results^{40,4}. Simple geometric order parameters, however, can still be suitable as integration variables for the determination of free energy difference between well defined reactant and product states distance. An additional problem is encountered in protic solutions; when the dissociating proton has reached a critical distance from its donor atom, the latter is vulnerable to reprotonation by a different solvent proton. This difficulty is circumvented if we enforce deprotonation by gradually decreasing the coordination number of the oxygen atom of a specific hydroxyl group.

In two separate series of coordination constrained CP-MD simulations, we have determined the free energy for deprotonation of an axial and an equatorial hydroxyl group of

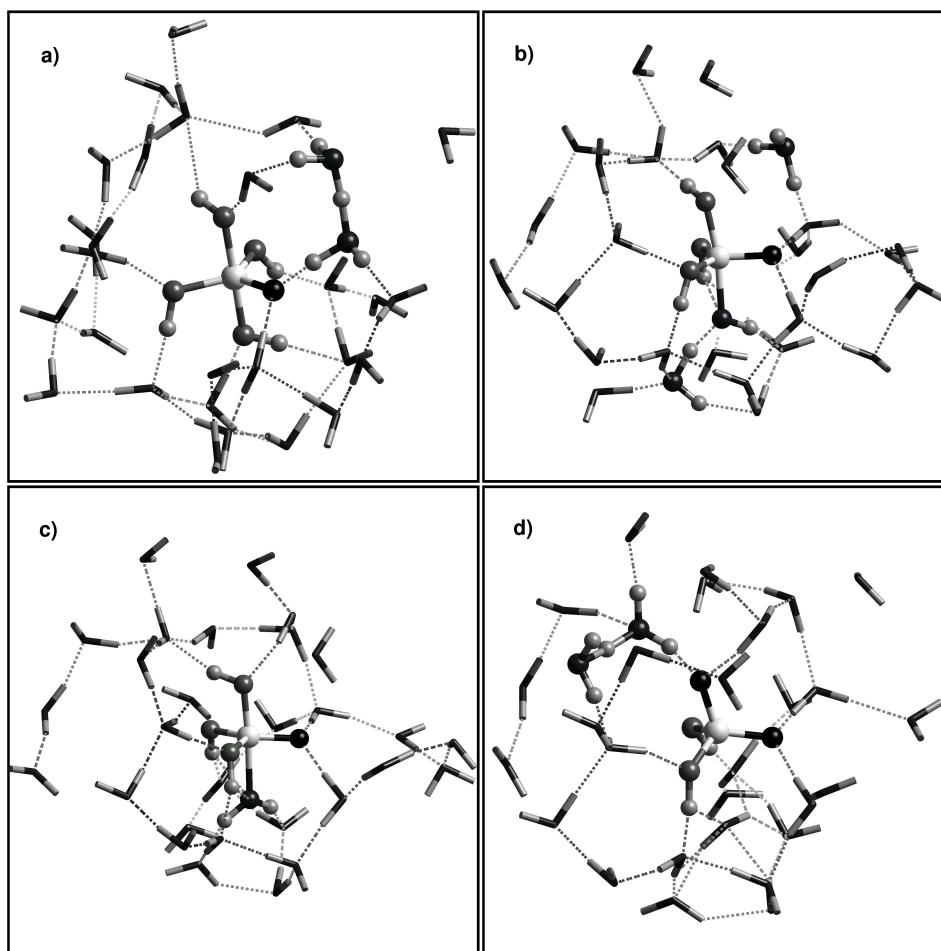


Figure 2. Sequence of snapshots illustrating a dehydration event: a) application of coordination constraints leads to the formation of H_5O_2^+ . b) a H_3O^+ ion breaks loose and subsequently donates a proton to the axial hydroxyl group of the phosphorane through a chain reaction of proton transfers. c) the axial hydroxyl group has been protonated and is about to break its PO bond. d) the remaining phosphoric acid loses a proton to the solvent.

the trigonally bipyramidal $\text{P}(\text{OH})_5$ respectively. The value of n (cf. Eq. (17)) was gradually decreased towards zero starting near its equilibrium value of approximately unity, thus transferring a proton from the hydroxyl group $\text{O}_\text{D}\text{H}$ to the acceptor solvent water molecule $\text{H}_2\text{O}_\text{A}$. For each, fixed, value of the coordination number a trajectory of roughly 2 to 3 ps length was computed. In Figure 1a the blue-moon corrected average constraint force (Eq. (15)), i.e. the mean force, is plotted as a function of the coordination number for both the axial and the equatorial site. The equilibrium value of n was determined to be 0.98 in both cases from the trajectory of unconstrained $\text{P}(\text{OH})_5$. In the region just below the equilibrium value of n we observe a steep rise of the mean force reaching a maximum at $n = 0.925$. As expected, the restoring force resisting proton abstraction is

significantly higher for an axial proton than for an equatorial proton, the mean forces being 286.6 kJ/mol and 214.6 kJ/mol, respectively. With further decreasing coordination number the mean force is seen to fall off towards zero and we observe the formation of a H_5O_2^+ ion (Zundel ion) (Figure 2a). The smallest value of n for which a meaningful trajectory could be obtained was 0.2 in the case of the equatorial site and 0.3 for the axial position. For smaller n the H_5O_2^+ ion breaks up into a hydronium ion and a water molecule H_2O_A which has accepted the proton from the phosphorane and donated another proton to its nearest neighbor. The newly created H_3O^+ ion is now free to attack the negatively charged phosphorane. Figure 2b-c illustrates how an axial hydroxyl group of the phosphorane is protonated turning it into a H_2O^+ group initially. The corresponding PO bond is subsequently broken and a neutral water molecule leaves the phosphorane molecule. Since we are now dealing with phosphoric acid, H_3PO_4 , in water, it is not surprising that we observe spontaneous deprotonation within femtoseconds (Figure 2d). This dehydration process is in competition with the deprotonation reaction we intend to enforce by using coordination constraints and makes it impossible to calculate any trajectories at very small values of n without dissociation of the phosphorane.

The free energy difference ΔA between two values of the coordination number can be obtained by integration of the mean force. In Figure 1b the resulting free energy curves for axial and equatorial hydroxyl groups are plotted taking equilibrium as the reference point. The equatorial curve ends at a free energy difference of 42.7 kJ/mol for $n = 0.2$ well below the axial curve ending at $\Delta A = 60.2$ kJ/mol for $n = 0.3$. Following the procedure outlined in Refs. 15, 16 we have then determined the $\text{p}K_a$ values to be 9.8 and 14.2 for equatorial and axial hydroxyl groups, respectively. Very recent experimental estimates based on a bond length – $\text{p}K_a$ correlation⁴¹ give 8.62 ± 1.87 and 13.5 ± 1.5 for the equatorial and axial OH groups of tetracyclohexyloxyhydroxyphosphorane¹⁵. The value of 7.9 calculated by Lopez et al.³⁹ from an empirical reaction field model for the equatorial OH group of ethylene phosphorane is substantially lower than our estimate. As we have pointed out above, however, explicit treatment of the solvent may make a crucial difference in the current context. Moreover, we should bear in mind that all three $\text{p}K_a$ estimates compared here have been obtained for *different* (although in many respects certainly similar) phosphoranes.

2.2.2 Targeted Molecular Dynamics

A detailed description of Targeted MD (TMD) including a discussion of the mechanical and statistical properties of the constraint is available in the literature^{9–11}. Here, we simply provide a brief outline including the most important equations.

In TMD, the reaction coordinate is defined as the target distance

$$D(\mathbf{R}) = \sqrt{\sum_{K=1}^N \frac{M_K}{M} (\mathbf{R}_K - \mathbf{T}_K)^2} \quad (18)$$

between the instantaneous reactant structure $\mathbf{R} = \mathbf{R}(t)$ and the fixed product (target) structure $\mathbf{T} = \{\mathbf{T}_K\}$, where M is the total mass of the system. In order to eliminate any translational or rotational contributions, the target distance $D(\mathbf{R})$ is minimized by superimposing the centres-of-mass of the two configurations, $\mathbf{R}$ and $\mathbf{T}$, followed by rotation of

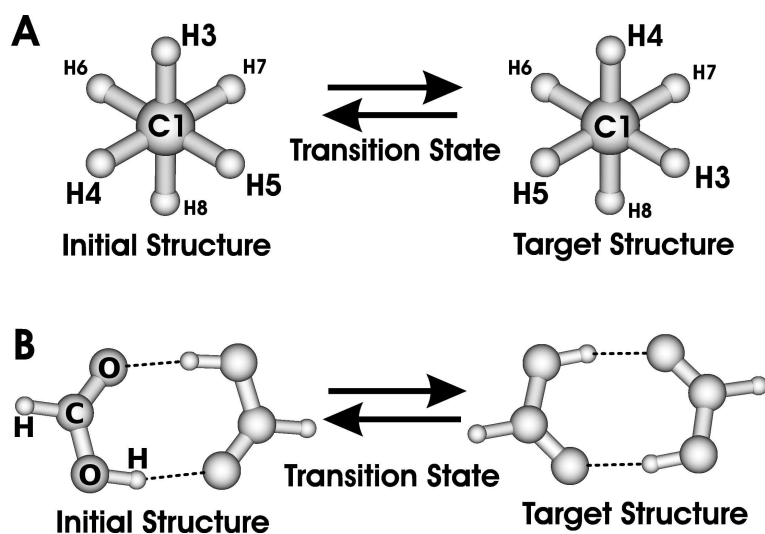


Figure 3. (A) Initial and target ethane structures using the staggered conformation (rotation by 120°) as the target structure (see text for more details). (B) Initial and target FAD structures.

the target structure. The remaining distance, called the target distance, is a measure of the structural root mean square distance between the two molecular conformations. During the course of the molecular dynamics simulation, this distance is gradually reduced towards zero and the initial reactant structure is driven towards the target structure.

A particularly attractive feature of this approach is the fact that the time averaged constraint force (Eq. (15)) can be obtained from a TMD run at a fixed value of D as the average Lagrange multiplier, i.e.

$$\left\langle \frac{\partial \mathcal{H}}{\partial \xi} \right\rangle_{\xi'}^{\text{cond.}} = \langle \lambda \rangle_{\xi'} \quad (19)$$

without the need to correct for metric tensor effects.

TMD using classical force fields has recently been employed with considerable success to study conformational transitions and folding in proteins^{42–47}. In the following, however, we will present results from the first TMD study based on Car-Parrinello MD (T-CP-MD). In this study, two test cases were considered, rotation of ethane about the C–C bond and double proton transfer (DPT) in formic acid dimer (FAD). Figure 3 provides a pictorial representation of the initial and target structures for the two systems investigated by T-CP-MD.

The average constraint force for the staggered-to-eclipsed-to-staggered conformational transition (rotation by 120°) in ethane is shown in Figure 4 as a function of the target distance including the associated free energy profile. The initial target distance between the two staggered conformations is $D = 0.88$ a.u. corresponding to a rotational angle about the C–C axis of 120° . At this initial D , the average constraint force is approximately zero. As D decreases, the constraint force increases as the constraint pushes the system towards the energetically unfavourable eclipsed conformation (rotation by 60°). Having

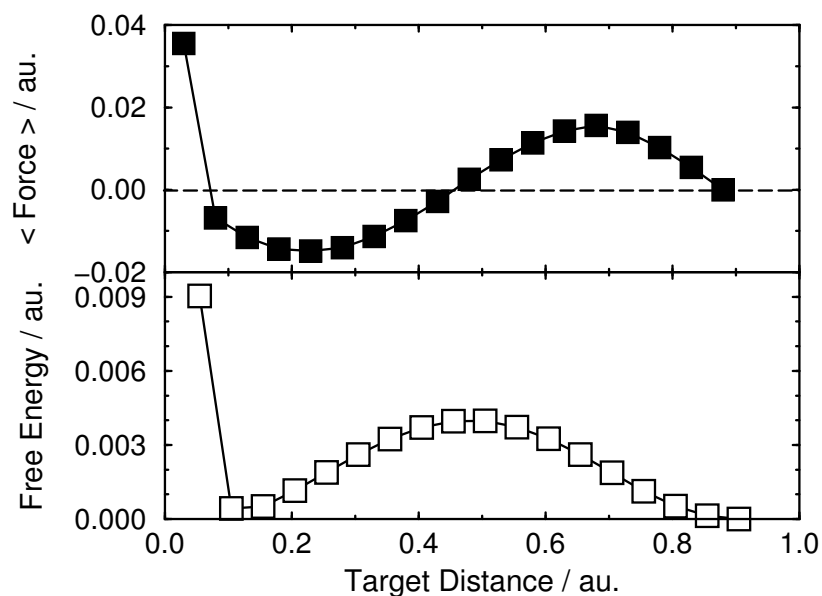


Figure 4. Variation of the average constraint force (top) and free energy profile (bottom) as a function of D for the staggered-eclipsed-staggered transition in ethane using T-CP-MD; see Figure 3(A).

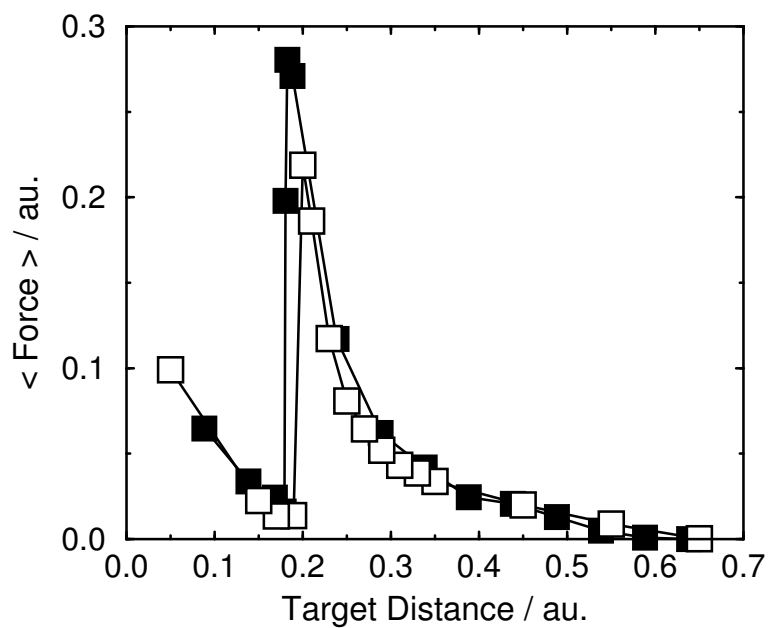


Figure 5. Variation in the average constraint force as a function of D for FAD. Filled squares represent the average constraint force for the 'full' TMD, while the open squares show the result for the partial target structure involving only the four oxygen atoms and the two protons involved in DPT. After the DPT event ($D = 0.2$ a.u.) the constraint loses control of the system.

reached a maximum, the average constraint force then returns to zero at the transition state ($D = 0.43$ a.u.). At this point, the system is sitting exactly on the top of the potential energy barrier. Further reduction of D results in negative values of the average constraint force. This arises from the fact that, having overcome the energy barrier, the system would preferentially move directly to the eclipsed (target) structure. The constraint however does not allow this, and acts in such a way as to “hold” the system away from the target structure, resulting in a negative average constraint force. As D gets very close to zero (in this case for $D < 0.1$ a.u.), the average force adopts exceedingly large positive values. This is because the phase space volume available to the system decreases with D resulting in an increasing centrifugal component of the constraint force¹⁰. In other words, when the available phase space becomes so small that it confines the vibrations of the system, the average constraint force increases dramatically as a result of this “entropy loss”. The free energy profile resembles a Gaussian form, with a maximum at approximately the midpoint of the reaction coordinate. Indeed the resulting free energy profile is very similar to that obtained from the MEP.

For the study of DPT in FAD, we performed two series of T-CP-MD simulations. In the first series of simulations we used a ‘full’ target constraint including all atoms of the molecule, whilst in the second simulation we employed a “partial” target constraint, only using for the evaluation of D those atoms that are predominantly involved in the DPT reaction, i.e. the four oxygen atoms and two protons constituting the two H-bonds. The resulting average constraint force curves are depicted in Figure 5. In both cases, we see a very different force curve compared to that obtained in the ethane simulation. The average constraint force increases slowly to a maximum as the DPT event is reached. Just after DPT, the average constraint force rapidly decreases to zero. At very small D , the average constraint force begins to increase again, due to the effect of the target constraint entropy. It is immediately apparent that in FAD, the constraint “loses control” of the system after the DPT reaction has occurred. We consider that this is partially due to the fact that FAD is much more flexible than ethane and partially due to the fact that the TMD constraint is not “focussed” directly on the protons involved in DPT, that is the mass-weighted target distance is dominated by the heavy atom frame.

3 Metadynamics

Another common strategy to accelerate barrier crossings consists in modifying the potential energy landscape in such a way as to “fill up” potential minima^{48–50, 17, 51, 52}. In this section, we will describe a particular method of this type which has been named metadynamics^{20–24, 53}.

Prior to the simulation one has to define a number, n , of reaction coordinates $\xi_i(\mathbf{R})$ characterizing the transitions between several local minima of the potential energy surface. For each reaction coordinate ξ_i an associated dynamical variable Ξ_i is introduced, component of the vector Ξ . The Lagrangian of the extended system is

$$\tilde{\mathcal{L}} = \mathcal{L} + \frac{1}{2} \sum_i^n \mu_i \dot{\Xi}_i^2 - \frac{1}{2} \sum_i^n k_i [\xi_i(\mathbf{R}) - \Xi_i]^2 - V(t, \Xi) \quad (20)$$

where $\mathcal{L}$ is the standard Lagrangian underlying the MD, in the simplest case (Eq. (1)), μ_i and k_i are the fictitious mass and the coupling constant associated with the reaction

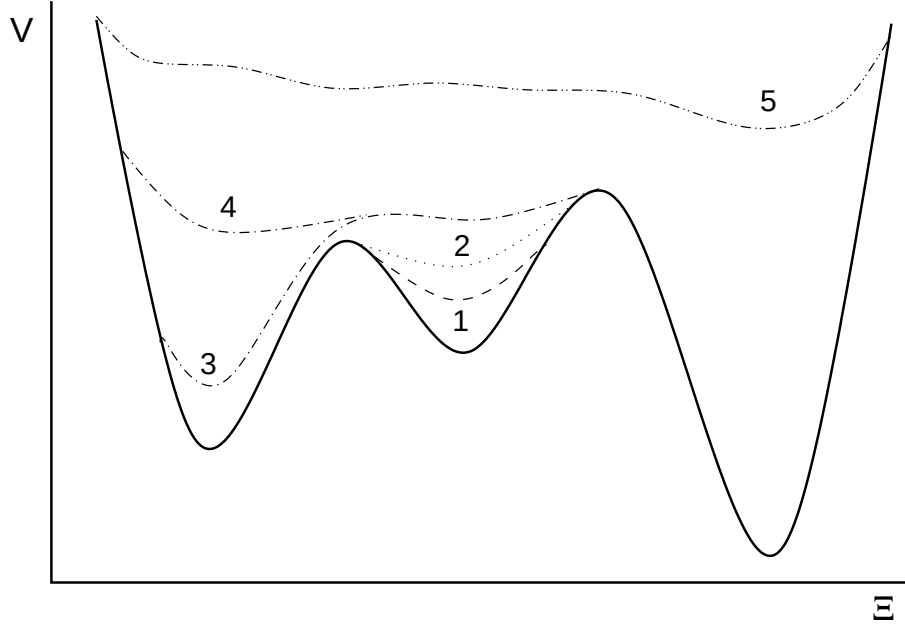


Figure 6. Schematic illustration of the metadynamics history-dependent potential $V(t, \Xi)$ as a function of time (curves labelled 1–5 in the order of increasing simulation time) added to the potential energy surface (solid line). The simulation started in the central minimum before the system is driven to the left minimum and finally to the minimum on the right.

coordinate ξ_i . The history-dependent biasing potential is defined as

$$V(t, \Xi) = \sum_{t_k < t} W_k \exp \left[-\frac{(\Xi - \Xi^k)^2}{2(\Delta \Xi^\perp)^2} \right] \exp \left[-\frac{[(\Xi^{k+1} - \Xi^k)(\Xi - \Xi^k)]^2}{2(\Delta \Xi^\parallel)^4} \right] \quad (21)$$

with $\Xi = \Xi(t)$ and $\Xi^k = \Xi(t_k)$; $\Delta \Xi_k^\parallel = |\Xi^{k+1} - \Xi^k|$ and $\Delta \Xi^\perp$ determine the width of the Gaussian in the direction perpendicular and parallel to the motion of Ξ . In earlier publications on metadynamics^{21,22}, the prefactor W_k was updated according to

$$W_k = \lambda \sum_i^n (\Xi_i^{k+1} - \Xi_i^k) \langle k_i [\xi_i(\mathbf{R}) - \Xi_i] \rangle \quad (22)$$

counterbalancing the restoring force contribution from the free energy surface. The average $\langle \rangle$ is taken over the time interval $\Delta t = t_{k+1} - t_k$ (typically 10–100 times larger than the MD time step) between two updates of the potential $V(t, \Xi)$ and λ is chosen smaller than unity²¹. In some recent works, a constant prefactor $W_k = W$ was found to be more advantageous^{53,24}.

As the simulation time approaches infinity, the history-dependent potential $V(t, \Xi)$ has filled all the potential wells of the free energy surface and represents the negative of the latter, i.e., using the same notation as above (cf. Eq. (8)),

$$\lim_{t \rightarrow \infty} V(t, \Xi) = -W(\Xi) + \text{const.} \quad (23)$$

A schematic one-dimensional illustration of the time evolution of the biasing potential $V(t, \Xi)$ for a typical potential energy surface can be seen in Figure 6. The Gaussian-shaped potential hills (Eq. (21)) drive the system away from previously visited regions of configuration space. This occurs in an indirect way by propagating the fictitious dynamical variables Ξ_i to which the actual reaction coordinates $\xi_i(\mathbf{R})$ are coupled through the third (harmonic) term on the rhs of Eq. (20). The parameters μ_i , k_i , W , and $\Delta\Xi^\perp$ have to be chosen carefully to ensure efficient and, at the same time, accurate exploration of the free energy surface. The width $\Delta\Xi^\perp$ of the Gaussian hills can be estimated from fluctuations of the Ξ_i in the unbiased potential, i.e. $V(t, \Xi) = 0$. The height of the individual hills, W , should be lower than kT . The fictitious masses, μ_i , and the spring constants, k_i , have to be chosen in such a way as to keep the Ξ_i close to the reaction coordinates $\xi_i(\mathbf{R})$ while maintaining adiabatic separation of fictitious and nuclear degrees of freedom.

Acknowledgements

I am grateful to Dominik Marx, Michiel Sprik, Jürgen Schlitter, Phineus Markwick, and Nisanth Nair for helpful discussions.

References

1. R. Car and M. Parrinello. *Phys. Rev. Lett.*, 55:2471, 1985.
2. D. Marx and J. Hutter. In J. Grotendorst, editor, *Modern Methods and Algorithms of Quantum Chemistry*. NIC, Jülich, 2000.
www.theochem.rub.de/go/cprev.html.
3. D. Chandler. In B. J. Berne, G. Ciccotti, and D. F. Coker, editors, *Classical and Quantum Dynamics in Condensed Phase Simulations*. World Scientific, Singapore, 1998.
4. P. G. Bolhuis, D. Chandler, C. Dellago, and P. L. Geissler. *Ann. Rev. Phys. Chem.*, 53:291, 2002.
5. L. R. Pratt. *J. Chem. Phys.*, 85:5045, 1986.
6. P. L. Geissler, C. Dellago, D. Chandler, J. Hutter, and M. Parrinello. *Science*, 291:2121, 2001.
7. A. Curioni, M. Sprik, W. Andreoni, H. Schiffer, J. Hutter, and M. Parrinello. *J. Am. Chem. Soc.*, 119:7218, 1997.
8. E. J. Meijer and M. Sprik. *J. Am. Chem. Soc.*, 120:6345, 1998.
9. J. Schlitter, M. Engels, P. Krüger, E. Jacobi, and A. Wollmer. *Mol. Sim.*, 10:291, 1993.
10. J. Schlitter, W. Swegat, and T. Mülders. *J. Mol. Mod.*, 7:171, 2001.
11. J. Schlitter and M. Klähn. *Mol. Phys.*, 101:3439, 2003.
12. P. Markwick, N. L. Doltsinis, and D. Marx. *J. Chem. Phys.*, 122:054112, 2005.
13. I. Shavitt. *Mol. Phys.*, 94:3, 1998. and references therein.
14. M. Sprik. *Chem. Phys.*, 258:139, 2000.
15. J. E. Davies, N. L. Doltsinis, A. J. Kirby, C. D. Roussev, and M. Sprik. *J. Am. Chem. Soc.*, 124:6594, 2002.
16. N. L. Doltsinis and M. Sprik. *Phys. Chem. Chem. Phys.*, 5:2612, 2003.

17. E. M. Müller, A. de Meijere, and H. Grubmüller. *J. Chem. Phys.*, 116:897, 2002.
18. J. VandeVondele and U. Röthlisberger. *J. Phys. Chem. B*, 106:203, 2002.
19. L. Rosso and M. E. Tuckerman. *Mol. Sim.*, 28:91, 2002.
20. A. Laio and M. Parrinello. *Proc. Natl. Acad. Sci. USA*, 99:12562, 2002.
21. M. Iannuzzi, A. Laio, and M. Parrinello. *Phys. Rev. Letters*, 90:238302, 2003.
22. A. Stirling, M. Iannuzzi, A. Laio, and M. Parrinello. *ChemPhysChem*, 5:1558, 2004.
23. A. Laio, A. Rodriguez-Forte, F. L. Gervasio, M. Ceccarelli, and M. Parrinello. *J. Phys. Chem. B*, 109:6714, 2005.
24. B. Ensing, A. Laio, M. Parrinello, and M. L. Klein. *J. Phys. Chem. B*, 109:6676, 2005.
25. J. P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen. *J. Comp. Phys.*, 23:327, 1977.
26. E. A. Carter, G. Ciccotti, J. T. Hynes, and R. Kapral. *Chem. Phys. Lett.*, 156:472, 1989.
27. M. Sprik and G. Ciccotti. *J. Chem. Phys.*, 109:7737, 1998.
28. G. Ciccotti and M. Ferrario. *Mol. Sim.*, 30:787, 2004.
29. G. Ciccotti and J. P. Ryckaert. *Comp. Phys. Rep.*, 4:345, 1986.
30. S. W. Deleeuw, J. W. Perram, and H. G. Petersen. *J. Stat. Phys.*, 61:1203, 1990.
31. H. C. Andersen. *J. Comp. Phys.*, 52:24, 1983.
32. G. Ciccotti and M. Ferrario. *J. Mol. Liq.*, 89:1, 2000.
33. I. Coluzza, M. Sprik, and G. Ciccotti. *Mol. Phys.*, 101:2885, 2003.
34. T. Mülders, P. Krüger, W. Swegat, and J. Schlitter. *J. Chem. Phys.*, 104:4869, 1996.
35. H. Langer, N. L. Doltsinis, and D. Marx. *ChemPhysChem*, 6:1734, 2005.
36. F. H. Westheimer. *Acc. Chem. Res.*, 1:70, 1968.
37. M. Oivanen, S. Kuusela, and H. Lonnberg. *Chem. Rev.*, 98:961, 1998.
38. D. M. Perreault and E. V. Anslyn. *Angew. Chem. Int. Ed. Engl.*, 36:432, 1997.
39. X. Lopez, M. Schneider, A. Dejaegere, and M. Karplus. *J. Am. Chem. Soc.*, 124:5010, 2002.
40. C. Dellago, P. G. Bolhuis, F. S. Csajka, and D. Chandler. *J. Chem. Phys.*, 108:1964, 1998.
41. R. D. Amos, N. C. Handy, P. G. Jones, A. J. Kirby, J. K. Parker, J. M. Percy, and D. M. Su. *J. Chem. Soc., Perkin Trans.*, 2:549, 1992.
42. B. Wroblowski, J.F. Diaz, J. Schlitter, and Y. Engelborghs. *Prot. Eng.*, 10:1163, 1997.
43. O. Roche and M.J. Field. *Prot. Eng.*, 12:285, 1999.
44. P. Fererra, J. Apostolakis, and A. Calfisch. *J. Phys. Chem. B.*, 104:4511, 2000.
45. P. Fererra, J. Apostolakis, and A. Calfisch. *Proteins*, 39:252, 2000.
46. P. Kruger, S. Verheyden, P.J. Declerck, and Y. Engelborghs. *Prot. Sci.*, 10:798, 2001.
47. Y.K. Kong, Y.F. Shen, T.E. Warth, and J.P. Ma. *Proc. Natl. Acad. Sci. USA*, 99:5999, 2002.
48. H. Grubmüller. *Phys. Rev. E*, 52:2893, 1995.
49. J. VandeVondele and U. Röthlisberger. *J. Chem. Phys.*, 113:4863, 2000.
50. J.-C. Wang, S. Pal, and K.-A. Fichthorn. *Phys. Rev. B*, 63:085403, 2001.
51. J. A. Rahman and J. C. Tully. *Chem. Phys.*, 285:277, 2002.
52. D. Hamelberg, J. Mongan, and J. A. McCammon. *J. Chem. Phys.*, 120:11919, 2004.
53. K. Jug, N. N. Nair, and T. Bredow. *Phys. Chem. Chem. Phys.*, 7:2616, 2005.

Molecular Dynamics Beyond the Born-Oppenheimer Approximation: Mixed Quantum–Classical Approaches

Nikos L. Doltsinis

Chair of Theoretical Chemistry
Ruhr-Universität Bochum
44780 Bochum, Germany
E-mail: nikos.doltsinis@theochem.rub.de

This contribution takes a closer look at the foundations of conventional molecular dynamics simulations such as the Born-Oppenheimer approximation and the treatment of atomic nuclei according to the laws of classical mechanics. Regimes of validity of the adiabatic approximation are defined and models that take into account nonadiabatic effects in situations where the Born-Oppenheimer approximation breaks down are introduced. We focus on two mixed quantum-classical methods that differ only in the way the forces on the — classical — atomic nuclei are determined from the solutions to the time-independent electronic Schrödinger equation. In the Ehrenfest approach, the system moves on a single potential energy surface obtained by weighted averaging over all adiabatic states, whereas the 'surface hopping' method allows transitions between pure adiabatic potential energy surfaces according to their weights. In both cases, the weights are the squares of the coefficients of the total electronic wavefunction expanded in terms of the adiabatic state functions.

1 Introduction

Molecular dynamics (MD), in the literal sense, is the simultaneous motion of a number of atomic nuclei and electrons forming a molecular entity. Strictly speaking, a complete description of such a system requires solving the full time-dependent Schrödinger equation including both electronic and nuclear degrees of freedom. This, however, is a formidable computational task which is in fact altogether unfeasible, at present, for systems consisting of more than three atoms and more than one electronic state^{1,2}. In order to study the dynamics of the vast majority of chemical systems, several approximations, therefore, have to be imposed.

Firstly, it is assumed in MD that the motions of slow and fast degrees of freedom are separable (adiabatic or Born-Oppenheimer approximation). In the molecular context this means that the electron cloud adjusts instantly to changes in the nuclear configuration. As a consequence, nuclear motion evolves on a *single* potential energy surface (PES), associated with a *single* electronic quantum state, which is obtained by solving the time-independent Schrödinger equation for a series of fixed nuclear geometries. In practice, most MD simulations are performed on a *ground state* PES.

Moreover, in addition to making the Born-Oppenheimer approximation, MD treats the atomic nuclei as *classical* particles whose trajectories are computed by integrating Newton's equations of motion.

MD has been applied with great success to study a wide range of systems from biomolecules to condensed phases^{3,4}. Its underlying approximations, on the other hand,

break down in many important physical situations and extensions of the method are needed for those scenarios. An accurate description of hydrogen motion, for instance, requires quantum mechanical treatment. Processes such as charge-transfer reactions and photochemistry are inherently *nonadiabatic*, i.e., they involve (avoided) crossings of different electronic states rendering the Born-Oppenheimer approximation invalid.

Critical assessment of the adiabatic approximation as well as discussion of nonadiabatic extensions will be the subject of the present paper.

Since our focus here is on potential applicability to large-scale systems, we shall retain the classical treatment of the nuclei and only describe the electrons quantum mechanically. We will use the term semiclassical for such mixed quantum-classical models. Both expressions can be frequently found in the literature.

Out of the great many semiclassical approaches to nonadiabatic dynamics that have been proposed two “standard” methods different in philosophy have emerged as the most popular ones. One extreme is the Ehrenfest method^{2,5-10}, where the nuclei move on *one* effective PES which is an average of all adiabatic states involved weighted by their populations (therefore also called mean-field method). The other extreme is the surface hopping approach^{11,12,8,9,13,14}, where the nuclei evolve on pure adiabatic PESs, but switches between adiabatic states are allowed when their populations change.

This article is organised as follows. In Section 2, the Born-Oppenheimer approximation is introduced. Starting from the full time-dependent Schrödinger equation, the uncoupled nuclear equations of motion are derived. Section 3 deals with the semiclassical approach replacing the nuclear wavefunction by a classical trajectory. This will form the basis of all nonadiabatic methods presented in later sections. Conditions for the validity of the Born-Oppenheimer approximation are discussed qualitatively. Two of the most commonly employed nonadiabatic dynamics methods are described in Section 4, namely the Ehrenfest and the surface hopping methods. The section closes by presenting a recent implementation of the surface hopping technique within the framework of Car-Parrinello MD⁵²⁻⁵⁴ together with an application to the cis-trans photoisomerisation of formalimine as a case study⁵².

2 Born-Oppenheimer Approximation

A complete, non-relativistic, description of a system of N atoms having the positions $\mathbf{R} = (\mathbf{R}_1, \mathbf{R}_2, \dots, \mathbf{R}_K, \dots, \mathbf{R}_N)$ with n electrons located at $\mathbf{r} = (\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_K, \dots, \mathbf{r}_n)$ is provided by the time-dependent Schrödinger equation

$$\mathcal{H}\Xi(\mathbf{r}, \mathbf{R}; t) = i\hbar \frac{\partial}{\partial t} \Xi(\mathbf{r}, \mathbf{R}; t) \quad , \quad (1)$$

with the total Hamiltonian

$$\mathcal{H}(\mathbf{r}, \mathbf{R}) = \mathcal{T}(\mathbf{R}) + \mathcal{T}(\mathbf{r}) + \mathcal{V}(\mathbf{R}) + \mathcal{V}(\mathbf{r}, \mathbf{R}) + \mathcal{V}(\mathbf{r}) \quad , \quad (2)$$

being the sum of kinetic energy of the atomic nuclei,

$$\mathcal{T}(\mathbf{R}) = -\frac{\hbar^2}{2} \sum_{K=1}^N \frac{\nabla_K^2}{M_K} \quad , \quad (3)$$

kinetic energy of the electrons,

$$\mathcal{T}(\mathbf{r}) = -\frac{\hbar^2}{2m_e} \sum_{k=1}^n \nabla_k^2 \quad , \quad (4)$$

internuclear repulsion,

$$\mathcal{V}(\mathbf{R}) = \frac{e^2}{4\pi\epsilon_0} \sum_{K=1}^{N-1} \sum_{L>K}^N \frac{Z_K Z_L}{|\mathbf{R}_K - \mathbf{R}_L|} \quad , \quad (5)$$

electronic – nuclear attraction,

$$\mathcal{V}(\mathbf{r}, \mathbf{R}) = -\frac{e^2}{4\pi\epsilon_0} \sum_{K=1}^N \sum_{k=1}^n \frac{Z_K}{|\mathbf{r}_k - \mathbf{R}_K|} \quad , \quad (6)$$

and interelectronic repulsion,

$$\mathcal{V}(\mathbf{r}) = \frac{e^2}{4\pi\epsilon_0} \sum_{k=1}^{n-1} \sum_{l>k}^n \frac{1}{|\mathbf{r}_k - \mathbf{r}_l|} \quad . \quad (7)$$

Here, M_K and Z_K denote the mass and atomic number of nucleus K ; m_e and e are the electronic mass and elementary charge, and ϵ_0 is the permittivity of vacuum. The nabla operators ∇_K and ∇_k act on the coordinates of nucleus K and electron k , respectively.

Defining the electronic Hamiltonian (fixed-nuclei approximation of $\mathcal{H}$) as

$$\mathcal{H}_{\text{el}}(\mathbf{r}, \mathbf{R}) = \mathcal{T}(\mathbf{r}) + \mathcal{V}(\mathbf{R}) + \mathcal{V}(\mathbf{r}, \mathbf{R}) + \mathcal{V}(\mathbf{r}) \quad , \quad (8)$$

we can rewrite the total Hamiltonian as

$$\mathcal{H}(\mathbf{r}, \mathbf{R}) = \mathcal{T}(\mathbf{R}) + \mathcal{H}_{\text{el}}(\mathbf{r}, \mathbf{R}) \quad . \quad (9)$$

Let us suppose the solutions of the time-independent (electronic) Schrödinger equation,

$$\mathcal{H}_{\text{el}}(\mathbf{r}, \mathbf{R}) \phi_i(\mathbf{r}, \mathbf{R}) = E_i(\mathbf{R}) \phi_i(\mathbf{r}, \mathbf{R}) \quad , \quad (10)$$

are known. Furthermore, the spectrum of $\mathcal{H}_{\text{el}}(\mathbf{r}, \mathbf{R})$ is assumed to be discrete and the eigenfunctions orthonormalised:

$$\int_{-\infty}^{\infty} \phi_i^*(\mathbf{r}, \mathbf{R}) \phi_j(\mathbf{r}, \mathbf{R}) d\mathbf{r} \equiv \langle \phi_i | \phi_j \rangle = \delta_{ij} \quad . \quad (11)$$

The total wavefunction Ξ can be expanded in terms of the eigenfunctions of $\mathcal{H}_{\text{el}}$ since these form a complete set:

$$\Xi(\mathbf{r}, \mathbf{R}; t) = \sum_j \phi_j(\mathbf{r}, \mathbf{R}) \chi_j(\mathbf{R}, t) \quad . \quad (12)$$

Insertion of this ansatz into the time-dependent Schrödinger equation (1) followed by multiplication from the left by $\phi_i^*(\mathbf{r}, \mathbf{R})$ and integration over the electronic coordinates leads to a set of coupled differential equations:

$$[\mathcal{T}(\mathbf{R}) + E_i(\mathbf{R})] \chi_i + \sum_j \mathcal{C}_{ij} \chi_j = i\hbar \frac{\partial}{\partial t} \chi_i \quad , \quad (13)$$

where the coupling operator $\mathcal{C}_{ij}$ is defined as

$$\mathcal{C}_{ij} \equiv \langle \phi_i | \mathcal{T}(\mathbf{R}) | \phi_j \rangle - \sum_K \frac{\hbar^2}{M_K} \langle \phi_i | \nabla_K | \phi_j \rangle \nabla_K \quad . \quad (14)$$

The diagonal term $\mathcal{C}_{ii}$ represents a correction to the (adiabatic) eigenvalue E_i of the electronic Schrödinger equation (10). In the case that all coupling operators $\mathcal{C}_{ij}$ are negligible, the set of differential Eqs. (13) becomes uncoupled:

$$[\mathcal{T}(\mathbf{R}) + E_i(\mathbf{R})] \chi_i = i\hbar \frac{\partial}{\partial t} \chi_i \quad . \quad (15)$$

This means that the nuclear motion proceeds without changes of the quantum state of the electron cloud and, correspondingly, the wavefunction (Eq. (12)) is reduced to a single term (adiabatic approximation):

$$\Xi(\mathbf{r}, \mathbf{R}; t) \approx \phi_i(\mathbf{r}, \mathbf{R}) \chi_i(\mathbf{R}, t) \quad . \quad (16)$$

For a great number of physical situations the Born-Oppenheimer approximation can be safely applied. On the other hand, there are many important chemical phenomena like, for instance, charge transfer and photoisomerisation reactions, whose very existence is due to the inseparability of electronic and nuclear motion. Inclusion of nonadiabatic effects will be the subject of the following sections.

3 Mixed Quantum–Classical Approach

Further simplification of the problem can be achieved by describing nuclear motion by classical mechanics and only the electrons quantum mechanically. In this so-called mixed quantum–classical (sometimes referred to as semiclassical) approach^{15,16}, the atomic nuclei follow some trajectory $\mathbf{R}(t)$ while the electronic motion is captured by some time-dependent total wavefunction $\Phi(\mathbf{r}; t)$ satisfying the time-dependent electronic Schrödinger equation,

$$\mathcal{H}_{\text{el}}(\mathbf{r}, \mathbf{R}(t)) \Phi(\mathbf{r}; t) = i\hbar \frac{\partial}{\partial t} \Phi(\mathbf{r}; t) \quad . \quad (17)$$

Again, the total wavefunction is written as a linear combination of adiabatic eigenfunctions $\phi_i(\mathbf{r}, \mathbf{R})$ (solutions of the time-independent Schrödinger equation (10)):

$$\Phi(\mathbf{r}; t) = \sum_j a_j(t) \phi_j(\mathbf{r}, \mathbf{R}) e^{-\frac{i}{\hbar} \int E_j(\mathbf{R}) dt} \quad . \quad (18)$$

Insertion of this ansatz into the time-dependent electronic Schrödinger equation (17) followed by multiplication from the left by $\phi_i^*(\mathbf{r}, \mathbf{R})$ and integration over the electronic coordinates leads to a set of coupled differential equations:

$$\dot{a}_i = - \sum_j a_j C_{ij} e^{-\frac{i}{\hbar} \int (E_j - E_i) dt} \quad , \quad (19)$$

where

$$C_{ij} \equiv \langle \phi_i | \frac{\partial}{\partial t} | \phi_j \rangle \quad (20)$$

are the nonadiabatic coupling elements. Integration of Eqs. (19) yields the expansion coefficients $a_i(t)$ whose square modulus, $|a_i(t)|^2$, can be interpreted as the probability of finding the system in the adiabatic state i at time t .

We now want to develop a condition for the validity of the Born-Oppenheimer approximation based on qualitative arguments. For this purpose, we shall consider a two-state system. To illustrate the problem, Figure 1 shows the avoided crossing between the covalent and ionic potential energy curves of NaCl^{17,18}. As we can see, the adiabatic wavefunctions ϕ_1 and ϕ_2 change their character as the bond length is varied. The characteristic length, l , over which ϕ_1 and ϕ_2 change significantly clearly depends on the nuclear configuration $\mathbf{R}$; in the vicinity of the NaCl avoided crossing, for instance, the character of the wavefunctions varies rapidly, whereas at large separations it remains more or less constant.

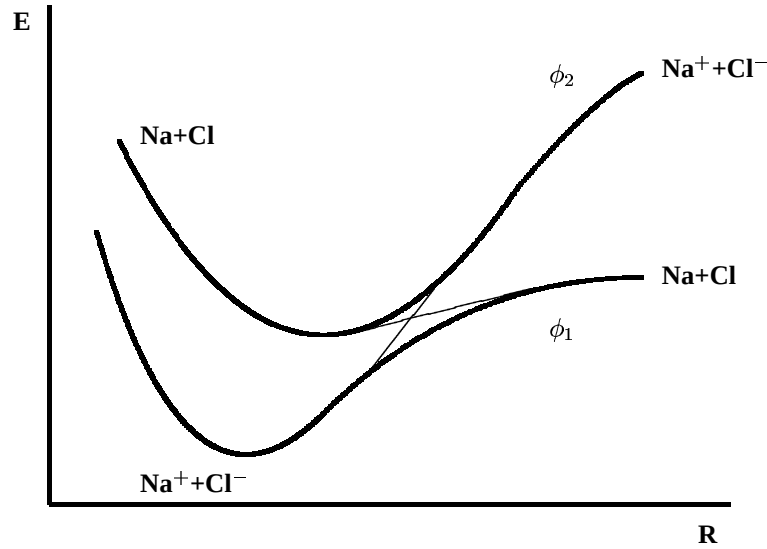


Figure 1. Avoided crossing between the covalent and ionic adiabatic potential curves of NaCl (thin lines: crossing of diabatic states).

Division of the characteristic length l by the velocity of the nuclei, $\dot{R} = |\dot{\mathbf{R}}|$, at a particular configuration $\mathbf{R}$ defines the time the system needs to travel the distance l around $\mathbf{R}$:

$$\text{passage time } \tau_p = \frac{l}{\dot{R}} \quad . \quad (21)$$

In order for the Born-Oppenheimer approximation to be valid, the electron cloud has to adjust instantly to the nuclear changes. The time scale characteristic of electronic motion can be obtained from the relation

$$\Delta E = |E_1 - E_2| = \hbar\omega \quad (22)$$

by taking the inverse transition frequency:

$$\tau_e = \frac{1}{\omega} = \frac{\hbar}{\Delta E} \quad . \quad (23)$$

The ratio

$$\xi = \frac{\tau_p}{\tau_e} = \frac{\Delta E l}{\hbar \dot{R}} \quad (24)$$

is the so-called Massay parameter. For values $\xi \gg 1$, i.e. large energy gaps ΔE and small velocities $\dot{R}$, nonadiabatic effects are negligible. In this case, if the system is prepared in some pure adiabatic state i ($|a_i|^2 = 1$) at time $t = 0$, the rhs of Eq. (19) will be zero at all times and the wavefunction expansion (Eq. (18)) can be replaced by a single term:

$$\Phi(\mathbf{r}; t) = \phi_i(\mathbf{r}, \mathbf{R}) e^{-\frac{i}{\hbar} \int E_i(\mathbf{R}) dt} \quad . \quad (25)$$

The atomic nuclei are then propagated by solving Newton's equations

$$M_K \ddot{\mathbf{R}}_K = \mathbf{F}_K(\mathbf{R}) \quad , \quad (26)$$

where

$$\mathbf{F}_K(\mathbf{R}) = -\nabla_K E_i(\mathbf{R}) \quad (27)$$

is the force on atom K .

4 Approaches to Nonadiabatic Dynamics

4.1 Mean-Field (Ehrenfest) Method

As we have discussed in the previous section, nonadiabaticity involves changes in the adiabatic state populations $|a_i|^2$ with changing nuclear configuration. Clearly, such a distortion of the electron cloud will, in turn, influence the nuclear trajectory. Although there are situations in which the impact of electronic nonadiabaticity on nuclear motion is negligible (e.g. for high energy collisions or small energy separations between adiabatic states), for many chemical systems it is of prime importance to properly incorporate electronic–nuclear feedback^{8,9}.

The simplest way of doing this is to replace the adiabatic potential energy surface E_i in Eq. (27) by the energy expectation value

$$E^{\text{eff}} = \langle \Phi | \mathcal{H}_{\text{el}} | \Phi \rangle = \sum_i |a_i|^2 E_i \quad , \quad (28)$$

where we have used Eq. (18). Thus, the atoms evolve on an effective potential representing an average over the adiabatic states weighted by their state populations $|a_i|^2$ (as illustrated in Figure 2). The method is therefore referred to as mean-field (also known as Ehrenfest) approach.

It is instructive to derive an expression for the nuclear forces either from the gradient of Eq. (28) or using the Hellmann-Feynman theorem

$$\mathbf{F}_K = -\langle \Phi | \nabla_K \mathcal{H}_{\text{el}} | \Phi \rangle \quad . \quad (29)$$

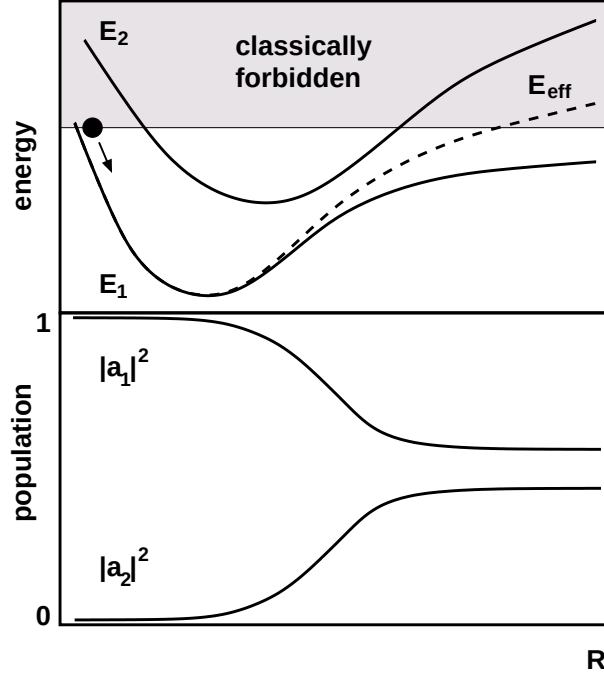


Figure 2. Top: avoided crossing between two adiabatic PES, E_1 and E_2 , and effective potential, E_{eff} , on which the nuclei are propagated in the Ehrenfest method. In the asymptotic region (right) E_{eff} contains contributions from classically forbidden regions of E_2 . Bottom: corresponding adiabatic state populations $|a_1|^2$ and $|a_2|^2$. The system is prepared in state 1 initially with zero kinetic energy. Upon entering the coupling region state 2 is increasingly populated.

Opting for the latter, we start by writing down the relation

$$\nabla_K \langle \phi_i | \mathcal{H}_{\text{el}} | \phi_j \rangle = \nabla_K E_i \delta_{ij} \quad (30)$$

$$= \langle \nabla_K \phi_i | \mathcal{H}_{\text{el}} | \phi_j \rangle + \langle \phi_i | \nabla_K \mathcal{H}_{\text{el}} | \phi_j \rangle + \langle \phi_i | \mathcal{H}_{\text{el}} | \nabla_K \phi_j \rangle \quad (31)$$

$$= \langle \phi_i | \nabla_K \mathcal{H}_{\text{el}} | \phi_j \rangle + (E_j - E_i) \mathbf{d}_{ji} \quad , \quad (32)$$

where we have defined the nonadiabatic coupling vectors, $\mathbf{d}_{ji}$, as

$$\mathbf{d}_{ji} = \langle \phi_j | \nabla_K | \phi_i \rangle \quad , \quad (33)$$

and used Eq. (10) together with the hermiticity of $\mathcal{H}_{\text{el}}$:

$$\langle \phi_i | \mathcal{H}_{\text{el}} | \nabla_K \phi_j \rangle = \langle \nabla_K \phi_j | \mathcal{H}_{\text{el}} | \phi_i \rangle^* = \langle \nabla_K \phi_j | E_j \phi_i \rangle^* = E_i \mathbf{d}_{ij}^* = -E_i \mathbf{d}_{ji} \quad . \quad (34)$$

Note that

$$\mathbf{d}_{ji}^* = -\mathbf{d}_{ij} \quad , \quad (35)$$

because

$$\nabla_K \langle \phi_i | \phi_j \rangle = \nabla_K \delta_{ij} = 0 \quad (36)$$

$$= \langle \nabla_K \phi_i | \phi_j \rangle + \langle \phi_i | \nabla_K \phi_j \rangle = \mathbf{d}_{ji}^* + \mathbf{d}_{ij} \quad . \quad (37)$$

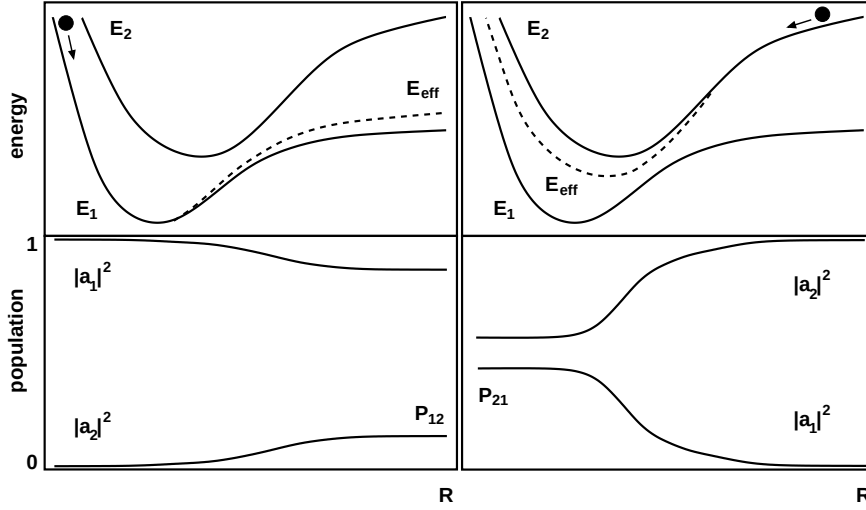


Figure 3. Top left: forward path effective potential, E_{eff} , for two weakly coupled adiabatic PES, E_1 and E_2 . Bottom left: state occupations for a system initially prepared in state 1. The final value of $|a_2|^2$ is equal to the transition probability P_{12} . Top right: backward path effective potential, E_{eff} , for two weakly coupled adiabatic PES, E_1 and E_2 . Bottom left: state occupations for a system initially prepared in state 2. The final value of $|a_1|^2$ is equal to the transition probability P_{21} .

Equating the rhss (right hand sides) of Eqs. (30) and (32) one obtains after rearranging,

$$\langle \phi_i | \nabla_K \mathcal{H}_{\text{el}} | \phi_j \rangle = \nabla_K E_i \delta_{ij} - (E_j - E_i) \mathbf{d}_{ji} \quad . \quad (38)$$

The nuclear forces (Eq. (29)) are thus given by

$$\mathbf{F}_K = - \sum_i |a_i|^2 \nabla_K E_i + \sum_{i,j} a_i^* a_j (E_j - E_i) \mathbf{d}_{ji} \quad . \quad (39)$$

Eq. (39) illustrates the two contributions to the nuclear forces; the first term is simply the population-weighted average force over the adiabatic states, while the second term takes into account nonadiabatic changes of the adiabatic state occupations. We would like to point out here that the nonadiabatic contributions to the nuclear forces are in the direction of the nonadiabatic coupling vectors $\mathbf{d}_{ji}$.

The Ehrenfest method has been applied with great success to a number of chemical problems including energy transfer at metal surfaces¹⁹. However, due to its mean-field character the method has some serious limitations. A system that was initially prepared in a pure adiabatic state will be in a mixed state when leaving the region of strong nonadiabatic coupling. In general, the pure adiabatic character of the wavefunction cannot be recovered even in the asymptotic regions of configuration space. In cases where the differences in the adiabatic potential energy landscapes are pronounced, it is clear that an average potential will be unable to describe all reaction channels adequately. In particular, if one is interested in a reaction branch whose occupation number is very small, the average path is likely to diverge from the true trajectory. Furthermore, the total wavefunction may contain significant contributions from adiabatic states that are energetically inaccessible (see Figure 2).

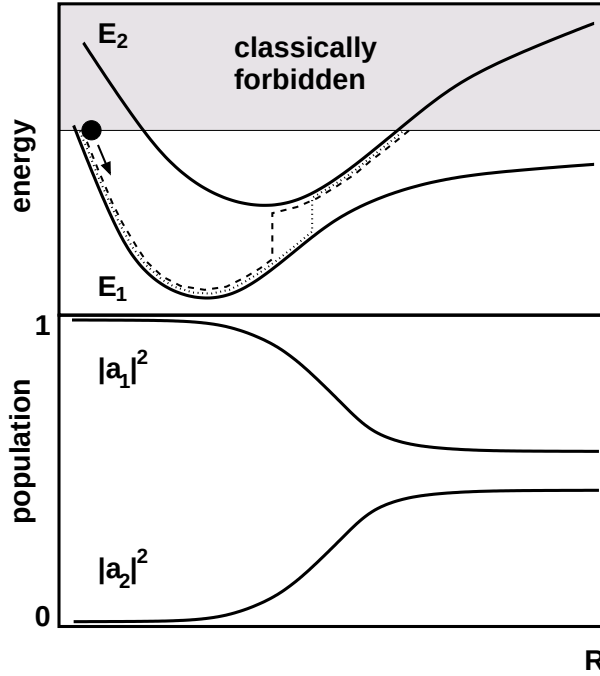


Figure 4. Top: avoided crossing between two adiabatic PES, E_1 and E_2 , and two typical forward surface hopping trajectories. Nonadiabatic transitions are most likely to occur in the coupling region. Bottom: corresponding adiabatic state populations $|a_1|^2$ and $|a_2|^2$. The system is prepared to be in state 1 initially with zero kinetic energy. Upon entering the coupling region state 2 is increasingly populated.

Figure 3 illustrates another severe drawback of the mean-field approach. The principle of microscopic reversibility demands that the forward path probability, $P_{12}^{\text{for}} = |a_2^{\text{final}}|^2$ for a system that was initially prepared in state 1 to end up in state 2 must be equal to the backward path probability, $P_{21}^{\text{back}} = |a_1^{\text{final}}|^2$ for a system that was initially prepared in state 2 to end up in state 1. One can easily think of situations, like the one depicted in Figure 3, for which the effective potentials for the forward and backward paths are very different, resulting also in different populations, $|a_i|^2$. The Ehrenfest method, therefore, violates microscopic reversibility.

It should be noted that the expansion of the total wavefunction in terms of (adiabatic) basis functions (Eq. (18)) is not a necessary requirement for the Ehrenfest method; the wavepacket Φ can be propagated numerically using Eq. (17). However, projection of Φ onto the adiabatic states facilitates interpretation. Knowledge of the expansion coefficients, a_i , is also the key to refinements of the method such as the surface hopping technique.

4.2 Surface Hopping

We have argued above that after exiting a well localised nonadiabatic coupling region it is unphysical for nuclear motion to be governed by a mixture of adiabatic states. Rather it would be desirable that in asymptotic regions the system evolves on a pure adiabatic PES.

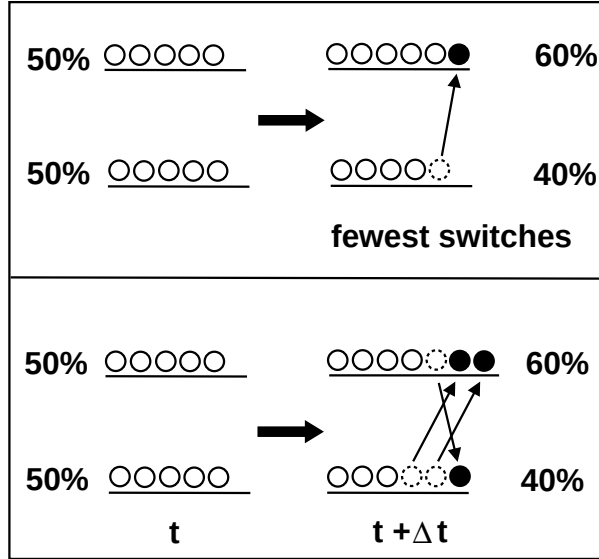


Figure 5. A two-state system with each state being equally (50%) populated at time t . At time $t + \Delta t$ the lower and the upper state are populated by 40% and 60% of ensemble members, respectively. The top panel shows how this distribution can be achieved with the minimum number of transitions, whereas the bottom panel shows one alternative route involving a larger number of transitions.

This idea is fundamental to the surface hopping approach. Instead of calculating the 'best' (i.e., state-averaged) path like in the Ehrenfest method, the surface hopping technique involves an ensemble of trajectories. At any moment in time, the system is propagated on some pure adiabatic state i , which is selected according to its state population $|a_i|^2$. Changing adiabatic state occupations can thus result in nonadiabatic transitions between different adiabatic PESs (see Figure 4). The ensemble averaged number of trajectories evolving on adiabatic state i at any time is equal to its occupation number $|a_i|^2$.

In the original formulation of the surface hopping method by Tully and Preston¹¹, switches between adiabatic states were allowed only at certain locations defined prior to the simulation. Tully¹² later generalized the method in such a way that nonadiabatic transitions can occur at any point in configuration space. At the same time, an algorithm — the so-called fewest switches criterion — was proposed which minimises the number of surface hops per trajectory whilst guaranteeing the correct ensemble averaged state populations at all times. The latter is important because excessive surface switching effectively results in weighted averaging over the adiabatic states much like in the case of the Ehrenfest method.

We shall now derive the fewest switches criterion. Out of a total of N trajectories, N_i will be in state i at time t ,

$$N_i(t) = \rho_{ii}(t)N \quad . \quad (40)$$

Here we have introduced the density matrix notation

$$\rho_{ij}(t) = a_i^*(t)a_j(t) \quad . \quad (41)$$

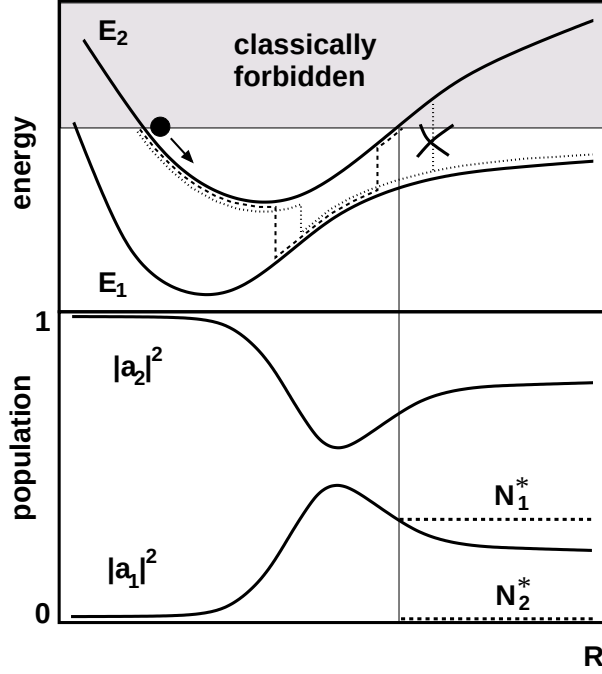


Figure 6. Top: avoided crossing between two adiabatic PES, E_1 and E_2 , and two typical forward surface hopping trajectories. Nonadiabatic transitions are most likely to occur in the coupling region. The cross indicates a classically forbidden transition; no switch is carried out in this case. Bottom: corresponding adiabatic state populations $|a_1|^2$ and $|a_2|^2$. The system is prepared in state 2 initially with zero kinetic energy. Upon entering the coupling region state 1 is increasingly populated. Upon exiting the coupling region, state population 1 decreases. For configurations $\mathbf{R}$ for which E_2 is in the classically forbidden region, the percentages of trajectories in state i , N_i^* , are unequal to $|a_i|^2$; N_2^* is zero whereas N_1^* remains constant.

At a later time $t' = t + \delta t$ the new occupation numbers are

$$N_i(t') = \rho_{ii}(t')N \quad (42)$$

Let us suppose that $N_i(t') < N_i(t)$ or $\delta N = N_i(t) - N_i(t') > 0$. Then the minimum number of transitions required to go from $N_i(t)$ to $N_i(t')$ is δN hops from state i to any other state and zero hops from any other state to state i (see Figure 5). The probability $P_i(t, \delta t)$ for a transition out of state i to any other state during the time interval $[t, t + \delta t]$ is then given by

$$P_i(t, \delta t) = \frac{\delta N}{N_i} = \frac{\rho_{ii}(t) - \rho_{ii}(t')}{\rho_{ii}} \approx -\frac{\dot{\rho}_{ii}\delta t}{\rho_{ii}}, \quad (43)$$

where we have used

$$\dot{\rho}_{ii} \approx \frac{\rho_{ii}(t') - \rho_{ii}(t)}{\delta t}. \quad (44)$$

The lhs of Eq. (44) can be written as

$$\dot{\rho}_{ii} = \frac{d}{dt}(a_i^* a_i) = \dot{a}_i^* a_i + a_i^* \dot{a}_i = (\dot{a}_i^* a_i)^* + a_i^* \dot{a}_i = 2\Re(a_i^* \dot{a}_i). \quad (45)$$

Inserting Eq. (19) into Eq. (45) we obtain

$$\dot{\rho}_{ii} = -2\Re \left(\sum_j \rho_{ij} C_{ij} e^{-\frac{i}{\hbar} \int (E_j - E_i) dt} \right) . \quad (46)$$

Substituting expression (46) into Eq. (43) the probability P_i can be rewritten as follows

$$P_i(t, \delta t) = \frac{2\Re \left(\sum_j \rho_{ij} C_{ij} e^{-\frac{i}{\hbar} \int (E_j - E_i) dt} \right) \delta t}{\rho_{ii}} . \quad (47)$$

Since the probability, P_i , for a switch from state i to any other state must be the sum over all states of the probabilities, P_{ij} , for a transition from state i to a specific state j ,

$$P_i(t, \delta t) = \sum_j P_{ij}(t, \delta t) , \quad (48)$$

it follows from Eq. (47) that

$$P_{ij}(t, \delta t) = \frac{2\Re \left(\rho_{ij} C_{ij} e^{-\frac{i}{\hbar} \int (E_j - E_i) dt} \right) \delta t}{\rho_{ii}} . \quad (49)$$

A transition from state i to state k is now invoked if

$$P_i^{(k)} < \zeta < P_i^{(k+1)} , \quad (50)$$

where ζ ($0 \leq \zeta \leq 1$) is a uniform random number and $P_i^{(k)}$ is the sum of the transition probabilities for the first k states,

$$P_i^{(k)} = \sum_j^k P_{ij} . \quad (51)$$

In order to conserve total energy after a surface hop has been carried out, the atomic velocities have to be rescaled. The usual procedure is to adjust only the velocity components in the direction of the nonadiabatic coupling vector $d_{ik}(\mathbf{R})$ (Eq. (33))¹². We can qualitatively justify this practice by our earlier observation that the nonadiabatic contribution to the Ehrenfest forces also are in the direction of the nonadiabatic coupling vector $d_{ik}(\mathbf{R})$ (see Eq. (39)). Certainly, such discontinuities in nuclear velocities must be regarded as a flaw of the surface hopping approach. In most physical scenarios, however, nonadiabatic surface switches take place only at relatively small potential energy separations so that the necessary adjustment to the nuclear velocities is reasonably small. Nevertheless, a severe limitation of the method is presented by its inability to properly deal with situations in which the amount of kinetic energy is insufficient to compensate for the difference in potential energy (so-called classically forbidden transitions). Tully's original suggestion not to carry out a surface hop while retaining the nuclear velocities in such cases has been demonstrated²⁰ to be more accurate than later proposals to reverse the velocity components in the direction of the nonadiabatic coupling vector $d_{ik}(\mathbf{R})$ ^{21,22}. The example presented in Figure 6 illuminates how classically forbidden transitions cause divergence between the target occupation numbers, $|a_i|^2$, and the actual percentages of trajectories

evolving in state i , N_i^* .

It should be noted that surface hopping exhibits a large degree of electronic coherence through continuous integration of Eqs. (19) along the entire trajectory. On the one hand, this enables the method to reproduce quantum interference effects¹² such as Stueckelberg oscillations¹⁵. On the other hand, due to treating nuclei classically, dephasing of the electronic degrees of freedom may be too slow, a shortcoming shared by the surface hopping and the Ehrenfest method alike. A number of semiclassical approaches incorporating decoherence have, therefore, been proposed^{23–29}. Some of these alternative methods attempt to combine the advantages of surface hopping (mainly, pure adiabatic states in asymptotic regions) with those of the mean-field method (no discontinuities in potential energy, no disallowed transitions) by employing an effective potential whilst enforcing gradual demixing of the total wavefunction away from the coupling regions^{27–29}.

4.3 Car-Parrinello Surface Hopping

So far we have assumed that a number of adiabatic potential energy surfaces (at least two) have been obtained by solving the time-independent Schrödinger equation (10) in some unspecified manner. Instead of precalculating the entire PESs, it is advantageous to compute the electronic energies and nuclear gradients “on the fly” as the system is propagated along the trajectory. A popular method in this context has been the Diatomics-in-Molecules (DIM)^{30–42} method which cheaply provides the required electronic eigenvalues and atomic forces for a multitude of molecular valence states simultaneously through diagonalisation of the Hamiltonian matrix. However, although the DIM method works remarkably well for some simple systems such as cationic rare-gas clusters^{43–46}, it is not generally applicable to more complex systems.

For ground state calculations, density functional theory^{47–49} based *ab initio* MD in the spirit of Car and Parrinello⁵⁰ has become the method of choice to study large molecules and condensed phase systems. Recently, Car-Parrinello simulations have become possible also in the first excited singlet state using a restricted open-shell Kohn-Sham (ROKS) approach⁵¹. We summarize here a recently developed Tully-style¹² trajectory surface hopping method coupling nonadiabatically the S_0 ground state and the S_1 excited state accessible within the Car-Parrinello framework^{52–54}. This approach has been applied to a variety of problems including photoisomerisation⁵², excited state intramolecular proton transfer⁵⁵, and the photochemistry of nucleobases^{56, 57, 55, 58}.

4.3.1 Restricted Open-Shell Kohn-Sham Method

Let us first take a brief look at the ROKS method for the S_1 state. Starting from a closed-shell ground state, S_0 , consider an excitation of an electron out of the HOMO into the LUMO. The resulting two unpaired spins can be arranged in four different ways, as illustrated in Figure 7, parallel spins forming triplet determinants and antiparallel spins being equal mixtures of singlet and triplet determinants. The S_1 singlet wavefunction, ϕ_1 , is constructed as

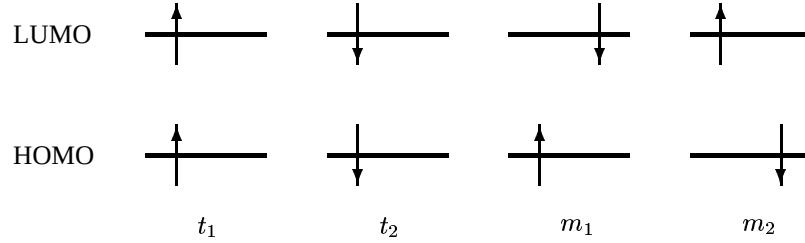


Figure 7. Four possible spin configurations upon excitation of one electron out of the highest occupied molecular orbital (HOMO) to the lowest unoccupied molecular orbital (LUMO). The two parallel spin configurations, t_1 and t_2 form triplet determinants, while the two antiparallel configurations, m_1 and m_2 form mixed determinants with equal singlet and triplet contributions.

$$\begin{aligned}\phi_1 &= \frac{1}{\sqrt{2}} \{ |m_1\rangle + |m_2\rangle \} \\ &= \frac{1}{\sqrt{2}} \left\{ |\varphi_1^{(1)} \bar{\varphi}_1^{(1)} \varphi_2^{(1)} \bar{\varphi}_2^{(1)} \cdots \varphi_l^{(1)} \bar{\varphi}_{l+1}^{(1)} \rangle \right. \\ &\quad \left. + |\varphi_1^{(1)} \bar{\varphi}_1^{(1)} \varphi_2^{(1)} \bar{\varphi}_2^{(1)} \cdots \varphi_{l+1}^{(1)} \bar{\varphi}_l^{(1)} \rangle \right\} ,\end{aligned}\quad (52)$$

where the “ket” notation signifies Slater determinants made up of Kohn-Sham orbitals, $\varphi_i^{(1)}$ (spin up) and $\bar{\varphi}_i^{(1)}$ (spin down); $l = \frac{n}{2}$ is half the number of electrons. It has been shown by Ziegler et al.⁵⁹ that the S_1 energy, $E(S_1)$, can be written as the difference between twice the energy of the mixed determinant, $E(m)$, and the energy of the triplet determinant, $E(t)$,

$$E(S_1) = 2E(m) - E(t) . \quad (53)$$

Within the ROKS scheme, a *single* set of orbitals $\{\varphi_i^{(1)}\}$ is determined that minimises the energy functional,

$$E[\{\varphi_i^{(1)}\}] = 2\langle m | \mathcal{H}^{\text{KS}} | m \rangle - \langle t | \mathcal{H}^{\text{KS}} | t \rangle - \sum_{i,j=1}^{l+1} \lambda_{ij} \left\{ \langle \varphi_i^{(1)} | \varphi_j^{(1)} \rangle - \delta_{ij} \right\} , \quad (54)$$

where $\mathcal{H}^{\text{KS}}$ is the Kohn-Sham Hamiltonian⁴⁹ and the λ_{ij} are Lagrange multipliers taking care of the orthonormality of the orbitals.

Due to this optimisation the entire set of orbitals $\{\varphi_i^{(1)}\}$ will, in general, differ from the set of orbitals $\{\varphi_i^{(0)}\}$ that define the ground state wavefunction, ϕ_0 ,

$$\phi_0 = |\varphi_1^{(0)} \bar{\varphi}_1^{(0)} \varphi_2^{(0)} \bar{\varphi}_2^{(0)} \cdots \varphi_l^{(0)} \bar{\varphi}_l^{(0)} \rangle . \quad (55)$$

As a consequence the two state functions, ϕ_0 and ϕ_1 , are nonorthogonal giving rise to the overlap matrix elements, S_{ij} ,

$$S_{01} = S_{10} \equiv S , \quad S_{ii} = 1 . \quad (56)$$

4.3.2 S_0 – S_1 Surface Hopping

Inserting ansatz (Eq. (18)) using the above basis functions, ϕ_0 and ϕ_1 , into Eq. (17) and replacing $\mathcal{H}_{\text{el}}$ with $\mathcal{H}^{\text{KS}}$ we obtain after integration over the electronic coordinates following multiplication by ϕ_i^* from the left

$$\sum_j a_j p_j (H_{ij} - E_j S_{ij}) = i\hbar \left\{ \sum_j \dot{a}_j p_j S_{ij} + \sum_j a_j p_j C_{ij} \right\} , \quad (57)$$

where the Hamiltonian matrix elements are given by

$$H_{ii} = \langle \Phi_i | \mathcal{H}^{\text{KS}} | \Phi_i \rangle = E_i , \quad (58)$$

$$H_{01} = H_{10} = E_0 S , \quad (59)$$

and the phase factor has been abbreviated as

$$p_j \equiv e^{-\frac{i}{\hbar} \int E_j dt} . \quad (60)$$

We should stress here that the discrepancy between Eqs. (57) and (19) arises purely because ϕ_1 is *not* an eigenfunction of $\mathcal{H}^{\text{KS}}$.

For $i = 0$, Eq. (57) thus becomes

$$a_1 p_1 S(E_0 - E_1) = i\hbar \{ \dot{a}_0 p_0 + \dot{a}_1 p_1 S + a_1 p_1 C_{01} \} , \quad (61)$$

and for $i = 1$

$$0 = \dot{a}_0 p_0 S + \dot{a}_1 p_1 + a_0 p_0 C_{10} . \quad (62)$$

Solving equations (61) and (62) for $\dot{a}_0$ and $\dot{a}_1$ one finds

$$\dot{a}_0 = \frac{1}{S^2 - 1} \left[i a_1 \frac{p_1}{p_0} S(E_0 - E_1) + a_1 C_{01} \frac{p_1}{p_0} - a_0 C_{10} S \right] , \quad (63)$$

$$\dot{a}_1 = \frac{1}{S^2 - 1} \left[a_0 C_{10} \frac{p_0}{p_1} - a_1 C_{01} S - i a_1 S^2 (E_0 - E_1) \right] . \quad (64)$$

We integrate these two coupled differential equations numerically using a fourth order Runge-Kutta scheme⁶⁰. It is computationally attractive to work with the nonadiabatic coupling elements, C_{ij} (Eq. (20)), instead of the nonadiabatic coupling vectors, $\mathbf{d}_{ji}$ (Eq. (33)), since the orbital velocities are readily available within the Car-Parrinello method.

If both electronic state functions were eigenfunctions of the Kohn-Sham Hamiltonian, $|a_0|^2$ and $|a_1|^2$ would be their respective occupation numbers. A look at the normalisation integral of the total wavefunction Φ ,

$$\langle \Phi | \Phi \rangle = |a_0|^2 + |a_1|^2 + 2S \Re \left(a_0^* a_1 \frac{p_1}{p_0} \right) \equiv 1 , \quad (65)$$

shows that the definition of state populations in this basis is ambiguous. We therefore expand the total wavefunction Φ in terms of an orthonormal set of auxiliary wavefunctions, ϕ'_0 and ϕ'_1 :

$$\Phi = d_0 \phi'_0 + d_1 \phi'_1 = b_0 \phi_0 + b_1 \phi_1 , \quad (66)$$

where

$$\langle \phi'_i | \phi'_j \rangle = \delta_{ij} \quad (67)$$

and

$$b_j = a_j p_j \quad . \quad (68)$$

Since Φ is normalised, the squares of our new expansion coefficients add up to unity and thus have the meaning of state populations in the orthogonal basis:

$$|d_0|^2 + |d_1|^2 = 1 \quad . \quad (69)$$

The orthonormal wavefunctions ϕ'_0 and ϕ'_1 can be expressed in terms of ϕ_0 and ϕ_1 as

$$\phi'_0 = c_{00}\phi_0 + c_{10}\phi_1 \quad , \quad (70)$$

$$\phi'_1 = c_{01}\phi_0 + c_{11}\phi_1 \quad , \quad (71)$$

$\mathbf{c}_0 = \begin{pmatrix} c_{00} \\ c_{10} \end{pmatrix}$ and $\mathbf{c}_1 = \begin{pmatrix} c_{01} \\ c_{11} \end{pmatrix}$ being solutions of the eigenvalue problem

$$\mathbf{H}\mathbf{C} = \mathbf{S}\mathbf{C}\mathbf{E} \quad . \quad (72)$$

Using the Hamiltonian matrix elements of Eqs. (58) and (59) and the overlap matrix of Eq. (56), one obtains the eigenvalues

$$e_0 = E_0 \quad (73)$$

and

$$e_1 = \frac{E_1 - S^2 E_0}{1 - S^2} \quad (> E_1, \text{ if } E_0 < E_1) \quad . \quad (74)$$

The corresponding eigenvectors are

$$\mathbf{c}_0 = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad (75)$$

and

$$\mathbf{c}_1 = \begin{pmatrix} -S \\ 1 \end{pmatrix} \quad (76)$$

leading to the orthonormal wavefunctions

$$\phi'_0 = \phi_0 \quad , \quad (77)$$

$$\phi'_1 = \frac{1}{\sqrt{1 - S^2}} [-S\phi_0 + \phi_1] \quad . \quad (78)$$

Inserting Eqs. (77) and (78) into Eq. (66) we determine the expansion coefficients to be

$$d_0 = b_0 + b_1 S \quad (79)$$

$$d_1 = b_1 \sqrt{1 - S^2} \quad . \quad (80)$$

The state occupation numbers are thus

$$|d_0|^2 = |b_0|^2 + S^2|b_1|^2 + 2S \Re(b_0^* b_1) \quad , \quad (81)$$

$$|d_1|^2 = (1 - S^2)|b_1|^2 \quad (82)$$

or alternatively

$$|d_0|^2 = |a_0|^2 + S^2|a_1|^2 + 2S \Re\left(a_0^* a_1 \frac{p_1}{p_0}\right) \quad (83)$$

$$|d_1|^2 = (1 - S^2)|a_1|^2 \quad . \quad (84)$$

We are now in a position to apply Tully's fewest switches criterion (Eq. (49)) using the coefficients d_i to construct the density matrix (Eq. (41)).

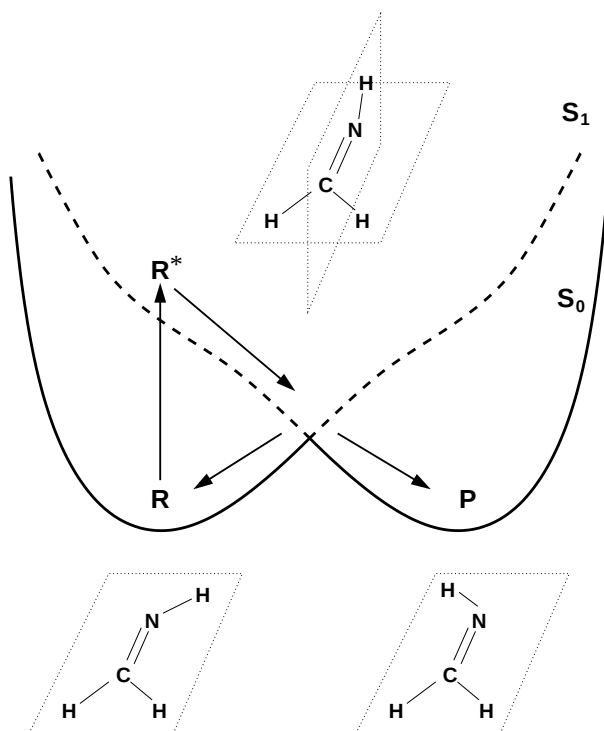


Figure 8. Schematic view of the photoreaction pathways of formaldimine. S_0 and S_1 energy curves are plotted against the reaction coordinate whose main contributor is the NH twist angle. The reactant R is vertically excited from the ground state into the S_1 state to form R^* . The system then falls into a conical intersection where relaxation to the ground state occurs. The reaction can proceed to either of the equivalent isomers, R and P. Formation of the photoproduct P corresponds to photoisomerisation.

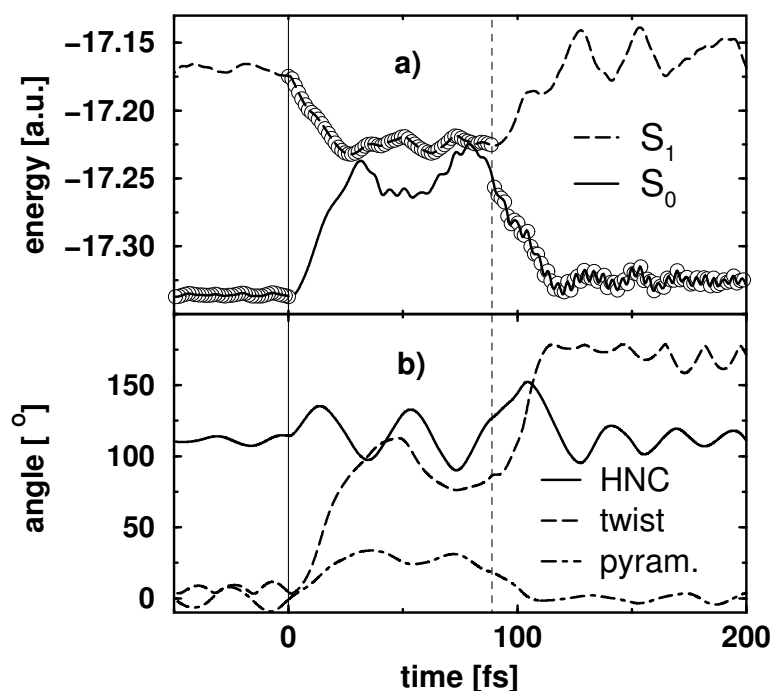


Figure 9. a) Time evolution of S_0 and S_1 energies following photoexcitation in the case of a successful $R \rightarrow P$ reaction. The solid and dashed vertical lines indicate the moment of the photoexcitation and the nonadiabatic transition to the ground state, respectively. The open circles denote the PES on which the nuclei are being propagated. b) Corresponding time evolution of the pyramidalisation, HNC and the HN twist angles. The HN bond is seen to flip from 0° to 180° resulting in the photoproduct. For HNC angles around 106.5° at orthogonal twist geometry, the energy gap is seen to be minimal.

4.3.3 Example: Photoisomerisation of Formaldimine

Figure 8 shows a schematic view of the photoreaction pathways of formaldimine. The reactant, R, is excited vertically from the ground state minimum into the S_1 state to form R^* . Subsequently, the system moves along the reaction coordinate, which predominantly involves an out-of-plane twist of the NH bond, into a conical intersection located at orthogonal twist geometry. In this region of strong nonadiabatic coupling a transition to the ground state occurs leading either to the photoisomerisation product, P, or back to the reactant R. As starting configurations for our nonadiabatic CP-MD calculations, we have picked 100 initial conditions at random from a standard ground state run at 300 K, in order to sample the canonical ensemble. For each of the two possible outcomes, i.e. $R \rightarrow P$ and $R \rightarrow R$, a typical trajectory is analysed in Figures. 9 and 10⁶¹. Figure 9a shows the evolution of the S_0 and S_1 energies as a function of time for a reactive trajectory leading to the photoproduct P. A photochemical laser excitation of the system is mimicked by vertical $S_0 \rightarrow S_1$ excitation at $t = 0$ (see circles in Figure 9a). The system is seen to quickly move down into the S_1 potential well dramatically reducing the energy gap to the ground state. As illustrated by Figure 9b, the main contribution to the S_1 energy reduction

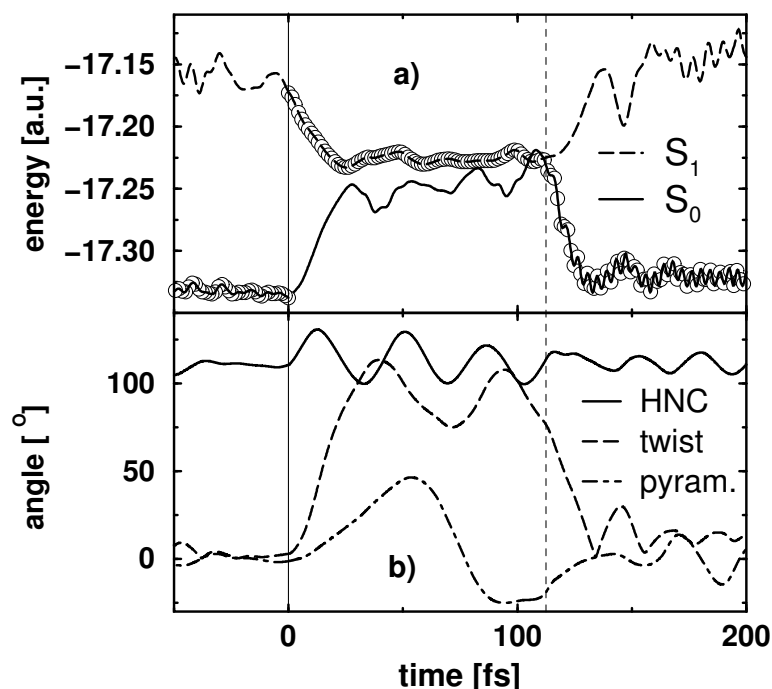


Figure 10. a) Time evolution of S_0 and S_1 energies following photoexcitation for a non-reactive $R \rightarrow R$ event (all symbols as in Figure 9.) b) Corresponding time evolution of the pyramidalisation, HNC and the HN twist angles. The HN bond is seen to orthogonalise initially and later flip back to 0° .

is due to the NH twist angle changing from near planarity (0°) to orthogonality (90°). Near the minimum of the S_1 PES, where the nonadiabatic coupling is strongest, a nonadiabatic transition to the S_0 state occurs leading to rapid widening of the energy gap accompanied by a change in the twist angle from around 90° to near 180° . This behaviour is in accord with static MRCI predictions^{52,62}. Additional insight can be gained by analysing the role of the HNC and pyramidalisation angles. It becomes clear that pyramidalisation is prerequisite for the two surfaces to cross. Similarly, small HNC angles minimise the energy gap, maximising the nonadiabatic transition probability. In the case of the non-reactive $R \rightarrow R$ event (Figure 10), the situation is initially very similar with the exception that the nonadiabatic surface hop occurs one HNC vibrational period later. Furthermore, the NH twist angle relaxes back to near 0° after initial orthogonalisation, signifying an unsuccessful photoisomerisation attempt. The two surfaces are seen to cross near orthogonal twist geometry with the HNC and pyramidalisation angles being close to a minimum and maximum, respectively.

It is possible, in principle, to determine the quantum yield of photoisomerisation by averaging over an ensemble of surface hopping trajectories. Since this would be beyond the scope of this article, we can only state here our non-converged result of 70 %.

Acknowledgements

I should like to thank Prof. D. Marx for his support.

References

1. V. Bonačić-Koutecký and R. Mitrić. *Chem. Rev.*, 105:11, 2005.
2. M. D. Hack and D. G. Truhlar. *J. Phys. Chem. A*, 104:7917, 2000.
3. M. P. Allen and D. J. Tildesley. *Computer Simulation of Liquids*. Clarendon Press, Oxford, 1987.
4. D. Frenkel and B. Smit. *Understanding Molecular Simulation. From Algorithms to Applications*. Academic Press, Boston, 1996.
5. P. Ehrenfest. *Z. Phys.*, 45:455, 1927.
6. H.-D. Meyer and W. H. Miller. *J. Chem. Phys.*, 70:3214, 1979.
7. D. A. Micha. *J. Chem. Phys.*, 78:7138, 1983.
8. J. C. Tully. In B. J. Berne, G. Ciccotti, and D. F. Coker, editors, *Classical and Quantum Dynamics in Condensed Phase Simulations*. World Scientific, Singapore, 1998.
9. J. C. Tully. In D. L. Thompson, editor, *Modern Methods for Multidimensional Dynamics Computations in Chemistry*. World Scientific, Singapore, 1998.
10. X. Li, J. C. Tully, H. B. Schlegel, and M. J. Frisch. *J. Chem. Phys.*, 123:084106, 2005.
11. J. C. Tully and R. K. Preston. *J. Chem. Phys.*, 55:562, 1971.
12. J. C. Tully. *J. Chem. Phys.*, 93:1061, 1990.
13. N. C. Blair and D. G. Truhlar. *J. Chem. Phys.*, 79:1334, 1983.
14. P. J. Kuntz. *J. Chem. Phys.*, 95:141, 1991.
15. E. E. Nikitin. In H. Hartmann, editor, *Chemische Elementarprozesse*. Springer, Berlin, 1968.
16. E. E. Nikitin and L. Zülicke. *Theory of Chemical Elementary Processes*. Springer, Berlin, 1978.
17. L. Salem. *Electrons in Chemical Reactions: First Principles*. Wiley, New York, 1982.
18. L. Salem, C. Leforestier, G. Segal, and R. Wetmore. *J. Am. Chem. Soc.*, 97:479, 1975.
19. J. C. Tully, M. Gomez, and M. Head-Gordon. *J. Vac. Sci. Technol.*, A11:1914, 1993.
20. U. Müller and G. Stock. *J. Chem. Phys.*, 107:6230, 1997.
21. S. Hammes-Schiffer and J. C. Tully. *J. Chem. Phys.*, 101:4657, 1994.
22. D. F. Coker and L. Xiao. *J. Chem. Phys.*, 102:496, 1995.
23. F. Webster, P. J. Rossky, and R. A. Friesner. *Comp. Phys. Comm.*, 63:494, 1991.
24. F. J. Webster, J. Schnitker, M. S. Friedrichs, R. A. Friesner, and P. J. Rossky. *Phys. Rev. Lett.*, 66:3172, 1991.
25. E. R. Bittner and P. J. Rossky. *J. Chem. Phys.*, 103:8130, 1995.
26. E. R. Bittner and P. J. Rossky. *J. Chem. Phys.*, 107:8611, 1997.
27. M. D. Hack and D. G. Truhlar. *J. Chem. Phys.*, 114:2894, 2001.
28. M. D. Hack and D. G. Truhlar. *J. Chem. Phys.*, 114:9305, 2001.

29. Y. L. Volobuev, M. D. Hack, M. S. Topaler, and D. G. Truhlar. *J. Chem. Phys.*, 112:9716, 2000.
30. F. O. Ellison. *J. Am. Chem. Soc.*, 85:3540, 1963.
31. J. C. Tully. *J. Chem. Phys.*, 59:5122, 1973.
32. J. C. Tully. *J. Chem. Phys.*, 58:1396, 1973.
33. J. C. Tully. *J. Chem. Phys.*, 64:3182, 1976.
34. J. C. Tully and C. M. Truesdale. *J. Chem. Phys.*, 65:1002, 1976.
35. P. J. Kuntz and A. C. Roach. *J. Chem. Soc. Faraday Trans.*, 68:259, 1971.
36. E. Steiner, P. R. Certain, and P. J. Kuntz. *J. Chem. Phys.*, 59:47, 1973.
37. P. J. Kuntz. *J. Phys. B*, 19:1731, 1986.
38. A. C. Roach and P. J. Kuntz. *J. Chem. Phys.*, 84:822, 1986.
39. J. C. Tully. In G. A. Segal, editor, *Modern Theoretical Chemistry*, Vol. 7A. Plenum Press, New York, 1977.
40. P. J. Kuntz. In R. B. Bernstein, editor, *Atom-Molecule Collision Theory*. Plenum Press, New York, 1979.
41. P. J. Kuntz. In M. Baer, editor, *Theory of Chemical Reaction Dynamics*, Vol. 1. Chemical Rubber, Boca Raton, 1985.
42. P. J. Kuntz. In Z. B. Maksic, editor, *Theoretical Models of Chemical Bonding*, Part 2. Springer, Berlin, 1990.
43. N. L. Doltsinis and P. J. Knowles. *Chem. Phys. Lett.*, 325:648, 2000.
44. N. L. Doltsinis. *Mol. Phys.*, 97, 1999.
45. N. L. Doltsinis and P. J. Knowles. *Chem. Phys. Lett.*, 301:241, 1999.
46. N. L. Doltsinis, P. J. Knowles, and F. Y. Naumkin. *Mol. Phys.*, 96:749, 1999.
47. R. G. Parr and W. Yang. *Density Functional Theory of Atoms and Molecules*. Oxford University Press, Oxford, 1989.
48. P. Hohenberg and W. Kohn. *Phys. Rev. B*, 136:864, 1964.
49. W. Kohn and L. J. Sham. *Phys. Rev. A*, 140:1133, 1965.
50. R. Car and M. Parrinello. *Phys. Rev. Lett.*, 55:2471, 1985.
51. I. Frank, J. Hutter, D. Marx, and M. Parrinello. *J. Chem. Phys.*, 108:4060, 1998.
52. N. L. Doltsinis and D. Marx. *Phys. Rev. Lett.*, 88:166402, 2002.
53. N. L. Doltsinis. In J. Grotendorst, D. Marx, and A. Muramatsu, editors, *Quantum Simulations of Complex Many-Body Systems: From Theory to Algorithms*. NIC, FZ Jülich, 2002.
www.fz-juelich.de/nic-series/volume10/doltsinis.pdf.
54. N. L. Doltsinis and D. Marx. *J. Theor. Comp. Chem.*, 1:319–349, 2002.
55. N. L. Doltsinis. *Mol. Phys.*, 102:499, 2004.
56. H. Langer and N. L. Doltsinis. *Phys. Chem. Chem. Phys.*, 5:4516, 2003.
57. H. Langer and N. L. Doltsinis. *Phys. Chem. Chem. Phys.*, 6:2742, 2004.
58. H. Langer, N. L. Doltsinis, and D. Marx. *ChemPhysChem*, 6:1734, 2005.
59. T. Ziegler, A. Rauk, and E. J. Baerends. *Theor. Chim. Acta*, 43:261, 1977.
60. W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in Fortran 77*, volume 1. Cambridge University Press, 1999.
61. For a graphical animation see
www.theochem.ruhr-uni-bochum.de/go/surfhop.html.
62. V. Bonačić-Koutecký and Josef Michl. *Theor. Chim. Acta*, 68:45, 1985.

Surface Phase Diagrams from Ab Initio Thermodynamics

Bernd Meyer

Chair of Theoretical Chemistry
Ruhr-Universität Bochum
44780 Bochum, Germany
E-mail: bernd.meyer@theochem.rub.de

A practical scheme is presented which allows to compare the relative stability of surface structures with different compositions. It is based on a combination of ab initio total energy calculations with a thermodynamic description of the surface. The result is a surface phase diagram of the most stable surface compositions and structures as a function of temperature and pressure. Thus, the predictive power of first-principles calculations can be extended to more relevant temperature and pressure conditions.

1 Introduction

In atomistic studies of multi-component compounds one often faces the problem that one would like to compare the relative stability of configurations which differ in their atomic composition. For example, let us consider a binary compound AB where we have introduced a surface. For a fixed crystallographic orientation we can create various surface terminations which differ in their chemical composition, for example, a stoichiometric or a A-rich termination as sketched in Figure 1. Which of the two possibilities is more stable? In a total energy electronic structure calculation the two surface terminations have to be represented, for example in a slab setup, by atomic configurations with a different number of A and B atoms. How can we compare the total energy of the two calculations? Only the energy of a primitive AB unit cell is a well defined quantity but we can not subdivide this energy into contributions originating exclusively from the A or the B atom.

It is obvious that we can only determine the relative stability of two different structures if we compare the total energies of configurations with the same number of A and B atoms. Therefore, if we change the composition of a structure we have to take appropriate reference phases into account from where we take additional atoms or where removed atoms are disposed. The choice of the reference phases, however, is not unique. As shown for our schematic example in Figure 1, the A-rich surface termination can be created either by taking A atoms from a A-bulk environment or by removing B atoms into a surrounding gas phase. This indicates that the answer to the question which of the two surface configurations is more stable can not be unique — the answer will depend on the experimental conditions. If we assume a B-poor environment, for example in an ultrahigh vacuum experiment at high temperatures, then it might be quite favorable to desorb B atoms from the surface (since we gain a lot of entropy) so that the A-rich surface will be the most stable one. Under B-rich conditions, on the other hand, at a high partial pressure of B molecules and low temperatures, probably the stoichiometric or even a B-rich surface structure will be favored.

The most general way to take these environmental conditions into account is to combine the total energy electronic structure calculations with a thermodynamic description of

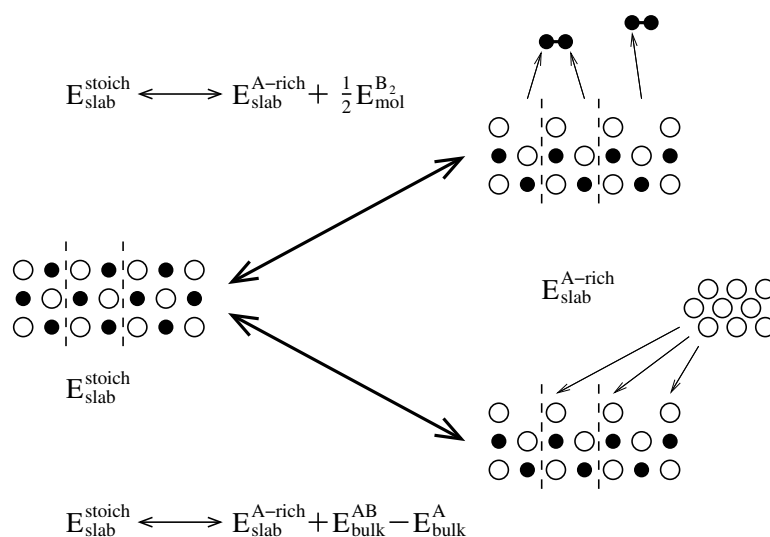


Figure 1. Schematic illustration of two surface terminations for an AB compound with different composition (left: stoichiometric surface, right: surface with excess of A atoms). A typical supercell in a slab calculation is indicated by dashed lines.

the surface (see Section 2). The general procedure will be that we introduce appropriate chemical potentials for the A and B atoms and an expression for the Gibbs free surface energy^{1,2}. Depending on the chemical potentials we then determine the surface structures with the lowest free energy. This information is used to construct a phase diagram for the surface. If we assume that the surface is in thermodynamic equilibrium with given particle reservoirs, for example a surrounding gas phase, we can finally relate the chemical potentials to specific temperature and pressure conditions. Thus, as a second major advantage, we are able to extend the zero temperature and zero pressure results of our total energy calculations quite easily to experimentally much more relevant temperature and pressure ranges, thereby bridging the gap between UHV-like conditions to ambient environments or temperatures and gas phase pressures that are typically applied in chemical reactions.

The problem of how to compare the total energy of calculations for configurations which differ in composition is also encountered when studying atomic point defects in multi-component compounds. Furthermore, if we assume that the homogeneity of an ordered compound is conserved when raising the temperature then it also has to be taken into account that inevitably various types of atomic defects have to be generated simultaneously by thermal excitations. Suppose, for instance, that only vacancies on the sublattice of the A atoms would be generated. Then, with increasing temperature, more and more A atoms would be removed from the A-atom sublattice of the bulk and deposited at inner or outer surfaces, i.e., the homogeneity of the compound would be destroyed. The only way to avoid this problem is to generate simultaneously for example vacancies on the B-sublattice or antistructure atoms with such a statistical weight that on the average the structure of the material is conserved. How this problem can be solved by assuming thermodynamic equilibrium and by combining total energy calculations with a grand canonical

statistical approach to derive defect concentrations and defect formation energies can be found in Refs. 3–10.

2 Gibbs Free Surface Energy

In this section we describe the thermodynamic formalism which can be used to calculate surface phase diagrams of the most stable compositions and structures of surface terminations of multi-component compounds as a function of external environmental conditions. As input the total energies of the different atomic configurations of the surfaces are needed which can be provided by any total energy electronic structure method. This formalism has been applied successfully in many surface studies^{11–23}. The cited articles represent only a small selection. More detailed descriptions of the surface phase diagram technique are included in Refs. 13, 20, 23.

We start from the basic assumption that we have a surface in thermodynamic equilibrium with separate reference phases for each atomic species i at a given temperature T and pressure p . The surface can exchange any amount of atoms with each reference phase (by adsorption and desorption processes) without changing the temperature or pressure. However, the overall number of atoms N_i^{tot} in the combined system will be constant. This situation is sketched schematically for one atomic species in Figure 2. The appropriate thermodynamic potential for a situation with fixed temperature T , pressure p and particle numbers N_i^{tot} is the Gibbs free energy. $G_{\text{slab}}(T, p, \{N_i\})$ and $G^{(i)}(T, p, N_i^{\text{ref}})$ are the Gibbs free energy of the surface (which we assume to be modeled by a slab) and of the reference phases i , respectively. $\{N_i\}$ are the number of atoms of the different species in the slab, and N_i^{ref} are the number of remaining atoms in the reference phases. Together they hold

$$N_i + N_i^{\text{ref}} = N_i^{\text{tot}} \quad . \quad (1)$$

In thermodynamic equilibrium the most stable structure and composition of the surface is the one that minimizes the total Gibbs free energy of the combined system:

$$G_{\text{slab}}(T, p, \{N_i\}) + \sum_i G^{(i)}(T, p, N_i^{\text{ref}}) = \min \quad . \quad (2)$$

This condition can be stated in a slightly different way. First, we make use of the Euler equation

$$G^{(i)}(T, p, N_i^{\text{ref}}) = (N_i^{\text{tot}} - N_i) \mu_i(T, p) \quad (3)$$

to replace the Gibbs free energies of the reference phases by chemical potentials $\mu_i(T, p)$. In contrast to the usual convention in macroscopic thermodynamics we define here the chemical potentials per atom rather than per mole. Since we are at thermodynamic equilibrium the chemical potentials in the reference phases and the surface box will be the same. This can be shown easily by taking the derivatives of Eq. (2) with respect to the particle numbers which have to vanish for a minimum. By subtracting the total Gibbs free energy of the atoms $\sum_i N_i^{\text{tot}} \mu_i(T, p)$ from Eq. (2) and normalizing the result to an energy per unit area by dividing through the surface area A of the slab surface unit cell we can define the Gibbs free surface energy¹

$$\gamma(T, p) = \frac{1}{A} \left(G_{\text{slab}}(T, p, \{N_i\}) - \sum_i N_i \mu_i(T, p) \right) \quad . \quad (4)$$

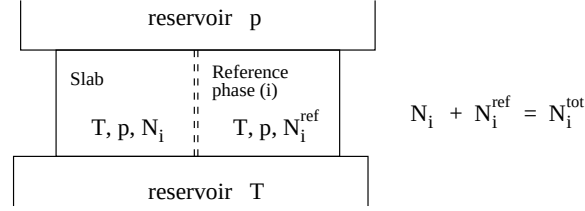


Figure 2. Schematic illustration of a surface in contact with a reference phase of atomic species i at a given temperature T and pressure p . The vertical dashed line symbolizes that the surface and the reference phase can exchange atoms by adsorption and desorption processes.

(Note that sometimes a factor of $1/2$ appears in Eq. (4) to take the two equivalent surfaces of a slab into account). It follows directly from Eq. (2) that the most stable surface configuration also minimizes the Gibbs free surface energy $\gamma(T, p)$. In thermodynamic equilibrium the chemical potentials are determined uniquely by the temperature T and the pressure p . Therefore, the optimum surface structure and composition at given T and p can be determined by a unconstrained minimization of Eq. (4) with respect to the particle numbers N_i . However, since we want to determine G_{slab} with rather time consuming total energy electronic structure calculations this is not very practical to do. Instead we will take a different approach: We calculate the surface free energy of a series of model surfaces with different structures and compositions as a function of the chemical potentials. For given chemical potentials we predict which surface structure is the most stable one by searching for the surface model with the lowest surface free energy. In a second step, the chemical potentials are then related to actual temperature and pressure conditions by assuming that the surface is in thermodynamic equilibrium with the reference phases.

For simplicity we restrict the discussion from now on explicitly to binary AB compounds. The generalization to compounds with A_xB_y composition or with more atomic species is quite obvious. In this case we have to consider only two chemical potentials: μ_A and μ_B . Those two chemical potentials are, however, not independent. Since we can always assume that the surface is in thermodynamic equilibrium with its own bulk material they have to add up to the Gibbs free energy of a bulk unit cell:

$$\mu_A(T, p) + \mu_B(T, p) = G_{\text{bulk}}^{\text{AB}}(T, p) \quad . \quad (5)$$

Thus, we can eliminate one of the chemical potentials from Eq. (4):

$$\gamma(T, p) = \frac{1}{A} \left(G_{\text{slab}}(T, p, N_A, N_B) - N_A G_{\text{bulk}}^{\text{AB}}(T, p) + (N_A - N_B) \mu_B(T, p) \right) \quad . \quad (6)$$

Furthermore, since we are mostly interested in the *relative* stability of surfaces with different composition, it is useful to introduce a reference surface structure with Gibbs free energy $G_{\text{slab}}^0(T, p)$ and particle numbers N_A^0 , N_B^0 and relate all energies relative to the reference surface. We define the change of the Gibbs free surface energy $\Delta\gamma(T, p)$ as the difference of Eq. (4) for a slab with a chosen surface structure and the reference surface

$$\begin{aligned} \Delta\gamma(T, p) = \frac{1}{A} \left(G_{\text{slab}}(T, p, \Delta N_A, \Delta N_B) - G_{\text{slab}}^0(T, p) - \Delta N_A G_{\text{bulk}}^{\text{AB}}(T, p) \right. \\ \left. + (\Delta N_A - \Delta N_B) \mu_B(T, p) \right) \end{aligned} \quad (7)$$

where we have introduced the changes in the particle numbers $\Delta N_A = N_A - N_A^0$ and $\Delta N_B = N_B - N_B^0$. The difference $\Delta\gamma(T, p)$ is negative if a surface is more stable than the reference structure and positive otherwise.

3 Gibbs Free Energies vs Total Energies

In principle we have to calculate now the Gibbs free energy of many slabs with different structure and composition to evaluate Eq. (7), including the contributions coming from changes in volume and in entropy. The entropy term may be calculated, for example, by evaluating the vibrational spectra in a quasiharmonic approximation²⁴, but in practice this is computationally very demanding. As is apparent from Eq. (7), only the *difference* of the Gibbs free energy of two surface structures enters the expression for $\Delta\gamma(T, p)$. In Ref. 20 it was shown that the vibrational contributions to the entropy usually cancel to a large extent and that the influence of volume changes are even smaller. Therefore it is quite common to neglect all entropy and volume effects for the slabs and the bulk phases and to replace the Gibbs free energies G by the internal energies E as they are obtained from the total energy calculations:

$$\Delta\gamma(T, p) = \frac{1}{A} \left(E_{\text{slab}}(\Delta N_A, \Delta N_B) - E_{\text{slab}}^0 - \Delta N_A E_{\text{bulk}}^{\text{AB}} + (\Delta N_A - \Delta N_B) \mu_B(T, p) \right) \quad (8)$$

However, via the chemical potentials we keep the full temperature and pressure dependence of our reference phases, which can be, for example, surrounding gas phases.

4 Limits for the Chemical Potentials

Finally we have to determine meaningful ranges in which we can vary the chemical potentials. First, there are upper bounds for all chemical potentials which are given by the total energy E_0 of their most stable phases as pure elements¹³ (neglecting again volume and entropy effects):

$$\mu_A \leq E_0^A \quad , \quad \mu_B \leq E_0^B \quad . \quad (9)$$

If we assume that the chemical potentials exceed this limit, then we could lower the surface energy (Eq. (8)) simply by condensating the elemental phases on the slab surfaces. These upper bounds are usually taken as the zero point of energy and the chemical potentials are related relative to the total energies of the elements:

$$\Delta\mu_A = \mu_A - E_0^A \quad , \quad \Delta\mu_B = \mu_B - E_0^B \quad . \quad (10)$$

Second, through Eq. (5) we immediately also introduce lower bounds for the chemical potentials. Using the energy of formation E_f of our compound AB:

$$E_f = \left| E_{\text{bulk}}^{\text{AB}} - E_0^A - E_0^B \right| \quad (11)$$

the allowed range for the chemical potentials is given by

$$-E_f \leq \Delta\mu_A \leq 0 \quad , \quad -E_f \leq \Delta\mu_B \leq 0 \quad . \quad (12)$$

Finally we have to relate the chemical potentials to actual temperature and pressure conditions. Accurate ab initio calculations for these relations are quite demanding. On the other hand, the temperature and pressure dependence of the chemical potentials for the different elements are well known from experiment, and it is much easier (and more accurate!) to take these values from thermochemical reference tables²⁵.

If one of the reference phases is a gas phase then it is often a very good approximation just to assume an ideal gas behavior. For example, for the pressure dependence of the chemical potential we can make use of the well-known thermodynamic expression for ideal gases:

$$\Delta\mu(T, p) = \tilde{\mu}(T, p^0) + k_B T \ln(p/p^0) \quad (13)$$

in which p^0 is the pressure of a reference state. The temperature dependence of the chemical potentials $\tilde{\mu}(T, p^0)$ at the reference pressure p^0 again can be taken from thermochemical reference tables²⁵.

Another possibility is that a surface is in thermodynamic equilibrium with the atomic defects of the bulk phase²⁶. In this case, the chemical potentials are related to the defect concentrations of the bulk. For example, if vacancies are the dominant thermal defects the chemical potential is given by^{5,9,10}

$$\Delta\mu(T, p) = \tilde{\mu}_0 - k_B T \ln(c_V/c_0) \quad (14)$$

c_V is the vacancy concentration and c_0 is the occupancy of the respective sublattice.

5 Restricted Thermodynamics

In the multi-component case beyond binary compounds various reactions between the different species are possible which will lead to further conditions for the chemical potentials. However, in some cases this is not desired. For example, if H_2 and O_2 are present, they would react to form H_2O in full thermodynamic equilibrium. This can be avoided by strictly keeping separate reservoirs for H and O and ignoring the formation of H_2O . This is justified by arguing that the reaction barrier for the formation of H_2O is high enough that the reaction plays no role on the time scales of interest.

6 Example: Polar O-Terminated ZnO(000 $\bar{1}$) Surface

We will briefly illustrate the surface phase diagram technique for the example of the polar O-terminated ZnO(000 $\bar{1}$) surface which has been studied in more detail in Ref. 23. The ideal truncated-bulk structure of this surface, which we have taken as our reference surface, is illustrated in Figure 3. To keep the example as simple as possible we allow only O atoms to adsorb and desorb from the surface. Hence, we assume that the surface is in thermodynamic equilibrium with an O_2 gas phase. Since the number of Zn atoms remains constant, the Gibbs free surface energy is given by

$$\Delta\gamma(T, p) = \frac{1}{A} \left(E_{\text{slab}}(N_V) - E_{\text{slab}}^{\text{ideal}} + \frac{1}{2} E_{\text{mol}}^{O_2} + N_V \Delta\mu_O(T, p) \right) \quad (15)$$

N_V is the number of oxygen vacancies that have been created in the surface unit cell.

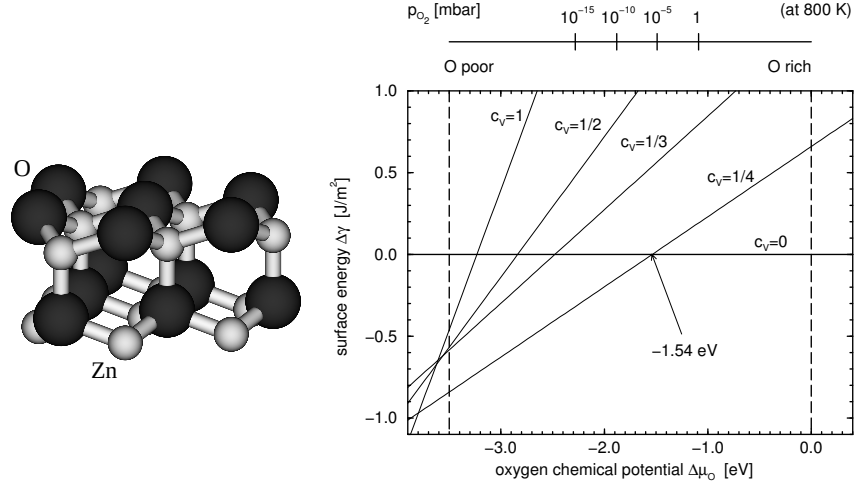


Figure 3. Surface phase diagram of the polar O-terminated ZnO(0001) surface in thermodynamic equilibrium with an O₂ gas phase. The surface free energy $\Delta\gamma$ for surfaces with different concentrations of O vacancies c_v is plotted as a function of the oxygen chemical potential $\Delta\mu_O$. In the top x axis, the chemical potential $\Delta\mu_O(T, p)$ has been translated into a pressure scale for the fixed temperature of $T = 800$ K. Vertical dashed lines indicate the allowed range of $\Delta\mu_O$: for higher chemical potentials O₂ will condensate on the surface and for lower values of $\Delta\mu_O$ metallic bulk Zn can form.

For a series of surface models with different vacancy concentrations the total energy $E_{\text{slab}}(N_V)$ has been calculated using a first-principles density functional theory approach²⁷. From slabs with (1×1) , (1×2) , (1×3) and (2×2) surface unit cells one O atom has been removed, thereby creating vacancy concentrations c_v of 1, 1/2, 1/3 and 1/4. In Figure 3 we have plotted the change of the Gibbs free surface energy $\Delta\gamma$ for the four defect structures relative to the defect-free ideal surface termination as a function of the oxygen chemical potential $\Delta\mu_O$. For low O chemical potentials (i.e. O poor reducing conditions) we find that the defective surface with 1/4 of the O atoms missing is the most stable configuration. Translating the chemical potential into temperature and pressure conditions (see upper x -axis in Figure 3) we see that this will be the most stable surface at typical UHV conditions of surface science experiments. However, at oxygen rich conditions (high pressure and low temperature, i.e. O rich oxidizing conditions), exceeding a chemical potential of -1.54 eV, the ideal, defect-free surface becomes the most stable structure. The other surfaces with 1, 1/2 and 1/3 vacancy concentrations are higher in energy for all chemical potentials and will therefore not be present in thermodynamic equilibrium.

It should be emphasized that plots like Figure 3 strictly allow only to rule out surface structures which are higher in energy than other surface models. Since we can only perform calculations for a limited set of surface models, it is always possible that a not yet considered structure with a lower energy exists. For example, since we use periodically repeated surface unit cells of a specific size in our DFT approach, all our defect structures are perfectly ordered. It is very well possible, however, that disordered or even incommensurate structures might lead to a lower energy.

References

1. J.W. Cahn, in: *Interfacial Segregation*, Eds. W.C. Johnson and J.M. Blakely, (American Society for Metals, Ohio, 1977).
2. M.W. Finnis, Phys. Stat. Sol. (a) **166**, 397 (1998).
3. S.M. Foiles and M.S. Daw, J. Mater. Res. **2**, 5 (1987).
4. C.L. Fu, Y.Y. Ye, M.H. Yoo, and K.-M. Ho, Phys. Rev. B **48**, 6712 (1993).
5. J. Mayer, C. Elsässer, and M. Fähnle, Phys. Stat. Sol. (b) **191**, 283 (1995).
6. J. Mayer and M. Fähnle, Acta Mater. **45**, 2207 (1997).
7. J. Mayer and M. Fähnle, Scr. Mater. **37**, 131 (1997).
8. M. Fähnle, G. Bester, and B. Meyer, Scr. Mater. **39**, 1071 (1998).
9. M. Hagen and M.W. Finnis, Phil. Mag. A **77**, 447 (1998)).
10. B. Meyer and M. Fähnle, Phys. Rev. B **59**, 6072 (1999).
11. E. Kaxiras, K.C. Pandey, Y. Bar-Yam, and J.D. Joannopoulos, Phys. Rev. Lett. **56**, 2819 (1986).
12. E. Kaxiras, Y. Bar-Yam, J.D. Joannopoulos, and K.C. Pandey, Phys. Rev. B **35**, 9625 (1987).
13. G.-X. Qian, R.M. Martin, and D.J. Chadi, Phys. Rev. B **38**, 7649 (1988).
14. J. Padilla and D. Vanderbilt, Phys. Rev. B **56**, 1625 (1997).
15. X.-G. Wang, W. Weiss, S.K. Shaikhutdinov, M. Ritter, M. Petersen, F. Wagner, R. Schlögl, and M. Scheffler, Phys. Rev. Lett. **81**, 1038 (1998).
16. A. Pojani, F. Finocchi, and C. Noguera, Surf. Sci. **442**, 179 (1999).
17. I.G. Batyrev, A. Alavi, and M.W. Finnis, Faraday Discuss. **114**, 33 (1999).
18. I.G. Batyrev, A. Alavi, and M.W. Finnis, Phys. Rev. B **62**, 4698 (2000).
19. X.-G. Wang, A. Chaka, and M. Scheffler, Phys. Rev. Lett. **84**, 3650 (2000).
20. K. Reuter and M. Scheffler, Phys. Rev. B **65**, 035406 (2001).
21. K. Reuter and M. Scheffler, Phys. Rev. Lett. **90**, 046103 (2003).
22. G. Kresse, O. Dulub, and U. Diebold, Phys. Rev. B **68**, 245409 (2003).
23. B. Meyer, Phys. Rev. B **69**, 045416 (2004).
24. W. Frank, U. Breier, C. Elsässer, and M. Fähnle, Phys. Rev. Lett. **77**, 518 (1996).
25. D.R. Stull and H. Prophet, *JANAF Thermochemical Tables*, 2nd ed., (US National Bureau of Standards, Washington DC, 1971).
26. M.W. Finnis, Psi-k Newsletter (1999); <http://psi-k.dl.ac.uk>.
27. B. Meyer, F. Lechermann, C. Elsässer, and M. Fähnle, FORTRAN 90 program for mixed-basis pseudopotential calculations for crystals, (Max-Planck Institut für Metallforschung, Stuttgart).

Spin-Polarized DFT Calculations and Magnetism

Rudolf Zeller

Institute of Solid State Research
Forschungszentrum Jülich
52425 Jülich, Germany
E-mail: ru.zeller@fz-juelich.de

The main aim of this presentation is to give an introduction into spin-polarized density-functional-theory (DFT) calculations as a method to predict and understand the band magnetism of itinerant electrons at zero temperature. It will be pointed out that, for instance the ferromagnetism of Fe, Co, and Ni is correctly predicted by such calculations without adjustable parameters and that the results can be explained in terms of a simple Stoner model. For practical applications it will be discussed why the convergence of density-functional self-consistency iterations can be very slow for magnetic systems and how this problem can be alleviated. As illustrating examples of recent applications in computational nanoscience, results for magnetic monatomic rows on noble metal substrates will be presented. Finally it will be argued that the magnetism in certain materials, for instance in antiferromagnets, does not arise from rigid-band shifts as in the Stoner model, but from covalent interactions between the electronic states of the participating atoms.

1 Introduction

Spin-polarized calculations within the framework of density-functional theory (DFT) are a powerful tool to describe the magnetism of itinerant electrons in solid state materials. Such calculations are not only the basis for a quantitative theoretical determination of spin magnetic moments, but can also be used to understand the basic mechanisms, which lead to the occurrence of magnetism in solid state materials. Since the original development of spin-density-functional theory by von Barth and Hedin¹ and Pant and Rajagopal² thousands of spin-polarized DFT calculations have been performed and published and it is thus entirely impossible to cover even the most relevant part of them. Therefore, the restricted aim of these lecture notes is the consideration of some aspects of zero-temperature spin magnetism in conceptually simple transition-metal solid state systems.

The tendency toward magnetism is determined by a competition between exchange and kinetic energy effects. Whereas the parallel alignment of the electronic spins leads to a gain of exchange energy, the alignment also causes a loss of kinetic energy. Contrary to the atoms, which usually are magnetic, most solid state systems are non-magnetic, since the gain in exchange energy is dominated by the loss in kinetic energy, which arises from the delocalization of the valence electrons in a solid. Only if these electrons are sufficiently localized, magnetism occurs as, for instance, in the elemental metals Fe, Co, Ni and Cr. These metals will serve here as examples to illustrate the predictive and explaining power of spin-polarized DFT calculations. The tendency toward magnetism is considerably enhanced in “lower-dimensional” systems like metallic surfaces and interfaces, multilayers, ultrathin films and wires, and magnetic clusters deposited on surfaces. These magnetic systems have received considerable experimental and theoretical attention in recent years and have been extensively studied by spin-polarized DFT calculations. Out of these calculation, I will consider here one typical “one-dimensional” example, the magnetism of

nanowires on metal substrates and present spin-density-functional results, which have been obtained by the Korringa-Kohn-Rostoker (KKR) method, which is introduced elsewhere in this Winter School on Computational Nanoscience.

The outline of my lecture notes is as follows. After an introduction to the underlying theory of spin-polarized density-functional calculations, it will be discussed how, in terms of the Stoner model, these calculations provide a basic description of the ferromagnetism of the elemental metals. This model will then be used to illustrate, why the self-consistency iterations for the solution of the density-functional equations in magnetic systems converge often rather slowly and how this can be avoided. Then follows a presentation of some recent results for magnetic monatomic wires. Finally, it is shown that the magnetism in certain solid state systems, in particular in the antiferromagnets, can be understood in terms of covalent interactions between the atoms, but not in terms of the Stoner model based on rigid band shifts.

2 Spin-Density-Functional Theory

If an external magnetic field is applied to an electronic system, it generally couples both to the electron spin and to the electronic orbital current. A framework for the description of spin coupling is provided by spin-density-functional theory, to which these lecture notes are restricted. For the description of orbital magnetism by density-functional approaches I refer to a recent Letter³ and the references therein.

The basic variables of spin-density-functional theory are the scalar electronic density $n(\underline{r})$ and the vector of the magnetization density $\underline{m}(\underline{r})$. Instead of these four variables, alternatively the 2×2 spin density matrix $n^{\alpha\beta}(\underline{r})$ can be used, where the spin indices α and β can have two values, either $+$ for spin up (majority spin) or $-$ for spin down (minority spin). The notation $\uparrow$ instead of $+$ and $\downarrow$ instead of $-$ is also often used in the literature. The connection between $n(\underline{r})$, $\underline{m}(\underline{r})$ and $n^{\alpha\beta}(\underline{r})$ is given by

$$n(\underline{r}) = \sum_{\alpha} n^{\alpha\alpha}(\underline{r}) \quad \underline{m}(\underline{r}) = \sum_{\alpha\beta} \underline{\sigma}^{\alpha\beta} n^{\alpha\beta}(\underline{r}) \quad (1)$$

and by

$$n^{\alpha\beta}(\underline{r}) = \frac{1}{2} (n(\underline{r})\delta^{\alpha\beta} + m_x(\underline{r})\sigma_x^{\alpha\beta} + m_y(\underline{r})\sigma_y^{\alpha\beta} + m_z(\underline{r})\sigma_z^{\alpha\beta}) . \quad (2)$$

Here 2×2 matrices are denoted by upper Greek indices and $\underline{\sigma} = (\sigma_x, \sigma_y, \sigma_z)$ is a vector consisting of the Pauli spin matrices

$$\sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} , \quad \sigma_y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} , \quad \sigma_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} . \quad (3)$$

The Hohenberg-Kohn-Sham spin-density functional^{1,2} is given by

$$E[n^{\alpha\beta}(\underline{r})] = T_s[n^{\alpha\beta}(\underline{r})] + \frac{e^2}{2} \int \int \frac{n(\underline{r})n(\underline{r}')}{|\underline{r} - \underline{r}'|} d\underline{r} d\underline{r}' \quad (4)$$

$$+ \sum_{\alpha\beta} \int V_{ext}^{\alpha\beta}(\underline{r}) n^{\alpha\beta}(\underline{r}) d\underline{r} + E_{xc}[n^{\alpha\beta}(\underline{r})]$$

as a sum of the kinetic energy T_s of non-interacting electrons, the electron-electron interaction in the Hartree approximation, the interaction energy with the external potential $V_{ext}^{\alpha\beta}$ and the exchange-correlation energy. The external potential $V_{ext}^{\alpha\beta}$ contains the electrostatic Coulomb potential of the nuclei and perhaps the potential arising from a magnetic field. The minimum of the functional (Eq. (4)) is obtained by inserting the ground-state spin-density matrix and yields the ground-state energy. The variation of Eq. (4) with respect to $n^{\alpha\beta}(\underline{r})$ cannot be performed directly, since the explicit form of the density functional $T_s[n^{\alpha\beta}(\underline{r})]$ is not known. This problem can be solved with the help of single-particle wavefunctions (orbitals) $\varphi_i^\alpha(\underline{r})$, which allow to represent the kinetic energy functional $T_s[n^{\alpha\beta}(\underline{r})]$ and the spin-density matrix $n^{\alpha\beta}(\underline{r})$ as

$$T_s[n^{\alpha\beta}(\underline{r})] = \sum_{\alpha i} \int \varphi_i^{\star\alpha}(\underline{r}) \left(-\frac{\hbar^2}{2m} \nabla_{\underline{r}}^2 \varphi_i^\alpha(\underline{r}) \right) d\underline{r} \quad (5)$$

and as

$$n^{\alpha\beta}(\underline{r}) = \sum_i \varphi_i^{\star\alpha}(\underline{r}) \varphi_i^\beta(\underline{r}), \quad (6)$$

where the sum over i includes all occupied orbitals. The last two equations provide an implicit representation of the kinetic energy in terms of the spin-density matrix and the minimum of Eq. (4) can now be found by variation of $E[n^{\alpha\beta}(\underline{r})]$ with respect to the single-particle wavefunctions. This leads to the Kohn-Sham equations

$$-\frac{\hbar^2}{2m} \nabla_{\underline{r}}^2 \varphi_i^\alpha(\underline{r}) + \sum_{\beta} V_{eff}^{\alpha\beta}(\underline{r}) \varphi_i^\beta(\underline{r}) = \epsilon_i \varphi_i^\alpha(\underline{r}), \quad (7)$$

where the ϵ_i represent Lagrange parameters, which guarantee that the wavefunctions are normalized as $\sum_{\alpha} (\varphi_i^\alpha, \varphi_i^\alpha) = 1$. The potential $V_{eff}^{\alpha\beta}$, which is defined as

$$V_{eff}^{\alpha\beta}(\underline{r}) = \delta^{\alpha\beta} e^2 \int \frac{n(\underline{r}')}{|\underline{r} - \underline{r}'|} d\underline{r}' + V_{ext}^{\alpha\beta}(\underline{r}) + V_{xc}^{\alpha\beta}(\underline{r}), \quad (8)$$

represents an effective one-particle potential. Useful approximations for the exchange-correlation energy functional can be given in terms of the eigenvalues $n^+(\underline{r})$ and $n^-(\underline{r})$ of the spin-density matrix $n^{\alpha\beta}(\underline{r})$. This matrix can be diagonalized as

$$\sum_{\alpha'\beta'} U^{\alpha\alpha'}(\underline{r}) n^{\alpha'\beta'}(\underline{r}) U^{\beta\beta'}(\underline{r}) = \delta^{\alpha\beta} n^\alpha(\underline{r}), \quad (9)$$

where $U^{\alpha\beta}(\underline{r})$ are spin-1/2 rotation matrices and $n^\alpha(\underline{r})$ the eigenvalues. Note that these matrices and eigenvalues generally depend on the position $\underline{r}$. In many applications, for instance in ferromagnetic and antiferromagnetic solids, a common magnetization axis exists for all atoms. The z axis can then be chosen globally along the direction of the magnetic field and the spin-1/2 rotation matrices in Eq. (9) are independent of $\underline{r}$. This has the simplifying consequence that the energy and all other physical observables are functionals of the density and of the magnitude of the magnetization density $m(\underline{r}) = |\underline{m}(\underline{r})|$ rather than of the vector $\underline{m}(\underline{r})$. For a discussion of the more general situation of non-collinear magnetism, where the magnetization axis changes within the system, I refer to the presentation by Bihlmayer elsewhere in this Winter School on Computational Nanoscience.

In terms of the spin up and spin down orbitals $\varphi_i^+(\underline{r})$ and $\varphi_i^-(\underline{r})$, the spin densities $n^+(\underline{r})$ and $n^-(\underline{r})$ can be represented as

$$n^\pm(\underline{r}) = \sum_i |\varphi_i^\pm(\underline{r})|^2 \quad (10)$$

and the Kohn-Sham equations (7) for the orbitals can be written as

$$\left[-\frac{\hbar^2}{2m} \nabla_{\underline{r}}^2 + V_{eff}^\pm(\underline{r}) \right] \varphi_i^\pm(\underline{r}) = \epsilon_i^\pm \varphi_i^\pm(\underline{r}) \quad (11)$$

with the effective potential

$$V_{eff}^\pm(\underline{r}) = e^2 \int \frac{n(\underline{r}')}{|\underline{r} - \underline{r}'|} d\underline{r}' + V_{ext}^\pm(\underline{r}) + V_{xc}^\pm(\underline{r}) . \quad (12)$$

In an externally applied magnetic field H , the external potential $V_{ext}^\pm$ contains a field term $-(\pm\mu_B H)$, where the negative sign means that the majority (spin up: +) electrons are energetically favored compared to the minority (spin down: -) electrons. The exchange-correlation potential $V_{xc}^\pm(\underline{r})$, which is defined by the functional derivative of the exchange-correlation energy as

$$V_{xc}^\pm(\underline{r}) = \frac{\delta E_{xc}[n^+(\underline{r}), n^-(\underline{r})]}{\delta n^\pm(\underline{r})} , \quad (13)$$

can have different values for the two spin directions even without an external magnetic field. This is, for instance, realized in the ferromagnetic transition metals Fe, Co and Ni.

The spin-density-functional theory presented above is exact in principle, however the functionals E_{xc} and $V_{xc}^\pm$, in which all complications of the many-electron system are hidden, are not known and must be approximated. Useful approximation like the local-spin-density approximation^{1,4,5} (LSDA), which depends locally on the spin densities, and the generalized gradient approximation (GGA)^{6,7}, which also depends on the gradients of the spin densities, have been developed and shown to be rather accurate for many applications.

3 Stoner Model for Ferromagnetism

It has independently been shown by several authors⁸⁻¹¹ that the spin-density formalism can be used to derive a Stoner model, which includes correlation effects, and to calculate the relevant Stoner parameter I . For this derivation it is convenient to switch from the spin densities $n^+(\underline{r})$ and $n^-(\underline{r})$ to the electron density

$$n(\underline{r}) = n^+(\underline{r}) + n^-(\underline{r}) \quad (14)$$

and the magnetization density

$$m(\underline{r}) = n^+(\underline{r}) - n^-(\underline{r}) \quad (15)$$

as variables to describe the system. Usually the magnetization $m(\underline{r})$ is a small parameter compared to the density $n(\underline{r})$. This means that the exchange-correlation potential can be expanded in terms of $m(\underline{r})$, which leads to the approximation

$$V_{xc}^\pm(\underline{r}) = V_{xc}^o(\underline{r}) \mp m(\underline{r}) \tilde{V}(n(\underline{r})) , \quad (16)$$

where the higher order terms in $m(\underline{r})$ are neglected and where $V_{xc}^o(\underline{r})$ is the exchange-correlation potential for non-spin-polarized electrons. The average value of $\tilde{V}(n(\underline{r}))$ is positive such that a more attractive potential acts on the majority electrons (with spin $+$) and a less attractive potential on the minority electrons (with spin $-$). Compared to Eq. (16) the Stoner model¹²⁻¹⁴ uses the approximation

$$V_{xc}^\pm(\underline{r}) = V_{xc}^o(\underline{r}) \mp \frac{1}{2}IM, \quad (17)$$

where the potential difference is simulated by $\underline{r}$ independent constants. Here M is chosen as the local moment

$$M = \int_{\Omega} m(\underline{r}) d\underline{r}, \quad (18)$$

which is given as an integral of the magnetization density $m(\underline{r})$ over the atomic unit cell Ω , and I is the exchange integral (= Stoner parameter). In ferromagnetic materials the moment M has the same value for all atoms. While the constant changes of the potential in (17) do not affect the wavefunctions $\varphi_i^\pm(\underline{r})$, which are identical to the non-magnetic ones, all eigenvalues $\epsilon_i^\pm$ are shifted by a constant amount $\mp \frac{1}{2}IM$. In periodic crystals the wavefunctions and eigenvalues are characterized by wavevectors $\underline{k}$ and band indices ν such that

$$\varphi_{\underline{k}\nu}^\pm(\underline{r}) = \varphi_{\underline{k}\nu}^o(\underline{r}) \quad \epsilon_{\underline{k}\nu}^\pm = \epsilon_{\underline{k}\nu}^o \mp \frac{1}{2}IM \quad (19)$$

holds. The constant shift of the eigenvalues $\epsilon_{\underline{k}\nu}^\pm$ describes an exchange-splitting $\Delta = IM$ between the spin up and spin down bands and leads to the following representation

$$n^\pm(E) = \sum_{\nu} \int_{BZ} \delta(E - \epsilon_{\underline{k}\nu}^\pm) d\underline{k} = n^o(E \pm \frac{1}{2}IM) \quad (20)$$

for the spin densities of states in terms of the density of states $n^o(E)$ of the non-magnetic system, where the integral in Eq. (20) is over the Brillouin zone (BZ). A good approximation of the constant shift $\pm \frac{1}{2}IM$ can be seen in Figure 3, where realistic densities of states are shown, which were obtained by density-functional calculations in the local-spin-density approximation. From Eq. (20) one obtains the number N of electrons per atom and the moment M in the unit cell by integration over all occupied states below the Fermi energy E_F as

$$N = \int^{E_F} \left[n^o(E + \frac{1}{2}IM) + n^o(E - \frac{1}{2}IM) \right] dE \quad (21)$$

and as

$$M = \int^{E_F} \left[n^o(E + \frac{1}{2}IM) - n^o(E - \frac{1}{2}IM) \right] dE. \quad (22)$$

Since $n^o(E)$ is determined by the non-magnetic calculation and N by the condition of charge neutrality, the two equations (21) and (22) can be used to calculate the two unknown variables E_F and M . Equation (21) implicitly defines $E_F = E_F(M)$ as function of M . The moment M can then be determined by the equation

$$M = F(M), \quad (23)$$

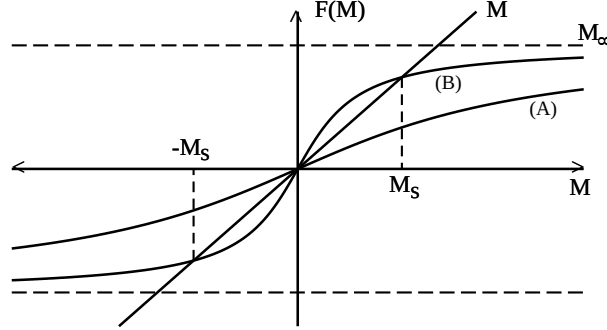


Figure 1. Graphical solution for the Stoner model. The intersection of $F(M)$ with the straight line M determines the solution M_s . The intersection at $M = 0$ is always a trivial solution.

where the function $F(M)$ is defined as

$$F(M) = \int^{E_F(M)} \left[n^o(E + \frac{1}{2}IM) - n^o(E - \frac{1}{2}IM) \right] dE. \quad (24)$$

From the definition (Eq. (24)) one can easily derive that the function $F(M)$ satisfies the four conditions

$$\begin{aligned} F(M) &= -F(-M), & F(0) &= 0, \\ F(\pm\infty) &= \pm M_\infty, & F'(M) &> 0, \end{aligned}$$

where the last condition arises from the fact that the density $n^o(E)$ is always positive. Here M_∞ is the saturation magnetization for full spin-polarization, when all majority states are occupied and all minority states are empty. This situation corresponds to the atomic limit, where according to Hund's first rule the spin moment is maximal.

Two different possibilities for $F(M)$ are shown in Figure 1. For the function denoted by (A), the equation $M = F(M)$ has only the trivial non-magnetic solution $M = 0$. For the function denoted by (B) three solutions exist, $M = \pm M_S$ with a finite spontaneous magnetization and $M = 0$. This non-magnetic solution is however unstable. From Figure 1 it is clear that a solution with finite magnetization always exists, if the slope of $F(M)$ at $M = 0$ is larger than one. Thus $F'(0) > 1$ is a sufficient condition for ferromagnetic solutions. From Eq. (24) the derivative $F'(M)$ can be calculated as

$$\begin{aligned} F'(M) &= \frac{I}{2} \left[n^o(E_F + \frac{1}{2}IM) + n^o(E_F - \frac{1}{2}IM) \right] \\ &+ \left[n^o(E_F + \frac{1}{2}IM) - n^o(E_F - \frac{1}{2}IM) \right] \frac{dE_F}{dM}. \end{aligned} \quad (25)$$

For $M = 0$ the last equation simplifies to

$$F'(0) = In^o(E_F), \quad (26)$$

and the sufficient condition for ferromagnetic solutions $F'(0) > 1$ can be written as

$$In^o(E_F) > 1. \quad (27)$$

This is the famous Stoner criterion, which shows that a large exchange integral I and a large density of states $n^o(E_F)$ are properties, which are favorable for ferromagnetism.

The Stoner model can be extended to include the effects of an external magnetic field H . This allows to determine the spin susceptibility X , which according to $M = XH$ describes the behavior for small magnetic fields. Instead of Eq. (17) the relevant potential then is given by

$$V^\pm(\mathbf{r}) = V^o(\mathbf{r}) \mp \frac{1}{2}IM \mp \mu_B H \quad (28)$$

and instead of $M = F(M)$ the equation

$$M = F\left(M + \frac{2\mu_B H}{I}\right) \quad (29)$$

must be solved self-consistently. Linearizing (Eq. (29)) around M_0 , which would be the magnetic moment without field, leads to

$$\begin{aligned} \Delta M = M - M_0 &= F\left(M + \frac{2\mu_B H}{I}\right) - F(M_0) \\ &\approx F'(M_0) \left(\Delta M + \frac{2\mu_B H}{I} \right) \\ &\approx \frac{1}{1 - F'(M_0)} F'(M_0) \frac{2\mu_B H}{I}. \end{aligned} \quad (30)$$

For $M_0 = 0$ the derivative $F'(M_0 = 0)$ is given by Eq. (26) and one obtains from Eq. (30) the result

$$M = \frac{1}{1 - In^o(E_F)} n^o(E_F) 2\mu_B H. \quad (31)$$

With $X = M/H$ the spin susceptibility is then given by

$$X = \frac{1}{1 - In^o(E_F)} n^o(E_F) 2\mu_B = SX^o, \quad (32)$$

where $X^o = n^o(E_F) 2\mu_B$ is the Pauli spin susceptibility, which is obtained for non-interacting electrons if the exchange interaction is neglected. The exchange interaction leads to a Stoner enhancement factor $S = [1 - In^o(E_F)]^{-1}$, which diverges for $In^o(E_F) = 1$. The non-magnetic state is stable for $In^o(E_F) < 1$, whereas the ferromagnetic state is stable for $In^o(E_F) > 1$. Ferromagnetic behavior is therefore favored, if the exchange integral I is large, and even more important, if the density of states $n^o(E_F)$ at the Fermi energy E_F is large in the non-magnetic calculation.

The density of states usually shows a rather detailed structure. However, in a simple approximation one can assume that the density of states scales inversely to the bandwidth W . This inverse relation is exact for a constant density of states, since the integral of the density of states over the bandwidth

$$\int_W n_l^o dE = 2l + 1 \quad (33)$$

is determined by the number of states for a given quantum number l , from which $n_l^o = (2l + 1)/W$ follows. For delocalized electrons the bandwidth is large and the density of states is small, whereas for more localized electrons the bandwidth is smaller and

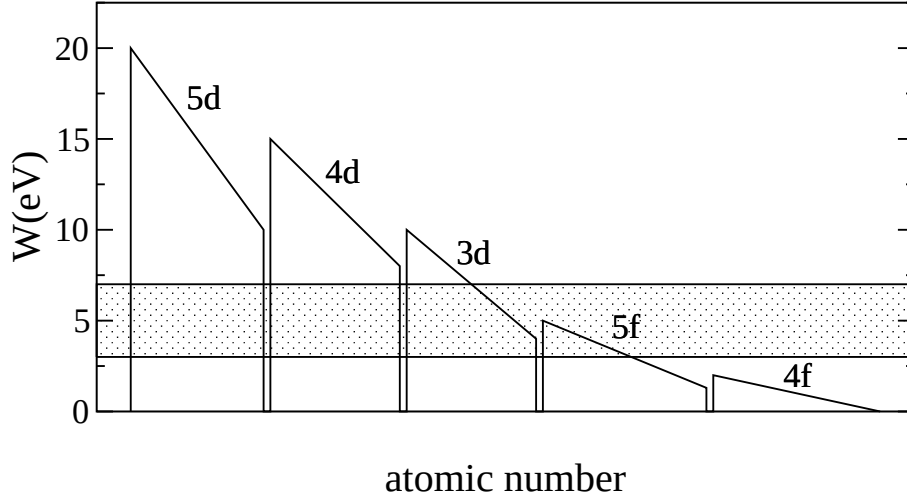


Figure 2. Schematic illustration for the bandwidth W of the transition metals ($3d$, $4d$ and $5d$ electrons), the rare earth metals ($4f$ electrons) and the actinides ($5f$ electrons). The shaded rectangle represents the region favorable for itinerant magnetism. Above the rectangle magnetism is suppressed and below the rectangle localized (atomic) magnetism is preferred.

the density of states larger. In the atomic limit the bandwidth goes to zero, the Stoner criterion is always satisfied and the magnetic moment attains its maximum according to Hund's first rule. Figure 2 shows a schematic representation for the bandwidths of the $3d$, $4d$ and $5d$ transition metals, the rare earth metals and the actinides. The $5f$ electrons of the early actinides and the $3d$ electrons of the late transition metals from Cr to Ni have a tendency for band magnetism, whereas the late actinides and the rare earth metals show localized magnetism with almost atomic moments in good agreement with Hund's rules.

The exchange integral I is an intra-atomic, element specific quantity, which in simplest approximation is independent of the local environment, which means independent of the structure and the site of a given atom, e.g., a surface, bulk or impurity atom. According to Gunnarsson¹⁰ and Janak¹⁵ the trend for the exchange integrals of the $3d$, $4d$ and $5d$ transition-metal series is

$$I_{3d} > I_{4d} > I_{5d}$$

which means that magnetism in the $3d$ elements is more likely than in the $4d$ and $5d$ elements. The same tendency for magnetism arises from the trend

$$n_{3d}^o > n_{4d}^o > n_{5d}^o$$

for the density of states, which corresponds to the bandwidth behavior

$$W_{3d} < W_{4d} < W_{5d} .$$

This behavior can be explained by the larger hybridization between d orbitals of larger quantum numbers because of their larger spatial extent as a consequence of the orthogonality condition, which these orbitals must satisfy with respect to the energetically lower lying orbitals of d and of s and p character.

metal	$n^o(E_F)[Ry^{-1}]$	$I[Ry]$	$In^o(E_F)$	$S = X/X_0$
Na	6.2	0.067	0.42	1.71
Al	5.6	0.045	0.25	1.34
Cr	9.5	0.028	0.27	1.36
Mn	21.	0.030	0.63	2.74
Fe	42.	0.034	1.43	-2.34
Co	27.	0.036	0.97	38.2
Ni	55.	0.037	2.04	-0.98
Cu	3.9	0.027	0.11	1.12
Pd	31.	0.025	0.78	4.46

Table 1. Densities of states $n^o(E_F)$ at the Fermi energy, exchange integral I , their product $In^o(E_F)$, and the Stoner enhancement factor S for some selected metals. These result have been obtained in density-functional calculations by the KKR method (from Ref. 15).

metal	$M_{LSDA}[\mu_B/atom]$	$M_{spin}[\mu_B/atom]$	$M[\mu_B/atom]$
Fe	2.15	2.12	2.22
Co	1.56	1.57	1.71
Ni	0.59	0.55	0.61

Table 2. Magnetic moments M_{LSDA} for Fe, Co and Ni calculated with the local-spin-density approximation in comparison with the experimental values for the spin-only moments M_{spin} and for the total moments M including the orbital contributions (from Ref. 17).

Quantitative results of non-spin-polarized density-functional calculations¹⁵ in local-spin-density approximation for the density of states at E_F and for the exchange integral I are given in Table 1 for some selected metals. The results in Table 1 show that Fe and Ni satisfy the Stoner criterion $In^o(E_F) > 1$ and that Co with $In^o(E_F) = 0.97$ almost satisfies this criterion. It is discussed in Ref. 15 that the calculated values in Table 1 provide a lower bound to the susceptibility and that the Co value $In^o(E_F) = 0.97$ does not contradict the observed ferromagnetism for this metal. As a matter of fact in spin-polarized calculations^{16,17}, which do not use the Stoner model and do not rely on an estimate of I , the ferromagnetic state of Co has been found to be more stable than the non-magnetic state. Thus these early spin-density-functional calculations are consistent with the observed ferromagnetism of Fe, Co and Ni. The fcc metal Pd also has a large Stoner factor and is almost magnetic. The experimental Stoner factor for Pd is even approximately twice as large as the one given in Table 1 in agreement with the fact that the tendency for magnetism is underestimated by the approximation for I used in Ref. 15.

Figure 3 shows densities of states obtained by spin-density-functional calculations for bcc Fe and fcc Co. The densities of states have rather similar shapes for both spin directions and are only shifted with respect to each other by the exchange splitting $\Delta = IM$. Thus Eq. (20) is satisfied in good approximation and the Stoner model can be well applied for these metals. Fcc Ni has a similar density of states as Co, but with a smaller exchange splitting. The majority d states for Co and Ni are fully occupied, whereas in the minority states 1.7 electrons are missing for Co and 0.6 electrons for Ni. This leads to moments of

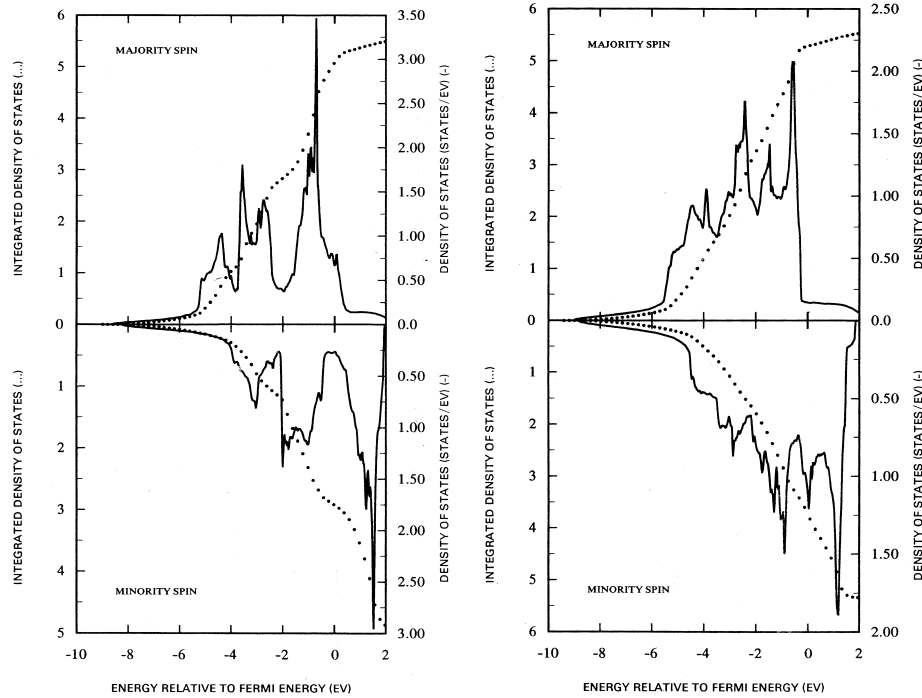


Figure 3. Density of states for bcc Fe (left picture) and fcc Co (right picture) from spin-density-functional calculations. The density of states for majority-spin electrons is plotted upwards and for minority-spin electrons downwards. States for negative energies (below E_F) are occupied, states for positive energies (above E_F) are unoccupied. The dotted curves represent the integrated density of states (from Ref. 17).

$1.7 \mu_B$ per Co atom and $0.6 \mu_B$ per Ni atom. The moment for Fe is calculated as $2.2 \mu_B$. Table 2 shows that the calculated moments agree remarkably well with the experimental values, in view of the fact that the only input to these calculations are the atomic number and the crystal structures. Whereas the majority d states for Co and Ni are completely occupied, there are some empty majority d states in Fe. This difference has been historically characterized as *strong* magnetism for Co and Ni and *weak* magnetism for Fe, since empty d states make the magnetism less robust against environmental changes or, for instance, against pressure. However, it has been pointed out by Malozemoff *et al.*¹⁸ that the pseudogap near the Fermi level in the minority d density of states for Fe also has a large stabilizing effect on the magnetism of Fe.

Figure 4 displays results of band structure calculations for fcc Co. The band structure is shown for high symmetry directions in the Brillouin zone. For each spin direction a separate band structure is obtained with a more or less constant band splitting. (In the Stoner model the splitting is independent of $\mathbf{k}$ according to Eq. (19).) The splitting in Figure 4 is only significant for the d bands. The parabolic s bands are barely split. Figure 4 shows that the majority d bands are completely below E_F in agreement with the fully occupied d density of states of Figure 3.

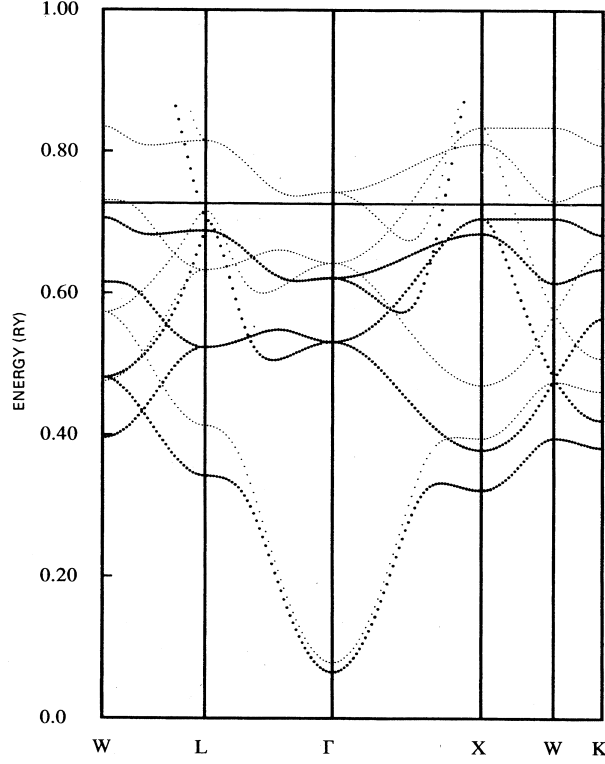


Figure 4. Band structure for fcc Co from spin-density-functional calculations. Bands for majority electrons are indicated by large points and for minority electrons by small points (from Ref. 17).

4 Self-Consistency Iterations

The density-functional equations (6-8) or (10-12) must be solved self-consistently, since the density is given in terms of the orbitals, which depend on the effective potential, which again depends on the density. The resulting nonlinear self-consistency problem, which may be written in the form

$$n(\underline{r}) = F_{\underline{r}}[n] \quad (37)$$

can only be solved by iteration

$$n_{i+1}(\underline{r}) = F_{\underline{r}}[n_i] \quad (38)$$

since other methods are not available. Very often, however, this standard iteration process diverges and leads to density oscillations of increasing amplitude. Then the iteration process must be modified to damp these oscillations. A damping is obtained if the iteration (Eq. (38)) is replaced by a modified iteration

$$n_{i+1} = \alpha F_{\underline{r}}[n_i] + (1 - \alpha)n_i \quad (39)$$

with a suitably chosen mixing parameter α . It has been shown¹⁹ that the convergence behavior of (38) is determined by the eigenvalues λ of a matrix $f(\underline{r}, \underline{r}')$, which is defined as the functional derivative of $F_{\underline{r}}[n]$ with respect to $n(\underline{r}')$. It has also been shown¹⁹ that these eigenvalues are real and that they must satisfy $\lambda < 1$, where this inequality arises from the stability condition for the solution of $n = F_{\underline{r}}[n]$.

The convergence of the iteration process (Eq. (38)) requires that the eigenvalues λ are in the range $-1 < \lambda < 1$, which is, however, often not satisfied, since large negative eigenvalues can exist. On the other hand, the iteration process (Eq. (39)) requires that the eigenvalues λ are in the range defined by

$$-1 < \alpha\lambda + 1 - \alpha < 1. \quad (40)$$

This condition can always be satisfied if $\alpha > 0$ is chosen small enough. By rewriting $\alpha\lambda + 1 - \alpha < 1$ as $\alpha\lambda < \alpha$, it is clear that the right inequality in Eq. (40) is satisfied for positive α because of the stability condition $\lambda < 1$. By rewriting $-1 < \alpha\lambda + 1 - \alpha$ as $\alpha(1 - \lambda) < 2$ or due to $1 - \lambda > 0$ as

$$\alpha < \frac{2}{1 - \lambda}, \quad (41)$$

where the right-hand side is positive because of $\lambda < 1$, it is clear that the left inequality in Eq. (40) can be satisfied if α is chosen small enough. Thus the largest α , for which the iteration process still converges, is determined by the negative eigenvalue λ with the largest absolute value. Unfortunately, this value can be rather large leading to rather small α 's, in particular if charge transfer occurs between the atoms in the considered system. Also intra-atomic charge transfers, for instance between the $4f$ and $5d$ shells of rare-earth metals, can lead to rather small values of suitable α 's. The main disadvantage of small values for α is that only a small amount of the output density $F_{\underline{r}}[n_i]$ is mixed with the input density n_i in Eq. (39). Thus the iterations only slowly move away from the first guess n_0 for input density and many iterations are required to find the self-consistent solution. It is well known that more sophisticated mixing schemes can be used to speed the iterations, for instance the Anderson²⁰ and Broyden²¹ mixing schemes. A comprehensive discussion of these schemes can be found in Refs. 22 and 23.

The slow convergence of the density iterations (Eq. (39)) due to a small mixing parameter α is a particular problem in spin-density functional calculations, if the system is near the threshold between a magnetic or non-magnetic state. This can be illustrated by the Stoner model. Figure 5 shows the iteration process $M_{i+1} = F(M_i)$ for the solution of $M = F(M)$ for two possible situations, if a magnetic solution exists (upper picture) and if the system is almost magnetic with $F'(0) \approx 1$ (lower picture). The iterations $M_{i+1} = F(M_i)$ always converge, since $F'(M_s) < 1$ is valid at the solution M_s , but near the threshold many iterations are necessary to approach the solution $M_s = 0$. If now, as required by the density iterations (Eq. (39)), the iteration process is modified into the much slower iteration process $M_{i+1} = \alpha F(M_i) + (1 - \alpha)M_i$, the convergence is extremely slow and an enormous number of iterations is required.

This problem can be solved to some extent by the observation¹⁹ that near $M_s = 0$ the convergence of the density and the convergence of the magnetization are practically decoupled because of $m(\underline{r}) \ll n(\underline{r})$. This decoupling allows to introduce two different

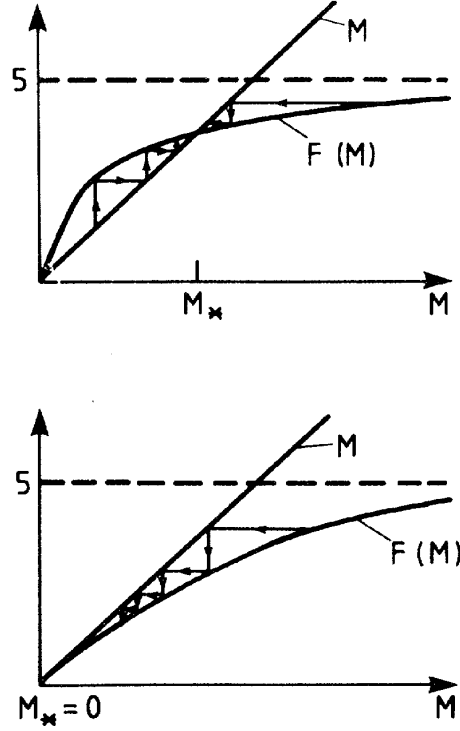


Figure 5. Schematic representation of the iteration process $M_{i+1} = F(M_i)$ for a magnetic system (above) and a non-magnetic system (below). Both iteration processes converge, since $F'(M) < 1$. The convergence is illustrated by the straight lines, where the arrows indicate the direction of the iteration process. Lines with vertical arrows represent the mapping from M to $F(M)$ and horizontal lines the choice of $F(M)$ as the input M for the next iteration (from Ref. 19).

mixing parameters α_n and α_m for the density and the magnetization

$$\begin{aligned} n_{i+1} &= \alpha_n F_n[n_i, M_i] + (1 - \alpha_n)n_i \\ m_{i+1} &= \alpha_m F_m[n_i, M_i] + (1 - \alpha_m)m_i, \end{aligned} \quad (42)$$

where α_m can be chosen much larger than α_n , which considerably improves the convergence near the threshold $F'(0) = 1$. It should be noted that for the Anderson and Broyden mixing schemes a choice of different mixing parameters is not useful, since these schemes are based on a simultaneous minimization of the errors of the density and the magnetization in a multidimensional space and this minimization automatically adapts to the different iteration behavior of the density and the magnetization. On the other hand, contrary to Eq. (39) which converges to a stable solution, the Anderson and Broyden mixing schemes can also converge to unstable solutions. This means that these schemes can easily miss a magnetic solution near the threshold $F'(0) \approx 1$ by improperly converging to the unstable non-magnetic solution.

5 Nanowires

Whereas atoms are usually magnetic, the occurrence of magnetism in the elemental solids is rather an exception. For instance, none of the sp bonded elements and only five of 30 d bonded transition metals are magnetic in their bulk crystalline phase: Co and Ni are ferromagnetic, Cr is antiferromagnetic, and Mn and Fe are ferromagnetic or antiferromagnetic depending on their crystal structure. The main reason preventing magnetism is the over-compensation of the gain in exchange energy by a loss in kinetic energy which occurs for small values of the density of states $n^o(E_F)$ at the Fermi level. Since the exchange integral I is more or less independent of the local environment of a atom in the solid, magnetism in alloys, at surfaces, for impurities in bulk systems and for adatoms at surfaces occurs mainly because of larger values of $n^o(E_F)$ arising from smaller bandwidths. The reduced bandwidth in these system is a consequence of reduced atomic coordination numbers, which leads to less hybridization between the electrons. This can be explained by simple nearest-neighbor tight-binding models, where the bandwidth scales like $W \sim \sqrt{N_c}$. For instance, the coordination number N_c of an atom in a fcc crystal is $N_c = 12$, for an atom in the (100) surface of a fcc crystal it is $N_c = 8$, and at the $(100) \times (111)$ and $(100) \times (110)$ step edges at terraces on the fcc (100) surface (see Figure 6), the coordination numbers are $N_c = 7$ and $N_c = 6$.

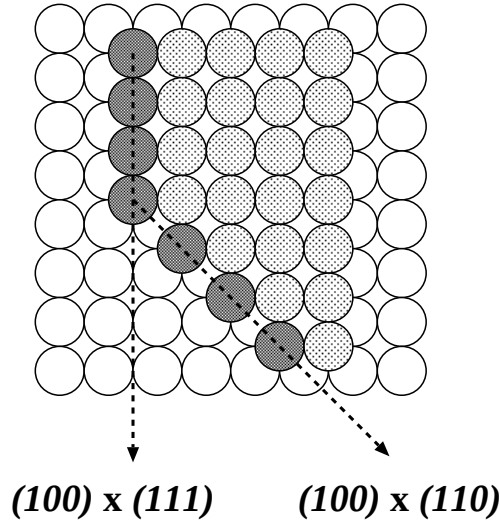


Figure 6. Step edges at the fcc (100) terrace. The step notation from Ref. 24 is used.

From these coordination numbers one expects the bandwidths scale like

$$W_{(100) \times (110)} : W_{(100) \times (111)} : W_{(100)} : W_{fcc} = 0.71 : 0.76 : 0.82 : 1.00$$

and that the densities of states scale like

$$n_{(100) \times (110)} : n_{(100) \times (111)} : n_{(100)} : n_{fcc} = 1.41 : 1.31 : 1.22 : 1.00 .$$

Thus the tendency toward magnetism should be larger at surfaces than in bulk crystals, larger for atoms in step edges than for atoms embedded in flat surfaces and larger for adatoms on surfaces than for the atoms in step edges. Even smaller coordination numbers appear in free-standing monolayers and monowires and an even larger tendency toward magnetism can be expected for these systems. Of course, the experimental realization of such free-standing systems seems to be difficult, but when such layers or wires are supported on non-magnetic substrates, the tendency toward magnetism can still be rather high, if the hybridization with substrate electronic states is small. Examples for such systems are transition-metal layers and wires on noble metal substrates. For a discussion of magnetic properties of thin layers I refer to recent reviews^{25–27} and the references therein. Spin-polarized DFT calculations for magnetic nanowires have been performed by Weinert and Freeman²⁸ already in 1983 and by many groups in the last few years^{29–50}, where this list of references is certainly not complete. Since a discussion of these studies is beyond the scope of the present lecture, I will present here only some of our own results³³ for $4d$ monatomic wires on the Ag surface. These results have been obtained by the screened Korringa-Kohn-Rostoker (KKR) method^{51–54}, which is well suited for such calculations because of its linear $O(N)$ scaling as discussed elsewhere in this Winter School on Computational Nanoscience.

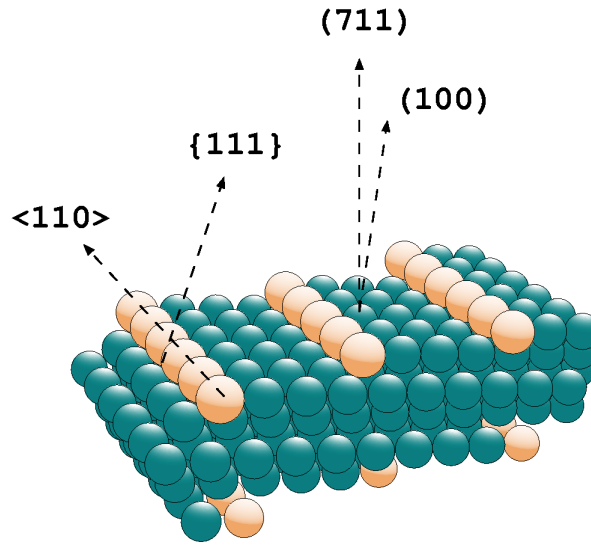


Figure 7. The fcc (711) vicinal surface with decorated steps as an example of periodically repeated monatomic wires, here at the $(100) \times (111)$ step edge (from Ref. 33).

In our calculations we have exploited the concept of step decoration on vicinal surfaces. These are surfaces with high Miller indices, for instance the fcc (711) surface shown in Figure 7. The same concept of step decoration is also exploited in the experimental preparation of extended ultrathin nanowires, where the arrays of parallel steps on these vicinal surfaces are used as templates along which the deposited material can nucleate. The darker

spheres in Figure 7 represent the Ag atoms and the lighter spheres the monatomic rows of $4d$ transition-metal atoms. The same geometry can also be used to describe monatomic rows on the middle of the terraces, if some rows of Ag atoms on the right-hand side of the transition-metal row are removed.

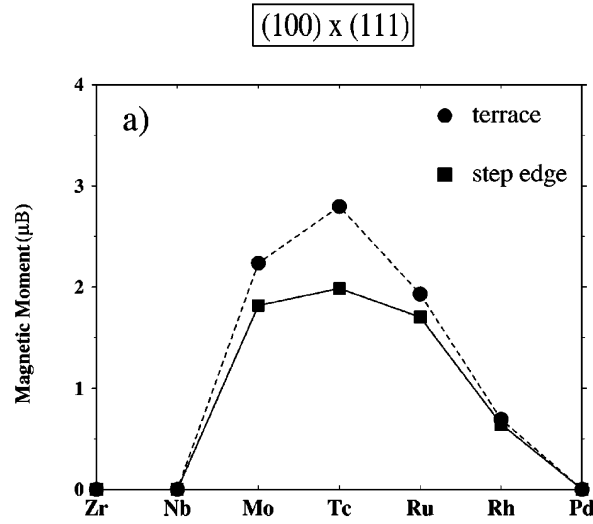


Figure 8. Magnetic moments of $4d$ monatomic rows on the fcc Ag(711) vicinal surface. Results denoted by circles are for rows in terrace and results denoted by squares for rows in step edge positions (from Ref. 33).

The calculated magnetic moments per atom for $4d$ magnetic monatomic wires on fcc (711) and fcc (410) Ag vicinal surfaces are shown in Figs. 8 and 9. A direct comparison between the magnetic moments for the two different step orientations shows that the magnetic moments for $(100) \times (111)$ wires are smaller than for the $(100) \times (110)$ wires for all elements except for Rh. The main reason for this difference is the increased hybridization of the $4d$ states between the atoms in the more closely packed $(100) \times (111)$ wires (see Figure 6), which leads to larger bandwidths and smaller densities of states $n^o(E_F)$. For Rh the situation is different. Whereas in non-magnetic calculations the local density of states for the Nb, Mo, Tc, and Ru wires shows a peak near the Fermi level, which due to broadening is reduced in height leading to a lower value of $n^o(E_F)$ in the close-packed rows, the non-magnetic density of states for the Rh wires has a peak below the Fermi level such that broadening leads to an increase of $n^o(E_F)$. Pictures for these non-magnetic densities of states can be found in Ref. 33.

Another feature evident in Figs. 8 and 9 is that the rows at the step edges (solid lines) have smaller moments than the rows on the terraces (dashed lines). This can be explained by an increase in the hybridization between the extended $4d$ orbitals of the row atoms and the sp like valence electrons of the substrate Ag atoms. Since the extent of the $4d$ orbitals is larger at the beginning of the $4d$ series, this effect is more pronounced at the beginning of the $4d$ series. The importance of the hybridization with the Ag substrate has also been

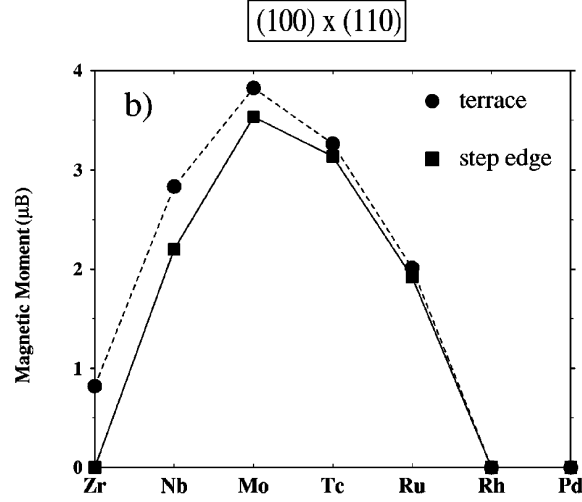


Figure 9. Magnetic moments of 4d monatomic rows on the fcc Ag(410) vicinal surface. Results denoted by circles are for rows in terrace and results denoted by squares for rows in step edge positions (from Ref. 33).

observed in spin-polarized DFT calculations for 4d dimers⁵⁵, where the magnetism of free dimers from the beginning of the 4d series (Y, Zr, Nb) is completely quenched, if they are adsorbed on the Ag(100) surface.

In addition to the ferromagnetic solutions shown in Figs. 8 and 9, we have also found antiferromagnetic solutions for Nb, Mo, and Tc wires, where the atoms in the rows have moments of alternating sign. In general, our results are in good agreement to similar calculation by Spisak and Hafner³⁹, who have used a quite different computational approach, the Vienna *ab initio* simulation package^{56,57} (VASP) in a projector augmented-wave (PAW) representation. Both methods, KKR and VASP, have shown that antiferromagnetism is preferred for Nb and Mo wires, that the ferromagnetic and antiferromagnetic states for Tc wires are almost degenerated in energy and that ferromagnetism is preferred in Ru and Rh wires. Compared to the KKR calculations, the VASP calculations have found somewhat larger moments and larger magnetic energy differences. In addition, the VASP calculations have found a metastable ferromagnetic solution for Nb wires and a metastable antiferromagnetic solution for Ru wires. These differences are probably related to the inclusion of gradient corrections to the exchange-correlation potential in Ref. 39 and to the spherical potential approximation used in Ref. 33.

6 Beyond the Stoner Model

The explanation of spin magnetism in terms of the Stoner model as discussed above is restricted to ferromagnetic materials, where all atoms have identical moments as in the elemental ferromagnets, and to ferromagnetic systems like magnetic overlayers, wires and impurities in contact to non-magnetic system, if the magnetic polarization of the non-magnetic atoms is small and its contribution can be neglected. For instance, it has been

shown⁵⁸ that a Stoner criterion $In_{loc}^o(E_F) > 1$ very similar to Eq. (27) can be used to characterize the behavior of magnetic impurities in non-magnetic materials. Here $n_{loc}^o(E)$ is the local density of states obtained by integrating the electron distribution over the impurity cell and neglecting contributions from the non-magnetic host atoms. The exchange integrals (Stoner parameters) for impurities are rather similar to the ones given in Table 1. This discussion for impurities is easily generalized for magnetic monolayers⁵⁹ and monowires³³, which are supported on non-magnetic substrates or free-standing in vacuum.

However, in magnetic systems, which consist of atoms with different moments, like magnetic alloys, antiferromagnets, magnetic monolayers and monowires on magnetic or almost magnetic substrates, or magnetic impurities in magnetic or almost magnetic hosts, the discussion within the Stoner model can fail. The failure arises in systems, where the main assumption of Stoner model, which is the rigid-band shift of the spin-dependent densities of states, is not satisfied. In these systems the magnetism cannot be explained by the Stoner model, but alternatively by spin-dependent changes in the covalent interaction⁶⁰ between the electronic states of neighboring atoms.

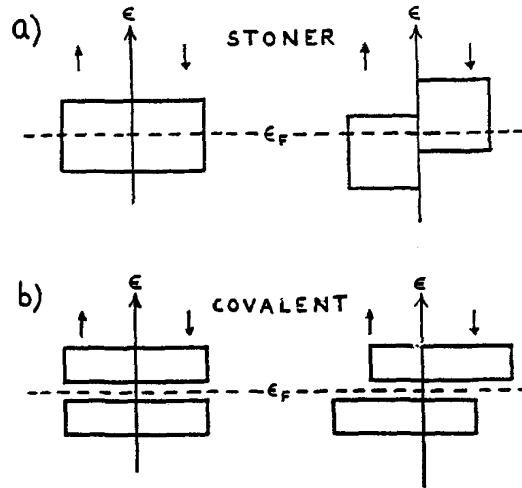


Figure 10. Schematic changes in the majority ($\uparrow$) and minority ($\downarrow$) spin densities of states, which lead to magnetism (from Ref. 60).

The distinction between "Stoner" or "covalent" magnetism is illustrated in Figure 10. On the left the densities of states are given for the non-magnetic and on the right for the magnetic situation. Whereas in the Stoner model the densities of states are shifted and magnetism occurs because a high density of states at the Fermi level makes the non-magnetic state unstable, for covalent magnetism the density of states at the Fermi level does not play a special role and magnetism is connected with a spectral-weight shift of the densities of states. The physical mechanism for the spectral-weight shift can be seen in its simplest form in the states of a diatomic molecule as illustrated in Figure 11.

When the states of the individual atoms interact, bonding and anti-bonding hybrids are formed. When the interacting states are degenerated ($\epsilon_1 = \epsilon_2$), the spectral weight is

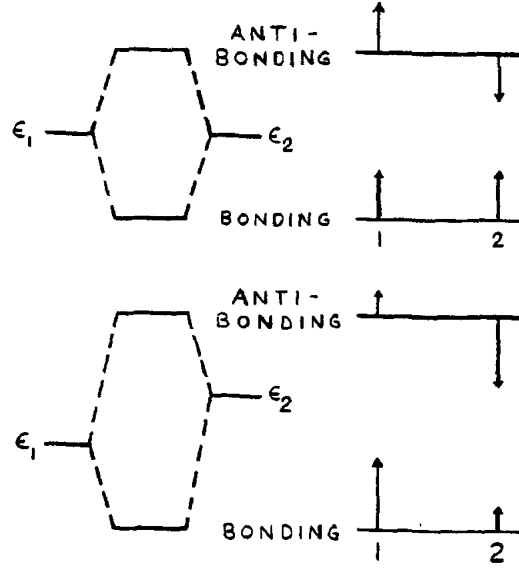


Figure 11. Changes in energy levels (left) and wavefunction weights (right) of covalent bonds caused by a relative displacement of the individual atomic levels ϵ_1 and ϵ_2 (from Ref. 60).

evenly distributed in both hybrids (upper right-hand portion of Figure 11). When one of the interacting states is displaced relative to the other ($\epsilon_1 < \epsilon_2$), for instance as a consequence of a larger moment on atom 1, the spectral weight is shifted within each hybrid (lower right-hand portion of Figure 11). The bonding hybrid obtains more weight on atom 1 with the lower energy level ϵ_1 and the anti-bonding hybrid obtains more weight on atom 2 with the higher energy level ϵ_2 . This spectral-weight shift leads to a charge transfer and is responsible for the covalent magnetism. Note that Figure 11 refers to the interaction of states of only one spin direction, for the other spin direction a similar picture applies with $\epsilon_1 > \epsilon_2$. The combined effect of both hybridizations leads to the densities of states shown in the lower right-hand portion of Figure 10.

The different hybridization behavior between ferromagnetically and antiferromagnetically aligned atoms has also been explained in detail for spin-polarized DFT calculations for interacting impurity pairs in Cu and Ag⁶¹. In antiferromagnetic systems the special situation occurs that, although the atoms are chemically equivalent with the same exchange integral I , they are magnetically different with alternating moments M and $-M$. The essential quantity, which decides upon the stability of the antiferromagnetic state, is again the spin-dependent susceptibility, which in first order perturbation theory connects the moment with the magnetic field

$$M = X_{AF} H, \quad (45)$$

where AF denotes antiferromagnetism and where the fictive field is assumed to point into the same direction as the moments thus representing an alternating staggered field. The

magnetic field and the magnetic moments lead to potentials

$$V_1^\pm = V^o + \Delta V_1^\pm = V^o \pm \left(\frac{1}{2}IM + \mu_B H \right) \quad (46)$$

and

$$V_2^\pm = V^o + \Delta V_2^\pm = V^o \mp \left(\frac{1}{2}IM + \mu_B H \right), \quad (47)$$

where $V_1^\pm$ stands for the potentials on the atoms in the sublattice with odd indices and $V_2^\pm$ for the potentials on the atoms in the sublattice even indices. The self-consistency conditions for the moment of atom $i = 0$ is given by

$$M = \int^{E_F(M)} [n_0^+(E) - n_0^-(E)] dE \quad (48)$$

as above in Eq. (22), but with the difference that $n_0^\pm(E)$ cannot be obtained by a simple shift from the non-magnetic density of states. By using basic properties of the Green function of the one-particle Schrödinger equation, which are given in the appendix, the moment M in (48) can be calculated as

$$\begin{aligned} M &= -\frac{1}{\pi} \text{Im} \int^{E_F(M)} [G_{00}^+(E) - G_{00}^-(E)] dE \\ &= -\frac{1}{\pi} \sum_i \text{Im} \int^{E_F(M)} G_{0i}^o(E) [\Delta V_i^+ - \Delta V_i^-] G_{i0}^o(E) dE + O(\Delta V_i^3). \end{aligned} \quad (49)$$

Linear in ΔV the last equation can be written as

$$M = X_{AF}^o \left(\frac{IM}{2\mu_B} + H \right), \quad (50)$$

where $\Delta V_i^\pm$ from Eqs. (46) and (47) was used and where the susceptibility X_{AF}^o for non-interacting electrons is defined as

$$X_{AF}^o = \frac{2\mu_B}{\pi} \text{Im} \int^{E_F(M)} \left[\sum_{\text{even } i} G_{0i}^o(E) G_{i0}^o(E) - \sum_{\text{odd } i} G_{0i}^o(E) G_{i0}^o(E) \right] dE. \quad (51)$$

This susceptibility can also be written as

$$X_{AF}^o = \frac{2\mu_B}{\pi} \text{Im} \int^{E_F(M)} \left[\sum_{\text{all } i} G_{0i}^o(E) G_{i0}^o(E) - 2 \sum_{\text{odd } i} G_{0i}^o(E) G_{i0}^o(E) \right] dE. \quad (52)$$

Here by use of Eqs. (64) and (62) the first term can be expressed by the density of states $n^o(E_F)$ and X_{AF}^o can be written as

$$X_{AF}^o = 2\mu_B [a(E_F) - n^o(E_F)], \quad (53)$$

where the function $a(E_F)$ is defined second term in Eq. (52). If Eq. (50) is solved for M , one obtains

$$M = X_{AF} H = \left(1 - \frac{IX_{AF}^o}{2\mu_B} \right)^{-1} X_{AF}^o H = S_{AF} X_{AF}^o H, \quad (54)$$

which shows that the spin susceptibility X_{AF} , similarly as in ferromagnets, is enlarged compared to X_{AF}^o by an enhancement factor S_{AF} . The criterion for a stable antiferromagnetic state is

$$\frac{IX_{AF}^o}{2\mu_B} > 1. \quad (55)$$

By using Eq. (53) the criterion can be written as

$$I[a(E_F) - n^o(E_F)] > 1 \quad (56)$$

which shows that a high density of states $n^o(E_F)$ at the Fermi energy is unfavorable for antiferromagnetism.

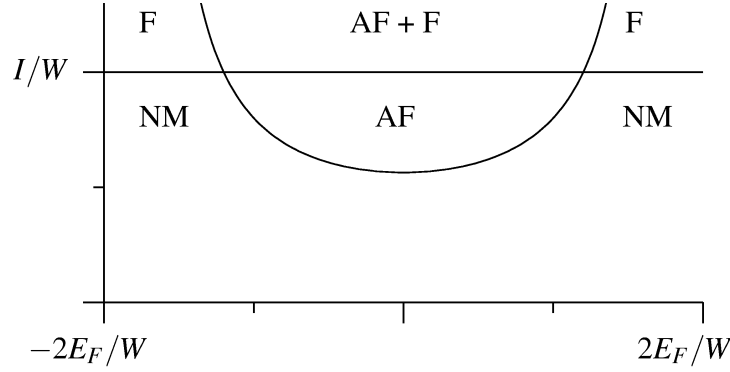


Figure 12. Phase diagram for a model with a rectangular density of states. The stability region for the non-magnetic state is indicated by NM, the stability regions for the ferromagnetic or antiferromagnetic state are indicated by F or AF.

For a simple model with a rectangular density of states $n^o(E) = 1/W$ for $-W/2 < E < W/2$ and $n^o(E) = 0$ for $|E| > W/2$, it is possible to determine $a(E_F)$ approximately from the imaginary and real part of the Green function $G^o(E)$ and to derive the phase diagram shown in Figure 12. The imaginary part of the Green function for this model is given by $\text{Im}G_{00}^o(E) = -\pi/W$ for $-W/2 < E < W/2$ and vanishes for $|E| > W/2$. The real part follows from the Kramers-Kronig relation (Eq. (62)) as

$$\text{Re}G_{00}^o(E) = -\frac{1}{W} \ln \left| \frac{E - W/2}{E + W/2} \right|. \quad (57)$$

To evaluate $a(E_F)$ approximately, the main contribution arising from atom $i = 0$ is taken into account and the contributions $G_{0i}^o(E)G_{i0}^o(E)$ of all other atoms $i \neq 0$ are neglected, since they decay with increasing distance between the atoms i and 0. From $\text{Im}(G_{00}^o)^2 = 2\text{Re}G_{00}^o\text{Im}G_{00}^o$ the result

$$a(E_F) - n_0^o(E_F) = \frac{1}{W} \left[\frac{4}{W} \int_{-W/2}^{E_F} \ln \left| \frac{E - W/2}{E + W/2} \right| dE - 1 \right] \quad (58)$$

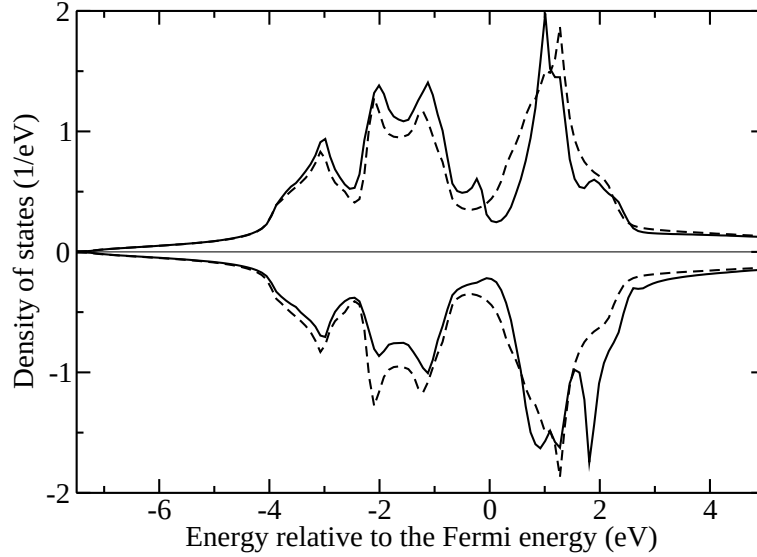


Figure 13. Density of states for one atom in antiferromagnetic bcc Cr (solid curves) compared to the density of states for one atom in nonmagnetic bcc Cr (dashed lines) The density of states for majority-spin electrons is plotted upwards and for minority-spin electrons downwards. The density of states for the other atom are obtained by interchanging spin up and spin down curves.

follows for $|E_F| < W/2$. The stability criterion (Eq. (56)) can be written as $I > [a(E_F) - n_0^o(E_F)]^{-1}$ or as

$$\frac{I}{W} > \left[2 \int_{-1}^{2E_F/W} \ln \left| \frac{x-1}{x+1} \right| dx - 1 \right]^{-1}, \quad (59)$$

which is obtained from Eq. (58) by the substitution $x = 2E/W$. The right-hand side of Eq. (59) defines a function $f(2E_F/W)$, which is shown as the bended curve in Figure 12. Above this curve the antiferromagnetic state is more stable and above the straight line at I/W the ferromagnetic state is more stable than the non-magnetic state. For $E_F = 0$ at center of the band, the function $f(2E_F/W)$ attains its minimum with the value $[4 \ln 2 - 1]^{-1} = 0.564... < 1$. This shows that a position of the Fermi energy E_F near to the band center makes antiferromagnetism more favorable than ferromagnetism.

For the $3d$ metal Cr, the Fermi energy is in the middle of the d band and furthermore lies in a pseudogap with a reduced $n^o(E_F)$ such that antiferromagnetism should be even more preferred for Cr than for the model with the phase diagram of Figure 12. To illustrate the situation for Cr, I have performed self-consistent density-functional calculations in the local-spin-density approximation⁴ by the full-potential KKR method as described in Ref. 62 using $l_{max} = 3$ for t matrices and Green functions and $l_{max} = 6$ for the density and the potential. To calculate the densities of states shown in Figure 13, I have used a lattice constant of 0.294 nm, which is two percent larger than the experimental value 0.288 nm. This leads to a moment of $1.08 \mu_B$ and increases the differences shown in Figure 13 compared to calculations for the experimental lattice constant with a moment of $0.39 \mu_B$.

The calculated density of states clearly shows the spectral-weight shift for the occupied states. Thus Figure 13 illustrates that the mechanism of covalent interaction discussed above (compare Figure 10) explains the antiferromagnetism in Cr.

Note, however, that the real ground-state of Cr is not the antiferromagnetic state, but an incommensurate spin-density wave state. The consideration of this state is beyond the scope of my presentation and I refer to Refs. 63 and 64 for a discussion about the applicability of spin-polarized DFT calculations for the incommensurate spin-density wave state of Cr.

Appendix

Green-Function Properties

The Green function $G(E)$ for a Hamilton operator $\mathcal{H}$ is defined by the operator equation

$$G = \frac{1}{E - \mathcal{H}}, \quad (60)$$

where the energy E is chosen as a complex quantity. For real energies it is necessary to perform a limiting process. Then the real energy E is replaced by $E + i\epsilon$ and all equations are understood in the sense that the limit $\epsilon \rightarrow 0+$ must be performed in the end. The relation

$$\lim_{\epsilon \rightarrow 0+} \frac{1}{E + i\epsilon - \mathcal{H}} = P \frac{1}{E - \mathcal{H}} - i\pi\delta(E - \mathcal{H}), \quad (61)$$

where P denotes the principal value, establishes the connection between the imaginary part of the Green function and the density of states $n(E) = \delta(E - \mathcal{H})$ and also the Kramers-Kronig relation between the imaginary and the real part of the Green function

$$\text{Im}G(E) = -\pi n(E), \quad \text{Re}G(E) = -\frac{1}{\pi}P \int_{-\infty}^{\infty} \frac{1}{E - E'} \text{Im}G(E') dE'. \quad (62)$$

Another useful relation for the Green function is given by

$$\frac{dG(E)}{dE} = \frac{d}{dE} \frac{1}{E - \mathcal{H}} = -\frac{1}{E - \mathcal{H}} \frac{1}{E - \mathcal{H}} = -G(E)G(E), \quad (63)$$

which leads to

$$G_{00}(E_F) = \int^{E_F} \frac{dG_{00}(E)}{dE} dE = - \int^{E_F} \sum_i G_{0i}(E) G_{i0}(E) dE, \quad (64)$$

where explicitly the position dependence was denoted. The Green functions for the potentials $V^\pm(\mathbf{r}) = V^o(\mathbf{r}) + \Delta V^\pm(\mathbf{r})$ are connected with the Green function for the non-magnetic state with potential $V^o(\mathbf{r})$ by the Dyson equation

$$G^\pm = G^o + G^o \Delta V^\pm G^\pm, \quad (65)$$

which follows from the fact that the inverse operators $(G^o)^{-1}$, $(G^\pm)^{-1}$ can be written as

$$(G^o)^{-1} = E - \mathcal{H}^o, \quad (G^\pm)^{-1} = E - \mathcal{H}^\pm. \quad (66)$$

where the Hamilton operators according to $\mathcal{H}^\pm = \mathcal{H}^o + \Delta V^\pm$ differ by $\Delta V^\pm = V^\pm - V^o$. This leads to $(G^\pm)^{-1} = (G^o)^{-1} - \Delta V^\pm$, which solved for $G^\pm$ gives equation (65). For small potential differences the expansion of Eq. (65) is given by

$$G^+ - G^- = G^o(\Delta V^+ - \Delta V^-)G^o + \dots \quad (67)$$

References

1. U. von Barth and L. Hedin, *A local exchange-correlation potential for the spin polarized case: I*, J. Phys. C **5**, 1629–1642 (1972).
2. M. M. Pant and A. K. Rajagopal, *Theory of inhomogeneous magnetic electron gas*, Solid State Commun. **10**, 1157–1160 (1972).
3. H. Eschrig, M. Sargolzaei, K. Koepnick, and M. Richter, *Orbital polarization in the Kohn-Sham-Dirac theory*, Europhys. Lett. **72**, 611–617 (2005).
4. S. H. Vosko, L. Wilk, and M. Nusair, *Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis*, Can. J. Phys. **58**, 1200–1211 (1980).
5. J. P. Perdew and A. Zunger, *Self-interaction correction to density-functional approximations for many-electron systems*, Phys. Rev. B **23**, 5048–5079 (1981).
6. J. P. Perdew, J. A. Chevary, S. H. Vosko, K. A. Jackson, M. R. Pederson, D. J. Singh, and C. Fiolhais, *Atoms, molecules, solids, and surfaces: Applications of the generalized gradient approximation for exchange and correlation*, Phys. Rev. B **46**, 6671–6687 (1992); Phys. Rev. B **48**, 4978 (1993).
7. J. P. Perdew, K. Burke, and M. Ernzerhof, *Generalized Gradient Approximation Made Simple*, Phys. Rev. Lett. **77**, 3865–3868 (1996).
8. S. H. Vosko, J. P. Perdew, and A. H. MacDonald, *Ab Initio Calculation of the Spin Susceptibility for the Alkali Metals Using the Density-Functional Formalism*, Phys. Rev. Lett. **35**, 1725–1728 (1975).
9. U. K. Poulsen, J. Kollar, and O. K. Andersen, *Magnetic and cohesive properties from canonical bands*, J. Phys. F **6**, L241–L247 (1976).
10. O. Gunnarsson, *Band model for magnetism of transition metals in the spin-density-functional formalism*, J. Phys. F **6**, 587–606 (1976).
11. O. Gunnarsson, *Band magnetism in the spin-density-functional formalism*, J. Appl. Phys. **49**, 1399–1404 (1978).
12. E. C. Stoner, *Collective Electron Specific Heat and Spin Paramagnetism in Metals*, Proc. Roy. Soc. A **154**, 656–678 (1936).
13. E. C. Stoner, *Collective Electron Ferromagnetism*, Proc. Roy. Soc. A **165**, 372–414 (1938).
14. E. C. Stoner, *Collective Electron Ferromagnetism II: Energy and Specific Heat*, Proc. Roy. Soc. A **169**, 339–371 (1939).
15. J. F. Janak, *Uniform susceptibilities of metallic elements*, Phys. Rev. B **16**, 255–262 (1977).
16. J. F. Janak, *Itinerant ferromagnetism in fcc cobalt*, Solid State Commun. **25**, 53–55 (1978).
17. V. L. Moruzzi, J. F. Janak, and A. R. Williams, *Calculated Electronic Properties of Metals* (Pergamon Press, New York, 1978).

18. A. P. Malozemoff, A. R. Williams, and V. L. Moruzzi, “*Band gap theory*” of strong ferromagnetism: Application to concentrated crystalline and amorphous Fe- and Co-metalloid alloys, *Phys. Rev. B* **29**, 1620–1631 (1984).
19. P. H. Dederichs and R. Zeller, *Self-consistency iterations in electronic-structure calculations*, *Phys. Rev. B* **28**, 5462–5472 (1983).
20. D. G. Anderson, *Iterative Procedures for Nonlinear Integral Equations*, *J. Assoc. Comput. Mach.* **12**, 547–560 (1965).
21. C. G. Broyden, *A Class of Methods for Solving Nonlinear Simultaneous Equations*, *Math. Comput.* **19**, 577–593 (1965).
22. S. Blügel, Ph.D. thesis, *First Principles Calculations of the Electronic Structure of Magnetic Overlayers on Transition Metal Surfaces* (Rheinisch-Westfälische Technische Hochschule Aachen), Berichte der Kernforschungsanlage Jülich Nr. 2197 (1987).
23. V. Eyert, *A Comparative Study on Methods for Convergence Acceleration of Iterative Vector Sequences*, *J. Comp. Phys.* **124**, 271–285 (1996).
24. M. A. Van Hove and G. A. Somorjai, *A new microfacet notation for high-Miller-index surfaces of cubic materials with terrace, step and kink structures*, *Surf. Sci.* **92**, 489–518 (1980).
25. T. Asada, G. Bihlmayer, S. Handschuh, S. Heinze, Ph. Kurz, and S. Blügel, *First-principles theory of ultrathin magnetic films*, *J. Phys.: Condensed Matter* **11**, 9347–9363 (1999).
26. M. Donath, *Magnetic order and electronic structure in thin films*, *J. Phys.: Condensed Matter* **11**, 9421–9436 (1999).
27. P. Pouloupoulos and K. Baberschke, *Magnetism in thin films*, *J. Phys.: Condensed Matter* **11**, 9495–9515 (1999).
28. M. Weinert and A. J. Freeman, *Magnetism of linear chains*, *J. Magn. Magn. Mater.* **38**, 23–33 (1983).
29. N. Zabala, M. J. Puska, and R. M. Nieminen, *Spontaneous Magnetization of Simple Metal Nanowires*, *Phys. Rev. Lett.* **80**, 3336–3339 (1998); *Phys. Rev. Lett.* **82**, 3000 (1999).
30. O. A. Starykh and D. L. Maslov, *Comment on “Spontaneous Magnetization of Simple Metal Nanowires”*, *Phys. Rev. Lett.* **82**, 2999 (1999).
31. D. I. Bazhanov, W. Hergert, V. S. Stepanyuk, A. A. Katsnelson, P. Rennert, K. Kokko, and C. Demangeat, *One-dimensional magnetism of Rh chains on the Ag(001) surface*, *Phys. Rev. B* **62**, 6415–6420 (2000).
32. S. Okada and A. Oshiyama, *Magnetic ordering of Ga wires on Si(100) surfaces*, *Phys. Rev. B* **62**, R13286–R13289 (2000).
33. V. Bellini, N. Papanikolaou, R. Zeller, and P. H. Dederichs, *Magnetic 4d monoatomic rows on Ag vicinal surfaces*, *Phys. Rev. B* **64**, 094403 (2001).
34. M. Eisenbach, B. L. Gyorffy, G. M. Stocks, and B. Ujfalussy, *Magnetic anisotropy of monoatomic iron chains embedded in copper*, *Phys. Rev. B* **65**, 144424 (2002).
35. D. Spisak and J. Hafner, *Fe nanowires on vicinal Cu surfaces: Ab initio study*, *Phys. Rev. B* **65**, 235405 (2002).
36. A. Ayuela, H. Raebiger, M. J. Puska, and R. M. Nieminen, *Spontaneous magnetization of aluminum nanowires deposited on the NaCl(100) surface*, *Phys. Rev. B* **66**, 035417 (2002).
37. M. Komelj, C. Ederer, J. W. Davenport, and M. Fähnle, *From the bulk to monatomic*

- wires: *An ab initio study of magnetism in Co systems with various dimensionality*, Phys. Rev. B **66**, 140407 (2002).
38. D. Spisak and J. Hafner, *Magnetism of monoatomic wires on vicinal surfaces*, Comput. Mater. Sci. **27**, 138–150 (2003).
 39. D. Spisak and J. Hafner, *Magnetism of ultrathin wires suspended in free space and absorbed on vicinal surfaces*, Phys. Rev. B **67**, 214416 (2003).
 40. J. Hong and R. Q. Wu, *First principles calculations of magnetic anisotropy energy of Co monatomic wires*, Phys. Rev. B **67**, 020406 (2003).
 41. A. Bergara, J. B. Neaton, and N. W. Ashcroft, *Ferromagnetic Instabilities in Atomically Thin Lithium and Sodium Wires*, Int. J. Quantum Chem. **91**, 239–244 (2003).
 42. A. Delin and E. Tosatti, *Magnetic phenomena in 5d transition metal nanowires*, Phys. Rev. B **68**, 144434 (2003).
 43. A. Delin and E. Tosatti, *Emerging magnetism in platinum nanowires*, Surf. Sci. **566–568**, 262–267 (2004).
 44. A. Delin, E. Tosatti, and R. Weht, *Magnetism in Atomic-Size Palladium Contacts and Nanowires*, Phys. Rev. Lett. **92**, 057201 (2004).
 45. A. Delin and E. Tosatti, *The electronic structure of 4d transition-metal monatomic wires*, J. Phys.: Condensed Matter **16**, 8061–8074 (2004).
 46. M. M. Rahman, M. Kisaku, T. Kishi, D. Matsunaka, W. A. Dino, H. Nakanishi, and H. Kasai, *Ab initio study of magnetic and electronic properties of Fe-filled single-walled carbon nanotubes*, J. Phys.: Condensed Matter **16**, S5755–S5758 (2004).
 47. B. Lazarovits, B. Ujfalussy, L. Szunyogh, G. M. Stocks, and P. Weinberger, *Ab initio study of canted magnetism of finite atomic chains at surfaces*, J. Phys.: Condensed Matter **16**, S5833–S5840 (2004).
 48. T. J. Yang, Y.-J. Zhao, and A. J. Freeman, *Magnetism and electronic structure of Fe chains and nano-wires*, J. Magn. Magn. Mater. **272–276**, 1648–1649 (2004).
 49. C. Jo and J. I. Lee, *Structures and magnetic properties of Fe_n ($n=1-4$) nanowires*, Phys. Status Solidi b **241**, 1427–1430 (2004).
 50. Y. Mokrousov, G. Bihlmayer, and S. Blügel, *Full-potential linearized augmented plane-wave method for one-dimensional systems: Gold nanowire and iron monowires in a gold tube*, Phys. Rev. B **72**, 045402 (2005).
 51. R. Zeller, P. H. Dederichs, B. Ujfalussy, L. Szunyogh, and P. Weinberger, *Theory and convergence properties of the screened Korringa-Kohn-Rostoker method*, Phys. Rev. B **52**, 8807–8812 (1995).
 52. R. Zeller, *Evaluation of the screened Korringa-Kohn-Rostoker method for accurate and large-scale electronic-structure calculations*, Phys. Rev. B **55**, 9400–9408 (1997).
 53. K. Wildberger, R. Zeller, and P. H. Dederichs, *Screened KKR-Green’s-function method for layered systems*, Phys. Rev. B **55**, 10074–10080 (1997).
 54. N. Papanikolaou, R. Zeller, and P. H. Dederichs, *Conceptual improvements of the KKR method*, J. Phys.: Condensed Matter **16**, 2799–2823 (2002).
 55. K. Wildberger, V. S. Stepanyuk, P. Lang, R. Zeller, and P. H. Dederichs, *Magnetic Nanostructures: 4d Clusters on Ag(001)*, Phys. Rev. Lett. **75**, 509–512 (1995).
 56. G. Kresse and J. Furthmüller, *Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set*, Phys. Rev. B **54**, 11169–11186 (1996).
 57. G. Kresse and J. Furthmüller, *Efficiency of ab-initio total energy calculations for met-*

- als and semiconductors using a plane-wave basis set*, Comput. Mater. Sci. **6**, 15–50 (1996).
58. J. Deutz, P. H Dederichs, and R. Zeller, *Local density of states of impurities in Al*, J. Phys. F **11**, 1787–1800 (1981).
 59. S. Blügel, *Ferromagnetism of 4d-metal monolayers on Ag, Au and Pd(001) surfaces*, Europhys. Lett. **18**, 257–262 (1992).
 60. A. R. Williams, R. Zeller, V. L. Moruzzi, C. D. Gelatt, Jr., and J. Kübler, *Covalent magnetism: An alternative to the Stoner model*, J. Appl. Phys. **52**, 2067–2069 (1981).
 61. A. Oswald, R. Zeller, P. J. Braspenning, and P. H Dederichs, *Interaction of magnetic impurities in Cu and Ag*, J. Phys. F **15**, 193–212 (1985).
 62. M. Asato, A. Settels, T. Hoshino, T. Asada, S. Blügel, R. Zeller, and P. H. Dederichs, *Full-potential KKR calculations for metals and semiconductors*, Phys. Rev. B **60**, 5202–5210 (1999).
 63. R. Hafner, D. Spisak, R. Lorenz, and J. Hafner, *Magnetic ground state of Cr in density-functional theory*, Phys. Rev. B **65**, 184432 (2002).
 64. S. Cottenier, B. De Vries, J. Meersschant, and M. Rots, *What density-functional theory can tell us about the spin-density wave in Cr*, J. Phys.: Condensed Matter **14**, 3275–3283 (2002).

Non-Collinear Magnetism: Exchange Parameter and T_C

Gustav Bihlmayer

Institute for Solid State Research
Forschungszentrum Jülich,
52425 Jülich, Germany
E-mail: g.bihlmayer@fz-juelich.de

The need for a reliable description of the ground state properties of complex magnetic systems as well as for the prediction of finite-temperature properties, has led to significant progress in the field of magnetic first-principles calculations. This article reviews the basic concepts of vector-spin density functional theory (DFT), *ab initio* spin dynamics, and constrained DFT. We describe the calculation of exchange parameters from a mapping of DFT results to model Hamiltonians, the determination of the magnetic ground state, and the estimation of critical temperatures from these exchange parameters. Finally, effects of spin-orbit coupling and orbital magnetism, relevant in low-dimensional magnetic systems, will be discussed.

1 Introduction

Up to now, in most *ab initio* calculations of magnetic systems only ferromagnetic – or some antiferromagnetic – states were considered. In contrast to these collinear magnetic configurations, many alloys, compounds, and even elements show non-collinear ground-states like conical or flat spin-spirals or commensurate superpositions of several spiral spin-density waves. Even in systems with collinear magnetic structures non-collinearity occurs, e.g. at domain walls or in (thermally) excited systems. To access such states from first-principles, vector-spin density functional theory (DFT) has to be applied, which treats the magnetization density as a vector field (and not as a scalar field, as in collinear DFT calculations).

From such calculations it is possible to follow several directions: like in molecular-dynamics calculations, spin-dynamics allows to study the magnetic degrees of freedom either exploring the ground state or excited state properties (like critical temperatures). Or the magnetic interactions are mapped onto a model (in the simplest case a classical Heisenberg model) and then this model is studied using parameters obtained by *ab initio* calculations. In both cases we introduce a discretization of the (vector) magnetization density: in spin-dynamics, the evolution of discrete spins, i.e. vectors attached to certain (atomic) positions, is monitored. Mapping the *ab initio* results to a model Hamiltonian which contains interactions between spins also requires that it is possible to assign a definite spin to an atom. I.e. it should be possible in the vicinity of an atom ν , e.g. within some sphere centered at the nucleus, to write the magnetization density, $\mathbf{m}(\mathbf{r})$, as

$$\mathbf{m}(\mathbf{r}) = M_\nu \hat{\mathbf{e}}_\nu \quad (1)$$

where M_ν is the magnetization and $\hat{\mathbf{e}}_\nu$ is the magnetization direction. Vector-spin DFT calculations allow to estimate whether Eq. (1) is a good approximation or not (cf. Figure 1).

In this contribution we will start with an outline of vector-spin DFT (Section 2) and the determination of stationary states (Section 2.1 and 2.2) as well as constrained DFT for

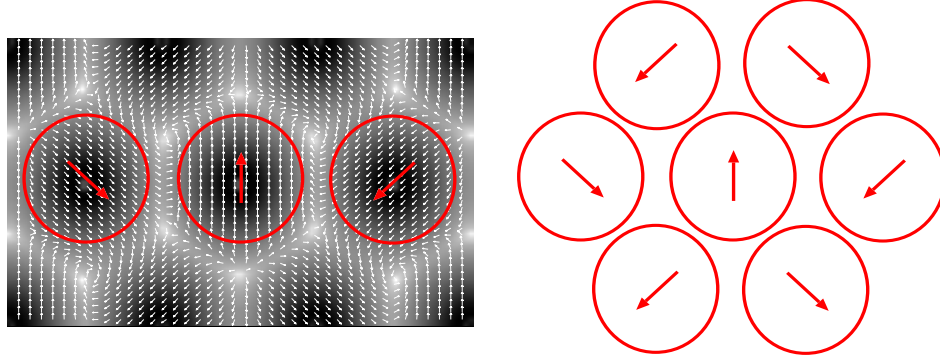


Figure 1. Left: ground state magnetization density of a hexagonal Cr monolayer with the Cu(111) in-plane lattice constant; the absolute value of the magnetization is shown in greyscale, the local directions are marked by small arrows. The average magnetization direction around an atom is indicated as red arrows. Right: schematic picture of the magnetic structure (Néel state) of the hexagonal Cr monolayer.

the calculation of non-stationary states (Section 2.3). The following sections will discuss the mapping of *ab initio* results, that have been obtained by vector-spin DFT, on models like the Heisenberg model (Section 3.1) and the determination of exchange parameters and critical temperatures (Section 3.2 and 3.3). A final chapter is then devoted to the magnetic anisotropy (Section 4) well as the relativistic effect that is mainly responsible for this anisotropy in most systems (Section 4.1). We conclude with a discussion of its influence on the ordering temperatures (Section 4.2) and the orbital magnetic moments (Section 4.3).

2 Vector-Spin Density Functional Theory

In 1964 Hohenberg and Kohn¹ worked out two central theorems that form the basis of density functional theory (DFT): For a system of N particles (e.g. electrons) moving in an external potential $v(\mathbf{r})$ (caused by e.g. nuclei) in a non-degenerate ground state (i) the many-body wavefunction Ψ and $v(\mathbf{r})$ are uniquely determined by the particle density distribution $n(\mathbf{r})$ and (ii) there exists an energy functional of this density, $E[n(\mathbf{r})]$, which is stationary with respect to variations of the ground-state density. These two theorems allow – at least in principle – the determination of the ground-state density and energy of a N -particle system. Extracting the classical Coulomb interaction energy, such a Hohenberg-Kohn energy functional takes the form

$$E[n(\mathbf{r})] = \int v(\mathbf{r})n(\mathbf{r})d\mathbf{r} + \frac{1}{2} \int \int \frac{n(\mathbf{r})n(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r}d\mathbf{r}' + G[n(\mathbf{r})] \quad (2)$$

where the functional $G[n(\mathbf{r})]$ has to be approximated.

In the Kohn-Sham theory², the kinetic energy T_0 of a non-interacting electron gas in its ground state with a density distribution $n(\mathbf{r})$ is further extracted from $G[n(\mathbf{r})]$, so that a new functional

$$E_{xc}[n(\mathbf{r})] = G[n(\mathbf{r})] - T_0[n(\mathbf{r})] \quad (3)$$

remains to be determined. E_{xc} is now called exchange-correlation energy functional, since without E_{xc} our energy functional E would yield just the Hartree energy. If we take into account that particle conservation, i.e. $N = \int n(\mathbf{r}) d\mathbf{r}$ has to be ensured, we can formulate the stationarity of E in Eq. (2) with respect to variations of the ground-state density as

$$v(\mathbf{r}) + \int \frac{n(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r}' + \frac{\delta T_0}{\delta n(\mathbf{r})} + \frac{\delta E_{xc}}{\delta n(\mathbf{r})} - \lambda = 0 \quad (4)$$

where the Lagrange parameter λ ensures the particle conservation. Expressing the kinetic energy of the non-interacting particles via their wavefunctions, ϕ_i , we can recast Eq. (4) in the form of an effective single-particle Schrödinger equation:

$$\left[-\frac{\hbar^2}{2m} \nabla^2 + v(\mathbf{r}) + \int \frac{n(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r}' + \frac{\delta E_{xc}}{\delta n(\mathbf{r})} \right] \phi_i(\mathbf{r}) = \epsilon_i \phi_i(\mathbf{r}) \quad (5)$$

which has to be solved self-consistently since $n(\mathbf{r}) = \sum_{i=1}^N |\phi_i(\mathbf{r})|^2$. The last term of the Hamiltonian is called the exchange-correlation potential.

In 1972 von Barth and Hedin extended this concept to spin-polarized systems³, replacing the scalar density by a hermitian 2×2 matrix $\underline{n}(\mathbf{r})$. If $\psi_\alpha(\mathbf{r})$ is the field operator for a particle of spin α , a component of the spin-density matrix can be defined as

$$n_{\alpha\beta}(\mathbf{r}) = \langle \Psi | \psi_\beta^\dagger(\mathbf{r}) \psi_\alpha(\mathbf{r}) | \Psi \rangle. \quad (6)$$

The potential matrix corresponding to this spin-density matrix is denoted as $\underline{v}(\mathbf{r})$ and replaces the scalar potential. Then, we can write Eq. (5) in the form

$$\left[\left(-\frac{\hbar^2}{2m} \nabla^2 + \sum_\alpha \int \frac{n_{\alpha\alpha}(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r}' \right) \underline{\mathbb{I}} + \underline{v}(\mathbf{r}) + \frac{\delta E_{xc}}{\delta \underline{n}(\mathbf{r})} \right] \begin{pmatrix} \phi_i^{(+)} \\ \phi_i^{(-)} \end{pmatrix} = \epsilon_i \begin{pmatrix} \phi_i^{(+)} \\ \phi_i^{(-)} \end{pmatrix} \quad (7)$$

where $\underline{\mathbb{I}}$ is a 2×2 unit matrix and the exchange-correlation potential is now also a 2×2 matrix. In terms of the Kohn-Sham wavefunctions, the density matrix can now be written as

$$n_{\alpha\beta}(\mathbf{r}) = \sum_{i=1}^N \phi_i^{*\alpha}(\mathbf{r}) \phi_i^\beta(\mathbf{r}) \quad \text{where} \quad \alpha, \beta = (+), (-). \quad (8)$$

Using the Pauli matrices, $\underline{\sigma}$, the density matrix can be decomposed into a scalar and a vectorial part, corresponding to the charge and magnetization density:

$$\underline{n}(\mathbf{r}) = \frac{1}{2} (n(\mathbf{r}) \underline{\mathbb{I}} + \underline{\sigma} \cdot \mathbf{m}(\mathbf{r})) = \frac{1}{2} \begin{pmatrix} n(\mathbf{r}) + m_z(\mathbf{r}) & m_x(\mathbf{r}) - i m_y(\mathbf{r}) \\ m_x(\mathbf{r}) + i m_y(\mathbf{r}) & n(\mathbf{r}) - m_z(\mathbf{r}) \end{pmatrix}. \quad (9)$$

Likewise, the potential matrices can be written in terms of a scalar potential and the magnetic field, $\mathbf{B}(\mathbf{r})$:

$$\underline{v}(\mathbf{r}) = v(\mathbf{r}) \underline{\mathbb{I}} + \mu_B \underline{\sigma} \cdot \mathbf{B}(\mathbf{r}) \quad \text{and} \quad \underline{v}_{xc}(\mathbf{r}) = v_{xc}(\mathbf{r}) \underline{\mathbb{I}} + \mu_B \underline{\sigma} \cdot \mathbf{B}_{xc}(\mathbf{r}) \quad (10)$$

where $\mu_B = \frac{e\hbar}{2mc}$ is the Bohr magneton.

2.1 Collinear and High Symmetry States

Supposing that the potential matrices in Eq. (10) are diagonal (i.e. the magnetic and exchange fields point in z direction), Eq. (7) decouples into two equations of the type of Eq. (5):

$$\begin{aligned} \left(-\frac{\hbar^2}{2m} \nabla^2 + v_{\text{Coul}}(\mathbf{r}) + v(\mathbf{r}) + B_z(\mathbf{r}) + v_{\text{xc}}^{(+)}(\mathbf{r}) \right) \phi_i^{(+)}(\mathbf{r}) &= \epsilon_i^{(+)} \phi_i^{(+)}(\mathbf{r}) \\ \left(-\frac{\hbar^2}{2m} \nabla^2 + v_{\text{Coul}}(\mathbf{r}) + v(\mathbf{r}) - B_z(\mathbf{r}) + v_{\text{xc}}^{(-)}(\mathbf{r}) \right) \phi_i^{(-)}(\mathbf{r}) &= \epsilon_i^{(-)} \phi_i^{(-)}(\mathbf{r}) \end{aligned} \quad (11)$$

where v_{Coul} denotes now the classical Coulomb potential and $v_{\text{xc}}^{(+,-)}$ the exchange-correlation potential that arises from the functional derivative of the exchange-correlation energy with respect to the spin-up (+) or spin-down (−) part of the diagonal density matrix.

Systems that can be described by Eq. (11) are all kinds of magnetic materials that assume a collinear magnetic order, e.g. ferromagnetic, antiferromagnetic or ferrimagnetic states. Like the density is a property that can – at least in principle – be obtained exactly in DFT, the spin density is a property that is well defined in spin-polarized DFT:

$$\mathbf{m}(\mathbf{r}) = -\mu_B \sum_{\alpha,\beta} \psi_{\alpha}^{+}(\mathbf{r}) \boldsymbol{\sigma}_{\alpha\beta} \psi_{\beta}(\mathbf{r}). \quad (12)$$

The integral spin moment, M , for a collinear system is then (in units of μ_B) simply

$$M_{\text{spin}} = \int \mathbf{m}(\mathbf{r}) d\mathbf{r} = \int (n^{(+)}(\mathbf{r}) - n^{(-)}(\mathbf{r})) d\mathbf{r}. \quad (13)$$

How well this quantity corresponds to experimental values depends of course on the quality of the exchange-correlation potential that is used for an actual calculation. Some examples of results obtained in the local spin density approximation (LSDA) and generalized gradient approximation (GGA) of (spin)-moments of elemental ferromagnets are given in Table 1.

Property	source	Fe (bcc)	Co (fcc)	Ni (fcc)	Gd (hcp)
M_{spin}	LSDA	2.15	1.56	0.59	7.63
M_{spin}	GGA	2.22	1.62	0.62	7.65
M_{spin}	experiment	2.12	1.57	0.55	
$M_{\text{tot.}}$	experiment	2.22	1.71	0.61	7.63

Table 1. Magnetic moments (in μ_B per atom) of ferromagnetic elements in the bulk. The experimentally determined total magnetization, $M_{\text{tot.}}$, consists of spin- and orbital moment contributions. The LSDA results for Fe, Co and Ni are taken from Moruzzi et al.⁴, the GGA values from Shallcross and coworkers⁵ where also experimental values are quoted. The calculated Gd data is from Kurz et al.⁶, the experimental one from White and coworkers⁷.

If we come back to our original assumption, that the magnetization density in the vicinity of some atom ν should be expressible by Eq. (1), then the total energy of a magnetic system as a function of its magnetic structure can be described as a functional $E[\{\mathbf{e}_{\nu}\}]$ of

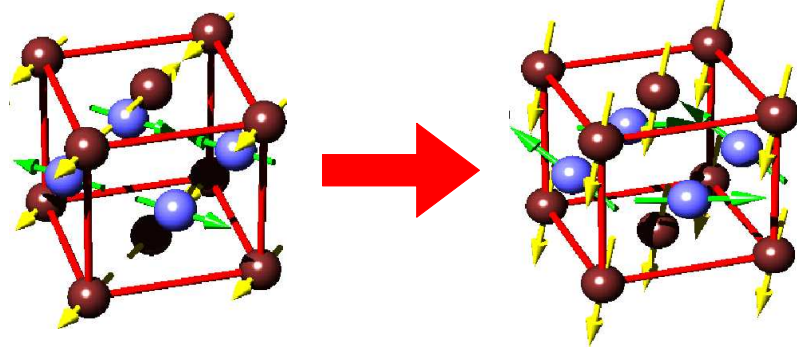


Figure 2. Determination of the magnetic ground state of ordered FeMn: the magnetic structure of the disordered alloy is a $2q$ -state (left). In an ordered alloy a more complex magnetic arrangement is obtained (right) by “relaxation” of the local spin directions.

the directions of the magnetic moments at the atoms ν in the magnetic unit cell. In this context collinear states ($\hat{\mathbf{e}}_\nu$ is identical for all atoms) are special solutions where $E[\{\hat{\mathbf{e}}_\nu\}]$ has a local or global maximum or minimum. Therefore, they constitute an important class of magnetic configurations that are often realized in magnetic materials. Unlike in non-spinpolarized DFT it is, however, in practical calculations not guaranteed that the solution obtained, $\underline{n}(\mathbf{r})$, is really the ground state, and often several metastable solutions can be obtained.

At this point we should notice, that we relied on the assumption that our functional $E[\{\hat{\mathbf{e}}_\nu\}]$ is invariant with respect to a uniform rotation of all magnetization directions, $\hat{\mathbf{e}}_\nu$. This was implicitly assumed when we arbitrarily (or, better, for convenience) selected in Eq. (11) the z direction as global magnetization axis. Indeed, in absence of an external field (or in its presence, as long as it is oriented in z direction) this implies no loss in generality, if v_{xc} is isotropic in space. If we start from a Schrödinger-Pauli like theory, there is indeed no term that could couple the spin-space to the lattice. Only if a spin-orbit coupling term (from a Dirac type theory) or – in some cases – dipolar interaction is included, a preferential direction for the collinear magnetization exists. This will be discussed in Section 4.

2.2 Spin-Dynamics, Magnetic Torque

If one is interested in the magnetic ground state of a system of given chemical composition and atomic positions, the final goal is to minimize the functional $E[\{\hat{\mathbf{e}}_\nu\}]$. The dimensionality of this problem will of course depend on the size of the unit cell chosen (some multiple of the chemical unit cell) and this minimization will involve the tricky task of determining the absolute minimum on a high-dimensional total energy surface. In analogy to molecular dynamics, i.e. the problem of minimizing the energy as a function of the atomic positions, we introduce here a spin dynamics, where the magnetic orientations, $\hat{\mathbf{e}}_\nu$, take the role of the variables.

Any vector-spin DFT calculation has to start with a reasonably chosen spin configuration in a prescribed unit cell. On a simple level, one can “relax” the directions of the magnetization at the atoms in the same way a relaxation of the atomic structure (e.g. at a

surface) is done. The magnetization directions, $\hat{\mathbf{e}}_\nu$, will then normally change to reduce the total energy (cf. Figure 2). The final magnetic state, that will be reached, will in general depend on the starting point of the calculation and a more elaborate technique will be needed to avoid being trapped in some local minimum of $E[\{\hat{\mathbf{e}}_\nu\}]$.

To this end we have to develop an equation of motion for the magnetization of an atom. To keep things simple, we will focus on the case, where the magnetization stays collinear within the vicinity of the atom. Let us start from the Hamiltonian of Eq. (7) and assume that the external potential matrix, $\underline{v}(\mathbf{r})$, has been chosen to be diagonal, and the exchange-correlation potential is separated into diagonal and off-diagonal parts. Following Antropov et al.^{8,9} we use

$$\frac{\delta E_{xc}}{\delta \underline{n}(\mathbf{r})} = v_{xc} \underline{1} - \underline{\sigma} \cdot \mathbf{B}_{xc} \quad \text{and} \quad \Phi = \begin{pmatrix} \phi^{(+)} \\ \phi^{(-)} \end{pmatrix},$$

and set up a time-dependent analogon of Eq. (7):

$$i \frac{d\Phi}{dt} = [H_d - \underline{\sigma} \cdot \mathbf{B}(\mathbf{r}, t)] \Phi, \quad (14)$$

where H_d is the Hamiltonian that contains now only diagonal parts.

We will now separate the evolution of the magnetization into fast (value of the magnetization) and slow (direction of the magnetization) degrees of freedom. The former part will be described quantum-mechanically, while the latter is treated on a semiclassical level. Let us assume a field of spin-rotation matrices, $\underline{U}(\mathbf{r}, t)$, that transforms our spin coordinate system such, that the local z -axis is always in the direction of the $\mathbf{B}$ -field. Using transformed basis functions, $\chi = \underline{U}\Phi$, we can reformulate Eq. (14) as

$$i \frac{d}{dt} \chi(\mathbf{r}, t) = \left[H_d - \sum_{\nu} \underline{\sigma}_{z,\nu} B_{z,\nu}(\mathbf{r}, t) \right] \chi(\mathbf{r}, t). \quad (15)$$

where $B_{z,\nu}(\mathbf{r}, t)$ is $\mathbf{B}(\mathbf{r}, t)$ at an atom ν in a locally diagonal form. At a given time, t , the time-independent version of this equation can be solved, provided that $\underline{U}(\mathbf{r}, t)$ is known. Since $\underline{U}$ follows the magnetization, we have to determine now an equation of motion for the magnetization.

The equation of motion of the magnetization, $\mathbf{m}(\mathbf{r}, t)$, can be obtained by multiplying Eq. (14) from the left with $\Phi^* \underline{\sigma}$ and adding the complex conjugate equation. Comparing this to the time derivative of Eq. (12) and using the relation $\underline{\sigma}(\underline{\sigma} \cdot \mathbf{B}) = \mathbf{B} - i \underline{\sigma} \times \mathbf{B}$ we get

$$\frac{d\mathbf{m}(\mathbf{r}, t)}{dt} = 2\mathbf{m} \times \mathbf{B} + \frac{i}{2} \nabla (\Phi^* \underline{\sigma} \cdot \nabla \Phi - c.c.). \quad (16)$$

The second term on the right side is complicated and describes longitudinal changes of the magnetization, which we will not consider on this level. Omitting this term, Eq. (16) describes the precession of the magnetization direction at an atom under the influence of the magnetic field generated by the atom itself and other atoms of the crystal.

Returning once more to Eq. (1), we can simplify Eq. (16) and write the evolution of the magnetization direction in atom ν as

$$\frac{d\hat{\mathbf{e}}_\nu}{dt} = -\frac{2}{\mu_B} \hat{\mathbf{e}}_\nu \times \mathbf{I}_\nu \quad (17)$$

where $\mathbf{I}_\nu = \mu_B \mathbf{B}$. If we explicitly also want to take into account the effect of other fields acting onto a magnetization direction, e.g. stemming from the spin-orbit interaction (magnetic anisotropy) or dipole-dipole interaction, these fields can be added to $\mathbf{I}_\nu$ in Eq. (17): $\mathbf{I}_\nu \rightarrow \mathbf{I}_\nu + \mathbf{I}_{\text{SO}} + \mathbf{I}_{\text{d-d}}$. More general expressions of Eq. (17), suitable for spin-dynamics with finite temperatures included, can be found in Ref. 9.

The next question, that has to be answered, is how to determine the fields $\mathbf{I}_\nu$, given i.e. a certain set of magnetization directions $\{\hat{\mathbf{e}}_\nu\}$ that gives the torque on a selected magnetic moment¹⁰. This problem can be solved in constrained vector-spin density functional theory, as introduced in the next section.

2.3 Non-Stationary States: Constrained (VS)DFT

In general, an arbitrary magnetic configuration given by a set of local (atomic) magnetization directions $\{\hat{\mathbf{e}}_\nu\}$ is not an extremum or a stationary solution of the total energy functional $E[\underline{n}(\mathbf{r})]$. Exceptions are high symmetry states, like collinear magnetic states, a certain class of spin-spiral states (see Section 3.2) and particular linear superpositions of several spin-spiral states. The constrained density functional theory developed by Dedrichs *et al.*¹¹ provides the necessary generalization to deal with arbitrary magnetic configurations, i.e. configurations where the orientations of the local moments are constrained to non-equilibrium directions. We define a generalized energy functional $\tilde{E}[\underline{n}(\mathbf{r})|\{\hat{\mathbf{e}}_\nu\}]$, where we ensure that the average magnetization in an atom, $\langle \mathbf{m} \rangle_\nu$, points in the direction $\hat{\mathbf{e}}_\nu$. This condition, $\hat{\mathbf{e}}_\nu \times \langle \mathbf{m} \rangle_\nu = 0$, is introduced by an Lagrange multiplier, $\boldsymbol{\lambda}$, such that¹²

$$\begin{aligned} \tilde{E}[\underline{n}(\mathbf{r})|\{\hat{\mathbf{e}}_\nu\}] &= E[\underline{n}(\mathbf{r})] + \sum_\nu \boldsymbol{\lambda}^\nu \cdot (\hat{\mathbf{e}}_\nu \times \langle \mathbf{m} \rangle_\nu) \\ &= E[\underline{n}(\mathbf{r})] + \mu_B \sum_\nu \mathbf{B}_c^\nu \cdot \langle \mathbf{m} \rangle_\nu. \end{aligned} \quad (18)$$

Here, we recast the Lagrange multiplier in the form of a magnetic field, $\mathbf{B}_c^\nu$. This is the constraining field in the atom ν that fixes the local (integrated) magnetic moment $\mathbf{M}^\nu$ parallel to the prescribed direction $\hat{\mathbf{e}}_\nu$. $\mathbf{M}^\nu$ is the magnetization density

$$\langle \mathbf{m}(\mathbf{r}) \rangle_\nu = \mathbf{M}^\nu = \int_{MT^\nu} \mathbf{m}(\mathbf{r}) d^3r \quad (19)$$

averaged over the sphere where Eq. (1) holds. Thus $\mathbf{B}_c^\nu$ ensures that the local moments have no components $\mathbf{M}_\perp^\nu$ normal to the directions $\hat{\mathbf{e}}_\nu$, $\hat{\mathbf{e}}_\nu^\perp$, for any atom.

The effective B -field, $\mathbf{B}_{\text{eff}}^\nu$, that enters the muffin-tin part of the Hamiltonian is given by (here, absence of external fields is assumed for simplicity)

$$\mathbf{B}_{\text{eff}}^\nu(\mathbf{r}) = B_{\text{xc}}^\nu[\underline{n}(\mathbf{r})] \hat{\mathbf{e}}_\nu + B_c^\nu \hat{\mathbf{e}}_\nu^\perp = B_{\text{eff}}^\nu(\mathbf{r}) \hat{\mathbf{e}}_B^\nu(\mathbf{r}). \quad (20)$$

Even if $\mathbf{B}_{\text{xc}}^\nu$ is assumed to be collinear in the vicinity of the atom ν (pointing in the direction $\hat{\mathbf{e}}_\nu$), the effective $\mathbf{B}$ -field is again a continuous non-collinear vector field in the muffin-tin spheres, with pointwise local directions $\hat{\mathbf{e}}_B^\nu(\mathbf{r})$.

In an actual constrained local moment (CLM) calculation, $\underline{n}(\mathbf{r})$ and $\mathbf{B}_c^\nu$ have to be determined self-consistently. The density matrix is calculated in the usual self-consistency cycle. At the same time, the local constraint fields $\mathbf{B}_c^\nu$ have to be adjusted, until the constraint conditions, $\hat{\mathbf{e}}_\nu \times \langle \mathbf{m} \rangle_\nu = 0$, are fulfilled (cf. Figure 3). At the end of such a

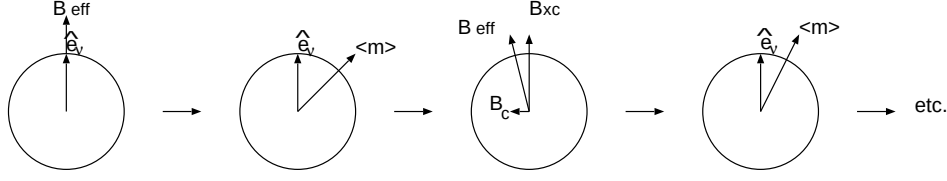


Figure 3. Determination of the constraint field: Initially, the effective $\mathbf{B}$ -field is parallel to the prescribed direction $\hat{\mathbf{e}}_\nu$ (left). The resulting magnetization, $\langle \mathbf{m} \rangle$, generally is not parallel to this direction. Therefore, a constraint field $\mathbf{B}_c$ is introduced, that points in opposite direction to the component of the magnetization that is perpendicular to $\hat{\mathbf{e}}_\nu$. Using this $\mathbf{B}_{\text{eff}}$, the direction of the magnetization is then adjusted towards $\hat{\mathbf{e}}_\nu$ (right).

calculation we obtain the self-consistent densities and a set of local constraint $\mathbf{B}$ -fields. The total energy of the system is given by the constrained energy functional, Eq. (18).

According to the the Hellmann-Feynman theorem we find that the change of the energy due to a change in magnetization direction, $d\hat{\mathbf{e}}_\nu$, is given by $dE = -\mu_B \mathbf{M}^\nu \cdot (\mathbf{B}_c^\nu \times d\hat{\mathbf{e}}_\nu)$. Therefore, the constraint field can be interpreted as a torque acting on the magnetic moment, in the spirit of the spin dynamics, formulated in the previous section. Thus, we have set up a formalism that allows us to find – at least in principle – the magnetic ground state of a system by spin-dynamics¹³. But CLM calculations can also be used in a different way: In the next section we will describe how they can be used to determine the exchange interactions in a system and utilize these results in models, like the classical Heisenberg model, to obtain information about the ground state, but also about excited states of a magnetic system.

3 Magnetism and Exchange Interactions

Using the methods and results described in the last sections, it is possible to simulate magnetic systems and determine their ground state quite accurately. Nevertheless, such a strategy might be not very satisfactory, since spin-dynamics simulations are computationally quite demanding, may depend on computational parameters like the unit-cell size, and finally do not offer very much insight into the underlying physics that leads to the magnetic ground state. Therefore, it is sometimes useful to map the results of DFT on a physical model that can help in the interpretation of the results (Section 3.1) and suggest possible magnetic ground state structures (Section 3.2). In some cases, it is even possible to obtain non-groundstate properties from these models (Section 3.3).

3.1 The Heisenberg Model and Its Extensions

Consider a system of N electrons localized on N lattice sites, such that every lattice site is occupied by exactly one electron. Each electron is localized in an orbital $f_n(\mathbf{r}, \sigma)$ with a spin index σ . The Hamiltonian of the system will consist of a kinetic energy part, the Coulomb repulsion, and some (scalar) potential, that keeps the electrons localized. If we are only interested in energy changes that come from a change of the spin-part of the wavefunctions (i.e. the magnetic structure of the lattice), the contributions of the kinetic energy and the scalar potential can be ignored. But the Coulomb interaction, together

with the Pauli exclusion principle, can differentiate between spin structures: suppose the Coulomb energy is given by

$$V = \frac{e^2}{2} \sum_{\sigma\sigma'} \int d\mathbf{r} d\mathbf{r}' \frac{\psi^*(\mathbf{r}', \sigma') \psi^*(\mathbf{r}, \sigma) \psi(\mathbf{r}, \sigma) \psi(\mathbf{r}', \sigma')}{|\mathbf{r} - \mathbf{r}'|} ; \quad \psi^*(\mathbf{r}, \sigma) = \sum_n c_{n\sigma}^* f_n^*(\mathbf{r}) \quad (21)$$

where $c_{n\sigma}^*$ is the creation operator of an electron of spin σ at site n . Then, generally,

$$V = \frac{e^2}{2} \sum_{\sigma\sigma'} \sum_{n_1 n_2 n_3 n_4} c_{n_1\sigma'}^* c_{n_2\sigma}^* c_{n_3\sigma} c_{n_4\sigma'} \int d\mathbf{r} \int d\mathbf{r}' \frac{f_{n_1}^*(\mathbf{r}') f_{n_2}^*(\mathbf{r}) f_{n_3}(\mathbf{r}) f_{n_4}(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|}. \quad (22)$$

A single term of this sum describes the interaction of an electron with spin σ at site 3 with an electron with spin σ' at site 4, so that finally a state with electrons with spins σ' and σ at sites 1 and 2 is reached. Since all electrons are localized, we have $(n_1, n_2) = (n_3, n_4)$. Introducing the number operators $n_{n\sigma} = c_{n\sigma}^* c_{n\sigma}$, we can write:

$$V = \frac{1}{2} \sum_{\sigma\sigma'} \sum_{nn'} (n_{n'\sigma'} n_{n\sigma} K_{nn'} + c_{n'\sigma'}^* c_{n\sigma}^* c_{n'\sigma} c_{n\sigma'} J_{nn'}) \quad (23)$$

$$K_{nn'} = e^2 \int d\mathbf{r} \int d\mathbf{r}' |f_n(\mathbf{r})|^2 |\mathbf{r} - \mathbf{r}'|^{-1} |f_{n'}(\mathbf{r}')|^2$$

$$J_{nn'} = e^2 \int d\mathbf{r} \int d\mathbf{r}' f_n^*(\mathbf{r}') f_n^*(\mathbf{r}) (\mathbf{r} - \mathbf{r}')^{-1} f_{n'}(\mathbf{r}) f_{n'}(\mathbf{r}').$$

The first term of Eq. (23) does not depend on the spin and can again be ignored. The second term, however, describes an exchange of spins between sites n and n' and $J_{nn'}$ is, therefore, called exchange coupling constant. Carrying out the spin summation in Eq. (23) this term becomes¹⁴:

$$V = - \sum_{nn'} J_{nn'} \left(\frac{1}{4} + \mathbf{s}_n \cdot \mathbf{s}_{n'} \right). \quad (24)$$

This term describes an attractive potential for electrons of the same spin at neighboring sites. It corrects the ordinary Coulomb repulsion described by K for the fact, that electrons of the same spin have to avoid each other due to the requirements of the Pauli principle.

In this picture, clearly ferromagnetism is favored. But modifications of the original assumptions (e.g. if the interaction between the sites is “mediated” by other orbitals), can also lead to antiferromagnetic couplings. In any case the Heisenberg Hamiltonian of the form

$$H = - \sum_{nn'} J_{nn'} \mathbf{s}_n \cdot \mathbf{s}_{n'} \quad (25)$$

can be used as a phenomenological starting point in the investigation of the magnetic interaction in a crystal. Although the Heisenberg model was introduced for magnetic insulators with localized moments, we can also apply Eq. (25) to metallic systems, as shown in Figure 4. In these hexagonal unsupported monolayers the behavior of the total energy as a function of the relative angle between the atoms can be described as cosine-like function, the exchange coupling constant being negative for Cr and Mn (preferring antiferromagnetic coupling) and positive for Fe (leading to a ferromagnetic ground state). The total

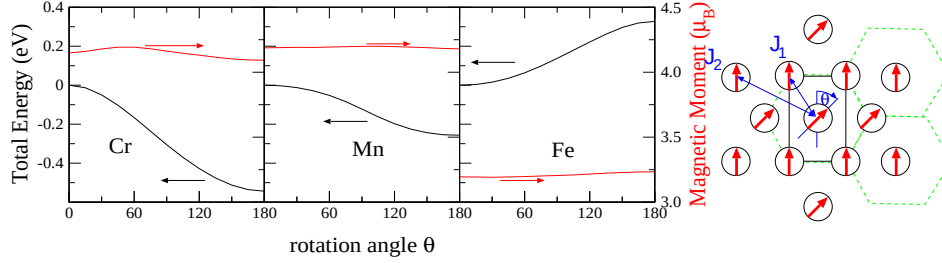


Figure 4. Total energy and magnetic moment of hexagonal monolayers of Cr, Mn, and Fe as a function of the angle of the magnetization in a two-atomic unit cell (right). $\theta = 0^\circ$ corresponds to a ferromagnetic state, $\theta = 180^\circ$ is a row-wise antiferromagnetic state. As lattice constant we chose the parameters of the Ag(111) surface. In the schematic picture of the hexagonal monolayer (right) the coupling to nearest neighbors (J_1) and next-nearest neighbors (J_2) is indicated.

energy has been calculated by a constrained DFT calculation as described above. We further see, that the magnetic moment does not change significantly as the spins are rotated, an important requirement for the application of the Heisenberg model.

From the right part of Figure 4 we can see, that rotating the local magnetic moment direction of one atom in the two-atom unit cell of the hexagonal lattice will change the relative orientation of that atom to four nearest neighbors, but does not affect two of the nearest neighbor (NN) atoms. Likewise, only four of the six second-NN atoms will change the relative orientation to the original atom. This leads to an expression for the total energy in the classical Heisenberg model up to second-NN:

$$E = -S^2(J_1 + J_2)(2 + 4 \cos \theta) \quad (26)$$

if $\mathbf{S}$ is now the total spin moment treated as a classical vector. This means, from a constrained local moment calculation we can at least estimate the size of $(J_1 + J_2)$. It is not difficult to find other unit cells and rotations that allow the determination of other linear combinations of J_1 and J_2 , thereby separating the individual exchange coupling constants¹⁵.

Of course, the energies obtained from the CLM calculation contain contributions of all J_n and also from interactions that are not described by the Heisenberg model. Such terms can be classified with respect to an expansion of H in powers of $\mathbf{S}$, such that the Heisenberg model contains the terms up to second order (since, as discussed above, it is derived from a model taking into account pairwise interactions). The next highest terms (in absence of spin-orbit coupling effects) are then to fourth order in $\mathbf{S}$:

$$H_{\text{bi-qu.}} = - \sum_{ij} B_{ij} (\mathbf{S}_i \cdot \mathbf{S}_j)^2 \quad \text{and} \quad (27)$$

$$H_{4\text{-spin}} = - \sum_{ijkl} K_{ijkl} [(\mathbf{S}_i \mathbf{S}_j)(\mathbf{S}_k \mathbf{S}_l) + (\mathbf{S}_j \mathbf{S}_k)(\mathbf{S}_l \mathbf{S}_i) - (\mathbf{S}_i \mathbf{S}_k)(\mathbf{S}_j \mathbf{S}_l)].$$

The first term, the biquadratic interaction, comes from hopping processes between sites 1 and 2 like $1 \rightarrow 2 \rightarrow 1 \rightarrow 2 \rightarrow 1$, the second term, the 4-spin interaction from a hopping $1 \rightarrow 2 \rightarrow 3 \rightarrow 4 \rightarrow 1$. Terms of third order can only occur in the presence of spin-orbit

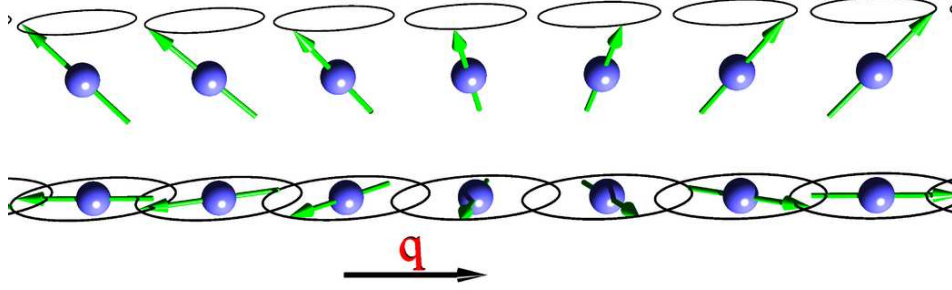


Figure 5. Spiral spin density wave (SSDW) or spin spiral propagating along a row of atoms with a wavevector $\mathbf{q}$. The opening angle is 45° in the upper and 90° in the lower example.

interaction. E.g. the so-called Dzyaloshinsky-Moriya interaction has the form:

$$H_{\text{DM}} = \mathbf{D} \cdot (\mathbf{S}_1 \times \mathbf{S}_2) \quad (28)$$

and comes from anisotropic exchange interaction¹⁶. All these different interaction terms can be extracted from a set of suitable *ab initio* calculations (possibly including spin-orbit interaction) and used to determine the magnetic ground state within the chosen model.

3.2 Spin Spirals and the Generalized Bloch Theorem

In a periodic crystal it is convenient to replace the quantities in Eq. (25) by their Fourier-transformed equivalents:

$$\mathbf{S}(\mathbf{q}) = \frac{1}{N} \sum_n \mathbf{S}_n e^{-i\mathbf{q}\mathbf{R}_n} \quad \text{and} \quad J(\mathbf{q}) = \sum_n J_{0n} e^{-i\mathbf{q}\mathbf{R}_n}. \quad (29)$$

Exploiting the translational invariance of the lattice, we can then rewrite Eq. (25) as

$$H = -N \sum_{\mathbf{q}} J(\mathbf{q}) \mathbf{S}(\mathbf{q}) \cdot \mathbf{S}(-\mathbf{q}) \quad (30)$$

where we have to ensure that the length of all spins $\mathbf{S}_n^2 = S^2$ is conserved on all sites n . This condition is fulfilled by solutions of the type¹⁶

$$\mathbf{S}_n = \sqrt{2}S (\hat{\mathbf{e}}_x \cos(\mathbf{q} \cdot \mathbf{R}_n) + \hat{\mathbf{e}}_y \sin(\mathbf{q} \cdot \mathbf{R}_n)) \quad (31)$$

where the unit vectors $\hat{\mathbf{e}}_x$ and $\hat{\mathbf{e}}_y$ just have to be perpendicular to each other, otherwise their directions are arbitrary. Eq. (31) describes a spiral spin density wave (SSDW) as shown in the lower half of Figure 5. These SSDWs are general solutions of the Heisenberg model. From Eq. (30) one can conclude that the SSDW with the lowest total energy will be the one with the propagation vector $\mathbf{Q}$ which maximizes $J(\mathbf{q})$. E.g. if $\mathbf{Q} = 0$ maximizes $J(\mathbf{q})$, the solution corresponds to the ferromagnetic state, if $\mathbf{Q} = \hat{\mathbf{e}}_z \frac{\pi}{a_z}$ and a_z is the lattice constant in z -direction, then the structure is layered antiferromagnetic in z -direction. Some other examples – for a hexagonal monolayer – are illustrated in Figure 6.

SSDWs can be described by the propagation vector of the spin-spiral $\mathbf{q}$, the rotation axis (which is, when no coupling between spin and lattice is present, not fixed with respect

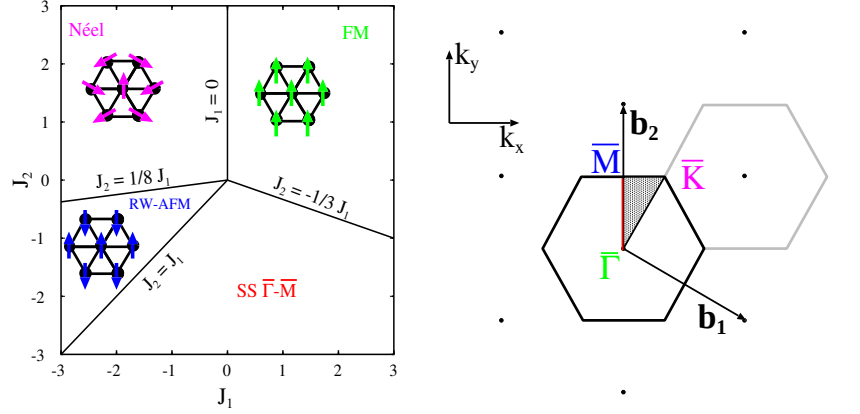


Figure 6. Phase diagram of the hexagonal lattice in next-nearest neighbor approximation of the classical Heisenberg model: two collinear, a ferromagnetic (FM) and a row-wise antiferromagnetic (RW-AFM) solutions can be obtained, and two non-collinear solutions, the Néel-state and a SSDW with $\mathbf{q}$ -vectors along the line $\bar{\Gamma} - \bar{M}$ of the Brillouin zone (right). The FM, RW-AFM and Néel-state correspond to SSDWs with $\mathbf{q}$ -vectors on the high-symmetry points $\bar{\Gamma}$, $\bar{M}$ and $\bar{K}$, respectively.

to the lattice) and the relative angle ϑ between the magnetic moment and the rotation axis. Upon translation by a lattice vector $\mathbf{R}$, the magnetic moment of an atom rotates by an angle $\varphi = \mathbf{q} \cdot \mathbf{R}$. Assuming a rotation around the z -axis (in absence of spin-orbit coupling this is not a loss of generality), the magnetic moment of an atom ν having the basis vector $\boldsymbol{\tau}^\nu$ in the unit cell n (with the origin at the lattice vector $\mathbf{R}_n$) points in the direction

$$\hat{\mathbf{e}}_{n\nu} = \begin{pmatrix} \cos(\mathbf{q} \cdot (\mathbf{R}_n + \boldsymbol{\tau}^\nu) + \xi^\nu) \sin \vartheta^\nu \\ \sin(\mathbf{q} \cdot (\mathbf{R}_n + \boldsymbol{\tau}^\nu) + \xi^\nu) \sin \vartheta^\nu \\ \cos \vartheta^\nu \end{pmatrix}. \quad (32)$$

For more than one magnetic atom in the unit-cell, in addition to its basis vector $\boldsymbol{\tau}_\nu$, an additional atom-dependent phase, ξ^ν , appears in the above equation. Nevertheless, the magnetic moments of all atoms rotate around the same axis.

SSDWs are sometimes also called frozen magnons, since a spin-spiral looks like a “snap shot” of a single magnon at a fixed time. Spin-spiral calculations can therefore be used to simulate the effect of temperature on a magnetic system in the adiabatic approximation, in particular at very low temperatures, when magnons with long wavelength dominate (cf. Section 3.3). But also at $T = 0$ many compounds and even elements show SSDW ground states. Some examples are shown in Figure 7.

A very elegant treatment of spin-spirals by first-principle calculations is possible if the generalized Bloch theorem^{17,18} is applied. However, this theorem can only be proved, when spin-orbit coupling is neglected. For this reason the spin-rotation axis will always be considered as parallel to the z -axis of the spin coordinate-frame. Thus, only the m_x and m_y components are rotated, while m_z does not change. Following Sandratskii¹⁸, we can define a generalized translation, $\mathcal{T}_n$, that combines a lattice translation, $\mathbf{R}_n$, and a spin rotation $\mathbf{U}$ that commutes with the Hamiltonian H . Applying such a generalized translation to $H\Phi$

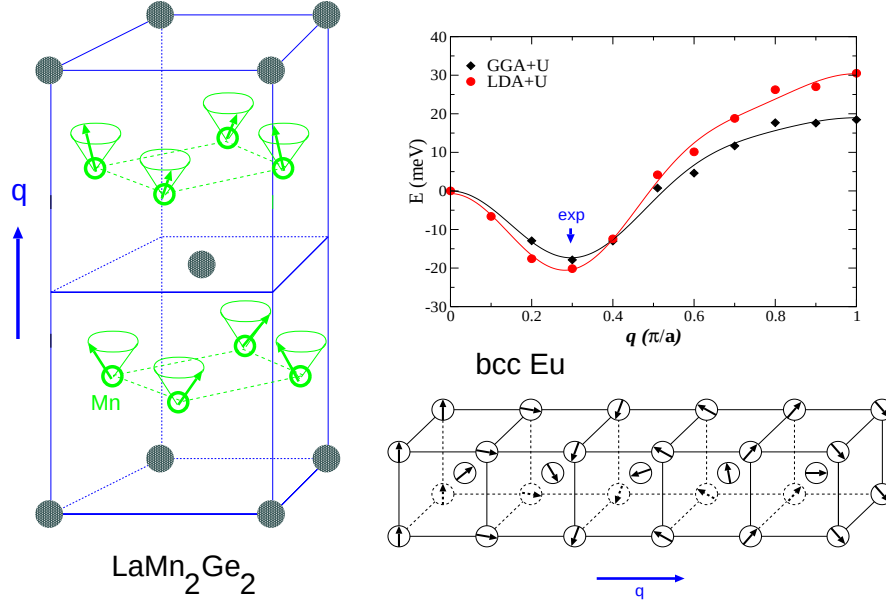


Figure 7. Examples of SSDW ground-states: in LaMn_2Ge_2 the spins on the Mn sublattice can be described by a conical spin-spiral with 58° semicone-angle and a $\mathbf{q}$ -vector of $(0, 0, 0.71)$ in units of $2\pi/c$ (left). Bulk bcc europium has a flat spiral in $[001]$ direction as ground state (right, bottom), the length of the $\mathbf{q}$ -vector is correctly reproduced in DFT calculations (right, top).

yields

$$\begin{aligned} \mathcal{T}_n H(\mathbf{r}) \Phi(\mathbf{r}) &= \mathbf{U}(-\mathbf{q}\mathbf{R}_n) H(\mathbf{r} + \mathbf{R}_n) \mathbf{U}^\dagger(-\mathbf{q}\mathbf{R}_n) \mathbf{U}(-\mathbf{q}\mathbf{R}_n) \Phi(\mathbf{r} + \mathbf{R}_n) \\ &= H(\mathbf{r}) \mathbf{U}(-\mathbf{q}\mathbf{R}_n) \Phi(\mathbf{r} + \mathbf{R}_n) \end{aligned} \quad (33)$$

where $\mathbf{U}(\mathbf{q}\mathbf{R}_n)$ is the spin 1/2 rotation matrix

$$\mathbf{U}(\mathbf{q}\mathbf{R}_n) = \begin{pmatrix} e^{-i\varphi/2} & 0 \\ 0 & e^{i\varphi/2} \end{pmatrix}, \quad \varphi = \mathbf{q} \cdot \mathbf{R}_n. \quad (34)$$

In analogy with the proof of Bloch's theorem¹⁹ it follows that the eigenstates can be chosen such that

$$\mathcal{T}_n \Phi(\mathbf{k}, \mathbf{r}) = \mathbf{U}(-\mathbf{q}\mathbf{R}_n) \Phi(\mathbf{k}, \mathbf{r} + \mathbf{R}_n) = e^{i\mathbf{k} \cdot \mathbf{R}_n} \Phi(\mathbf{k}, \mathbf{r}). \quad (35)$$

Since these eigenstates are labeled with the same Bloch vector $\mathbf{k}$ as the eigenstates of the translation operator without the spin rotation, the lattice periodic part of these states follows the chemical lattice, $\mathbf{R}_n$, i.e. we can calculate the spin spiral state in the chemical unit cell.

Without the application of the generalized Bloch theorem the investigation of such magnetic structures requires very large unit cells. Since the spin-spiral is the exact solution of the classical Heisenberg model at $T = 0$, it is believed that they cover a large and important part of the phase space of possible spin states. Thus among all possible magnetic states, spin-spirals are the next relevant class of spin states besides the high-symmetry magnetic states, i.e. the ferromagnetic, antiferromagnetic, or ferrimagnetic configurations.

3.3 Critical Temperatures, T_C , T_N

Let us now see, how temperature will influence the magnetic order in a ferromagnetic solid. Staying within the Heisenberg model, we will assume that the magnitude of the magnetic moments at the atoms will – in first approximation – not be changed, and discuss just their mutual orientation. At $T = 0$ the spin at a selected atom will be fixed in parallel direction to the spins at all other atoms by an effective field that will be proportional to $S \sum_n J_{0n} = S J_0$. At a finite temperature T , this field, acting on the spin at site 0 is reduced due to the thermal fluctuation on the sites n . The thermal average of the projection of the spin at site n on the spin at site 0 is denoted as $\langle S(\mathbf{R}_n) \rangle$. In the “mean field approximation” (MFA), it is assumed that the effective field at finite temperatures that acts on spin 0 is:

$$B_{\text{eff}} = \sum_n J_{0n} \langle S(\mathbf{R}_n) \rangle \quad (36)$$

In this model it is possible to calculate the temperature-dependence of the average magnetization of the solid and, specifically, the temperature where the average magnetization vanishes, the critical temperature. For a ferromagnet this temperature is called Curie temperature and in the MFA it is given by

$$T_C = \frac{2S(S+1)}{3k_B} J_0 \quad (37)$$

It has to be mentioned, that in most cases the MFA severely overestimates T_C (by about 20 to 50%, depending on the lattice). Nevertheless, it gives a simple estimate of the ordering temperature in systems, where the approximations of the Heisenberg model are reasonable. On the other hand some properties, like the – material independent – critical exponents, are in any case not reliably reproduced by the MFA.

On a more sophisticated level, the “random phase approximation” (RPA) can give quite reliable results. In contrast to the MFA, where the thermal averaging was done over the sites n that determine B_{eff} , here the Hamiltonian is first transformed into reciprocal space (Eq. (30)), and then the averaging is done over one of the Fourier components:

$$H = -N \sum_{\mathbf{q}} J(\mathbf{q}) \mathbf{S}(\mathbf{q}) \cdot \langle \mathbf{S}(-\mathbf{q}) \rangle \quad (38)$$

If the term $S(S+1)$ is included in the exchange coupling constants (as it is usually done, when the J ’s are determined from first-principles calculations), then the Curie temperature in the MFA and RPA can be expressed as

$$k_B T_C^{\text{MFA}} = \frac{2}{3} J_0 \quad k_B T_C^{\text{RPA}} = \frac{2}{3} \left(\sum_{\mathbf{q}} \frac{1}{J(\mathbf{q})} \right)^{-1} \quad (39)$$

From these expressions it is obvious, that calculating T_C in the RPA involves not more information than what is needed on a mean-field level, if the exchange coupling constants are calculated in reciprocal space by using the generalized Bloch theorem.

Also for antiferromagnets (or, generally spin-spiral states characterized by a vector $\mathbf{Q}$) expressions for the ordering temperature, the Néel temperature T_N , can be derived. In the MFA with $S(S+1)$ again included in J , this is given simply by

$$k_B T_N^{\text{MFA}} = \frac{2}{3} J(\mathbf{Q}) \quad (40)$$

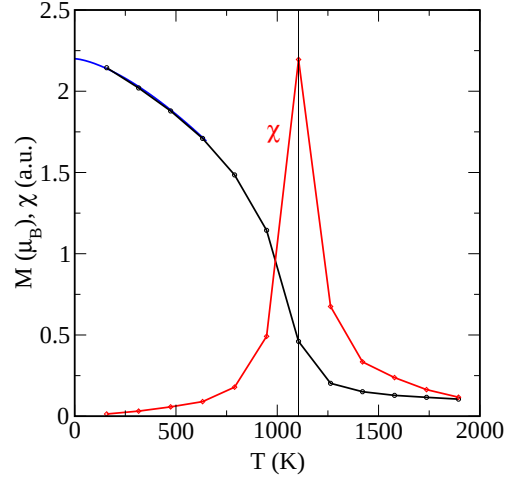


Figure 8. Magnetization, (M), and susceptibility (χ , in arbitrary units) of bulk Fe obtained by MC simulations with several hundred exchange coupling constants taken from *ab initio* calculations²². The $T = 0$ behavior of the magnetization is extrapolated by Bloch's law (blue curve). The finite magnetization at $T_C = 1105\text{K}$ is an artefact of the finite MC simulation super-cell (in this case $10 \times 10 \times 10$ bcc unit cells).

while a slightly more involved expression can be derived in the random phase approximation²⁰. Comparison of these results with experimental values gave reasonable results, e.g. for bcc europium Néel temperatures of 147 K and 110 K were obtained in MFA and RPA, respectively²⁰. These values have to be compared to the experimental T_N of 90.5 ± 0.5 K.

Although there exist several more methods to calculate critical temperatures from DFT results, we will outline here just one further possibility, which seems to be rather flexible and appropriate for many systems with different magnetic ground states: the Monte Carlo technique (MC) allows to study finite-temperature magnetic properties by implementation of a Heisenberg Hamiltonian (Eq. (25), possible with extensions like biquadratic terms or an uniaxial anisotropy (see below)), into a Metropolis algorithm²¹. Unit cells of different size are then studied so that finite-size effects can be eliminated. In these unit cells the evolution of the magnetic property in question (in our case the average magnetization) as a function of temperature can then be monitored (cf. Figure 8).

	T_C (K)				exp.
	MFA	MFA2	RPA	MC	
Fe (bcc)	1414	550, 1190	950	1060	1044 – 1045
Co (fcc)	1645	1120, 1350	1311	1080	1388 – 1390
Ni (fcc)	397	320, 820	350	510	624 – 631
Gd (hcp)	334				293

Table 2. Calculated and experimental Curie temperature T_C for some ferromagnetic materials. MFA and RPA data for Fe, Co and Ni taken from Padja et al.²³, MFA2 results and experimental values as quoted by Shallcross and coworkers⁵, while the MC results were obtained by Rosengaard and Johansson²⁴. Data for Gd can be found in the papers of Kurz et al.⁶ and Turek and coworkers²⁵.

Results of *ab initio* calculations of the Curie temperature of Fe, Co and Ni are presented in Table 2. From this table one can easily see that, compared to RPA, the MFA typically overestimates T_C by 25–50%. For Fe and Co RPA gives quite good estimates of the Curie temperature, while for Ni T_C is underestimated in both approximations. MC simulations work better for Ni and Fe, but give a too low T_C for Co. While we quoted here results for “simple” metals, it is nowadays possible to investigate in the same manner the temperature dependent properties of complex multicomponent systems, e.g. half-metallic Heusler alloys²⁶ or dilute magnetic semiconductors²⁷.

At low but finite temperatures, collective spin-wave excitations or magnons are excited in the ferromagnetic crystal. These magnons can again be characterized by their wave-vector $\mathbf{q}$. In the long wavelength limit, i.e. around $\mathbf{q} = 0$ the spin-wave dispersion behaves almost quadratically and can be described as Dq^2 . The spin stiffness, D , characterizes the magnetic properties of a ferromagnet at low temperatures and can also be calculated from the exchange coupling constants:

$$D = \frac{2}{3M} \sum_n J_{0n} R_{0n}^2. \quad (41)$$

Here, M is the magnetic moment of the ferromagnetic state. Some results of *ab initio* calculations are given in Table 3. Again, for Fe and Co agreement with experimental data

	D (meV Å ²)			
	th.(1)	th.(2)	th.(3)	exp.
Fe (bcc)	247	322, 313	250	280, 314, 330
Co (fcc)	502	480, 520	663	510, 580
Ni (fcc)	739	541, 1796	756	422, 550, 555

Table 3. Calculated and experimental spin-wave stiffness (D) for Fe, Co and Ni. The theoretical data was obtained in different approximations as described by Rosengaard and Johansson²⁴ [th.(1)], Shallcross and coworkers⁵ [th.(2)] and Padja et al.²³ [th.(3)], experimental data was taken as cited in these references.

is reasonable, but for Ni most methods fail to reproduce the experimental spin stiffness.

4 Magnetic Anisotropy

In the discussions above, we often assumed that the direction of the magnetic moment(s) is independent from the orientation of the crystal lattice. In many cases, e.g. in bulk crystals like Fe, Co or Ni, this assumption does not lead to qualitatively wrong results, but there are also systems, where the coupling of the spin to the lattice is essential. For example a monolayer of ferromagnetic material, deposited on a nonmagnetic and weakly interacting substrate, would show no long-range magnetic order without this coupling (cf. Section 4.2). But it is also clear that in this case the classical dipole-dipole interaction will be different for a magnetization that is perpendicular to this monolayer or parallel to the surface. For a ferromagnet with only one atom in the unit cell, the dipole energy is given

by

$$E_d = \frac{\mu_B^2}{2} \sum_{n,n'} \frac{m^2}{(\mathbf{R}_n - \mathbf{R}_{n'})^3} (1 - 3 \cos^2 \theta_{nn'}) \quad (42)$$

where $\theta_{nn'}$ is the angle between the direction of the magnetic moment $\mathbf{m}$ and the vector connecting $\mathbf{R}_n$ and $\mathbf{R}_{n'}$. In this way, the magnetization of a thin film is always coupled to the crystallographic directions. The fact, that it is more favorable to magnetize a magnetic material in a certain direction than in another direction is then called magnetic anisotropy.

This dipole-dipole interaction always depends on the shape of the macroscopic sample and the resulting magnetic anisotropy is, therefore, called shape anisotropy²⁸. But there are also other interactions which lead to magnetically anisotropic behavior, most importantly the spin-orbit coupling discussed in the next section.

4.1 Spin-Orbit Coupling

An electron, traveling with a velocity $\mathbf{v}$ on a classical trajectory around the nucleus, experiences an electric field $\mathcal{E}$ (from the screened nucleus) as a magnetic field, $\mathbf{B} = \frac{1}{c} \mathcal{E} \times \mathbf{v}$. This field will couple to the magnetic (spin) moment μ of the electron as $-\mu \cdot \mathbf{B}$.^a If we assume, that in a solid the crystal field forces the electron to move e.g. in a certain crystallographic plane, the electron spin will be aligned in a direction normal to this plane. In such a way, a uniaxial anisotropy can arise regardless of the shape of a crystal.

To treat this quantitatively on a quantum-mechanical basis, it is necessary to start from the Dirac equation. In the Schrödinger equation – even for a magnetic system – there is no term that could differentiate the various magnetization directions. But if we include a certain term from the Pauli equation (a two-component approximation to the Dirac equation³⁰) we get

$$\left[\frac{\hbar^2}{2m} \nabla^2 + \underline{v}(\mathbf{r}) - \frac{\mu_B}{2mc} \boldsymbol{\sigma} \cdot (\mathcal{E}(\mathbf{r}) \times \mathbf{p}) \right] \Phi_i = \epsilon_i \Phi_i. \quad (43)$$

It is this relativistic correction (factor $\frac{1}{c}$) that leads to the coupling between spin-space ($\boldsymbol{\sigma}$) and lattice ($\mathcal{E}(\mathbf{r})$). If we assume that the electric field is derived from a spherically symmetric potential, $V(r)$, (as occurs in the vicinity of an atomic nucleus) we can transform this term

$$\boldsymbol{\sigma} \cdot (\mathcal{E}(\mathbf{r}) \times \mathbf{p}) = \boldsymbol{\sigma} \cdot (\nabla V(r) \times \mathbf{p}) = \frac{1}{r} \frac{dV(r)}{dr} \boldsymbol{\sigma} \cdot (\mathbf{r} \times \mathbf{p}) = \frac{1}{r} \frac{dV(r)}{dr} (\boldsymbol{\sigma} \cdot \mathbf{L}) = \xi \boldsymbol{\sigma} \cdot \mathbf{L}, \quad (44)$$

where $\mathbf{L}$ is the orbital momentum operator. This term is called the spin-orbit coupling (SOC) term with the spin-orbit coupling constant ξ . Since the radial derivative of the potential in a crystal will be largest in the vicinity of a nucleus, we can expect that the major contribution to the spin-orbit interaction will come from this region. For an atom ν then r is the radial part of the vector $\mathbf{r}_\nu = \mathbf{r} - \boldsymbol{\tau}_\nu$. Furthermore, since for small r_ν the potential will be Coulomb-like ($V(r) = -\frac{Z}{r}$), its derivative $\frac{\partial V}{\partial r_\nu}$ is proportional to the

^a Although this interaction has the form of a Zeeman term (the interaction of the spin with an external magnetic field), due to kinematical effects this spin-orbit interaction is smaller by a factor of two. The origin of this effect is called Thomas-precession²⁹.

nuclear number of the atom, Z_ν . We thus expect that ξ will be large for heavy atoms, but small for lighter ones.

The magnetocrystalline anisotropy energy (MAE) will then result from the anisotropy of the spin-orbit interaction, i.e. it is the difference of total energies obtained from Hamiltonians including the spin-orbit coupling term with the magnetization pointing in two different directions. Physically, this difference arises due to the crystal field that forces the orbital motion of the electron into a preferred plane³¹. This energy difference is small (in the order of a few μeV 's) for bulk systems with high symmetry, e.g. cubic crystals like Fe or Ni. It is larger for crystals with a unique crystallographic axis, like hexagonal Co. But for lower dimensional systems, thin films or atomic wires, the magnetocrystalline anisotropy will essentially determine the magnetic properties, especially at finite temperatures.

4.2 Critical Temperatures in Low Dimensional Systems

In a bulk system, like Fe, the magnetic order is stabilized by exchange interactions against temperature induced fluctuations. The high Curie-temperature $T_C = 1045\text{K}$ of Fe suggests already that the involved energy scale is about 100 meV ($k_B = 8.617 \cdot 10^{-5}\text{ eV/K}$), and therefore five orders of magnitude larger than typical anisotropy energies in cubic systems. Although the exchange interaction is strong, it is rather short-ranged and decays exponentially with the distance. For two-dimensional systems, Mermin and Wagner³² showed that no spontaneous magnetization will occur for finite temperatures if the magnetic order is stabilized by an interaction that decays faster than $(\tau_i - \tau_j)^{-n}$ (where the τ 's are the atomic positions and n is a number).

In nature, strictly two-dimensional systems do not occur. Even if only one monolayer of magnetic atoms is deposited on a weakly interacting substrate (like Ag), there is always some interaction between the magnetic atoms via the polarized substrate atoms. But even if this interaction did not exist, there is still a possibility to stabilize magnetic order in this monolayer – via the magnetic anisotropy stemming either from the dipolar or the spin-orbit interaction. For a two-dimensional Heisenberg ferromagnet, Bander and Mills³³ demonstrated that there is indeed a critical temperature if a uniaxial anisotropy is present in the system. If the anisotropy constant is denoted as K and the transition temperature for the same Heisenberg ferromagnet in three dimensions is T_{3D} , the transition temperature of the two-dimensional system is given by:

$$T_{2D} = T_{3D} / \ln \left(\frac{3\pi k_B T_{3D}}{4K} \right). \quad (45)$$

In the above mentioned case of uniaxial anisotropy, the energy difference between the situations where a spin $\mathbf{S}$ is perpendicular or parallel to the film normal, chosen to be $\hat{\mathbf{e}}_z$, is given by anisotropy constant K (note, that the different directions in the film plane are assumed to be equivalent here). The uniaxial anisotropy can then be described as a magnetic field $\mathbf{B} = K\hat{\mathbf{e}}_z$, that acts on the spin $\mathbf{S}$. This field gives rise to a Zeeman-term that can be added to the Heisenberg Hamiltonian:

$$H = - \sum_{nn'} J_{nn'} \mathbf{S}_n \cdot \mathbf{S}_{n'} + g\mu_B \sum_n \mathbf{S}_n \cdot \mathbf{B}. \quad (46)$$

In model Hamiltonian calculations or Monte-Carlo simulations, this type of Hamiltonian can be used when inclusion of an uniaxial anisotropy term is required.

4.3 The Orbital Moment

In section 4.1 we have seen, that in crystalline systems also an orbital motion of the electron around the nucleus occurs, giving rise to a magnetic moment, the orbital moment. Compared to the situation in a free atom this motion is of course restricted by the crystal field that quenches the orbital moment. But spin-orbit coupling provides a mechanism that counteracts this orbital moment quenching and small moments (typically $0.1 - 0.2\mu_B$) can be found (compare Table 1).

The orbital magnetization can be defined in analogy to Eq. (12), expressed in single-particle wavefunctions ϕ_i :

$$\mathbf{m}(\mathbf{r}) = -\mu_B \sum_i \langle \phi_i | \mathbf{r} \times \mathbf{v} | \phi_i \rangle. \quad (47)$$

where $\mathbf{v}$ is the velocity operator. At a certain atom ν , the orbital moment M_ν^{orb} can then be obtained by an integration similar to Eq. (19):

$$M_\nu^{\text{orb}} = -\mu_B \sum_i \langle \phi_i | \mathbf{r} \times \mathbf{v} | \phi_i \rangle_\nu. \quad (48)$$

While this definition of the orbital moment poses no difficulties in periodic solids, we note here that the evaluation of the total orbital moment of a periodic crystal is more involved³⁴. In most cases, however, the atomic orbital moments and also the magnetic anisotropy energies, obtained in density functional theory calculations, are too small as compared to experiment. Practical methods that can overcome this deficiency have been discussed in the literature³⁵.

References

1. P. Hohenberg and W. Kohn, *Inhomogeneous Electron Gas*, Phys. Rev. **136**, B864 (1964).
2. W. Kohn and L. J. Sham, *Self-Consistent Equations Including Exchange and Correlation Effects*, Phys. Rev. **140**, A1133 (1965).
3. U. von Barth and L. Hedin, *A local exchange-correlation potential for the spin polarized case: I*, J. Phys. C: Solid State Phys. **5**, 1629 (1972).
4. V. L. Moruzzi, J. F. Janak, and A. R. Williams, *Calculated Electronic Properties of Metals*, (Pergamon, New York, 1978).
5. S. Shallcross, A. E. Kissavos, V. Meded, and A. V. Ruban, *An ab initio effective Hamiltonian for magnetism including longitudinal spin fluctuations*, Phys. Rev. B **72**, 104437 (2005).
6. Ph. Kurz, G. Bihlmayer, and S. Blügel, *Magnetism and electronic structure of hcp Gd and the Gd(0001) surface*, J. Phys.: Condens. Matter **14**, 6353 (2002).
7. H. W. White, B. J. Beaudry, P. Burgardt, S. Legvold, and B. N. Harmon, *Magnetic moments of ferromagnetic gadolinium alloys*, AIP Conf. Proc. **29**, 329 (1975).
8. V. P. Antropov, M. I. Katsnelson, M. van Schilfgaarde, and B. N. Harmon, *Ab Initio Spin Dynamics in Magnets*, Phys. Rev. Lett. **75**, 729 (1995).
9. V. P. Antropov, M. I. Katsnelson, B. N. Harmon, M. van Schilfgaarde, and D. Kusnezov, *Spin dynamics in magnets: Equation of motion and finite temperature effects*, Phys. Rev. B **54**, 1019 (1996).

10. G. M. Stocks, B. Újfalussy, X.-D. Wang, Y. Wang, D. M. C. Nicholson, W. A. Shelton, A. Canning, and B. L. Györfy, *Towards a constrained local moment model for first principles spin dynamics*, Phil. Mag. B **78**, 665 (1998).
11. P. H. Dederichs, S. Blügel, R. Zeller, and H. Akai, *Ground States of Constrained Systems: Application to Cerium Impurities*, Phys. Rev. Lett. **53**, 2512 (1984).
12. Ph. Kurz, F. Förster, L. Nordström, G. Bihlmayer, and S. Blügel, *Ab initio treatment of noncollinear magnets with the full-potential linearized augmented plane wave method*, Phys. Rev. B **69**, 024415 (2004).
13. B. Újfalussy, X.-D. Wang, D. M. C. Nicholson, W. A. Shelton, G. M. Stocks, Y. Wang, and B. L. Györfy, *Constrained density functional theory for first principles spin dynamics*, J. Appl. Phys. **85**, 4824 (1999).
14. P. W. Anderson, *Theory of Magnetic Exchange Interactions: Exchange in Insulators and Semiconductors*, Solid State Physics **14**, 99 (1963).
15. Ph. Kurz, G. Bihlmayer, K. Hirai, and S. Blügel, *Three-Dimensional Spin Structure on a Two-Dimensional Lattice: Mn/Cu(111)*, Phys. Rev. Lett. **86**, 1106 (2001).
16. K. Yosida, *Theory of Magnetism* (Springer, Berlin-Heidelberg, 1996).
17. C. Herring, in *Magnetism*, edited by G. Rado and H. Suhl (Academic, New York, 1966).
18. L. M. Sandratskii, *Energy band structure calculations for crystals with spiral magnetic structure*, Phys. Status Solidi B **136**, 167 (1986).
19. N. Ashcroft and N. Mermin, *Solid State Physics* (Saunders College, Philadelphia, 1976).
20. I. Turek, J. Kudrnovský, M. Diviš, P. Franek, G. Bihlmayer, and S. Blügel, *A first-principle study of electronic structure and exchange interactions in bcc europium*, Phys. Rev. B **68**, 224431 (2003).
21. N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, *Equation of State Calculations by Fast Computing Machines*, J. Chem. Phys. **21**, 1087 (1953).
22. M. Ležaić, Ph. Mavropoulos, and S. Blügel, unpublished result (2005).
23. M. Pajda, J. Kudrnovský, I. Turek, V. Drchal, and P. Bruno, *Ab initio calculations of exchange interactions, spin-wave stiffness constants, and Curie temperatures of Fe, Co, and Ni*, Phys. Rev. B **64**, 174402 (2001).
24. N. M. Rosengaard and Börje Johansson, *Finite-temperature study of itinerant ferromagnetism in Fe, Co, and Ni*, Phys. Rev. B **55**, 14975 (1997).
25. I. Turek, J. Kudrnovský, G. Bihlmayer, and S. Blügel, *Ab initio theory of exchange interactions and the Curie temperature of bulk Gd*, J. Phys.: Condens. Matter **15**, 2771 (2003).
26. E. Şaşıoğlu, L. M. Sandratskii, and P. Bruno, *Above-room-temperature ferromagnetism in half-metallic Heusler compounds NiCrP, NiCrSe, NiCrTe, and NiVAs: A first-principles study* J. Appl. Phys. **98**, 063523 (2005).
27. K. Sato, P. H. Dederichs, and H. Katayama-Yoshida, *Is High T_C Possible in (Ga,Mn)N?: Monte Carlo Simulation vs. Mean Field Approximation* J. Supercond. **18**, 33 (2005).
28. S. Blügel, *Magnetische Anisotropie und Magnetostriktion (Theorie)* in Magnetische Schichtsysteme, Publications by Research Centre Jülich: Matter and materials, Vol.2 (1999).

29. J. D. Jackson, *Classical electrodynamics*, (Wiley & Sons, 1962).
30. H. A. Bethe and E. E. Salpeter *Quantum Mechanics of One- and Two-Electron Systems* (New York, Plenum 1977).
31. J. Stöhr, *Exploring the microscopic origin of magnetic anisotropies with X-ray magnetic circular dichroism (XMCD) spectroscopy*, J. Magn. Magn. Mater. **200**, 470 (1999).
32. N. D. Mermin and H. Wagner, *Absence of Ferromagnetism or Antiferromagnetism in One- or Two-Dimensional Isotropic Heisenberg Models*, Phys. Rev. Lett. **17**, 1133 (1966).
33. M. Bander and D. L. Mills, *Ferromagnetism of ultrathin films*, Phys. Rev. B **38**, 12015 (1988).
34. T. Thonhauser, D. Ceresoli, D. Vanderbilt, and R. Resta, *Orbital Magnetization in Periodic Insulators*, Phys. Rev. Lett. **95**, 137205 (2005).
35. I. Yang, S. Y. Savrasov, and G. Kotliar, *Importance of Correlation Effects on Magnetic Anisotropy in Fe and Ni*, Phys. Rev. Lett. **87**, 216405 (2001).

Ab Initio Description of Electronic Transport

Daniel Wortmann

Institute of Solid State Research
Forschungszentrum Jülich
52425 Jülich, Germany
E-mail: d.wortmann@fz-juelich.de

Ab initio calculations of electronic transport became an important tool for the interpretation of quantum transport experiments on the atomic scale. Different theoretical concepts used in these kind of calculations are presented. The Landauer equation is introduced, its basic assumptions and the corresponding limits of applicability are discussed and compared to a calculation of the tunneling current based on perturbation theory. We also give a glimpse on the techniques used to treat open systems and systems out of equilibrium by means of the Green function approach.

1 Introduction

Calculating the electronic current flowing in a system due to some bias voltage is one of the most difficult problems in theoretical solid state physics. Different levels of approximations have been applied to the problem and theoretical models for many different aspects of the problem have been discussed. Among these, one of the most well known methods to treat electronic transport in solids is based on the Boltzmann formalism, which describes the time-dependent change in the electronic distribution function due to the applied field $\vec{E}$. It is a classical approximation as it considers the electrons to be moving freely between individual scattering events. Interference effects due to the quantum nature of the electrons are neglected. On the other hand, the very fact that the electrons move in a solid without scattering at every atomic site, i.e. the possibility to describe the electrons as (quasi) particles moving freely around is of course a quantum mechanical phenomenon. Very often, the Boltzmann equation is used in the so-called relaxation-time approximation, assuming a simple linear dependence of the scattering rate of the conducting electrons on the change Δf of the distribution function f that can be expressed in terms of a characteristic scattering time τ or alternatively in terms of a characteristic average length $L = \tau v$ between two scattering events. In this approximation, the Boltzmann formula takes the form

$$\frac{\partial}{\partial t} f + \vec{v} \frac{\partial}{\partial r} f - e \vec{E} \frac{1}{\hbar} \frac{\partial}{\partial k} f = -\frac{1}{\tau} \Delta f . \quad (1)$$

The average length between two such scattering events, the so-called mean-free path is the basic input parameter in the Boltzmann formula (Figure 1). One can derive approximations for different contributions to such scattering events and e.g. find the low-temperature dependence of the resistance of simple metals due to electron-phonon scattering as given in the famous Bloch T^5 law.

The approximations used in the Boltzmann approach restrict its applicability to the case in which the sample dimensions are much larger than the mean-free path and at the same time the scattering events can be viewed as independent from each other, i.e. quantum interference effects can be neglected. Such quantum interference can lead to a drastic variation of the resistance of a sample in an effect known as Anderson localization.

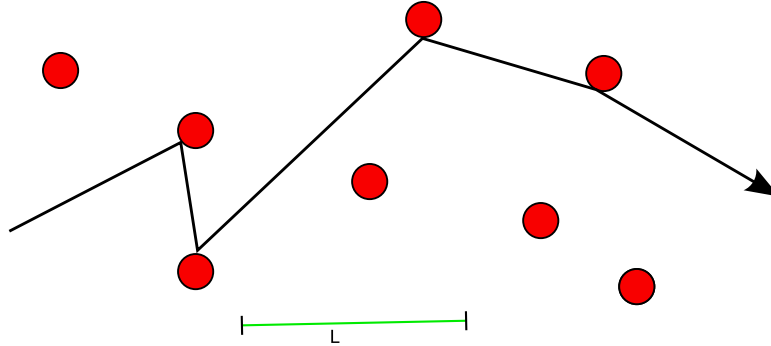


Figure 1. Scattering of electrons traveling in a solid. The average distance between such scattering events is the mean-free path L .

A more basic approach to the problem of electric conductance uses the idea of treating the current as a response of the quantum system to the applied electric field. In its most rigorous formulation this idea can be used in the linear response limit in which one assumes that the current depends linearly on the field. Using the corresponding time-dependent quantum mechanical formulation one obtains the so-called Kubo formalism which can be used to obtain the conductance of the system. While this approach is very general and would allow to include all kinds of scattering, it is not easy to apply it to realistic systems.

In the following, we will concentrate on calculating the electronic transport from a quantum mechanical point of view. In particular, we will consider systems with typical dimensions much smaller than the mean-free path due to scattering at structural impurities, by interface roughness, by phonons, magnons or other temperature dependent excitations present in real experiments. In the nanosize setups we consider, the resistance is due to the scattering of the electrons on the potential change in some scattering volume. A typ-

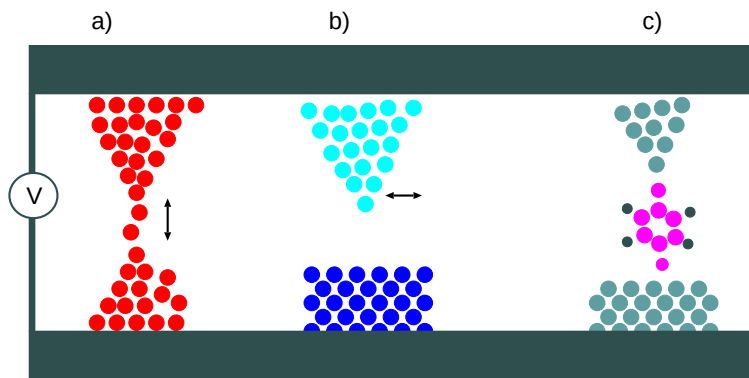


Figure 2. Different experimental setups used in nanoscale transport experiments. a) Metallic nanowires which might have different length, b) Scanning tunneling microscopy with a tip scanning the surface, c) Transport through single molecules.

ical experimental setup of such a system would be a break-junction with an atomic wire suspended between the two electrodes or a tunneling setup like a scanning tunneling microscope (STM) (Figure 2a,b)). The same approximation is also frequently applied to electron transport through single molecules (Figure 2c)) even though one could expect that scattering due to molecular vibrations will be important here.

This lecture will cover two main subjects. First, we will discuss the basic models to describe electron transport in a single-particle picture. This part will cover the famous Landauer approach to ballistic transport (Section 2.1) as well as a more specialized approach to the tunneling problem – Bardeen’s approach based on perturbation theory (Section 2.3). After introducing these models, we will focus on the specific problems one faces when trying to apply these ideas using the density functional theory. In particular we will discuss how to deal with semi-infinite leads and open systems without periodicity (Section 3). Our discussion will close with a short summary of the basic ideas applied to treat a non-equilibrium situation with a finite applied bias voltage to a transport junction (Section 3.1).

2 Single-Particle View on Quantum Transport

Even when restricting the theoretical description to the electron system only, the transport process is actually a complicated many electron problem of a system in non-equilibrium. On the macroscopic scale one can already define some of the different quantities describing the system like current and charge density or the applied electric field. However, it is very difficult to track these quantities down to the microscopic scale due to the complicated thermodynamically averaging taking place. In the following we will only deal with the very restricted subset of phenomena that are due to quantum mechanical nature of the electrons and we will therefore consider systems which can be described by pure wavefunctions without any statistical averaging. Still, in this picture one would have to describe the electron transport by the time-dependent many-body wavefunction $\Psi(t)$. For example, one could consider the probability $P_{i,f}$ of the system changing its state from some initial multi-electron state Ψ_i into a final multi-electron state Ψ_f where the two states differ with respect to their charge distribution. Thus, this approach makes it necessary to calculate the many particle time-dependent wavefunctions of the entire system. This is a very difficult task which cannot be solved in general.

To overcome this fundamental obstacle we will switch to the single-electron picture of electron transport. Similar to the replacement of the many particle problem of determining the ground state of a many-body quantum mechanical system in density functional theory by a single-particle Kohn-Sham formalism we will treat the electronic transport as due to the transport of many independent single-electrons. Furthermore, we will assume that the single-particle states in the Kohn-Sham formulation actually describe these independent single-electrons. Of course, there can be no hope that this very simplistic model actually is able to catch all the essential physics of the transport process. However, in analogy to standard band structure calculations of solids in equilibrium in which in many cases the Kohn-Sham single particle eigenvalues and eigenstates can be successfully interpreted as the elementary excitations of the systems we will apply the same procedure to the electronic transport and assume that the effects of the atomic arrangement, of the electronic (self-consistent) charge density and single-particle potential on the current can actually be modeled by this approach.

Many effects restrict the validity of the single-electron approach. Most obvious might be electron-electron scattering effects of different conducting electrons, but also interactions with the lattice beyond the static approximation, i.e. electron-phonon scattering, screening and charging effects or many particle interactions in magnetic systems might limit the validity of the single-electron picture. Only if these processes are sufficiently weak, one can hope that the single-particle approximations in terms of the Kohn-Sham states will provide reasonable results. This corresponds to the limit discussed in the introduction in which the mean-free path is much larger than the system size. In the single-particle picture the description in terms of the Kohn-Sham wavefunction will hold only on length scales shorter than this length scale, since these processes not included in the model will lead to the scattering destroying the phase coherence between the single-electron states involved.

One should note, that in many cases this description of the electron current in terms of single-particle physics is a completely inappropriate point of view. As soon as quantum many-body effects come into play, qualitatively different phenomena can be observed. Examples of such effects are the Kondo-effect in which a two level quantum scatterer embedded in a metallic environment leads to conductance abnormalities at low temperatures or correlation effects like the Coulomb blockade which are not reproduced in standard density functional theory treatments.

Obviously our description does not include any processes by which the single particle energy of the states carrying the current is changed. Hence only elastic transport can be described. This kind of quantum transport with only elastic scattering included is frequently called ballistic transport in mesoscopic physics as in many aspects the electrons behave like classical particles moving in a “billiard” like fashion. However, one should be aware of underlying quantum mechanical picture of single-electron states by which the electrons are described. These single-particle states describe the movement of the electrons between scattering events which scatter them from one single particle state into another.

2.1 The Landauer Approach

Landauer¹ proposed a theory of the transport process which is well adapted to describe the ballistic transport. A very intuitive and simple derivation will be presented here. The Landauer equation can also be derived more rigorously starting from linear response theory. In the Landauer approach to transport one considers the region Ω in which the electrons travel ballistically to be attached to two reservoirs L and R (see Figure 3).

The conductance Γ of the region Ω is defined by the current I_{LR} divided by the potential difference between the two reservoirs. The current I_{LR} on the other hand is given by the current due to all electrons traveling from L to R minus the current due to the electrons traveling vice versa

$$I_{LR} = I_{L \rightarrow R} - I_{R \rightarrow L} . \quad (2)$$

To arrive at an equation for these currents, one can start with a simple one-dimensional model. The current from the left to the right is determined by all electrons leaving the left reservoir, entering the scattering region Ω , and leave this scattering region by passing into the right reservoir. If one now assumes a very simple picture of the region Ω in which its electronic structure is described by single band in which states with $k > 0$ propagate from

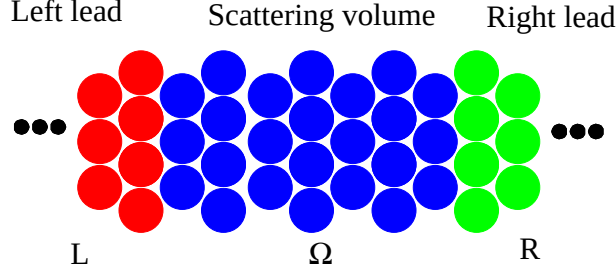


Figure 3. Basic setup of the Landauer approach to transport. The leads act as reservoirs and are assumed to be in thermal equilibrium.

the left to the right the current is given by an integral over all states with $k > 0$ up to the Fermi wave-vector k_F

$$I_{L \rightarrow R} = \int_0^{k_F} e v(k) dk, \quad (3)$$

where v denotes the group velocity of the state. Since

$$v = \frac{1}{\hbar} \frac{\partial E}{\partial k} \quad (4)$$

and converting the integral over k into an energy integration using the density of states $n(E)$,

$$\begin{aligned} I_{L \rightarrow R} &= \int_0^{\mu_L} \frac{e}{\hbar} \frac{\partial E}{\partial k} n(E) dE \\ &= \int_0^{\mu_L} \frac{e}{\hbar} \frac{\partial E}{\partial k} \frac{1}{\partial E / \partial k} \frac{1}{2\pi} dE \\ &= \int_0^{\mu_L} \frac{e}{\hbar} dE = \frac{e}{\hbar} \mu_L, \end{aligned} \quad (5)$$

where the energy integration has to be performed over all energies up to the Fermi energy (the chemical potential) of the left reservoir. This can be understood from the requirement that the electrons were assumed to be incoming from the left and therefore must be occupied in the reservoir.

Using the same derivation for the states incoming from the right reservoir one obtains

$$I_{LR} = \frac{e}{\hbar} (\mu_L - \mu_R). \quad (6)$$

Identifying the difference in the chemical potentials μ_L and μ_R with the applied voltage $eV = (\mu_L - \mu_R)$ one obtains the following interesting equation for the conductance

$$\Gamma = \frac{I_{LR}}{V} = \frac{e^2}{\hbar}. \quad (7)$$

This equation is truly remarkable since it states that each conducting band contributes the same to the conductance. Irrespectively of the density of states or the group velocity of the conducting states the conductance is always given by the fundamental quantum of conductance $\frac{e^2}{h}$. Indeed, as Eq. (5) shows, states with a low velocity and therefore a low current $j = ev$ are compensated by their higher density of states such that the conductance remains constant.

In the case of multiple bands, the derivation has to be modified by the inclusion of an extra sum over the different bands. Therefore in the general case of N conducting bands one obtains

$$\Gamma = \frac{e^2}{h} N . \quad (8)$$

The different „bands“ in this discussions are usually called „channels“. The argumentation presented so far did not care about the proper definition of these channels. These were simply assumed to form some kind of „band“ within Ω described by the usual formalism of a wave-vector k and a dispersion relation $E(k)$. Strictly speaking, since the system is not periodic, one cannot speak of Bloch states with some wave-vector having a component k in the direction of the current.

Since a key point in the discussion was the preparation of a state traveling from within the reservoirs through the region Ω one should clarify this idea. For such a state traveling to the right, one might assume the typical scattering problem. Within the left reservoir one considers a wavefunction being a Bloch state propagating towards the region Ω . „Propagate towards“ in this context should be understood as a state having a current flowing towards Ω . Within the reservoirs the resulting scattering state can be written in terms of reflected ψ_r and transmitted ψ_t states which are all solutions of the bulk Schrödinger equation in the reservoirs with the same energy as the incoming state ψ_{in} . The k values of these transmitted and reflected states have to be chosen such that the states „propagate away“ from Ω .

$$\psi(\vec{r}) = \begin{cases} \psi_{in}(\vec{r}) + \sum_n r_{in,n} \psi_r^n(\vec{r}) & \vec{r} \text{ in left reservoir} \\ \sum_{n'} t_{in,n'} \psi_t^{n'}(\vec{r}) & \vec{r} \text{ in right reservoir} . \end{cases} \quad (9)$$

Here the summations can be considered to be performed over all reflected Bloch states or all transmitted Bloch states. In principle, also states decaying away from the interfaces into the reservoirs must be included in this expansion. However, since these do not carry any current and by shifting the interface far enough into the reservoirs one can eliminate these decaying states.

Looking back to the derivation of the Landauer formula, an important change has to be made. While in Eq. (3) and Eq. (4) the summation over the incoming states and the evaluation of the current from their group velocities were all performed within the same single band picture, now one has to distinguish more carefully. The k integration in Eq. (3) has to be performed over the „in“ label of the expansion in Eq. (9). The sum over the velocities on the other hand is best performed in the right electrode. This is possible since current is conserved and can be very easily be done if all transmitted states and the incoming state are normalized to carry unit current. Using the orthogonality of the Bloch states one can perform the same steps as in Eq. (2) to (6) again to derive the more general Landauer equation

for ballistic transport in the presence of some scattering of the incoming electrons,

$$\Gamma = \frac{e^2}{h} \sum |t_{ij}|^2, \quad (10)$$

where i, j label the Bloch states in the reservoirs traveling from the left to the right.

Eq. (10) allows a simple interpretation of the transport in terms of the underlying quantum mechanical property of the transmission probability $P_{ij} = |t_{ij}|^2$ of an electron from the incoming Bloch state i into the transmitted Bloch state j . This interpretation makes the requirement of normalizing the incoming and transmitted Bloch states to unit current very clear, since in this normalization the direct interpretation of this probability is reasonably well defined and Eq. (10) can be seen as a simple generalization of Eq. (8).

2.2 Interpretation of the Landauer Formula

The Landauer formula (Eq. (10)) was the source of some confusion for quite some time after its first formulation^{2,3}. The most striking feature of the equation might be its limit for a perfectly transmitting region, i.e. for a region with $P_{ij} = |t_{ij}|^2 = 1$ for some set of i, j . For example if one would consider a perfect bulk crystal sandwiched between reservoirs of the same bulk material the expansion of Eq. (9) would collapse to

$$\psi(\vec{r}) = \begin{cases} \psi_{\text{in}}(\vec{r}) + \sum 0\psi_r & \vec{r} \text{ in left reservoir} \\ 1 \psi_t = \psi_{\text{in}}(\vec{r}) & \vec{r} \text{ in right reservoir} \end{cases}, \quad (11)$$

and one would rediscover Eq. (8) with N denoting the number of incoming Bloch states. At first glance, this means that the Landauer equation predicts a limited conductance of a system without any de-coherent scattering, i.e. of a perfect bulk crystal. In the same way the Landauer equation would also give a finite conductivity of a free electron gas. In this case a question which can always be asked only becomes more obvious to ask: How can a region with ballistic transport, i.e. without any dissipative processes, have a finite conductance? Since there is a voltage drop over the region and a current is flowing, some energy must dissipate. The key to the answer to this question lies in the definition of the reservoirs which were assumed to be in thermal equilibrium with some chemical potential μ attached to them. This is only possible, if there are actually dissipative processes in the reservoirs leading to the „thermalization“ of the „hot“ electrons being transferred across the region of ballistic transport.

The surprising result of a finite conductance in the case of a perfect crystal can now be interpreted in different ways. Either the setup described was not correct, since the reservoirs could not remain in thermal equilibrium and being perfect crystals like the region of ballistic transport at the same time or, which is actually very much the same, no finite voltage can be applied across such a system. The finite conductance of such a system with perfect ballistic transmission can now be interpreted as due the finite resistance at the interface between the reservoir and the ballistic region. This is also called the Sharvin-resistance of the system.

Another point to mention in the discussion of the physical significance of the Landauer equation is its formulation in terms of a two-terminal device. Both the current and the voltage drop are defined between the same two reservoirs. In many experiments, especially in mesoscopic physics a four point measurement is performed in which the current is driven between electrodes different than those between the voltage drop is measured.

Büttiker² presented a generalization of the Landauer equation to these multi-terminal cases. While this approach is very appropriate for mesoscopic physics, on the atomic scale multi-terminal arrangements are not the typical experimental arrangements and thus Eq. (10) will be sufficient. As depicted in Figure 4, additional resistances present in the current circuit are frequently eliminated in a four-point measurement, in which two additional potential probes are attached close to the scattering volume. However, for scattering volumes on the atomic scale, these geometries are not appropriate and thus we will restrict ourselves to simple two point geometries.

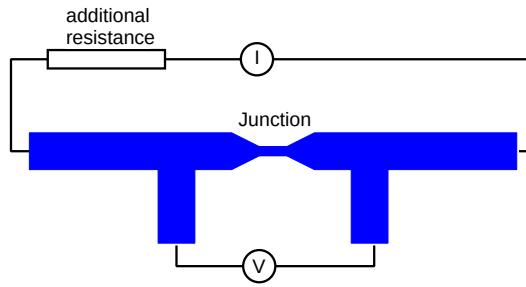


Figure 4. Four point geometry used experimentally to determine the voltage drop across a junction without including additional resistances present in the current circuit.

We will come back to this point when discussing the non-equilibrium transport later in this manuscript.

While the Landauer equation is valid in many cases reaching from systems with high conductivity to systems in the tunneling regime, one has to be careful in its application in some cases. Only states in which the incoming and transmitted waves can be described by Bloch states contribute to the tunneling current. This excludes any state which is localized within the region of ballistic transport to contribute. This corresponds to the fact that these states do not carry any current within the simple one-electron picture of transport chosen. In reality, there exist processes beyond this picture which lead to some coupling of these localized states to the otherwise orthogonal Bloch states in the reservoirs. For example the many-body electron-electron interaction, electron-phonon scattering, or structural defects not included in the description can provide such a coupling. Thus, while the Landauer approach will be correct for cases of high transmission through Bloch states, one could imagine that in the limit of a very low transmission probability other processes of transport across the ballistic region become important. In the one-electron picture these processes could be thought of as the transition of an electron from the reservoirs into some localized state of the reservoir, the transition of the electron from one side of the reservoir to the other and then the transition of the electron into a state of the other reservoir. The validity of the Landauer model is now limited by the transmission probability between the reservoir states and the localized state. If this probability becomes comparable to the probabilities $P_{ij} = |t_{ij}|^2$ the Landauer equation breaks down.

On the other hand, one can of course treat the other limit in which the transition probability between the two sides of the reservoir becomes very small and the details of the scattering processes needed to couple the states can be neglected. This limit can be suc-

cessfully described by theories for the quantum mechanical tunneling process.

2.3 The Bardeen Approach to Tunneling

The following description of the tunneling process is based on Bardeen's approach to tunneling which essentially applies time-dependent perturbation theory to the problem. Figure 5 shows the tunneling setup used in this approach. Two semi-infinite crystals are separated by a barrier region, which will be assumed to be a vacuum barrier for simplicity. If this vacuum barrier is sufficiently high and wide one can think the total setup to consist of two independent systems: one at the left (L) and one at the right (R) side.

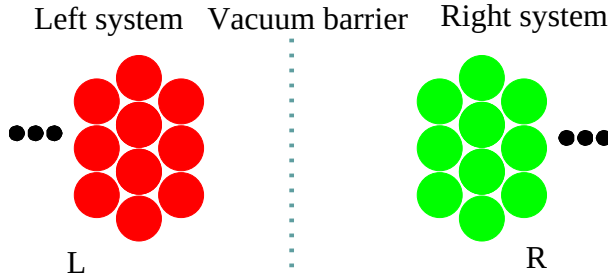


Figure 5. Tunneling setup used in Bardeen's approach to transport. The two semi-infinite crystals at the left and the right are separated by a vacuum barrier.

This total separation of the systems leads to two independent Schrödinger equations for the two sides

$$\begin{aligned}(T + U_L)\psi_L &= \epsilon_L \psi_L \\ (T + U_R)\psi_R &= \epsilon_R \psi_R\end{aligned}\tag{12}$$

where T denotes the operator of the kinetic energy of a single electron and U_L and U_R are the potentials of the left and right system respectively. The single-particle wavefunction $\psi(t)$ of the entire setup is determined by the total Hamiltonian $H = T + U_L + U_R$.

Now one can apply time-dependent perturbation theory to describe the tunneling of an electron across the vacuum barrier. Tunneling from the left to the right is assumed, the case of an electron tunneling vice versa may be treated completely analogously. The initial state of the tunneling process is localized in the left system. Therefore, there exists an eigenstate ψ_L^μ with $|\psi(t \rightarrow -\infty)\rangle = |\psi_L^\mu\rangle$. The time-dependence of the state $|\Psi(t)\rangle$ is governed by the Hamiltonian of the whole system.

$$i\hbar \frac{\partial}{\partial t} |\psi(t)\rangle = H |\psi(t)\rangle .\tag{13}$$

The tunneling probability is given by the overlap of this time-dependent wavefunction with a wavefunction $|\psi_R^\nu\rangle$ of the right system. Multiplying Eq. (13) from the left with $\langle\psi_R^\nu|$ leads to

$$\langle\psi_R^\nu| \left(i\hbar \frac{\partial}{\partial t} \right) |\psi(t)\rangle = \langle\psi_R^\nu| H |\psi(t)\rangle .\tag{14}$$

Using the Schrödinger equation for the left state one obtains

$$i\hbar \frac{\partial}{\partial t} \langle \psi_R^\nu | \psi \rangle = \langle \psi_R^\nu | H | \psi \rangle - \langle \psi_R^\nu | H_R | \psi \rangle = \langle \psi_R^\nu | U_L | \psi \rangle \quad (15)$$

Substituting $|\psi(t \rightarrow -\infty)\rangle = |\psi_L^\mu\rangle$ for $|\psi\rangle$ at the right hand side of Eq. (15) leads to first order perturbation theory

$$i\hbar \frac{\partial}{\partial t} \langle \psi_R^\nu | \psi \rangle = \langle \psi_R^\nu | U_L | \psi_L^\mu \rangle. \quad (16)$$

Even though this equation looks familiar one has to emphasize that this is not a result obtained by standard time-dependent perturbation theory. The states $|\psi_L\rangle$ and $|\psi_R\rangle$ are eigenstates of the Hamiltonians H_L and H_R respectively. Therefore, they do not form a complete orthogonal basis of the eigenspace of the total Hamiltonian $H = T + U_L + U_R$ and the matrix elements at the left side of Eq. (16) are not sufficient to determine the total time-dependence of $|\psi\rangle$. This is a basic weakness of Bardeen's approach. However many applications⁴⁻⁷ of this formalism have shown that Bardeen's approximation produces reliable results for systems which are well separated, i.e. systems where the overlap of the two wavefunctions ψ_R and ψ_L is small.

Since the potential U_L is not small in the left region, the question arises whether one is allowed to use perturbation theory at all. However, it can be seen from Eq. (16) that the quantity which in fact determines the strength of the perturbation of the initial state is $\langle \psi_R^\nu | U_L | \psi_L^\mu \rangle$. Since the final wavefunction $|\psi_R\rangle$ is localized in the right region in which the left potential U_L is very weak this perturbation might still be regarded as a small perturbation and thus time-dependent perturbation will lead to reasonable results.

By separating the time-dependence of the states $|\psi_L^\mu\rangle = e^{i\epsilon_\mu t} |\Psi_L^\mu\rangle$ and $|\psi_R^\nu\rangle = e^{i\epsilon_\nu t} |\Psi_R^\nu\rangle$, integrating Eq. (16) and performing the limit $t \rightarrow \infty$, one obtains an expression for the tunneling probability per time interval

$$P_{\mu\nu}^{LR} = \lim_{t \rightarrow \infty} \frac{1}{t} \frac{1}{\hbar^2} \int_0^t |\langle \psi_R^\nu | U_L | \psi_L^\mu \rangle|^2 \quad (17)$$

$$= \lim_{t \rightarrow \infty} \frac{4}{\hbar} \frac{\sin^2\left(\frac{\epsilon_\nu - \epsilon_\mu}{2\hbar} t\right)}{(\epsilon_\nu - \epsilon_\mu)^2 t} |M_{\mu\nu}^{RL}|^2, \quad (18)$$

where the matrix element $M_{\mu\nu}^{LR}$ is given by the stationary-state matrix element of the potential

$$M_{\mu\nu}^{LR} = \langle \psi_R^\nu | U_L | \psi_L^\mu \rangle. \quad (19)$$

Assuming a continuous range of energy levels ϵ_μ (or ϵ_ν) the limit of Eq. (18) can be evaluated directly⁸. One obtains

$$P_{\mu\nu}^{LR} = \frac{2\pi}{\hbar} \delta(\epsilon_\nu - \epsilon_\mu) |M_{\mu\nu}^{LR}|^2. \quad (20)$$

This result is similar to the well known 'Golden Rule' Fermi obtained for standard time-dependent perturbation theory. It describes elastic tunneling with energy $\epsilon_\nu = \epsilon_\mu$ only. Formally this condition is taken care of by the δ -function in Eq. (20).

To evaluate this matrix element one can introduce an additional approximation. He assumed the potential U_L to be zero in the right region of space. Similar the right potential

should be zero in the left region. More formal one assumes a separation surface S which separates the regions in which the two potentials differ from zero. This can be written down by the condition $U_L U_R = 0$ for any point in space. Figure 6 shows the setup as used in this additional approximation.

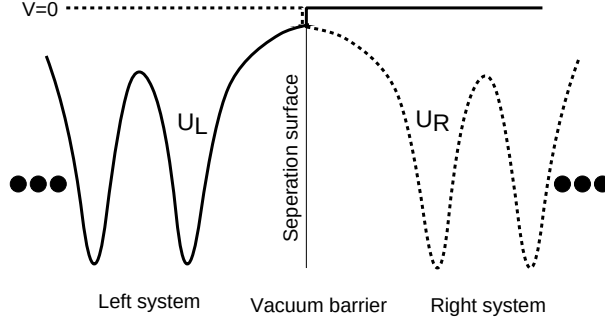


Figure 6. Potential used in the Bardeen approach to tunneling. The left (right) potential U_L (U_R) is then assumed to be zero in the right (left) region.

Of course, this approximation will become better if the potentials U_L and U_R are reasonably small at and beyond the separation surface. This will be the case if the separation surface is located far out in the vacuum.

Using the Schrödinger equation for the left wavefunction and having in mind that the potential U_L is zero in the right space one can now rewrite the matrix element as an integral over the left region only

$$M_{\mu\nu}^{LR} = \int_L \Psi_R^\nu(\vec{r})^* \left(\epsilon_\mu + \frac{\hbar^2}{2m} \vec{\nabla}^2 \right) \Psi_L^\mu(\vec{r}) dV \quad (21)$$

which can be written in a more symmetric form

$$\begin{aligned} M_{\mu\nu}^{RL} &= \int_L \left\{ \Psi_R^\nu(\vec{r})^* \epsilon_\nu \Psi_L^\mu(\vec{r}) + \Psi_R^\nu(\vec{r})^* \frac{\hbar^2}{2m} \vec{\nabla}^2 \Psi_L^\mu(\vec{r}) \right\} dV \\ &= \int_L \left\{ \Psi_R^\nu(\vec{r})^* \overleftarrow{(T + U_R)} \Psi_L^\mu(\vec{r}) + \Psi_R^\nu(\vec{r})^* \frac{\hbar^2}{2m} \vec{\nabla}^2 \Psi_L^\mu(\vec{r}) \right\} dV \\ &= -\frac{\hbar^2}{2m} \int_L \left\{ \Psi_L^\mu(\vec{r}) \vec{\nabla}^2 \Psi_R^\nu(\vec{r})^* - \Psi_R^\nu(\vec{r})^* \vec{\nabla}^2 \Psi_L^\mu(\vec{r}) \right\} dV \end{aligned} \quad (22)$$

In these transformations in the first step the eigenvalue ϵ_μ was substituted by ϵ_ν because energy conservation requires the calculation of matrix elements with $\epsilon_\mu = \epsilon_\nu$ only. In the second step the Schrödinger equation for the right state was used (the arrow indicates the wavefunction the operators acts on). The integration area is the left region. Since the potential U_R is assumed to be zero in this region, it was dropped in the last step. Using Greens theorem and the boundary condition that the right wavefunction is zero at infinite distance from the separation surface this integral can be transformed into an integral over

the separation surface

$$M_{\mu\nu}^{LR} = -\frac{\hbar^2}{2m} \int_S \left(\Psi_L^\mu(\vec{r}) \vec{\nabla} \Psi_R^\nu(\vec{r})^* - \Psi_R^\nu(\vec{r})^* \vec{\nabla} \Psi_L^\mu(\vec{r}) \right) dS. \quad (23)$$

So far only an expression for the probability of the transition of an electron from a left state into a right state was obtained.

Slightly modifying Eq. (20) this probability can be written as

$$P_{\mu\nu}^{LR} = \frac{2\pi}{\hbar} \delta(\epsilon_\mu^L - \epsilon_\nu^R - eV) |M_{\mu\nu}^{LR}|^2, \quad (24)$$

where the additional term eV is introduced to account for the bias voltage V applied between the two sides. To calculate the tunneling current one has to sum over all different possible left and right states and one has to keep in mind that the electrons might tunnel from the left to the right as well as vice versa. The total current therefore is given by

$$\begin{aligned} I &= I^{L \rightarrow R} - I^{R \rightarrow L} \\ &= e \sum_{\mu\nu} f(\epsilon_\mu)(1 - f(\epsilon_\nu + eV)) P_{\mu\nu}^{LR} - e \sum_{\mu\nu} (1 - f(\epsilon_\mu)) f(\epsilon_\nu + eV) P_{\nu\mu}^{RL} \\ &= e \sum_{\mu\nu} (f(\epsilon_\mu) - f(\epsilon_\nu + eV)) P_{\mu\nu}^{LR} \end{aligned} \quad (25)$$

where $f(\epsilon)$ denotes the Fermi-distribution function which is introduced to ensure that only tunneling from occupied to unoccupied states can occur. In Eq. (25) the symmetry of the tunneling probability $P_{\mu\nu}^{LR} = P_{\nu\mu}^{RL}$ which can easily be deduced from Eq. (23) was used. The sum in Eq. (25) has to be performed over all right states labeled by ν and all left states labeled by μ . No further assumption is made on the nature of these left and right states, i.e. both Bloch states and surface states decaying into the bulk contribute to the current and therefor this formula differs significantly from the Landauer formula.

2.4 Landauer Conductance Versus Bardeen's Tunneling

One might wonder which the difference will be between the results of the Bardeen formula of tunneling and of the Landauer approach. Of course, as stressed in Section 2.3 the results will differ significantly as soon as localized states are present in the vicinity of the barrier region. However, it remains to be clarified what should be expected in the absence of these states. Before the difference between these two transport formulas will be investigated for more realistic systems, it is instructive to look at a very simple model that one can treat analytically.

Figure 7 shows the setup chosen for this simple model, a rectangular barrier of height V_0 and width d between leads in which the electrons are described by free electron wavefunctions. In the Landauer picture of transport, the wavefunction is constructed as

$$\psi(x) = \begin{cases} \frac{1}{k} (\exp(ikx) + r \exp(-ikx)) & x < 0 \text{ in left region} \\ a \exp(-\kappa x) + b \exp(\kappa x) & 0 < x < d \text{ in the barrier} \\ \frac{1}{k} t \exp(ikx) & x > d \text{ in right region.} \end{cases} \quad (26)$$

The decay constant is given by $\kappa = \sqrt{2V_0 - k^2}$, the coefficients a, b and r, t can be determined by wavefunction matching, i.e. by the requirement that the wavefunction and its

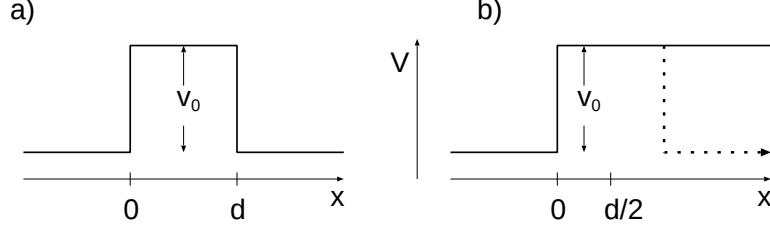


Figure 7. a) Setup of a one-dimensional rectangular barrier used for the analytic evaluation of the Landauer equation. b) Corresponding system for the Bardeen formula. The two sides are separated by extending the barrier to infinity. Only the left system of Bardeen's setup is shown, the right is constructed mirrored.

derivative are continuous at $x = 0$ and $x = d$. The normalization factor $\frac{1}{k}$ has been chosen such that the Bloch states carry unit current as required. Simple algebra reveals the well known formula

$$t = \frac{4i\kappa k e^{-ikd}}{(ik + \kappa)^2 e^{-\kappa d} + (k + i\kappa)^2 e^{\kappa d}}. \quad (27)$$

Inserting this expression into Eq. (10) one obtains the tunneling conductance of this simple barrier in the Landauer approach

$$\begin{aligned} \Gamma_L &= \frac{e^2}{h} \frac{(4\kappa k)^2}{(k^2 + \kappa^2)^2 (1 - e^{-2\kappa d})^2 + (4\kappa k)^2 e^{-2\kappa d}} e^{-2\kappa d} \\ &= \frac{e^2}{h} \left| \frac{4\kappa k}{k^2 + \kappa^2} \right|^2 e^{-2\kappa d} + O(e^{-4\kappa d}). \end{aligned} \quad (28)$$

The second term in this expressions denotes contributions of order $e^{-4\kappa d}$ which can be neglected for any sufficiently thick barrier. This accounts to neglecting the multiple scattering processes in which the electron is scattered multiple times at the interfaces.

If one wants to treat the same system using Bardeen's formula, one has to separate the two systems by extending the barrier to infinity as indicated in Figure 7b). The wavefunctions of the two systems are of course equal except for the transformation $x \leftrightarrow d - x$ and can be written, e.g. for the left side, as

$$\psi_L = \begin{cases} \exp(ikx) + \frac{ik + \kappa}{ik - \kappa} \exp(-ikx) & x \text{ in the leads} \\ \frac{2ik}{ik - \kappa} \exp(-\kappa x) & x \text{ in the barrier.} \end{cases} \quad (29)$$

This formula does not depend on the barrier thickness as it describes completely decoupled systems. The conductance in Bardeen's approach is now given by inserting a separation surface at $x = d/2$ and to evaluate the transition probabilities of Eq. (24) at energy ϵ_0 to obtain

$$\begin{aligned} P_{\text{Bardeen}} &= \frac{\pi \hbar^3}{2m^2} \left| 2 \frac{2ik}{ik - \kappa} e^{-\kappa d/2} \kappa \frac{2ik}{ik - \kappa} e^{-\kappa d/2} \right|^2 \delta(\epsilon - \epsilon_0) \\ &= \frac{\pi \hbar^3}{2m^2} k^2 \left| \frac{4\kappa k}{k^2 + \kappa^2} \right|^2 e^{-2\kappa d} \delta(\epsilon - \epsilon_0). \end{aligned} \quad (30)$$

Finally the summation over all states gives according to Eq. (25)

$$\Gamma_B = \frac{e^2}{h} \left| \frac{4\kappa k}{k^2 + \kappa^2} \right|^2 e^{-2\kappa d}. \quad (31)$$

Hence, both Eqs. (28) and (31) reveal the same conductance in the limit of large κd , i.e. for a small conductance. Figure 8 illustrates this equivalence for a set of parameters typical for an STM setup. It is seen that for barriers thicker than ~ 1 Å Eqs. (31) and (28) give practically identical results.

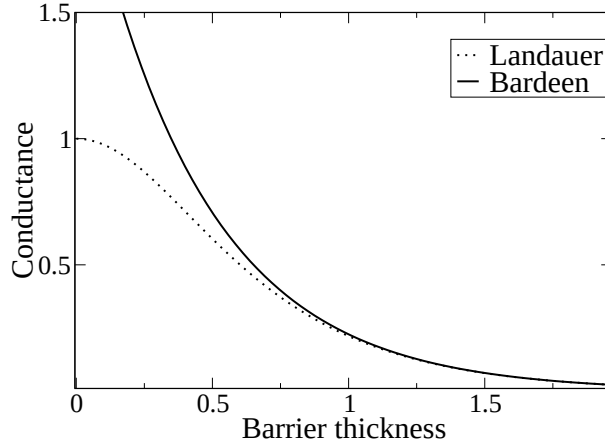


Figure 8. Conductance (in units of the conductance quantum e^2/h) through a rectangular barrier of 5eV barrier height between free electron leads ($\epsilon = 1$ eV) as a function of the barrier thickness (in Å).

2.5 Cu-Vacuum-Cu Tunneling

To demonstrate the two formulas for a more realistic, but still simple case we investigated the electron transmission with normal incidence ($k_{\parallel} = 0$) through a 8.3 Å vacuum barrier separating two Cu(111) surfaces. The calculation was performed within the DFT using the FLAPW-method, as implemented in the embedding version of the FLEUR-code⁹⁻¹¹.

Figure 9 compares the results of the Landauer formula and of Bardeen's approach. In contrast to the simple analytic model discussed before in this setup both localized surface states and delocalized propagating Bloch states can be found. In the energy range below ≈ -0.9 eV a band of Bloch electrons with $k_{\parallel} = 0$ can be found which due to their low energies have very low tunneling probabilities. This transmission probability is described by both formulas and similar to the analytic case discussed before both approaches lead to essentially identical results.

At -0.4 eV, however, a surface state on the Cu(111) surface can be found for $k_{\parallel} = 0$. In the case of the setup required for the Landauer equation these surface states will occur on both sides and hence these states will be split into two levels by hybridization. On the other hand, in Bardeen's approach, because of the complete separation of the two systems

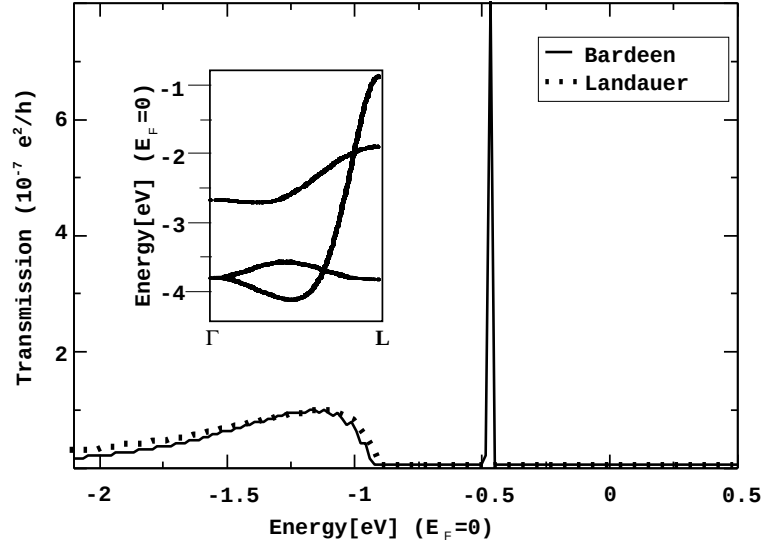


Figure 9. Conductance through a vacuum barrier separating two Cu(111) surfaces. Only electrons with normal incidence are considered. The peak width is due to a finite imaginary part in the energy and its height has been scaled down by a factor of five. The inset shows the normal incidence part of the Cu band structure relevant for the energy window in which transmission from Bloch states can be found from both approaches.

only a single surface state peak occurs and no hybridization splitting is seen in Figure 9. The finite width of this peak is due to the introduction of an imaginary contribution to the energy of $\delta = 0.03$ eV for illustrative purpose only. Such a small broadening of the peak is needed for a numerical integration of the total conductance as well but should not be confused with the modeling of the coupling of the surface state to the leads. For reasonably small imaginary parts the integrated conductance of the peak does not depend on the exact choice of the imaginary energy.

In the case of the setup required for the Landauer equation, one may estimate the electron hopping rate between the two levels from the energy splitting. Nevertheless, these states do not contribute to the conductance in the Landauer equation as no Bloch states are present in the leads at this energy. Bardeen's equation on the other hand contains a strong contribution from the surface state. In this very artificial situation with the same surface state on both sides of the barrier this effect can only be seen at exactly zero bias. However, the arguments presented here are equally valid for situations in which one encounters a localized state at one side of the barrier and a continuum at the other.

In order for electron hopping between two surface states to be measured as the current, they must be coupled to extended states in both leads, for example, via impurity-induced random potentials, electron-phonon interaction,¹² and electron-electron scattering.¹³ Applying Bardeen's equation to the surface-state conductance assumes implicitly that the transition between two localized states on both sides is the rate-limiting process.

The present results clearly indicate the limits of validity of the two different approaches. While the Landauer formula will be suitable in cases in which a relatively high conductance

is obtained by the coupling of Bloch states, it must be applied with care for tunneling setups in which states localized at interfaces might play a crucial role. On the other hand the Bardeen approach to tunneling is only suitable for exactly these situations and requires a high barrier leading to a low conductance.

3 DFT Calculations of Transport Setups

The most challenging problem when evaluating most of these transport equations starting from a DFT single-particle description is the particular geometry of the system. While most computer codes are build to treat either (i) periodic systems making use the periodic boundary conditions and the Bloch construction of the states or (ii) finite systems like molecules or clusters with vacuum boundary conditions, in a transport system one has to deal with an open system. At both sides of the region of interest in which the quantum scattering on some potential perturbation takes place, semi-infinite leads are attached. The system has no periodicity in this direction of space and solving the Kohn-Sham equation within the scattering volume will only be possible by taking the appropriate boundary condition into account which ensures that the effects of the semi-infinite leads are treated correctly.

To formalize the task, we want to treat this scattering volume only, i.e. we want to deal with the Kohn-Sham Hamiltonian limited to this finite volume H_Ω . Of course we cannot simply consider the equation

$$H_\Omega \psi_\Omega = \epsilon \psi_\Omega \quad (32)$$

since such a simple approach will completely miss out the effect of the attached leads. In fact the way we wrote Eq. (32) it is not even well defined. The Hamiltonian of the total system can not be written as

$$H_\Omega + H_L + H_R \quad (33)$$

containing contributions from the left H_L and the right H_R lead as well as the central region as such a description does not contain the coupling between these volumes. Hence the Kohn-Sham Hamiltonian can not simply be restricted to the center region without specifying the boundary condition imposed at the boundary of this region or equivalently, without considering the coupling to the leads and the properties of this leads.

In order to understand a scheme suitable to achieve the desired treatment of the barrier volume only, we write the total Hamiltonian of the system in matrix form

$$H = \begin{pmatrix} H_L & V_{L\Omega} & 0 \\ V_{\Omega L} & H_\Omega & V_{\Omega R} \\ 0 & V_{R\Omega} & H_R \end{pmatrix}, \quad (34)$$

where $V_{L\Omega}$ and $V_{R\Omega}$ denote the coupling of the left and right lead to the central scattering region. As the Kohn-Sham Hamiltonian is local no direct coupling between the leads needs to be considered and one obtains Eq. (34) by using any sufficiently localized basis set. Starting from this Hamiltonian the Green function of the scattering volume G_Ω can be obtained. The single-particle Green function is formally defined by

$$(H(\vec{r}) - \epsilon) G(\vec{r}, \vec{r}'; \epsilon) = \delta(\vec{r} - \vec{r}'), \quad (35)$$

or more concise in matrix notation

$$(H - \epsilon) G(\epsilon) = 1. \quad (36)$$

Our goal is the calculation of the Green function for the total Hamiltonian restricted to the scattering volume Ω . In analogy to the partitioning of the Hamiltonian we can write

$$\begin{pmatrix} H_L - \epsilon & V_{L\Omega} & 0 \\ V_{\Omega L} & H_\Omega - \epsilon & V_{\Omega R} \\ 0 & V_{R\Omega} & H_R - \epsilon \end{pmatrix} \begin{pmatrix} G_L & G_{L\Omega} & G_{LR} \\ G_{\Omega L} & G_\Omega & G_{\Omega R} \\ G_{RL} & G_{R\Omega} & G_R \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \quad (37)$$

Hence, the Green function G_Ω is defined by

$$V_{\Omega L} G_{L\Omega} + (H_\Omega - \epsilon) G_\Omega + V_{\Omega R} G_{R\Omega} = 1. \quad (38)$$

From the off-diagonal elements of Eq. (37) one obtains ($X = L$ or $X = R$)

$$\begin{aligned} G_{X\Omega} &= -(H_X - \epsilon)^{-1} V_{X\Omega} G_\Omega \\ &= -G_X^0 V_{X\Omega} G_\Omega, \end{aligned} \quad (39)$$

where the Green function $G_X^0 = (H_X - \epsilon)^{-1}$ of the isolated leads has been introduced. consequently one can rewrite Eq. (38) into

$$(H_\Omega - \epsilon - \Sigma_L - \Sigma_R) G_\Omega = 1. \quad (40)$$

This demonstrates that one can calculate the Green function in the scattering volume G_Ω by adding two additional terms $\Sigma_{L/R}$ to the Hamiltonian which take into account the effect of the attached semi-infinite leads. These so-called self-energies are given by

$$\Sigma_X = V_{\Omega X} (H_X - \epsilon)^{-1} V_{X\Omega} = V_{\Omega X} G_X^0 V_{X\Omega}. \quad (41)$$

and play a crucial role in the *ab initio* calculation of scattering setups. By means of the self-energy the coupling to the leads is described. Consequently, both the interaction between the leads and the scattering volume $V_{X\Omega}$ and the electronic structure of the leads H_X enter in its definition. Since these Σ are energy dependent, one can not easily cast Eq. (38) into an eigenvalue problem to obtain the single-electron wavefunctions but instead must express the quantities of interest in terms of the Green function.

The linear response conductivity of the Kohn-Sham system has been expressed by many authors in terms of the Green function. This approach yields

$$\Gamma = \frac{e^2}{h} \text{Tr} (G_\Omega \Im \Sigma_R G_\Omega^* \Im \Sigma_L). \quad (42)$$

One can prove^{14,9} that this expression is equivalent to the Landauer formula as given in Eq. (10). We will not repeat this derivation here but instead give some handwaving arguments which create understanding by analogies in the interpretation of the different constituents of the expression:

- As one can see from Eq. (41) the imaginary part of the self-energies of the two leads is due to the Green function of the leads G_X^0 . Hence the imaginary part of the self energy contains the information on the Bloch states in the leads and projects out the propagating states in the trace performed in Eq. (42) in analogy to the sum over asymptotic Bloch states in the original Landauer equation (Eq. (10)).

- As the self-energy describes the coupling to the leads they are highly localized at the surface between the scattering volume of the leads and in products of the form $\Im\Sigma_L G \Im\Sigma_R$ the two spacial arguments of the Green are basically placed on the two opposite sides of the scattering volume. Hence this Green function basically corresponds to the transmission coefficient in the original Landauer equation.

3.1 Non-Equilibrium Calculation

The formalism we discussed so far is restricted to the treatment of electrical conductivity in the linear response regime valid for the limit of zero bias voltage. This limit is definitely appropriate for systems with nearly perfect transmission. If there would be very little scattering of the incoming electrons, i.e. if the incoming Kohn-States do not couple to reflected states but are perfectly transmitted, it would be impossible to apply a large bias voltage. This can be understood very simply from Ohm's law which tells you that you can not have a potential drop if you have no resistance. Hence, in this highly conducting case the zero bias treatment is the only reasonable description as long as one does not go beyond our model of ballistic transport without scattering processes making the DFT description invalid.

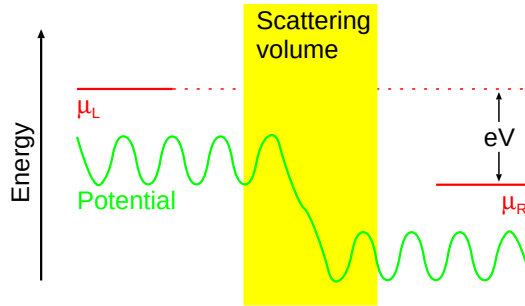


Figure 10. Potential and local Fermi-levels in a system out of equilibrium. The bias applied between the two leads gives rise to a potential drop which is confined to the scattering volume so that the leads stay in local equilibrium.

The situation is different in the case of a strong reflection of the electrons within the scattering volume. In this case a substantial potential drop occurs within the scattering volume and this effect can be treated. In the limit of very little transmission like in a tunneling setup, one can safely assume that the total resistance measured in an experiment is actual due to the tunnel barrier, i.e. to the effect of the scattering volume. In this case the potential is flat in the leads, and all the potential drop will be across the barrier.

The basic assumption one makes when calculating a conducting junction under non-equilibrium conditions is the neglect of the effect of the applied bias on the leads. This includes the assumption that the potential in the leads is “flat” i.e. the leads are assumed to have the bulk potential. Additionally, no effect of the current in the leads is considered, the electronic structure of the leads is unchanged from the situation of the bulk. These assumptions of course require that the current is reasonably small.

We will consider the bias voltage V as being applied to the left lead changing the local potential in the lead

$$V_L \rightarrow V_L + eV. \quad (43)$$

As the leads are assumed to be in a local equilibrium one can assign a local chemical potential $\mu_{(L/R)}$ to both leads. These different chemical potential levels are shifted according to the applied bias voltage

$$\mu_L = \mu_R + eV. \quad (44)$$

The basic task in the self-consistent DFT cycle for the system in this non-equilibrium situation is the calculation of the charge density in the scattering volume. This charge density can be viewed as being composed out of the charge due to electrons incident from the left lead N_L and the charge due to electrons incident from the right N_R ^a

$$N = N_L + N_R. \quad (45)$$

Due to the different Fermi levels in the leads two different Fermi distribution functions f have to be considered in the evaluation of these charges

$$N_L = \frac{1}{\pi} \int_{-\infty}^{\infty} f_L(\epsilon) A_L(\epsilon) d\epsilon \quad \text{and} \quad N_R = \frac{1}{\pi} \int_{-\infty}^{\infty} f_R(\epsilon) A_R(\epsilon) d\epsilon, \quad (46)$$

where $f_{L/R}(\epsilon) = f(\epsilon - \mu_{(L/R)})$ and $A_{L/R}$ denotes the spectral function due the left and right incoming states. This left and right spectral functions are key quantities in non-equilibrium calculations. They are given in terms of Kohn-Sham states $\psi_{L/R}$ which are incoming from the respective lead^b

$$A_{L/R}(\epsilon) = \pi \sum_i \psi_{i;L/R}^* \psi_{i;L/R} \delta(\epsilon_i - \epsilon). \quad (47)$$

The sum has to be performed over all possible incoming electron states. While some codes exist that use such an expression of the spectral function in terms of the wavefunctions most approaches employ a reformulation of the spectral function in terms of the Green function of the scattering volume and the self-energies of the leads.

$$A_{L/R} = G^* \Im \Sigma_{L/R} G. \quad (48)$$

This equation can be derived from the definition of the spectral function as the imaginary part of the Green function. Hence, assuming that the spectral function can be separated into a left/right part one can write

$$A_L + A_R = \frac{-i}{2} (G - G^*). \quad (49)$$

The Green function of the scattering volume is given in terms of the Hamiltonian and the self-energies of the leads (Eq. (38))

$$G^{-1} = \epsilon - H - \Sigma_L - \Sigma_R \quad (50)$$

^a Actually, this statement is not valid if states exist which are localized within the scattering volume and which decay exponentially into the leads. Such states impose a formidable problem if they are energetically located between the two Fermi levels. In this case the occupation of the localized state is not defined without considering further effects beyond the simple DFT approach.

^b These states are usually called scattering states.

and

$$G^{*-1} = \epsilon - H - \Sigma_L^* - \Sigma_R^*, \quad (51)$$

as the Hamiltonian is Hermitian. Subtracting Eqs. (50) and (51) and multiplying with G^* from the left and G from the right one obtains

$$\frac{-i}{2}(G - G^*) = G^* \Im \Sigma_L G + G^* \Im \Sigma_R G \quad (52)$$

from which one can easily identify Eq. (48).

We will finalize the discussion of this expression of the left/right spectral function with some comments:

- While we only presented a rough and handwaving derivation of these expressions, one can also obtain them most rigorously within the non-equilibrium Keldysh Green function technique. Such a derivation also allows to systematically improve the level of approximation by including further interactions in a perturbative treatment.
- The two expressions in terms of the Green functions and in terms of the wavefunctions are equivalent. However, in practical calculations the Green function formalism is more widely used.
- The presence of localized states in the scattering volume introduces a significant problem in non-equilibrium calculations. Since these states do not couple to the Bloch states of the left or right leads it is unclear how the occupancy of such states has to be chosen as soon as their energy will be between the two Fermi levels. In reality, the occupancy of these states is determined by the scattering processes beyond the DFT which provide a small coupling to the leads. However, since these processes are not modeled in the simple non-interacting DFT calculation, additional assumptions must be introduced to deal with localized states.
- The decomposition of the spectral function into a right and left part allows to interpret the Landauer equation in a slightly different way. Comparing Eqs. (42) and (48) one realizes that

$$\Gamma = \frac{e^2}{h} \text{Tr} (A_L \Im \Sigma_R) = \frac{e^2}{h} \text{Tr} (A_R \Im \Sigma_L). \quad (53)$$

Hence, the Landauer conductance is given by the projection of the left spectral function on the Bloch states of the right lead or vice versa.

3.2 Typical Flow of a Non-Equilibrium Calculation

We will finalize our discussion by summarizing the different tasks one has to perform in order to put the non-equilibrium scheme into practice. These tasks are summarized in Figure 11 as well. One of the most crucial parts is the calculation of the self-energy of the leads. When using the definition of Eq. (41) the most difficult problem in this step is the calculation of the Green function of the isolated lead. While the Green function of the infinite lead is relatively easy to obtain, the calculation of the Green function $G_{L/R}^O$ of the isolated, semi-infinite leads imposes a substantial problem due to the breaking of

translational symmetry. The missing hopping to the central scattering volume make the first layers of the lead substantially different from the bulk, while – at the same time – the potential should remain unchanged. Different approaches have been developed to solve this problem, most of which employ some kind of iterative calculation of the Green function or the self energy. In many codes this calculation of the self-energies is a time-consuming and difficult task.

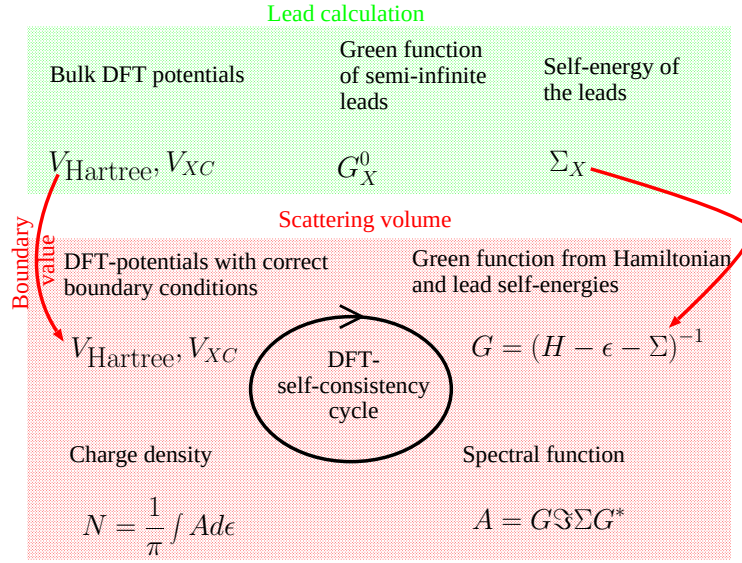


Figure 11. Flow of a self-consistent calculation using the non-equilibrium Green function formalism.

Once the self-energy of the leads is known, the self-consistent calculation of the charge density in the scattering volume can be started. The two major steps here are the calculation of the Hartree potential in the scattering volume and the subsequent calculation of the left and right spectral functions.

The applied bias enters the calculation of the Hartree potential. At the boundaries of the central volume, the Hartree potential has to fulfill a Dirichlet boundary condition imposed by the bulk Hartree potential of the leads. The bias shifts this potential on one side of the leads effectively imposing an electric field into the scattering volume. Obviously, the Hartree potential depends on the charge density within the scattering volume and hence this calculation is the starting point of a DFT self-consistency cycle.

The calculation of the Green function of the scattering volume including the self-energies of the lead is the next important task. Very often this is a very time consuming and memory intensive part as a possibly very large Hamiltonian of the scattering volume has to be inverted (see Eq. (38)) for a huge set of different energies and possibly different $\vec{k}_{\parallel}$ -values in a periodic junction. The Green function of the scattering volume then allows the calculation of the left and right spectral functions in accordance with Eq. (48). The DFT self-consistency circle will ensure that the charge redistribution due to the applied

bias is taken into account properly. For example, screening charges will be induced at the interface such that the metallic leads become free of the external electric field and the applied bias voltage, which will lead to a simply linear potential drop across all the scattering volume without considering the self-consistent redistribution of the charge, will be confined to a smaller volume which includes only the scattering potential.

4 Summary

Electronic structure calculations on the basis of the density functional theory allow to investigate the effect of details of the atomic structures on the transport in nanostructures. We discussed the approximations most frequently used in these calculations, the Landauer formula and Bardeen's tunneling approach and stressed their similarities as well as the differences in the treatment of the tunneling conductance. Finally, we introduced the basic ideas used to treat systems out of equilibrium in which a finite bias voltage is applied across a scattering region. The key formulas of the non-equilibrium Green function technique applied to the non-interacting DFT system were introduced and interpreted.

References

1. R. Landauer. *IBM Journal Res. Dev.*, 1:223, 1957.
2. M. Büttiker. Four-terminal phase coherent conductance. *Phys. Rev. Lett.*, 57:1761, 1986.
3. A.D. Stone and A. Szafer. What is measured when you measure a resistance? - The Landauer formula revisited. *IBM Journal Res. Dev.*, 32:384, 1988.
4. S. Heinze, S. Blügel, R. Pascal, M. Bode, and R. Wiesendanger. *Phys. Rev. B*, 58:16432, 1998.
5. S. Heinze, R. Abt, S. Blügel, G. Gilarowski, and H. Niehus. *Phys. Rev. Lett.*, 83:4808, 1999.
6. B. Voigtländer, V. Scheuch, H.P. Bonzel, S. Heinze, and S. Blügel. *Phys. Rev. B*, 55:R13444, 1997.
7. V. P. LaBella, H. Yang, D. W. Bullock, P. M. Thibabo, Peter Kratzer, and Matthias Scheffler. *Phys. Rev. Lett.*, 83:2989, 1999.
8. F. Schwabl. *Quantenmechanik*. Springer, 4 edition, 1993.
9. D. Wortmann, H. Ishida, and S. Blügel. *Phys. Rev. B*, 66:075113, 2002.
10. D. Wortmann, H. Ishida, and S. Blügel. *Phys. Rev. B*, 65:165103, 2002.
11. <http://www.flapw.de>.
12. A. Eiguen, B. Hellsing, E. V. Chulkov, and P. M. Echenique. *Phys. Rev. B*, 67:235423, 2003.
13. S. Link, H. A. Dürr, G. Bihlmayer, S. Blügel, W. Eberhardt, E. V. Chulkov, V. M. Silkin, and P. M. Echenique. *Phys. Rev. B*, 63:115420, 2001.
14. H. U. Baranger and A. D. Stone. *Phys. Rev. B*, 40:8169, 1989.

Introduction to Parallel Computing

Bernd Mohr

John von Neumann Institute for Computing
Central Institute for Applied Mathematics
Forschungszentrum Jülich
52425 Jülich, Germany
E-mail: b.mohr@fz-juelich.de

The major parallel programming models for scalable parallel architectures are the message passing model and the shared memory model. This article outlines the main concepts of these models as well as the industry standard programming interfaces MPI and OpenMP. To exploit the potential performance of parallel computers, programs need to be carefully designed and tuned. We will discuss design decisions for good performance as well as programming tools that help the programmer in program tuning.

1 Introduction

Although the performance of sequential computers increases incredibly fast, it is insufficient for a large number of challenging applications. Applications requiring much more performance are numerical simulations in industry and research as well as commercial applications such as query processing, data mining, and multi-media applications. Current hardware architectures offering high performance do not only exploit parallelism on a very fine grain level within a single processor but apply a medium to large number of processors concurrently to a single computation. High-end parallel computers currently (2005) deliver up to 280 Teraflop/s (10^{12} floating point operations per second) and are developed and exploited within the ASCI (Accelerated Strategic Computing Initiative) program of the Department of Energy in the USA.

This article concentrates on programming numerical applications on parallel computer architectures introduced in Section 1.1. Parallelization of those applications centers around selecting a decomposition of the data domain onto the processors such that the workload is well balanced and the communication between processors is reduced (Section 1.2)⁵.

The parallel implementation is then based on either the message passing or the shared memory model (Section 2). The standard programming interface for the message passing model is MPI (Message Passing Interface)^{9–11}, offering a complete set of communication routines (Section 3). OpenMP^{3,12} is the standard for directive-based shared memory programming and will be introduced in Section 4.

Since parallel programs exploit multiple threads of control, debugging is even more complicated than for sequential programs. Section 5 outlines the main concepts of parallel debuggers and presents TotalView¹³, the most widely available debugger for parallel programs.

Although the domain decomposition is key to good performance on parallel architectures, program efficiency also heavily depends on the implementation of the communication and synchronization required by the parallel algorithms and the implementation techniques chosen for sequential kernels. Optimizing those aspects is very system dependent and thus, an interactive tuning process consisting of measuring performance data and

applying optimizations follows the initial coding of the application. The tuning process is supported by programming model specific performance analysis tools. Section 6 presents basic performance analysis techniques.

1.1 Parallel Architectures

A *parallel computer* or *multi-processor system* is a computer utilizing more than one processor. A common way to classify parallel computers is to distinguish them by the way how processors can access the system's main memory because this influences heavily the usage and programming of the system.

In a *distributed memory architecture* the system is composed out of single-processor nodes with local memory. The most important characteristic of this architecture is that access to the local memory is faster than to remote memory. It is the challenge for the programmer to assign data to the processors such that most of the data accessed during the computation are already in the node's local memory. Two major classes of distributed memory computers can be distinguished:

No Remote Memory Access (NORMA) computers do not have any special hardware support to access another node's local memory directly. The nodes are only connected through a computer network. Processors obtain data from remote memory only by exchanging messages over this network between processes on the requesting and the supplying node. Computers in this class are sometimes also called **Network Of Workstations (NOW)** or **Clusters Of Workstations (COW)**.

Remote Memory Access (RMA) computers allow to access remote memory via specialized operations implemented by hardware, however the hardware does not provide a global address space, i.e., a memory location is not determined via an address in a shared linear address space but via a tuple consisting of the processor number and the local address in the target processor's address space.

The major advantage of distributed memory systems is their ability to scale to a very large number of nodes. Today (2005), systems with more than 130,000 nodes have been built. The disadvantage is that such systems are very hard to program.

In contrast, a *shared memory architecture* provides (in hardware) a global address space, i.e., all memory locations can be accessed via usual load and store operations. Access to a remote location results in a copy of the appropriate cache line in the processor's cache. Therefore, such a system is much easier to program. However, shared memory systems can only be scaled to moderate numbers of processors, typically 64 or 128. Shared memory systems are further classified according to the quality of the memory accesses:

Uniform Memory Access (UMA) computer systems feature one global shared memory subsystem which is connected to the processors through a central bus or memory switch. All of the memory is accessible to all processors in the same way. Such a system is also often called **Symmetrical Multi Processor (SMP)**.

Non Uniform Memory Access (NUMA) computers are more scalable by physically distributing the memory but still providing a hardware implemented global address space. Therefore access to memory local or close to a processor is faster than to remote memory. If such a system has additional hardware which also ensures that multiple copies of data stored in different cache lines of the processors is kept coherent, i.e.,

the copies always do have the same value, then it is called a **Cache-Coherent Non Uniform Memory Access (ccNUMA)** system. ccNUMA systems offer the abstraction of a shared linear address space resembling physically shared memory systems. This abstraction simplifies the task of program development but does not necessarily facilitate program tuning.

While most of the early parallel computers were simple single processor NORMA systems, today's large parallel systems are typically *hybrid systems*, i.e., shared memory NUMA nodes with a moderate number of processors are connected together to form a distributed memory cluster system.

1.2 Data Parallel Programming

Applications that scale to a large number of processors usually perform computations on large data domains. For example, crash simulations are based on partial differential equations that are solved on a large finite element grid and molecular dynamics applications simulate the behavior of a large number of particles. Other parallel applications apply linear algebra operations to large vectors and matrices. The elemental operations on each object in the data domain can be executed in parallel by the available processors.

The scheduling of operations to processors is determined by a *domain decomposition*⁶ specified by the programmer. Processors execute those operations that determine new values for elements stored in local memory (owner-computes rule). While processors execute an operation, they may need values from other processors. The domain decomposition has thus to be chosen so that the distribution of operations is balanced and the communication is minimized. The third goal is to optimize single node computation, i.e., to be able to exploit the processor's pipelines and the processor's caches efficiently.

A good example for the design decisions taken when selecting a domain decomposition is Gaussian elimination². The main structure of the matrix during the steps of the algorithm is outlined in Figure 1.

The goal of this algorithm is to eliminate all entries in the matrix below the main diagonal. It starts at the top diagonal element and subtracts multiples of the first row from the second and subsequent rows to end up with zeros in the first column. This operation is repeated for all the rows. In later stages of the algorithm the actual computations have to be done on rectangular sections of decreasing size. If the main diagonal element of the current row is zero, a pivot operation has to be performed. The subsequent row with the maximum value in this column is selected and exchanged with the current row.

A possible distribution of the matrix is to decompose its columns into blocks, one block for each processor. The elimination of the entries in the lower triangle can then be performed in parallel where each processor computes new values for its columns only. The main disadvantage of this distribution is that in later computations of the algorithm only a subgroup of the processors is actually doing any useful work since the computed rectangle is getting smaller.

To improve load balancing, a cyclic column distribution can be applied. The computations in each step of the algorithm executed by the processors differ only in one column.

In addition to load balancing also communication needs to be minimized. Communication occurs in this algorithm for broadcasting the current column to all the processors since it is needed to compute the multiplication factor for the row. If the domain decomposition

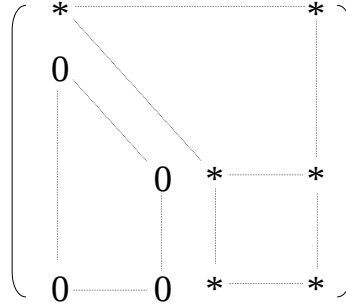


Figure 1. Structure of the matrix during Gaussian elimination.

is a row distribution, which eliminates the need to communicate the current column, the current row needs to be broadcast to the other processors.

If we consider also the pivot operation, communication is necessary to select the best row when a row-wise distribution is applied since the computation of the global maximum in that column requires a comparison of all values.

Selecting the best domain decomposition is further complicated due to optimizing single node performance. In this example, it is advantageous to apply BLAS3 operations for the local computations. These operations make use of blocks of rows to improve cache utilization. Blocks of rows can only be obtained if a block-cyclic distribution is applied, i.e., columns are not distributed individually but blocks of columns are cyclically distributed.

This discussion makes clear, that choosing a domain decomposition is a very complicated step in program development. It requires deep knowledge of the algorithm's data access patterns as well as the ability to predict the resulting communication.

2 Programming Models

Programming parallel computers is almost always done via the so-called *Single Program Multiple Data* (SPMD) model. SPMD means that the same program (executable code) is executed on all processors taking part in the computation, but it computes on different parts of the data which were distributed over the processors based on a specific domain decomposition. If computations are only allowed on specific processors, this has to be explicitly programmed by using conditional programming constructs (e.g., with `if` or `where` statements). There are two main programming models, *message passing* and *shared memory*, offering different features for implementing applications parallelized by domain decomposition.

The message passing model is based on a set of processes with private data structures. Processes communicate by exchanging messages with special send and receive operations. It is a natural fit for programming distributed memory machines but also can be used on shared memory computers. The domain decomposition is implemented by developing a code describing the local computations and local data structures of a single process. Thus, global arrays have to be split up and only the local part has to be allocated in a process. This handling of global data structures is called *data distribution*. Computations on the global arrays also have to be transformed, e.g., by adapting the loop bounds, to ensure that

only local array elements are computed. Access to remote elements has to be implemented via explicit communication, temporary variables have to be allocated, messages have to be constructed and transmitted to the target process.

The shared memory model is based on a set of threads that is created when parallel operations are executed. This type of computation is also called *fork-join parallelism*. Threads share a global address space and thus access array elements via a global index. The main parallel operations are *parallel loops* and *parallel sections*. Parallel loops are executed by a set of threads also called a *team*. The iterations are distributed among the threads according to a predefined strategy. This scheduling strategy implements the chosen domain decomposition. Parallel sections are also executed by a team of threads but the tasks assigned to the threads implement different operations. This feature can for example be applied if domain decomposition itself does not generate enough parallelism and whole operations can be executed in parallel since they access different data structures.

In the shared memory model, the distribution of data structures onto the node memories is not enforced by decomposing global arrays into local arrays, but the global address space is distributed onto the memories by the operating system. For example, the pages of the virtual address space can be distributed cyclically or can be assigned at first touch. The chosen domain decomposition thus has to take into account the granularity of the distribution, i.e., the size of pages, as well as the system-dependent allocation strategy.

While the domain decomposition has to be hard-coded into the message passing program, it can easily be changed in a shared memory program by selecting a different scheduling strategy for parallel loops.

Another advantage of the shared memory model is that automatic and incremental parallelization is supported. While automatic parallelization leads to a first working parallel program, its efficiency typically needs to be improved. The reason for this is that parallelization techniques work on a loop-by-loop basis and do not globally optimize the parallel code via a domain decomposition. In addition, dependence analysis, the prerequisite for automatic parallelization, is limited to access patterns known at compile time. The biggest disadvantage of this model is that it can only be used on shared memory computers.

In the shared memory model, a first parallel version is relatively easy to implement and can be incrementally tuned. In the message passing model instead, the program can be tested only after finishing the full implementation. Subsequent tuning by adapting the domain decomposition is usually time consuming.

3 MPI

The Message Passing Interface (MPI)⁹⁻¹¹ was developed between 1993 and 1997. It is a community standard which standardizes the calling interface for a communication and synchronization function library. It provides Fortran 77, Fortran 90, C and C++ language bindings. It includes routines for point-to-point communication, collective communication, one-sided communication, and parallel IO. While the basic point-to-point and collective communication primitives have already been defined since 1994 and implemented on all parallel computers, parallel IO and especially one-sided communication routines are part of MPI 2.0 and are not available on all machines. For a simple example see the appendix.

3.1 MPI Basic Routines

MPI consists of more than 320 functions. But realistic programs can already be developed based on no more than six functions:

MPI_Init initializes the library. It has to be called at the beginning of a parallel operation before any other MPI routines are executed.

MPI_Finalize frees any resources used by the library and has to be called at the end of the program.

MPI_Comm_size determines the number of processors executing the parallel program.

MPI_Comm_rank returns the unique process identifier.

MPI_Send transfers a message to a target process. This operation is a blocking send operation, i.e., it terminates when the message buffer can be reused either because the message was copied to a system buffer by the library or because the message was delivered to the target process.

MPI_Recv receives a message. This routine terminates if a message was copied into the receive buffer.

3.2 MPI Communicator

All communication routines depend on the concept of a *communicator*. A communicator consists of a process group and a communication context. The processes in the process group are numbered from zero to process count - 1. The process number returned by **MPI_Comm_rank** is the identification in the process group of the communicator which is passed as a parameter to this routine.

The communication context of the communicator is important in identifying messages. Each message has an integer number called a *tag* which has to match a given selector in the corresponding receive operation. The selector depends on the communicator and thus on the communication context. It selects only messages with a fitting tag and having been sent relative to the same communicator. This feature is very useful in building parallel libraries since messages sent inside the library will not interfere with messages outside if a special communicator is used in the library. The default communicator that includes all processes of the application is **MPI_COMM_WORLD**.

3.3 MPI Collective Operations

Another important class of operations are *collective operations*. Collective operations are executed by a process group identified via a communicator. All the processes in the group have to perform the same operation. Typical examples for such operations are:

MPI_Barrier synchronizes all processes. None of the processes can proceed beyond the barrier until all the processes started execution of that routine.

MPI_Bcast allows to distribute the same data from one process, the so-called *root* process, to all other processes in the process group.

MPI_Scatter also distributes data from a root process to a whole process group, but each receiving process gets different data.

MPI_Gather collects data from a group of processes at a root process.

MPI_Reduce performs a global operation on the data of each process in the process group. For example, the sum of all values of a distributed array can be computed by first summing up all local values in each process and then summing up the local sums to get a global sum. The latter step can be performed by the reduction operation with the parameter **MPI_SUM**. The result is delivered to a single target processor.

3.4 MPI IO

Data parallel applications make use of the IO subsystem to read and write big data sets. These data sets result from replicated or distributed arrays. The reasons for IO are to read input data, to pass information to other programs, e.g., for visualization, or to store the state of the computation to be able to restart the computation in case of a system failure or if the computation has to be split into multiple runs due to its resource requirements.

IO can be implemented in three ways:

1. Sequential IO

A single node is responsible to perform the IO. It gathers information from the other nodes and writes it to disk or reads information from disk and scatters it to the appropriate nodes. While the IO is sequential and thus need not be parallelized, the full performance of the IO subsystem might not be utilized. Modern systems provide high performance IO subsystems that are fast enough to support multiple IO requests from different nodes in parallel.

2. Private IO

Each node accesses its own files. The big advantage of this implementation is that no synchronization among the nodes is required and very high performance can be obtained. The major disadvantage is that the user has to handle a large number of files. For input the original data set has to be split according to the distribution of the data structure and for output the process-specific files have to be merged into a global file for post-processing.

3. Parallel IO

In this implementation all the processes access the same file. They read and write only those parts of the file with relevant data. The main advantages are that no individual files need to be handled and that reasonable performance can be reached. The disadvantage is that it is difficult to reach the same performance as with private IO. The parallel IO interface of MPI provides flexible and high-level means to implement applications with parallel IO.

Files accessed via MPI IO routines have to be opened and closed by collective operations. The open routine allows to specify hints to optimize the performance such as whether the application might profit from combining small IO requests from different nodes, what size is recommended for the combined request, and how many nodes should be engaged in merging the requests.

The central concept in accessing the files is the view. A view is defined for each process and specifies a sequence of data elements to be ignored and data elements to be read or written by the process. When reading or writing a distributed array the local information

can be described easily as such a repeating pattern. The IO operations read and write a number of data elements on the basis of the defined view, i.e., they access the local information only. Since the views are defined via runtime routines prior to the access, the information can be exploited in the library to optimize IO.

MPI IO provides blocking as well as nonblocking operations. In contrast to blocking operations, the nonblocking ones only start IO and terminate immediately. If the program depends on the successful completion of the IO it has to check it via a test function. Besides the collective IO routines which allow to combine individual requests, also non-collective routines are available to access shared files.

3.5 MPI Remote Memory Access

Remote memory access (RMA) operations (also called *1-sided communication*) allow to access the address space of other processes without participation of the other process. The implementation of this concept can either be in hardware, such as in the CRAY T3E, or in software via additional threads waiting for requests. The advantages of these operations are that the protocol overhead is much lower than for normal send and receive operations and that no polling or global communication is required for setting up communication.

In contrast to explicit message passing where synchronization happens implicitly, accesses via RMA operations need to be protected by explicit synchronization operations.

RMA communication in MPI is based on the *window concept*. Each process has to execute a collective routine that defines a window, i.e., the part of its address space that can be accessed by other processes.

The actual access is performed via *put* and *get* operations. The address is defined by the target process number and the displacement relative to the starting address of the window for that process.

MPI also provides special synchronization operations relative to a window. The `MPI_Win_fence` operation synchronizes all processes that make some address ranges accessible to other processes. It is a collective operation that ensures that all RMA operations started before the fence operation terminate before the target process executes the fence operation and that all RMA operations of a process executed after the fence operation are executed after the target process executed the fence operation. There are also more fine grained synchronization methods available in the form of the General Active Target Synchronization or via locks.

4 OpenMP

OpenMP^{3,12} is a directive-based programming interface for the shared memory programming model. It consists of a set of directives and runtime routines for Fortran 77 (published 1997), for Fortran 90 (2000), and a corresponding set of pragmas for C and C++ (1998).

Directives are special comments that are interpreted by the compiler. Directives have the advantage that the code is still a sequential code that can be executed on sequential machines (by ignoring the directives/pragmas) and therefore there is no need to maintain separate sequential and parallel versions.

Directives start and terminate parallel regions. When the master thread hits a parallel region a team of threads is created or activated. The threads execute the code in parallel

and are synchronized at the beginning and the end of the computation. After the final synchronization the master thread continues sequential execution after the parallel region. The main directives are:

!\$OMP PARALLEL DO specifies a loop that can be executed in parallel. The DO loop's iterations can be distributed among the set of threads according to various scheduling strategies including **STATIC(CHUNK)**, **DYNAMIC(CHUNK)**, and **GUIDED(CHUNK)**. **STATIC(CHUNK)** distribution means that the set of iterations are consecutively distributed among the threads in blocks of **CHUNK** size (resulting in block and cyclic distributions). **DYNAMIC(CHUNK)** distribution implies that iterations are distributed in blocks of **CHUNK** size to threads on a first-come-first-served basis. **GUIDED (CHUNK)** means that blocks of exponentially decreasing size are assigned on a first-come-first-served basis. The size of the smallest block is determined by **CHUNK** size.

!\$OMP PARALLEL SECTIONS starts a set of sections that are each executed in parallel by a team of threads.

!\$OMP PARALLEL introduces a code region that is executed redundantly by the threads. It has to be used very carefully since assignments to global variables will lead to conflicts among the threads and possibly to nondeterministic behavior.

!\$OMP DO is a work sharing construct and may be used within a parallel region. All the threads executing the parallel region have to cooperate in the execution of the parallel loop. There is no implicit synchronization at the beginning of the loop but a synchronization at the end. After the final synchronization all threads continue after the loop in the replicated execution of the program code.

The main advantage of this approach is that the overhead for starting up the threads is eliminated. The team of threads exists during the execution of the parallel region and need not be built before each parallel loop.

!\$OMP SECTIONS is also a work sharing construct that allows the current team of threads executing the surrounding parallel region to cooperate in the execution of the parallel sections.

Program data can either be shared or private. While threads do have their own copy of private data, only one copy exists of shared data. This copy can be accessed by all threads. To ensure program correctness, OpenMP provides special synchronization constructs. The main constructs are *barrier synchronization* enforcing that all threads have reached this synchronization operation before execution continues and *critical sections*. Critical sections ensure that only a single thread can enter the section and thus, data accesses in such a section are protected from race conditions. For example, a common situation for a critical section is the accumulation of values. Since an accumulation consists of a read and a write operation unexpected results can occur if both operations are not surrounded by a critical section. For a simple example of an OpenMP parallelization see the appendix.

5 Parallel Debugging

Debugging parallel programs is more difficult than debugging sequential programs not only since multiple processes or threads need to be taken into account but also because program behavior might not be deterministic and might not be reproducible. These problems are not

solved by current state-of-the-art commercial parallel debuggers. They only deal with the first problem by providing menus, displays, and commands that allow to inspect individual processes and execute commands on individual or all processes.

The widely used debugger is TotalView from Etnus¹³. It provides breakpoint definition, single stepping, and variable inspection via an interactive interface. The programmer can execute those operations for individual processes and groups of processes. TotalView also provides some means to summarize information such that equal information from multiple processes is combined into a single information and not repeated redundantly. It also supports MPI and OpenMP programs on many platforms.

6 Parallel Performance Analysis

Performance analysis is an iterative subtask during program development. The goal is to identify program regions that do not perform well. Performance analysis is structured into three phases:

1. Measurement

Performance analysis is done based on information on runtime events gathered during program execution. The basic events are, for example, cache misses, termination of a floating point operation, start and stop of a subroutine or message passing operation. The information on individual events can be summarized during program execution or individual trace records can be collected for each event.

Summary information has the advantage to be of moderate size while trace information tends to be very large. The disadvantage is that it is not fine grained; the behavior of individual instances of subroutines can for example not be investigated since all the information has been summed up.

2. Analysis

During analysis the collected runtime data are inspected to detect *performance problems*. Performance problems are based on *performance properties*, such as the existence of message passing in a program region, which have a condition for identifying it and a severity function that specifies its importance for program performance.

Current tools support the user in checking the conditions and the severity by a visualization of the program behavior. Future tools might be able to automatically detect performance properties based on a specification of possible properties. During analysis the programmer applies a threshold. Only performance properties whose severity exceeds this threshold are considered to be performance problems.

3. Ranking

During program analysis the severest performance problems need to be identified. This means that the problems need to be ranked according to the severity. The most severe problem is called the *program bottleneck*. This is the problem the programmer tries to resolve by applying appropriate program transformations.

Current techniques for performance data collection are *profiling* and *tracing*. Profiling collects summary data only. This can be done via *sampling*. The program is regularly

interrupted, e.g., every 10 ms, and the information is added up for the source code location which was executed in this moment. For example, the UNIX profiling tool *prof* applies this technique to determine the fraction of the execution time spent in individual subroutines.

A more precise profiling technique is based on *instrumentation*, i.e., special calls to a *monitoring library* are inserted into the program. This can either be done in the source code by the compiler or specialized tools, or can be done in the object code. While the first approach allows to instrument more types of regions, for example, loops and vector statements, the latter allows to measure data for programs where no source code is available. The monitoring library collects the information and adds it to special counters for the specific region.

Tracing is a technique that collects information for each event. This results, for example, in very detailed information for each instance of a subroutine and for each message sent to another process. The information is stored in specialized trace records for each event type. For example, for each start of a send operation, the time stamp, the message size and the target process can be recorded, while for the end of the operation, the time stamp and bandwidth are stored.

The trace records are stored in the memory of each process and are written to a trace file either when the buffer is filled up or when the program terminates. The individual trace files of the processes are merged together into one trace file ordered according to the time stamps of the events.

7 Summary

This article gave an overview of parallel programming models as well as programming tools. Parallel programming will always be a challenge for programmers. Higher-level programming models and appropriate programming tools only facilitate the process but do not make it a simple task.

While programming in MPI offers the greatest potential performance, shared memory programming with OpenMP is much more comfortable due to the global style of the resulting program. The sequential control flow among the parallel loops and regions matches much better with the sequential programming model all the programmers are trained for.

Although programming tools were developed over years, the current situation seems not to be very satisfying. Program debugging is done per thread, a technique that does not scale to larger numbers of processors. Performance analysis tools do also suffer scalability limitations and, in addition, the tools are complicated to use. The programmers have to be experts for performance analysis to understand potential performance problems, their proof conditions, and their severity. In addition they have to be experts for powerful but also complex user interfaces.

Future research in this area has to try to automate performance analysis tools, such that frequently occurring performance problems can be identified automatically. It is the goal of the IST working group APART on *Automatic Performance Analysis: Resources and Tools* to investigate base technologies for future, more intelligent tools¹. An important result of this work is a collection of performance problems for parallel programs that have been formalized with the ASL, the *APART Specification Language*⁴. This approach will lead to a formal representation of the knowledge applied in the manually executed performance analysis process and thus will make this knowledge accessible for automatic processing.

First automatic tools are already available: ParaDyn⁸ from the University of Wisconsin-Madison and KOJAK⁷ from the Forschungszentrum Jülich.

A second important trend that will effect parallel programming in the future is the move towards clustered shared memory systems. Clearly, a hybrid programming approach will be applied on those systems for best performance, combining message passing between the individual SMP nodes and shared memory programming in a node. This programming model will lead to even more complex programs and program development tools have to be enhanced to be able to help the user in developing these codes.

Appendix

This appendix provides three versions of a simple example of a scientific computation. It computes the value of π by numerical integration:

$$\pi = \int_0^1 f(x) dx \quad \text{with} \quad f(x) = \frac{4}{1+x^2}$$

This integral can be approximated numerically by the midpoint rule:

$$\pi \approx \frac{1}{n} \int_1^n f(x_i) \quad \text{with} \quad x_i = \frac{(i-0.5)}{n} \quad \text{for} \quad i = 1, \dots, n$$

Larger values of the parameter n will give us more accurate approximations of π . This is not, in fact, a very good way to compute π , but it makes a good example because it has the typical, complete structure of a numerical simulation program (initialization - loop-based calculation - wrap-up), and the whole source code fits on one page or slide.

To parallelize the example, each process/thread computes and adds up the areas for a different subset of the rectangles. At the end of the computation, all of the local sums are combined into a global sum representing the value of π .

MPI Version of Example Program

The following listing shows a Fortran90 implementation of the π numerical integration example parallelized with the help of MPI.

First, the MPI system has to be initialized (lines 7 to 9) and terminated (line 33) with the necessary MPI calls. Next, the input of parameters (line 11 to 14) and the output of results (lines 29 to 31) has to be restricted so that it is only executed by one processor. Then, the input has to be broadcasted to the other processors (line 16). The biggest (and most complicated) change is to program the distribution of work and data. The do-loop in line 20 has to be changed so that each processor only calculates and summarizes its part of the distributed computations. Finally, the reduce call in lines 26/27 collects the local sums and delivers the global sum to processor 0.

```

1 program main
2 include 'mpif.h'
3 integer          :: i, n, ierr, myrank, numprocs
4 double precision :: f, x, sum, pi, h, mypi
5 f(x) = 4.d0/(1.d0 + x*x)
6
7 call MPI_Init(ierr)
8 call MPI_Comm_rank(MPI_COMM_WORLD, myrank, ierr)
9 call MPI_Comm_size(MPI_COMM_WORLD, numprocs, ierr)
10
11 if ( myrank == 0 ) then
12     write(*,*) "number of intervals?"
13     read(*,*) n
14 end if
15
16 call MPI_Bcast(n, 1, MPI_INTEGER, 0, MPI_COMM_WORLD, ierr)
17
18 h = 1.0 d0 / n
19 sum = 0.0 d0
20 do i = myrank+1, n, numprocs
21     x = ( i - 0.5 d0)*h
22     sum = sum + f(x)
23 end do
24 mypi = h * sum
25
26 call MPI_Reduce(mypi, pi, 1, MPI_DOUBLE_PRECISION, &
27                MPI_SUM, 0, MPI_COMM_WORLD, ierr)
28
29 if ( myrank == 0 ) then
30     write(*, fmt="(A,F16.12)") "Value of pi is ", pi
31 endif
32
33 call MPI_Finalize(ierr)
34 end program

```

Sequential and OpenMP Version of Example Program

The following listing shows the corresponding implementation of the π integration example using OpenMP. As one can see, because of the need to explicitly program all aspects of the parallelization, the MPI version is almost twice as long as the OpenMP version. Although this is clearly more work, it gives a programmer much more ways to express and control parallelism. Also, the MPI version will run on all kinds of parallel computers, while OpenMP is restricted to the shared memory architecture.

As OpenMP is based on directives (which are plain comments in a non-OpenMP compilation mode), it is at the same time also a sequential implementation of the example.

```

1 program main
2 integer          :: i,n
3 double precision :: f,x,sum,pi,h
4 f(x) = 4.d0/(1.d0 + x*x)
5 write(*,*) "number_of_intervals?"
6 read(*,*) n
7
8 h = 1.0 d0 / n
9 sum = 0.0 d0
10 !$omp parallel do private(i,x) reduction(+:sum)
11 do i = 1, n
12     x = (i - 0.5 d0)*h
13     sum = sum + f(x)
14 end do
15 pi = h * sum
16
17 write(*, fmt="(A,F16.12)") "Value_of_pi_is_", pi
18 end program

```

The OpenMP directive in line 10 declares the following do-loop as parallel resulting in a concurrent execution of loop iterations. As the variables `i` and `x` are used to store values during the execution of the loop, they have to be declared private, so that each thread executing iterations has its own copy. The variable `h` is only read, so it can be shared. Finally, it is specified that there is a reduction (using addition) over the variable `sum`.

References

1. APART: IST Working Group on Automatic Performance Analysis Resources and Tools, <http://www.fz-juelich.de/apart/>.
2. D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*, Prentice-Hall (1989).
3. L. Dagum and R. Menon, *OpenMP: An Industry-Standard API for Shared-memory Programming*, IEEE Computational Science & Engineering, Vol. 5, No. 1, 46–55 (1998).
4. Th. Fahringer, M. Gerndt, B. Mohr, F. Wolf, G. Riley, J. Träff, *Knowledge Specification for Automatic Performance Analysis*, APART Technical Report, Forschungszentrum Jülich FZJ-ZAM-IB-2001-08 (2001).
5. I. Foster, *Designing and Building Parallel Programs*, Addison Wesley (1994).
6. G. Fox, *Domain Decomposition in Distributed and Shared Memory Environments*, International Conference on Supercomputing June 8-12, 1987, Athens, Greece, Lecture Notes in Computer Science 297, edited by C. Polychronopoulos (1987).
7. F. Wolf and B. Mohr, Automatic Performance Analysis of Hybrid MPI/OpenMP Applications. *Journal of Systems Architecture*, 49(10–11):421–439 (2003).
8. B. P. Miller, M. D. Callaghan, J. M. Cargille, J. K. Hollingsworth, R. B. Irvine, K. L. Karavanic, K. Kunchithapadam, and T. Newhall, The Paradyn Parallel Performance Measurement Tool, *IEEE Computer*, Vol. 28, No. 11, 37–46 (1995).

9. MPI Forum: *Message Passing Interface*, <http://www.mpi-forum.org>.
10. M. Snir, S. Otto, S. Huss-Lederman, D. Walker, and J. Dongarra, *MPI - the Complete Reference, Volume 1, The MPI Core*. 2nd ed., MIT Press (1998).
11. W. Gropp, S. Huss-Lederman, A. Lumsdaine, E. Lusk, B. Nitzberg, W. Saphir, and M. Snir, *MPI - the Complete Reference, Volume 2, The MPI Extensions*. MIT Press (1998).
12. OpenMP Forum: *OpenMP Standard*, <http://www.openmp.org>.
13. Etnus LLC.: *TotalView*, <http://www.etnus.com/>.

Parallel Linear Algebra Methods

Bernd Körfgen and Inge Gutheil

John von Neumann Institute for Computing
Central Institute for Applied Mathematics
Forschungszentrum Jülich
52425 Jülich, Germany

E-mail: {b.koerfgen, i.gutheil}@fz-juelich.de

This contribution presents two examples for the numerical treatment of partial differential equations using parallel algorithms / computers. The first example solves the Poisson equation in two dimensions; the second partial differential equation describes the physical process of vibration of a membrane. Both problems are tackled with different strategies: The Poisson equation is solved by means of the simple Jacobi algorithm and a suitable parallelization scheme is discussed; in the second case the parallel calculation is performed with the help of the ScaLAPACK library and the issue of data distribution is addressed.

1 Introduction

Numerical linear algebra is an active field of research which provided over the years many methods / algorithms for the treatment of standard problems like the solution of systems of linear equations, the factorization of matrices, the calculation of eigenvalues / eigenvectors etc.¹. The most suitable algorithm for a given linear algebra problem, e.g. arising in a scientific application, has to be determined depending on the properties of the system / matrix (see for instance Ref. 2) like:

- **symmetry**
- **definiteness** (positive, negative, . . .)
- **non-zero structure** (dense, sparse, banded)
- **real or complex** coefficients

and so on. Furthermore the scientist has to decide whether to use a **direct** solver, leading to a transformation of the original matrix and thus (for large problems) generating a need for huge **main memory**, or to use an **iterative** solver which works with the original matrix.

The same rationale holds for the more specialized field of **parallel linear algebra** methods. There the additional aspects originating from the parallel **computer architecture** have to be taken into account in order to choose a suitable algorithm. Several topics influencing the choice and even more the consequent implementation of these algorithms are^{3,4}:

- memory architecture (**shared-memory** vs. **distributed memory**)
- amount of **memory per process/processor**
- implemented **cache** structures

It is far beyond the scope of this contribution to give an overview of the available algorithms. Instead we refer to review articles like Refs. 5–7.

From a practical point of view another important decision is whether the user implements the linear algebra algorithm **himself** or relies on **available software / libraries**. A variety of well-known, robust packages providing high computational performance are on the market, which can be used as building blocks for an application software. Some **freely-available** libraries are:

- Basic Linear Algebra Subprograms (**BLAS**)⁸
- Linear Algebra Package (**LAPACK**)⁹
- Scalable LAPACK (**ScaLAPACK**)¹⁰
- (**P**)**ARPACK** - a (parallel) package for the solution of large eigenvalue problems¹¹
- Portable, Extensible Toolkit for Scientific computation (**PETSc**)¹²

Some of them like BLAS or LAPACK are **serial** software, which help to gain good single processor performance, but leave the task of **parallelization** of the high-level linear algebra computations, e.g. solution of the coupled linear equations, to the user; others, e.g. ScaLAPACK or PARPACK, contain implementations of **parallel solvers**. Thus these packages relieve the user of the parallelization, but still they rely on special data distribution schemes¹³ which require a specific organization of the application program. As a consequence the user has to handle the corresponding data distribution on his own, i.e. he has to parallelize his program at least partly. Nevertheless this might be a lot easier than to implement the full parallel linear algebra algorithm.

Since both strategies are preferable under certain circumstances, we will present in the following two simple physical problems where the **parallel numerical solution** will be demonstrated paradigmatically along the two different approaches:

In Section 2 the Poisson equation will be treated using a **parallel Jacobi solver** for the evolving system of linear equations.

In Section 3 the eigenvalue problem arising from the calculation of the vibration of a membrane is solved using a **ScaLAPACK routine**.

Of course, one would not use these solutions in real applications. Neither is the Jacobi algorithm a state-of-the-art method for the solution of a system of linear equations, nor is the eigensolver from ScaLAPACK the optimal choice for the given problem. Both examples result in a sparse matrix as will be shown in the following. ScaLAPACK contains solvers for full and banded systems, whereas (P)ARACK is a library based on the Arnoldi method which is very suitable for the calculation of a **few** eigenvalues for large **sparse** systems; thus (P)ARPACK would be the natural choice for this kind of problem.

Nevertheless due to the importance of ScaLAPACK for many application fields, e.g. **nanoscience**, and the simplicity of the Jacobi algorithm we present them as illustrative examples.

2 The Poisson Problem

In this section we discuss the numerical solution of the Poisson equation as an example for the approximate treatment of partial differential equations. We give a short outline of

the steps necessary to obtain a serial and later on parallel implementation of the numerical solver. Similar but more elaborate material on this topic can be found in the Refs. 14–16.

In a first step we discuss the discretization of the Poisson equation and introduce one simple solver for the evolving system of linear equations. Afterwards we focus on the parallelization of this algorithm.

2.1 Discretization of the Poisson Equation

The **Poisson equation** in two dimensions is given by

$$\Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = f(x, y) \quad , \quad (x, y) \in \Omega \subset \mathbb{R}^2 \quad (1)$$

where Ω is a domain in $\mathbb{R}^2$. For simplicity $u(x, y)$ shall be given on the boundary $\partial\Omega$ by a **Dirichlet boundary condition**

$$u(x, y) = g(x, y) \quad , \quad (x, y) \in \partial\Omega \quad (2)$$

The functions $f(x, y)$ and $g(x, y)$ are given and $u(x, y)$ is to be calculated.

Since the analytic solution of such a partial differential equation might not be feasible depending on the shape of the domain, the functions f, g etc., one often has to resort to the numerical solution of such a differential equation.

In the following we will develop a simple scheme how to calculate u approximately. For this we assume that the domain has a simple form: Ω is a rectangle (Figure 1).

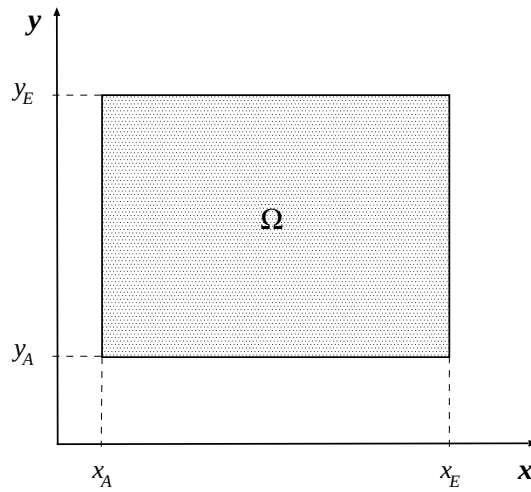


Figure 1. Rectangular domain in $\mathbb{R}^2$

In order to determine the approximate solution of the Poisson equation, u is calculated at certain points of the rectangle. We impose $\Omega = (x_A, x_E) \times (y_A, y_E)$ with an equidistant mesh (Figure 2), where (x_A, x_E) is divided into $(NI + 1)$ sub-intervals and (y_A, y_E) into $(NJ + 1)$ sub-intervals, $(NI, NJ \in \mathbb{N})$. The mesh width h is then given by

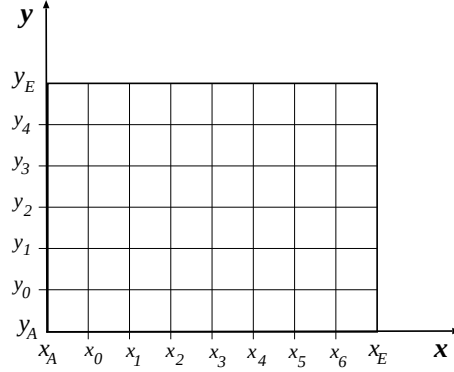


Figure 2. Mesh for $NI = 7$ and $NJ = 5$

$$h = \frac{(x_E - x_A)}{(NI + 1)} = \frac{(y_E - y_A)}{(NJ + 1)} \quad (3)$$

With this choice for the mesh the approximate solution will be calculated at the $NI \cdot NJ$ inner points of the domain (The outer points don't have to be calculated, because they are given by the Dirichlet boundary condition!).

As a next step the second derivatives are replaced by finite differences. For this purpose we use the following Taylor expansions of u at a point (x, y) :

$$u(x + h, y) = u(x, y) + hu_x(x, y) + \frac{h^2}{2!}u_{xx}(x, y) + \frac{h^3}{3!}u_{xxx}(x, y) \pm \dots \quad (4)$$

$$u(x - h, y) = u(x, y) - hu_x(x, y) + \frac{h^2}{2!}u_{xx}(x, y) - \frac{h^3}{3!}u_{xxx}(x, y) \pm \dots \quad (5)$$

Addition of both equations and division by h^2 gives

$$\frac{u(x - h, y) - 2u(x, y) + u(x + h, y)}{h^2} = u_{xx}(x, y) + O(h^2) \quad (6)$$

The result of the analogous procedure for the y -direction is

$$\frac{u(x, y - h) - 2u(x, y) + u(x, y + h)}{h^2} = u_{yy}(x, y) + O(h^2) \quad (7)$$

Using these finite differences the Poisson equation for the $NI \cdot NJ$ inner mesh points of the domain Ω is given by

$$u_{xx}(x_i, y_j) + u_{yy}(x_i, y_j) = f(x_i, y_j) \quad (8)$$

$$(i = 0, \dots, NI - 1 ; j = 0, \dots, NJ - 1)$$

By neglecting the discretization error $O(h^2)$ Eqs. (8) can be written as:

$$u_{i,j-1} + u_{i-1,j} - 4u_{i,j} + u_{i,j+1} + u_{i+1,j} = h^2 f_{i,j} \quad (9)$$

for $i = 0, \dots, NI - 1 ; j = 0, \dots, NJ - 1$. The unknowns

$$u_{i,j} := u(x_i, y_j) \quad (10)$$

have to be calculated from the $NI \cdot NJ$ coupled linear equations (9).

The approximation used here for $u_{xx} + u_{yy}$ is called **5-point stencil** (Figure 3). The name describes the numerical dependency between the points of the mesh. The lexico-

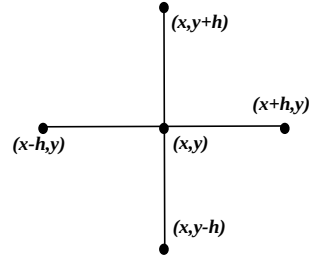


Figure 3. 5-point stencil

graphical numbering (Figure 4) of the mesh points

$$l = j \cdot NI + i + 1 \quad ; i = 0, \dots, NI - 1 \quad ; j = 0, \dots, NJ - 1 \quad (11)$$

and

$$u_l := u_{i,j} \quad (12)$$

allows a compact representation of the system of linear equations by means of a matrix. The coefficient matrix A is a block tridiagonal matrix:

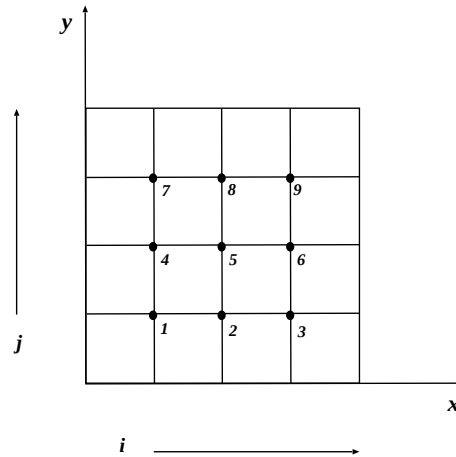


Figure 4. Lexicographical numbering of a 5×5 mesh (with 3×3 inner points)

$$A = \begin{pmatrix} A_1 & I & & & \\ I & A_2 & I & & \\ & \ddots & \ddots & \ddots & \\ & & I & A_{NJ-1} & I \\ & & & I & A_{NJ} \end{pmatrix} \in \mathbb{R}^{(NI \cdot NJ) \times (NI \cdot NJ)} \quad (13)$$

with $A_i, I \in \mathbb{R}^{NI \times NI}$; here I is the unit matrix and

$$A_i = \begin{pmatrix} -4 & 1 & & & \\ 1 & -4 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -4 & 1 \\ & & & 1 & -4 \end{pmatrix} \quad i = 1, \dots, NJ \quad (14)$$

This means the task to solve the Poisson equation numerically leads us to the problem to find the solution of a system of linear equations:

$$A \vec{u} = \vec{b} \quad (15)$$

with

$$A \in \mathbb{R}^{(NI \cdot NJ) \times (NI \cdot NJ)} \quad \text{and} \quad \vec{u}, \vec{b} \in \mathbb{R}^{(NI \cdot NJ)} \quad (16)$$

The right hand side $\vec{b}$ contains the $f_{i,j}$ of the differential equations as well as the Dirichlet boundary condition.

For the solution of these coupled linear equations many well-known numerical algorithms are available. We will focus here on the classic but very simple Jacobi algorithm.

2.2 The Jacobi Algorithm for Systems of Linear Equations

Suppose

$$A = D - L - U \quad (17)$$

is a decomposition of the matrix A , where D is the diagonal sub-matrix, $-L$ is the strict lower triangular part and $-U$ the strict upper triangular part. Then for the system of linear equations holds

$$A \vec{u} = \vec{b} \Leftrightarrow (D - L - U) \vec{u} = \vec{b} \Leftrightarrow D \vec{u} = (L + U) \vec{u} + \vec{b} \Leftrightarrow \quad (18)$$

$$\vec{u} = D^{-1}(L + U) \vec{u} + D^{-1} \vec{b} \quad \text{if } D^{-1} \text{ exists.} \quad (19)$$

From Eq. (19) follows the iteration rule (for D non-singular)

$$\vec{u}^{(k)} = D^{-1}(L + U) \vec{u}^{(k-1)} + D^{-1} \vec{b} \quad \text{with } k = 1, 2, \dots \quad (20)$$

This iterative procedure is known as **Jacobi** or **total-step method**. The second name is motivated by the fact that the next iteration is calculated **only** from the values of the unknowns of the last iteration. There are other schemes, e.g. Gauß-Seidel algorithm, which depend on old **and** the current iteration of the unknowns!

The corresponding pseudo code for the serial Jacobi algorithm is given here:

Jacobi algorithm

Choose an initial vector $\vec{u}^{(0)} \in \mathbb{R}^n$

For $k = 1, 2, \dots$

For $i = 1, 2, \dots, n$

$$u_i^{(k)} = \frac{1}{a_{ii}} \left(b_i - \sum_{\substack{j=1 \\ j \neq i}}^n a_{ij} u_j^{(k-1)} \right)$$

The Poisson equation (9) discretized with the **5-point stencil** results in the following iteration procedure

$$\begin{pmatrix} u_1 \\ \vdots \\ \vdots \\ \vdots \\ u_N \end{pmatrix}^{(k)} = -\frac{1}{4} \begin{pmatrix} A'_1 & -I & & & \\ -I & A'_2 & -I & & \\ & \ddots & \ddots & \ddots & \\ & & -I & A'_{NJ-1} & -I \\ & & & -I & A'_{NJ} \end{pmatrix} \begin{pmatrix} u_1 \\ \vdots \\ \vdots \\ \vdots \\ u_N \end{pmatrix}^{(k-1)} - \frac{1}{4} \begin{pmatrix} b_1 \\ \vdots \\ \vdots \\ \vdots \\ b_N \end{pmatrix} \quad (21)$$

with $N = NI \cdot NJ$ and

$$A'_i = \begin{pmatrix} 0 & -1 & & \\ -1 & 0 & -1 & \\ & \ddots & \ddots & \ddots \\ & & -1 & 0 & -1 \\ & & & -1 & 0 \end{pmatrix} \quad i = 1, \dots, NJ \quad (22)$$

This can be seen easily by application of the Jacobi matrix decomposition on the coefficient matrix given by Eqs. (13) and (14). The pseudo code for this special case is shown here

Jacobi algorithm for the Poisson equation

Choose initial vector $\vec{u}^{(0)} \in \mathbb{R}^N$

For $k = 1, 2, \dots$

For $j = 0, 1, \dots, NJ - 1$

For $i = 0, 1, \dots, NI - 1$

$$u_{i,j}^{(k)} = \frac{1}{4} \left(u_{i,j-1}^{(k-1)} + u_{i-1,j}^{(k-1)} + u_{i,j+1}^{(k-1)} + u_{i+1,j}^{(k-1)} - h^2 f_{i,j} \right)$$

2.3 Parallelization of the Jacobi Algorithm

The numerical treatment of the Poisson equation led us to the task to solve a system of linear equations. We introduced the Jacobi algorithm as a simple method to calculate this solution and presented the corresponding serial pseudo code. Now the next step is to discuss strategies **how to implement the Jacobi algorithm on a parallel computer**.

The important point about the Jacobi (total-step) algorithm one has to remember is that the calculation of the new iteration only depends on the values of the unknowns from the **last iteration** as can be seen for instance from Eq. (21). As a consequence the processors of a parallel computer can calculate the new iteration of the unknowns simultaneously, supposed each unknown is assigned to its own processor. This makes the parallelization of the Jacobi algorithm quite easy compared to other methods with more complicated dependencies between different iterations.

Usually the number of unknowns is much larger than the number of available processors. Thus some / many unknowns have to be assigned to one processor, i.e. for our example: the inner points of Ω (Figure 1) are distributed to the available processors. With other words the **domain Ω** is **decomposed** according to a suitable strategy.

The criteria for a “suitable” strategy are

- **load balance**, i.e. same / similar number of unknowns for each processor
- minimization of the **communication** between the processors, i.e. the dependency on unknowns stored on other processors (within one iteration step!) is reduced

For our example, the Poisson equation in two dimensions, a reasonable domain decomposition is shown in Figure 5: Each processor “owns” a domain of the same size, i.e. each

P_{13}	P_{14}	P_{15}	P_{16}
P_9	P_{10}	P_{11}	P_{12}
P_5	P_6	P_7	P_8
P_1	P_2	P_3	P_4

Figure 5. Domain decomposition of a square Ω with 16 processors

P_i “owns” the same number of points. Furthermore the ratio area to edges of each square and consequently the ratio between the number of inner points (no dependency on points “owned” by other processors) to the number of points near the boundary is rather good. This point can be seen even better from Figure 6.

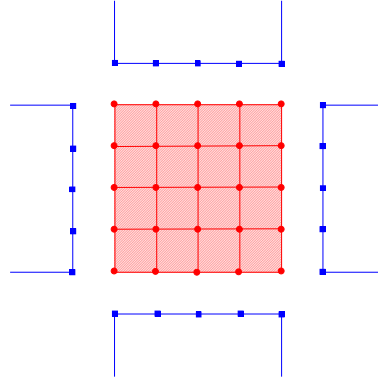


Figure 6. Dependency of the unknowns of processor P_i (red) on values stored on the neighbors (blue)

In Fig. 6 the points / corresponding unknowns of processor P_i are represented by red circles, whereas the blue squares depict the **ghost points**, i.e. points stored on other processors which are required for the calculation of the next iteration of the unknowns on processor P_i .

The dependencies / ghost points shown in Fig. 6 are a result of the 5-point stencil (see Fig. 3) originating from the Laplace operator in Eq. (1). Thus the domain decomposition of choice might differ for other differential equations or other discretization schemes, e.g. finite elements.

Due to the dependencies between the unknowns “owned” by different processors it is clear that the parallelization of the Jacobi algorithm has to introduce statements which will take care of the communication between the processors. One portable way to handle the communication is the widely used **Message Passing Interface (MPI)**¹⁷ library. The pseudo code of the parallel Jacobi algorithm is given here:

Parallel Jacobi algorithm

*Choose initial values for the **own mesh points** and the **ghost points***

Choose initial Precision (e.g. Precision = 10^{10})

While Precision > ε (e.g. $\varepsilon = 10^{-5}$)

1. *Calculate next iteration for the **own domain***
2. *Send the new iteration **on boundary of domain** to **neighboring processors***
3. *Receive the new iteration for the **ghost points***
4. *Calculate Precision = $\|A\vec{u}^{(k)} - \vec{b}\|$*

End While

The steps 2 and 3 show the extension of the serial Jacobi algorithm by *Send* and *Receive* statements. This is of course only one very simple way to implement such a communication with the four neighboring processors. In real applications one will look for more efficient communication patterns.

Step 4 requires implicitly **global communication**, because the vector $\vec{u}^{(k)}$ holding the approximate solution of the system of linear equations is distributed over all processors. As soon as the required precision of the solution is achieved the iteration stops.

3 Vibration of a Membrane

The vibration of a homogeneous membrane is governed by the time-dependent partial differential equation¹⁸

$$\frac{\partial^2 v}{\partial t^2} = \Delta v \quad (23)$$

In order to solve this equation we make a separation ansatz for the time and spatial variables:

$$v(x, y, t) = u(x, y) g(t) \quad (24)$$

By insertion of Eq. (24) into Eq. (23) one immediately obtains

$$g(t) \Delta u(x, y) = u(x, y) g''(t) \Leftrightarrow \quad (25)$$

$$\frac{\Delta u(x, y)}{u(x, y)} = \frac{g''(t)}{g(t)} \quad (26)$$

The left side of Eq. (26) is independent of t , the right side of x, y . Therefore both sides must be equal to a constant $-\lambda$

$$\frac{\Delta u(x, y)}{u(x, y)} = \frac{g''(t)}{g(t)} = -\lambda \Leftrightarrow \quad (27)$$

$$\Delta u(x, y) = -\lambda u(x, y) \quad \text{and} \quad g''(t) = -\lambda g(t) \quad (28)$$

The differential equation for $g(t)$ can be solved easily with the usual ansatz (a linear combination of trigonometric functions).

In the following we want to solve the spatial partial differential equation

$$\Delta u(x, y) = -\lambda u(x, y) \quad (29)$$

numerically. In section 2.1 we presented the discretization of the Poisson equation in two dimensions. In order to allow a re-use of the results derived there, we will calculate the solution of Eq. (29) for a **rectangular** membrane / domain.

Furthermore we choose for simplicity the Dirichlet boundary condition

$$u(x, y) = 0 \quad \text{for } (x, y) \in \partial\Omega \quad (30)$$

Using the same discretization for the Laplace operator and lexicographical numbering of the mesh points / unknowns as in section 2.1 one can see easily that Eq. (29) leads to the **eigenvalue problem**

$$A \vec{u} = -\lambda \vec{u} \quad (31)$$

where the matrix A is given by Eqs. (13) and (14).

In section 2 we presented a simple algorithm for the solution of the system of linear equations and discussed the parallelization by hand. For the eigenvalue problem we choose a different strategy: We make use of a widely used parallel library, namely the **ScaLAPACK** library.

3.1 Parallel Solution Using the ScaLAPACK Library

The largest and most flexible public domain library with linear algebra routines for distributed memory parallel systems up to now is ScaLAPACK¹⁰. Within the ScaLAPACK project many LAPACK routines were ported to distributed memory computers using MPI.

The basic routines of ScaLAPACK are the **PBLAS** (Parallel Basic Linear Algebra Subroutines)¹⁹. They contain parallel versions of the BLAS which are parallelized using **BLACS** (Basic Linear Algebra Communication Subprograms)²⁰ for communication and sequential BLAS for computation. Thus the PBLAS deliver very good performance on most parallel computers.

ScaLAPACK contains direct parallel solvers for dense linear systems (LU and Cholesky decomposition), linear systems with band matrices as well as parallel routines for the solution of linear least squares problems and for singular value decomposition.

Furthermore there are several different routines for the solution of the full symmetric eigenproblem. We will focus in the following on a **simple driver routine** using the QR-algorithm, which computes all eigenvalues and optionally all eigenvectors of the matrix.

Besides this there are other eigensolvers available which are implementations of other algorithms, e.g. a divide-and-conquer routine; an additional expert driver allows to choose a range of eigenvalues and optionally eigenvectors to be computed.

For performance and load balancing reasons ScaLAPACK uses a **two-dimensional block cyclic distribution** for full matrices (see ScaLAPACK Users' Guide)¹³:

First the matrix is divided into blocks of size $MB \times NB$, where MB and NB are the number of rows and columns per block, respectively. These blocks are then uniformly distributed across the $MP \times NP$ **rectangular processor grid** in a cyclic manner. As a result, each process owns a collection of blocks. Figure 7 shows the distribution of a (9×9) matrix subdivided into blocks of size (3×2) distributed across a (2×2) processor grid.

	0		1		0		1		0
0	a_{11}	a_{12}	a_{13}	a_{14}	a_{15}	a_{16}	a_{17}	a_{18}	a_{19}
	a_{21}	a_{22}	a_{23}	a_{24}	a_{25}	a_{26}	a_{27}	a_{28}	a_{29}
	a_{31}	a_{32}	a_{33}	a_{34}	a_{35}	a_{36}	a_{37}	a_{38}	a_{39}
1	a_{41}	a_{42}	a_{43}	a_{44}	a_{45}	a_{46}	a_{47}	a_{48}	a_{49}
	a_{51}	a_{52}	a_{53}	a_{54}	a_{55}	a_{56}	a_{57}	a_{58}	a_{59}
	a_{61}	a_{62}	a_{63}	a_{64}	a_{65}	a_{66}	a_{67}	a_{68}	a_{69}
0	a_{71}	a_{72}	a_{73}	a_{74}	a_{75}	a_{76}	a_{77}	a_{78}	a_{79}
	a_{81}	a_{82}	a_{83}	a_{84}	a_{85}	a_{86}	a_{87}	a_{88}	a_{89}
	a_{91}	a_{92}	a_{93}	a_{94}	a_{95}	a_{96}	a_{97}	a_{98}	a_{99}

Figure 7. Block cyclic 2D distribution of a (9×9) matrix subdivided into (3×2) blocks on a (2×2) processor grid. The numbers outside the matrix indicate processor row and column indices, respectively.

ScaLAPACK as a parallel successor of LAPACK attempts to leave the calling sequence of the subroutines unchanged as much as possible in comparison to the corresponding sequential subroutine from LAPACK. The user should have to change only a few parameters in the calling sequence to use ScaLAPACK routines instead of LAPACK routines.

Therefore ScaLAPACK uses so-called **descriptors**, which are integer arrays containing all necessary information about the **distribution of the matrix**. This descriptor appears in the calling sequence of the parallel routine instead of the leading dimension of the matrix in the sequential one.

For example the sequential simple driver **DSYEV** from LAPACK for the computation of **all eigenvalues** and (optionally) eigenvectors of a **real symmetric** ($N \times N$) matrix A has the following calling sequence²¹:

```
...  
CALL DSYEV( JOBZ, UPLO, N, A, LDA, W, WORK, LWORK, INFO )
```

where JOBZ and UPLO are characters indicating whether to compute eigenvectors, and whether the lower or the upper triangular part of the matrix A is provided. LDA is the leading dimension of A and W is the array of eigenvalues of A . The other variables are used as workspace and for error report.

The corresponding ScaLAPACK routine **PDSYEV** is called as follows¹³:

```
...  
      CALL PDSYEV ( JOBZ, UPLO, N, A, IA, JA, DESCA,  
$              W, Z, IZ, JZ, DESCZ, WORK, LWORK, INFO )  
...
```

As one can see the leading dimension LDA of the LAPACK call is substituted by the indices IA and JA and the descriptor DESCA. IA and JA indicate the start position of the **global matrix** (usually IA, JA = 1, but in cases where the global matrix is a sub-matrix of a larger matrix IA, JA $\neq$ 1 might occur), whereas DESCA contains all information regarding the distribution of the global matrix. The parameters IZ, JZ, and DESCZ provide the same information for Z, the matrix of the eigenvectors calculated by PDSYEV.

In order to use the ScaLAPACK routine the user has to distribute his system matrix in the way required by ScaLAPACK. Thus the user has to setup the **processor grid** by initializing MP, the number of processor rows, and NP, the number of processor columns. Furthermore one has to choose a suitable blocking of the matrix, i.e. MB and NB. For many routines, especially for the eigenvalue solvers and the Cholesky decomposition, MB=NB is mandatory. (Since MB and NB are crucial for the performance of the solver, one has to use these parameters with care.²²) Further details on the two-dimensional block cyclic distribution of the matrix A given by Eqs. (13) and (14) can be found in the appendix.

Once the matrix has been distributed to the processors, the calculation of the eigenvalues and corresponding eigenvectors for the vibration of the rectangular membrane (Eq. (31)) can be calculated easily by one call of the routine **PDSYEV**. Please note that the matrix of the eigenvectors Z, is distributed to the processors; thus if necessary, e.g. for output, it is again the task of the user to collect the different local data and to generate the global matrix.

4 Conclusion

In this contribution we presented two examples for the numerical treatment of partial differential equations using parallel algorithms / computers. Both problems were tackled with different strategies: The first example has been solved by means of the simple Jacobi algorithm and a suitable parallelization scheme was discussed. In the second case the parallel calculation has been performed with the help of the ScaLAPACK library.

The pros and cons of the different strategies are obvious. If a suitable parallel library is available and a reorganization of the application software according to the complex data distribution schemes of the libraries is possible, the parallel routines from the library will provide a robust numerical solution with fair or even good performance.

Otherwise the user has to choose a parallelization scheme which best fits his specific application problem and he has to implement the necessary algorithms himself; in order to improve the single processor performance it is still recommendable to use serial library routines, e.g. from BLAS or LAPACK, wherever possible!

Appendix

In section 3.1 some information on the **two-dimensional block cyclic distribution** of the data used by **ScaLAPACK** has been given. In this appendix we will discuss this issue in greater detail.

In Fig. 8 a code fragment is shown which distributes the $N \times N$ matrix given by Eqs. (13) and (14) according to the ScaLAPACK data scheme with block sizes $NB=MB$ to an $MP \times NP$ processor grid. Inclusion of this fragment into a parallel program allows the calculation of the eigenvalues and eigenvectors using the routine **PDSYEV**:

```
...  
      CALL PDSYEV ( JOBZ, UPLO, N, A, IA, JA, DESCA,  
$           W, Z, IZ, JZ, DESCZ, WORK, LWORK, INFO )  
...
```

Notice that in the sequential as well as in the parallel routine the matrix **A** is destroyed. The difference is that in the sequential case if the eigenvectors are requested **A** is overwritten by the eigenvectors whereas in the parallel case the eigenvectors are stored to a separate matrix **Z**.

The matrix **Z** has to be allocated with the same local sizes as **A** and **DESCZ** is filled with the same values as **DESCA**. The size **LWORK** of the local workspace **WORK** can be found in the ScaLAPACK Users' Guide.

The four nested loops in Fig. 8 show how local and global indices can be computed from block sizes, the number of rows and columns in the processor grid and the processor coordinates. The conversion of global to local indices and vice versa is supported by some auxiliary routines in the **TOOLS** sub-library of ScaLAPACK. Here the routine **NUMROC** is used to calculate the number of rows / columns stored on the corresponding processor.

There is also a sub-library **REDIST** of ScaLAPACK which allows the redistribution of any two-dimensional block cyclically distributed matrix to any other block cyclic two-dimensional distribution. Thus if **A** was column cyclically distributed or if the eigenvectors have to be column cyclically distributed for further computations they can be redistributed

```

1  ! Create the MP * NP processor grid
2  CALL BLACS_GRIDINIT(ICTXT,'Row-major',MP,NP)
3  ! Find my processor coordinates MYROW and MYCOL
4  ! NPROW returns the same value as MP,
5  ! NPCOL returns the same value as NP
6  CALL BLACS_GRIDINFO(ICTXT,NPROW,NPCOL,MYROW,MYCOL)
7  ! Compute local dimensions with routine NUMROC from TOOLS
8  ! N is dimension of the matrix
9  ! NB is block size
10 MYNUMROWS = NUMROC(N,NB,MYROW,0,NPROW)
11 MYNUMCOLS = NUMROC(N,NB,MYCOL,0,NPCOL)
12 ! Local leading dimension of A,
13 ! Number of local rows of A
14 MXLLDA = MYNUMROWS
15 ! Allocate only the local part of A
16 ALLOCATE(A(MXLLDA,MYNUMCOLS))
17 ! Fill the descriptors, P0 and Q0 are processor coordinates
18 ! of the processor holding global element A(1,1)
19 CALL DESCINIT(DESCA,N,N,NB,NB,P0,Q0,ICTXT,MXLLDA,INFO)
20 ! Fill the local part of the matrix with data
21 do j = 1, MYNUMCOLS, NB ! Fill the local column blocks
22   do jj=1,min(NB,MYNUMCOLS-j+1) ! all columns of one block
23     jloc = j-1 + jj ! local column index
24     jglob = (j-1)*NPCOL + MYCOL*NB + jj ! global column index
25     do i = 1, MYNUMROWS, NB ! local row blocks in column
26       do ii=1,min(NB,MYNUMROWS-i+1) ! rows in row block
27         iloc = i-1 + ii ! local row index
28         iglob = (i-1)*NPROW + MYROW*NB + ii ! global row index
29         A(iloc,jloc) = 0
30         If (iglob==jglob) A(iloc,jloc)=-4
31         If (iglob==jglob+1.and.mod(jglob,NI)/=0) &
32                               A(iloc,jloc)=1
33         If (jglob==iglob+1.and.mod(iglob,NI)/=0) &
34                               A(iloc,jloc)=1
35         If (iglob==jglob+NI) A(iloc,jloc)=1
36         If (jglob==iglob+NI) A(iloc,jloc)=1
37       enddo
38     enddo
39   enddo
40 enddo

```

Figure 8. Code fragment which distributes the matrix given by Eqs. (13) and (14) according to ScaLAPACK (It is assumed that $MB=NB=NI$ and $N=NI \cdot NJ$).

by such a routine, as a column cyclic distribution is nothing else but a block cyclic two-dimensional distribution to a $1 \times \text{NPR}$ (with NPR = number of processors) grid with block size 1.

References

1. J. J. Dongarra, I. S. Duff, D. .C. Sorensen, and H. A. van der Vorst, *Numerical Linear Algebra for High-Performance Computers*, SIAM, Philadelphia (1998).
2. <http://gams.nist.gov/Classes.html> and especially <http://gams.nist.gov/serve.cgi/Class/D/>
3. B. Wilkinson and M. Allen, *Parallel Programming: Techniques and Applications using networked Workstations and Parallel Computers*, Pearson, Upper Saddle River (2005).
4. A. Grama, A. Gupta, G. Karypis, and V. Kumar, *Introduction to Parallel Computing*, Pearson, Harlow (2003).
5. M. Bücker, *Iteratively Solving Large Sparse Linear Systems on Parallel Computers in Quantum Simulation of Complex Many-Body Systems: From Theory to Algorithms*, J. Grotendorst et al. (Ed.), John von Neumann Institute for Computing, NIC Series Vol. **10**, 521-548 (2002)
<http://www.fz-juelich.de/nic-series/volume10/buecker.pdf>
6. B. Lang, *Direct Solvers for Symmetric Eigenvalue Problems in Modern Methods and Algorithms of Quantum Chemistry*, J. Grotendorst (Ed.), John von Neumann Institute for Computing, NIC Series Vol. **3**, 231-259 (2000).
<http://www.fz-juelich.de/nic-series/Volume3/lang.pdf>
7. B. Steffen, *Subspace Methods for Sparse Eigenvalue Problems in Modern Methods and Algorithms of Quantum Chemistry*, J. Grotendorst (Ed.), John von Neumann Institute for Computing, NIC Series Vol. **3**, 307-314 (2000).
<http://www.fz-juelich.de/nic-series/Volume3/steffen.pdf>
8. <http://www.netlib.org/blas/>
9. <http://www.netlib.org/lapack/>
10. <http://www.netlib.org/scalapack/>
11. <http://www.caam.rice.edu/software/ARPACK/>
12. <http://www-unix.mcs.anl.gov/petsc/petsc-as/>
13. L. S. Blackford, J. Choi, A. Cleary et al., *ScaLAPACK Users' Guide*, SIAM, Philadelphia (1997).
14. G. Ahlefeld, I. Lenhardt, and H. Obermaier, *Parallele numerische Algorithmen*, Springer, Berlin (2002).
15. J. M. Ortega, *Introduction to parallel and vector solution of linear systems*, Plenum Press, New York (1988).
16. J. Zhu, *Solving partial differential equations on parallel computers*, World Scientific, Singapore (1994).
17. W. Gropp, E. Lusk, and A. Skjellum, *Using MPI : Portable Parallel Programming with the Message-Passing Interface*, MIT Press, Cambridge (1994).
18. R. Courant and D. Hilbert, *Methods of Mathematical Physics*, Volume I, Interscience Publishers, New York (1953).
19. http://www.netlib.org/scalapack/pblas_qref.html

20. J. J. Dongarra and R. C. Whaley, *A User's Guide to the BLACS*, LAPACK Working Note **94** (1997).
<http://www.netlib.org/lapack/lawns/lawn94.ps>
21. E. Anderson, Z. Bai, C. Bischof et al., *LAPACK Users' Guide, Second Edition*, SIAM, Philadelphia (1995).
22. I. Gutheil, *Basic Numerical Libraries for Parallel Systems* in Modern Methods and Algorithms of Quantum Chemistry, J. Grotendorst (Ed.), John von Neumann Institute for Computing, NIC Series Vol. **3**, 47-65 (2000).
<http://www.fz-juelich.de/nic-series/Volume3/gutheil.pdf>

Already published:

**Modern Methods and Algorithms of Quantum Chemistry -
Proceedings**

Johannes Grotendorst (Editor)

Winter School, 21 - 25 February 2000, Forschungszentrum Jülich

NIC Series Volume 1

ISBN 3-00-005618-1, February 2000, 562 pages

out of print

**Modern Methods and Algorithms of Quantum Chemistry -
Poster Presentations**

Johannes Grotendorst (Editor)

Winter School, 21 - 25 February 2000, Forschungszentrum Jülich

NIC Series Volume 2

ISBN 3-00-005746-3, February 2000, 77 pages

out of print

**Modern Methods and Algorithms of Quantum Chemistry -
Proceedings, Second Edition**

Johannes Grotendorst (Editor)

Winter School, 21 - 25 February 2000, Forschungszentrum Jülich

NIC Series Volume 3

ISBN 3-00-005834-6, December 2000, 638 pages

out of print

**Nichtlineare Analyse raum-zeitlicher Aspekte der
hirnelektrischen Aktivität von Epilepsiepatienten**

Jochen Arnold

NIC Series Volume 4

ISBN 3-00-006221-1, September 2000, 120 pages

**Elektron-Elektron-Wechselwirkung in Halbleitern:
Von hochkorrelierten kohärenten Anfangszuständen
zu inkohärentem Transport**

Reinhold Löwenich

NIC Series Volume 5

ISBN 3-00-006329-3, August 2000, 146 pages

**Erkennung von Nichtlinearitäten und
wechselseitigen Abhängigkeiten in Zeitreihen**

Andreas Schmitz

NIC Series Volume 6

ISBN 3-00-007871-1, May 2001, 142 pages

**Multiparadigm Programming with Object-Oriented Languages -
Proceedings**

Kei Davis, Yannis Smaragdakis, Jörg Striegnitz (Editors)
Workshop MPOOL, 18 May 2001, Budapest
NIC Series Volume 7
ISBN 3-00-007968-8, June 2001, 160 pages

**Europhysics Conference on Computational Physics -
Book of Abstracts**

Friedel Hossfeld, Kurt Binder (Editors)
Conference, 5 - 8 September 2001, Aachen
NIC Series Volume 8
ISBN 3-00-008236-0, September 2001, 500 pages

NIC Symposium 2001 - Proceedings

Horst Rollnik, Dietrich Wolf (Editors)
Symposium, 5 - 6 December 2001, Forschungszentrum Jülich
NIC Series Volume 9
ISBN 3-00-009055-X, May 2002, 514 pages

**Quantum Simulations of Complex Many-Body Systems:
From Theory to Algorithms - Lecture Notes**

Johannes Grotendorst, Dominik Marx, Alejandro Muramatsu (Editors)
Winter School, 25 February - 1 March 2002, Rolduc Conference Centre,
Kerkrade, The Netherlands
NIC Series Volume 10
ISBN 3-00-009057-6, February 2002, 548 pages

**Quantum Simulations of Complex Many-Body Systems:
From Theory to Algorithms- Poster Presentations**

Johannes Grotendorst, Dominik Marx, Alejandro Muramatsu (Editors)
Winter School, 25 February - 1 March 2002, Rolduc Conference Centre,
Kerkrade, The Netherlands
NIC Series Volume 11
ISBN 3-00-009058-4, February 2002, 194 pages

**Strongly Disordered Quantum Spin Systems in Low Dimensions:
Numerical Study of Spin Chains, Spin Ladders and
Two-Dimensional Systems**

Yu-cheng Lin
NIC Series Volume 12
ISBN 3-00-009056-8, May 2002, 146 pages

**Multiparadigm Programming with Object-Oriented Languages -
Proceedings**

Jörg Striegnitz, Kei Davis, Yannis Smaragdakis (Editors)
Workshop MPOOL 2002, 11 June 2002, Malaga
NIC Series Volume 13
ISBN 3-00-009099-1, June 2002, 132 pages

**Quantum Simulations of Complex Many-Body Systems:
From Theory to Algorithms - Audio-Visual Lecture Notes**

Johannes Grotendorst, Dominik Marx, Alejandro Muramatsu (Editors)
Winter School, 25 February - 1 March 2002, Rolduc Conference Centre,
Kerkrade, The Netherlands
NIC Series Volume 14
ISBN 3-00-010000-8, November 2002, DVD

Numerical Methods for Limit and Shakedown Analysis

Manfred Staat, Michael Heitzer (Eds.)
NIC Series Volume 15
ISBN 3-00-010001-6, February 2003, 306 pages

**Design and Evaluation of a Bandwidth Broker that Provides
Network Quality of Service for Grid Applications**

Volker Sander
NIC Series Volume 16
ISBN 3-00-010002-4, February 2003, 208 pages

**Automatic Performance Analysis on Parallel Computers with
SMP Nodes**

Felix Wolf
NIC Series Volume 17
ISBN 3-00-010003-2, February 2003, 168 pages

**Haptisches Rendern zum Einpassen von hochaufgelösten
Molekülstrukturdaten in niedrigaufgelöste
Elektronenmikroskopie-Dichteverteilungen**

Stefan Birmanns
NIC Series Volume 18
ISBN 3-00-010004-0, September 2003, 178 pages

Auswirkungen der Virtualisierung auf den IT-Betrieb

Wolfgang Gürich (Editor)
GI Conference, 4 - 5 November 2003, Forschungszentrum Jülich
NIC Series Volume 19
ISBN 3-00-009100-9, October 2003, 126 pages

NIC Symposium 2004

Dietrich Wolf, Gernot Münster, Manfred Kremer (Editors)
Symposium, 17 - 18 February 2004, Forschungszentrum Jülich
NIC Series Volume 20
ISBN 3-00-012372-5, February 2004, 482 pages

**Measuring Synchronization in Model Systems and
Electroencephalographic Time Series from Epilepsy Patients**

Thomas Kreutz

NIC Series Volume 21

ISBN 3-00-012373-3, February 2004, 138 pages

**Computational Soft Matter: From Synthetic Polymers to Proteins -
Poster Abstracts**

Norbert Attig, Kurt Binder, Helmut Grubmüller, Kurt Kremer (Editors)

Winter School, 29 February - 6 March 2004, Gustav-Stresemann-Institut Bonn

NIC Series Volume 22

ISBN 3-00-012374-1, February 2004, 120 pages

**Computational Soft Matter: From Synthetic Polymers to Proteins -
Lecture Notes**

Norbert Attig, Kurt Binder, Helmut Grubmüller, Kurt Kremer (Editors)

Winter School, 29 February - 6 March 2004, Gustav-Stresemann-Institut Bonn

NIC Series Volume 23

ISBN 3-00-012641-4, February 2004, 440 pages

**Synchronization and Interdependence Measures and their Applications
to the Electroencephalogram of Epilepsy Patients and Clustering of Data**

Alexander Kraskov

NIC Series Volume 24

ISBN 3-00-013619-3, May 2004, 106 pages

High Performance Computing in Chemistry

Johannes Grotendorst (Editor)

Report of the Joint Research Project:

High Performance Computing in Chemistry - HPC-Chem

NIC Series Volume 25

ISBN 3-00-013618-5, December 2004, 160 pages

**Zerlegung von Signalen in unabhängige Komponenten:
Ein informationstheoretischer Zugang**

Harald Stögbauer

NIC Series Volume 26

ISBN 3-00-013620-7, April 2005, 110 pages

Multiparadigm Programming 2003

Joint Proceedings of the

**3rd International Workshop on Multiparadigm Programming with
Object-Oriented Languages (MPOOL'03)**

and the

**1st International Workshop on Declarative Programming in the
Context of Object-Oriented Languages (PD-COOL'03)**

Jörg Striegnitz, Kei Davis (Editors)

NIC Series Volume 27

ISBN 3-00-016005-1, July 2005, 300 pages

**Integration von Programmiersprachen durch strukturelle Typanalyse
und partielle Auswertung**

Jörg Striegnitz

NIC Series Volume 28

ISBN 3-00-016006-X, May 2005, 306 pages

**OpenMolGRID - Open Computing Grid for Molecular Science
and Engineering**

Final Report

Mathilde Romberg (Editor)

NIC Series Volume 29

ISBN 3-00-016007-8, July 2005, 86 pages

All volumes are available online at <http://www.fz-juelich.de/nic-series/>.