

# A 6    **Linear Scaling Electronic Structure Methods**

S. Goedecker  
Institut für Physik  
Universität Basel

Reprinted excerpts and figures with permission from Stefan Goedecker, Rev. Mod. Phys. Vol. 71, 1085 (1999). Copyright 1999 by American Physical Society.

## **Contents**

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Locality in Quantum Mechanics</b>	<b>4</b>
<b>3</b>	<b>Basic strategies for <math>O(N)</math> scaling</b>	<b>12</b>
3.1	The Chebychev Fermi Operator Expansion . . . . .	14
3.2	The Rational Fermi Operator Expansion . . . . .	21
3.3	The Fermi Operator Projection Method . . . . .	22
3.4	The Density Matrix Minimization approach . . . . .	23
3.5	The Optimal Basis Density Matrix Minimization method . . . . .	26
<b>4</b>	<b>Comparison of the basic methods</b>	<b>29</b>
4.1	Small basis sets . . . . .	29
4.2	Large basis sets . . . . .	32

# 1 Introduction

The exact quantum mechanical equations for many-electron systems are highly intricate. Any attempt to solve these equations analytically for real systems is doomed to fail. Numerical methods such as Configuration Interaction based methods or Quantum Monte Carlo methods can in principle solve these many-electron equations but because of the extremely high numerical effort required, their applicability is rather limited in practice.

The bulk of all practical applications is therefore done within various independent-electron approximations such as the Hartree-Fock method, Density Functional methods or Tight Binding methods. Even these approximate quantum mechanical equations are still fairly complicated and in general not solvable by analytical methods. Finding efficient algorithms to solve the many-electron problem numerically within any of these approximations is imperative for the applicability of quantum mechanics to physics as well as to chemistry and materials science. Due to efforts in the past satisfactory algorithms are now available and computational electronic structure methods are making very important contributions to our understanding of matter at the microscopic level. The 1998 Nobel prize for W. Kohn and J. Pople is a landmark sign of the importance of this approach.

Due to the constant increase in computer power and due to algorithmic improvements the importance of computational methods is growing further. Whereas computational methods nowadays mainly supplement experimentally obtained information, they are expected to increasingly supersede this information.

This article will concentrate on methods that allow us to calculate the total energy within various independent-electron methods for large systems. Practically all physical observables can be obtained from the total energy, for instance in the form of derivatives with respect to certain external parameters. The reason why large systems containing many atoms are accessible with these algorithms is their linear scaling with respect to the number of atoms.

Traditional electronic structure algorithms calculate eigenstates associated with discrete energy levels. The reason for this is probably historical since the prediction of these experimentally observed levels was the first big success of quantum mechanics. The disadvantage of this approach is that it leads to a diagonalization problem which has a cubic scaling in the computational effort. Direct diagonalization, which was the standard approach in the early days of the computational electronic structure era, has a cubic scaling with respect to the size of the Hamiltonian matrix, i.e. with respect to the number of basis functions  $M_b$ . Iterative diagonalization schemes, preconditioned conjugate gradient minimizations and the Car-Parrinello method for molecular dynamics simulations were a big algorithmic progress because of their improved scaling behavior. Their scaling was not any more proportional to the cube of the the number of basis functions but grew only like  $M_b \log(M_b)$  if plane waves were used as a basis set. Nevertheless these methods still have a cubic scaling with respect to the number of atoms  $N_{at}$ , which comes from the orthogonality requirement of the wavefunctions. The reason why this orthogonalization step scales cubically can easily be seen. As the system grows, each wavefunction extends over a larger volume and has therefore to be represented by a larger basis set resulting in a longer vector. At the same time there are more such wavefunctions and each wavefunction has to be orthogonalized to all the others. Thus there are 3 factors that grow linearly, resulting in the postulated cubic behavior. The computer time  $T_{CPU}$  required to do the calculation is thus given by

$$T_{CPU} = c_3 N_{at}^3, \quad (1)$$

where  $c_3$  is a prefactor. It has to be pointed out that Equation (1) gives only the asymptotic scal-

ing behavior. Within Density Functional and Hartree Fock calculations there are other terms with a lower scaling which dominate for system sizes of less than a few hundred atoms due to their large prefactor. In the case of plane wave type calculations the Fast Fourier transformations necessary for the application of the potential to the wavefunctions consume most of the computational time for small systems, in the case of calculations using Gaussian type orbitals it is the calculation of the Hartree potential. This cubic scaling is a major bottleneck nowadays since in many problems of practical interest one has to do electronic structure calculations for systems containing many (a few hundred or more) atoms. Evidently, cubic scaling means that if one doubles the number of atoms in the systems the required computer time will increase by a factor of eight. By enlarging the system one therefore rapidly reaches the limits of the most powerful computers.

So called  $O(N)$  or low complexity algorithms are therefore a logical next step of algorithmic progress since they exhibit linear scaling with respect to the number of atoms

$$T_{CPU} = c_1 N_{at} . \quad (2)$$

These methods offer thus the potential to calculate very large systems. The prefactors  $c_1$  and  $c_3$  depend on the approximation used for the many-electron problem. For a Density Functional calculation with a large basis set the prefactors are of course much larger than for a Tight Binding calculation, where the number of degrees of freedom per atom is much smaller. The prefactor  $c_1$  depends also on what  $O(N)$  method is used, but in general the prefactor  $c_1$  is always larger than  $c_3$  assuming that the same independent-electron approximation is used both in the traditional and  $O(N)$  version. There is therefore a so called cross over point. For system sizes smaller than the cross over point the traditional cubic scaling algorithms are faster, for larger systems the  $O(N)$  methods win. Tight Binding calculations are an ideal test environment for  $O(N)$  algorithms. Because of their rather small memory and CPU requirements one can easily treat systems comprising of a very large number of atoms and venture into regions beyond the cross over point. Contrary to what one might naively think, the importance of  $O(N)$  algorithms will also increase as computers get faster. Whereas at present it is difficult to access the cross over region situated at some 100 atoms using the Density Functional framework, this will be easy with faster computers and  $O(N)$  algorithms will be the algorithms of choice.

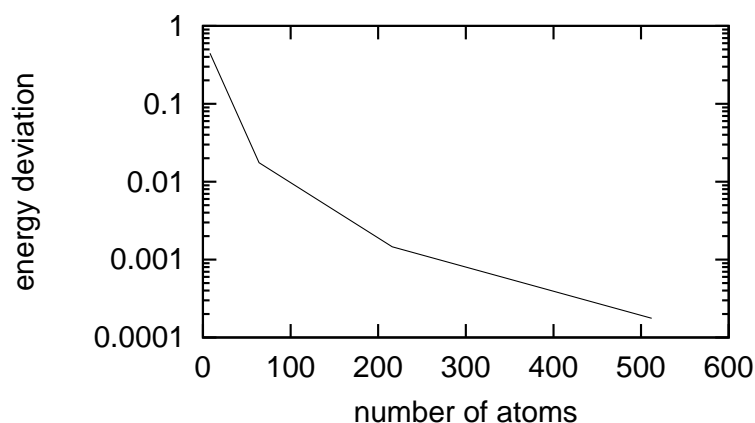
Even though  $O(N)$  algorithms contain many aspects of mathematics and computer science they have nevertheless deep roots in physics. Obtaining linear scaling is not possible by purely mathematical tricks but it is based on the understanding of the concept of locality in quantum mechanics. Conversely, the need of constructing  $O(N)$  algorithms was also an incentive to investigate locality questions more deeply, and has thus lead to a better understanding of this very fundamental concept. An algorithmic description of electronic structure in local terms can give a justification of the well established concepts of bonds and lone electron pairs in empirical chemistry.

Since  $O(N)$  algorithms are based on a certain subdivision of a big system into smaller subsystems, techniques developed in this context might also be helpful in reaching another important goal for treating large systems, namely combining electronic structure methods of different accuracy such as empirical Tight Binding and Density Functional theory in a single system.

## 2 Locality in Quantum Mechanics

Locality in Quantum Mechanics means that the properties of a certain observation region comprising one or a few atoms are only weakly influenced by factors that are spatially far away from this observation region. This fundamental characteristic of insulators is well established within independent-electron theories [1] and it can even be carried over into the many-electron framework [2].

Traditional chemistry is based on local concepts. Covalently bonded materials are described in terms of bonds and lone electron pairs. It is standard textbook knowledge that the properties of a bond are mainly determined by its immediate neighborhood. The decisive factors are what type of atoms and how many of them (the coordination number) are surrounding it. Second nearest neighbors and other more distant atoms have a very small influence. As an example let us look at the total energy of a hydrocarbon chain molecule  $C_nH_{2n+2}$ . In this case each  $CH_2$  subunit is from an energetical point of view practically an independent unit. As one adds one  $CH_2$  subunit, the energy increases by an amount which is nearly independent of the chain length. Already the insertion of a  $CH_2$  subunit into the smallest chain  $C_2H_6$  gives an energy gain which agrees within  $10^{-4}$  a.u. with the asymptotic value of the insertion energy for very long chains. This means that the electrons belonging to this inserted subunit already do not see any more the end of the chain for very short chain lengths. This example is a drastic illustration of a principle sometimes termed “nearsightedness” [3]. In other insulating materials the influence of the neighboring atoms decays slower. An example is shown in Figure (1), where the total energy per silicon atom is plotted as a function of the size of its crystalline environment.



**Fig. 1:** The deviation of the total energy per silicon atom from its asymptotic bulk value as a function of the size of the periodic volume in which it is embedded. The calculation was done with a Tight Binding scheme using exact diagonalization

Even in metallic systems, where the elementary bond concept is not any more valid, locality still exists. This is supported by the well known fact, that the total charge density in a metal is given with reasonable accuracy by the superposition of the atomic charge densities. Since atomic charge densities decay rapidly, this implies that the charge density at the midpoint of two neighboring atoms is mainly determined by the two closest atoms and very little by other more distant atoms. Another related example is given by V. Heine [1] who points out, that the magnetic moment of an iron atom, which is embedded in an iron-aluminum alloy differs by less than 5 % from the value for pure iron if the atoms are locally surrounded by only eight aluminum atoms.

This locality is not at all reflected in standard electronic structure calculations which are based on eigenorbitals extending over the whole system, making both the interpretation of the results more difficult and requiring unnecessary computational effort. The simplistic bond concepts of empirical chemistry are certainly not adequate for electronic structure calculations aiming at high accuracy. Nevertheless one might hope to incorporate some more general locality concepts into electronic structure calculation to make them both more intuitive and efficient. In the following we will therefore carefully examine the range of interactions in quantum mechanical systems.

Self-consistent electronic structure methods require essentially two steps. The calculation of the potential from the electronic charge distribution and the determination of the wavefunction for a given potential. In non-self-consistent calculations such as Tight Binding calculations, the first step is not needed.

The calculation of the potential consists usually of two parts, the exchange correlation potential, and the Coulomb potential. The exchange correlation potential is a purely local expression in Density Functional Theory and can therefore be calculated with linear scaling. In the Hartree Fock scheme one might first think that the exchange part is non-local, but a more profound examination reveals that it is local for insulators. The Coulomb potential on the other hand is very long range and needs proper treatment. A naive evaluation of the potential  $U$  arising from a charge distribution  $\rho$  by subdividing space into subvolumes  $\Delta V$  and summing over these subvolumes,

$$U(\mathbf{r}_i) = \sum_j \frac{\rho(\mathbf{r}_j)}{|\mathbf{r}_i - \mathbf{r}_j|} \Delta V ,$$

would result in a quadratic scaling since both indices  $i$  and  $j$  have to run over all grid points in the system. The Coulomb problem actually arises not only in the context of electronic structure calculations but also in classical calculations of Coulombic and gravitational systems such as galaxies of stars. Much effort has therefore been invested in this computational problem and several algorithms are known which solve the problem with linear scaling.

The more interesting and more difficult part is to assess the role of locality for a given external potential. The appropriate quantity to study this property is the density matrix. The one-particle density matrix  $F$  completely specifies our quantum mechanical system within the independent electron approximation and all quantities of interest can easily be calculated from it. The central quantities in any electronic structure calculation, the kinetic energy  $E_{kin}$ , the potential energy  $E_{pot}$  and the electronic charge density  $\rho$  are given by

$$E_{kin} = -\frac{1}{2} \int \nabla_{\mathbf{r}}^2 F(\mathbf{r}, \mathbf{r}')|_{\mathbf{r}=\mathbf{r}'} d\mathbf{r}' \quad (3)$$

$$E_{pot} = \int F(\mathbf{r}', \mathbf{r}') U(\mathbf{r}') d\mathbf{r}' \quad (4)$$

$$\rho(\mathbf{r}) = F(\mathbf{r}, \mathbf{r}) , \quad (5)$$

where  $U(\mathbf{r}')$  is the potential. A related quantity which will frequently be used throughout the article is the band structure energy  $E_{BS}$  defined as

$$E_{BS} = E_{kin} + E_{pot} \quad (6)$$

and the grand potential

$$\Omega = E_{BS} - \mu N_{el} , \quad (7)$$

where  $\mu$  is the chemical potential and  $N_{el}$  the number of electrons. Subtracting  $\mu N_{el}$  from  $E_{BS}$  leaves  $\Omega$  invariant under a constant potential offset. If one applies the shift ( $U(\mathbf{r}) \rightarrow U(\mathbf{r}) + const$ ) the potential energy will increase by  $N_{el} const$ . In order to conserve the total number of electrons,  $\mu$  also has to be shifted ( $\mu \rightarrow \mu + const$ ) and thus  $\Omega$  remains constant. Discretizing the Hamiltonian  $H$  which is the sum of the kinetic and potential energy as well as  $F$  with respect to a finite orthogonal basis  $\phi_i(\mathbf{r})$ ,  $i = 1, \dots, M_b$  one obtains

$$H_{i,j} = \int \phi_i^*(\mathbf{r}) \left( -\frac{1}{2} \nabla_{\mathbf{r}}^2 + U(\mathbf{r}) \right) \phi_j(\mathbf{r}) d\mathbf{r} \quad (8)$$

$$F_{i,j} = \int \int \phi_i^*(\mathbf{r}) F(\mathbf{r}, \mathbf{r}') \phi_j(\mathbf{r}') d\mathbf{r} d\mathbf{r}' \quad (9)$$

and the expressions for the central quantities become

$$E_{BS} = Tr[FH] \quad (10)$$

$$\Omega = Tr[F(H - \mu I)] \quad (11)$$

$$\rho(\mathbf{r}) = \sum_{i,j} F_{i,j} \phi_i(\mathbf{r}) \phi_j(\mathbf{r}), \quad (12)$$

where  $Tr$  denotes the trace. It follows from Equation (12) that the total number of electrons  $N_{el}$  in the system is given by

$$N_{el} = Tr[F]. \quad (13)$$

Evaluating the traces using the eigenfunctions  $\Psi_n$  of the Hamiltonian one obtains immediately the well known expressions for  $N_{el}$ ,  $E_{BS}$ ,  $\Omega$  and  $\rho$  within the context of conventional calculations which are based on diagonalization. Denoting the eigenvalues associated with the eigenfunctions  $\Psi_n$  by  $\epsilon_n$  one obtains

$$N_{el} = \sum_n f(\epsilon_n) \quad (14)$$

$$E_{BS} = \sum_n f(\epsilon_n) \epsilon_n \quad (15)$$

$$\Omega = \sum_n (f(\epsilon_n) - \mu) \epsilon_n = \sum_n f(\epsilon_n) \epsilon_n - \mu N_{el} \quad (16)$$

$$\rho(\mathbf{r}) = \sum_n f(\epsilon_n) \Psi_n^*(\mathbf{r}) \Psi_n(\mathbf{r}). \quad (17)$$

The function  $f$  is the the Fermi distribution

$$f(\epsilon) = \frac{1}{1 + \exp\left(\frac{\epsilon - \mu}{k_B T}\right)}, \quad (18)$$

where  $k_B$  is Boltzmann's constant and  $T$  the temperature. If we talk about temperature in this article, we always mean the electronic temperature since we are not considering the motion of the ionic degrees of freedom which might be associated with a different ionic temperature. In the expressions (14), (15), (16), and (17), as well as in the remainder of the whole article, we will use the convention, that all the subscripts indexing eigenvalues and eigenfunctions are combined orbital and spin indices, i.e. that we can put at most one electron in each orbital. This will eliminate bothering factors of 2. The usually relevant case of an unpolarized spin

restricted system can always easily be obtained by cutting into half all sums over these indices and multiplying by 2.

In terms of the Hamiltonian  $H$  the density matrix is defined as the following matrix functional

$$F = f(H) . \quad (19)$$

Since  $F$  is a matrix function of  $H$  it has the same eigenfunctions  $\Psi_n$  as  $H$

$$H\Psi_n = \epsilon_n\Psi_n \quad (20)$$

$$F\Psi_n = f(\epsilon_n)\Psi_n . \quad (21)$$

The density matrix can consequently be written as

$$F(\mathbf{r}, \mathbf{r}') = \sum_n f(\epsilon_n) \Psi_n^*(\mathbf{r}) \Psi_n(\mathbf{r}') , \quad (22)$$

where  $n$  runs over all the eigenstates of the Hamiltonian. From the functional form of the Fermi distribution it follows that the eigenvalues  $f(\epsilon_n)$  are always in the interval  $[0:1]$ . At zero temperature the density matrix of an insulating system containing  $N_{el}$  electrons will have  $N_{el}$  eigenvalues of value one, all others being zero. Thus the density matrix does not have full rank, but only rank  $N_{el}$ . Hence we can write it as

$$F(\mathbf{r}, \mathbf{r}') = \sum_{n=occ} \Psi_n^*(\mathbf{r}) \Psi_n(\mathbf{r}') , \quad (23)$$

where  $n$  runs now only over the  $N_{el}$  occupied states. It is easy to see that  $F(\mathbf{r}, \mathbf{r}')$  is a projection operator in this case

$$\int F(\mathbf{r}, \mathbf{r}'') F(\mathbf{r}'', \mathbf{r}') d\mathbf{r}'' = F(\mathbf{r}, \mathbf{r}') . \quad (24)$$

A new set of  $N_{el}$  eigenfunctions  $\Psi_n^{new}(\mathbf{r})$  can be obtained by any unitary transformation of all the  $N_{el}$  degenerate eigenfunctions  $\Psi_n(\mathbf{r})$  associated with eigenvalues one,

$$\Psi_n^{new}(\mathbf{r}) = \sum_{m=occ} U_{n,m} \Psi_m(\mathbf{r}) , \quad (25)$$

where  $U$  is a unitary  $N_{el}$  by  $N_{el}$  matrix. In the case of a crystalline periodic solids such a transformation can be used to generate the localized Wannier functions [4] from the extended eigenfunctions  $\Psi_n$ . We will refer to any set of orthogonal exponentially localized orbitals which can be used to represent the density matrix according to Equation (23) as Wannier functions. How to construct an optimally localized set of Wannier functions by the minimization of the total spread  $\sum_n \langle r^2 \rangle_n - \langle \mathbf{r} \rangle_n^2$  in a crystalline periodic solid has recently been shown by Marzari and Vanderbilt [5]. It has been well known in the chemistry community that sets of maximally localized orbitals give excellent insight into the bonding properties of systems. In addition to the spread criterion used by Marzari *et al.* there are still other criteria in common use in the chemistry community. They are all in a certain sense arbitrary, but usually lead to the same interpretation of the bonding properties.

The density matrix  $F(\mathbf{r}, \mathbf{r}')$  is a diagonally dominant operator, whose off-diagonal elements decay with increasing distance from the diagonal. The exact decay behavior depends on the

material. We will derive the decay properties within the theoretical framework of the description of periodic crystalline solids. For a periodic solid the density matrix is given by

$$\begin{aligned} F(\mathbf{r}, \mathbf{r}') &= \sum_n \frac{V}{(2\pi)^3} \int_{BZ} d\mathbf{k} f(\epsilon_n(\mathbf{k})) \Psi_{n,\mathbf{k}}^*(\mathbf{r}) \Psi_{n,\mathbf{k}}(\mathbf{r}') \\ &= \sum_n \frac{V}{(2\pi)^3} \int_{BZ} d\mathbf{k} f(\epsilon_n(\mathbf{k})) u_{n,\mathbf{k}}^*(\mathbf{r}) u_{n,\mathbf{k}}(\mathbf{r}') e^{i\mathbf{k}(\mathbf{r}' - \mathbf{r})}, \end{aligned} \quad (26)$$

where  $\Psi_{n,\mathbf{k}}(\mathbf{r}) = u_{n,\mathbf{k}}(\mathbf{r}) e^{i\mathbf{k}(\mathbf{r})}$  are the Bloch functions associated with the wave vector  $\mathbf{k}$  and band index  $n$ . The integral is taken over the Brillouin zone (BZ) and  $V$  is the volume of the real space primitive cell.

The Wannier functions  $W_n$  of the  $n$ -th band in an insulating crystal are defined in the usual way

$$W_n(\mathbf{r} - \mathbf{R}) = \frac{V}{(2\pi)^3} \int_{BZ} d\mathbf{k} e^{-i\mathbf{k}\mathbf{R}} \Psi_{n,\mathbf{k}}(\mathbf{r}). \quad (27)$$

The Wannier functions are not uniquely defined. One can construct a different set of Bloch functions by multiplying them with a phase factor,  $\Psi_{n,\mathbf{k}}(\mathbf{r}) \leftarrow e^{i\omega(\mathbf{k})} \Psi_{n,\mathbf{k}}(\mathbf{r})$ , where  $\omega(\mathbf{k})$  is an arbitrary function. This will obviously modify the Wannier functions. Further ambiguities arise in the case of degenerate bands. Because of these ambiguities in the construction of the Wannier functions it is advantageous to work with the density matrix where any phase factors cancel (Equation (26)) and where degeneracies do not cause any problems since one sums over all the occupied bands.

We will first discuss the decay properties of the density matrix in metallic systems. In this discussion we will assume that metals behave essentially like jellium and that exact results for jellium can be carried over to real metals.

The decay properties of the density matrix of a metallic system at zero temperature are well known (March). Because the integral in Equation (26) contains a discontinuity in the metallic case, the density matrix decays only algebraically with respect to the distance between  $\mathbf{r}$  and  $\mathbf{r}'$ . The decay is given by

$$F(\mathbf{r}, \mathbf{r}') \propto k_F \frac{\cos(k_F |\mathbf{r} - \mathbf{r}'|)}{|\mathbf{r} - \mathbf{r}'|^2}, \quad (28)$$

where the Fermi wave vector  $k_F$  is related to the valence electron density by  $\frac{N_{el}}{V} = \frac{k_F^3}{3\pi^2}$  in a non-spin-polarized system.

Introducing a finite electronic temperature  $T$  in a metal leads to a drastic change in this decay behavior. Instead of an algebraic decay one has a much faster exponential decay. As shown independently by Goedecker [8] and Ismail-Beigi and Arias [7], the decay at low temperatures is then given by

$$F(\mathbf{r}, \mathbf{r}') \propto k_F \frac{\cos(k_F |\mathbf{r} - \mathbf{r}'|)}{|\mathbf{r} - \mathbf{r}'|^2} \exp\left(-c \frac{k_B T}{k_F} |\mathbf{r} - \mathbf{r}'|\right), \quad (29)$$

where  $c$  is a constant on the order of 1. We thus find oscillatory behavior with an exponentially damped amplitude. The decay rate depends linearly on temperature and the oscillatory part is described by the wave vector  $k_F$ . In an insulator finite temperature plays no role as long as the thermal energy  $k_B T$  is much smaller than the gap, which is usually fulfilled.



Let us next discuss the important case of an insulator with a band gap  $\epsilon_{gap}$  at zero temperature. We will first present some numerical results, then we will put forward some arguments to explain the qualitative features of the density matrix and finally discuss in a more quantitative way the factors which determine the exact decay rate.

Numerical calculations of the density matrix or the related Wannier functions show an oscillatory behavior with a decaying amplitude. There is exactly one node per primitive cell and logarithmic plots of the amplitude clearly reveal an exponential decay. In the case of alkanes the decay of the density matrix calculated by the Hartree-Fock method has been studied and plotted on a logarithmic scale by Maslen *et al.* [9]. Interestingly, the decay depends also on the basis set used. Small low quality basis sets lead to a larger band gap and consequently to a faster decay of the density matrix.

Let us now make plausible the exponential decay of the density matrix. The demonstration is based on the fact, that one can express the Fourier components  $\epsilon_n(\mathbf{R})$  of the band energy  $\epsilon_n(\mathbf{k})$  through the Wannier functions  $W_n(\mathbf{r})$

$$\epsilon_n(\mathbf{R}) = \frac{V}{(2\pi)^3} \int_{BZ} \epsilon_n(\mathbf{k}) e^{-i\mathbf{k}\mathbf{R}} d\mathbf{k} = \frac{(2\pi)^3}{V} \int_{space} W_n^*(\mathbf{r}') H W_n(\mathbf{r}' - \mathbf{R}) d\mathbf{r}', \quad (30)$$

where  $\mathbf{R}$  is a Bravais lattice vector. Now it is known, that the band energy  $\epsilon_n(\mathbf{k})$  is an analytic function [4]. This is actually not surprising. The first and second derivatives of the band-structure have physical meaning since they are related to the electron velocity and effective mass. So it is to be expected that higher derivatives exist as well. Since the Fourier transform of an analytic function decays faster than algebraically there exists a decay constant  $\gamma$  and a normalization constant  $C$  such that

$$C e^{-\gamma R} \geq \epsilon_n(\mathbf{R}) = \frac{1}{V} \int_{space} W_n^*(\mathbf{r}') H W_n(\mathbf{r}' - \mathbf{R}) d\mathbf{r}' \quad (31)$$

It is reasonable to expect that  $H W_n(\mathbf{r})$  will behave similarly as  $W_n(\mathbf{r})$ . In particular we expect  $W_n(\mathbf{r})$  to be small whenever  $H W_n(\mathbf{r})$  is small. So we will just drop  $H$  in Equation (31). In addition we will define this modified integral not only for lattice vectors  $\mathbf{R}$  but for arbitrary vectors  $\mathbf{r}$  to obtain.

$$C e^{-\gamma r} \geq \frac{1}{V} \int_{space} W_n^*(\mathbf{r}') W_n(\mathbf{r}' - \mathbf{r}) d\mathbf{r}' \quad (32)$$

If Equation (32) holds, then one can use the mean value theorem to show that

$$\begin{aligned} C e^{-\gamma r} &\geq \frac{1}{V} \int_{space} W_n^*(\mathbf{r}') W_n(\mathbf{r}' - \mathbf{r}) d\mathbf{r}' \\ &= \frac{1}{V} \sum_{\mathbf{R}'} \int_{cell} W_n^*(\mathbf{r}' - \mathbf{R}') W_n(\mathbf{r}' - \mathbf{R}' - \mathbf{r}) d\mathbf{r}' \\ &= \sum_{\mathbf{R}'} W_n^*(\mathbf{s}(\mathbf{r}) - \mathbf{R}') W_n(\mathbf{s}(\mathbf{r}) - \mathbf{R}' - \mathbf{r}) \\ &= F(\mathbf{s}(\mathbf{r}), \mathbf{s}(\mathbf{r}) - \mathbf{r}) \end{aligned} \quad (33)$$

where the mean value  $\mathbf{s}(\mathbf{r})$  is a vector within the primitive cell. Assuming that the density matrix has the same order of magnitude within each cell one can neglect the dependence of  $\mathbf{s}$  on  $\mathbf{r}$  to obtain the final result

$$C e^{-\gamma r} \geq F(\mathbf{s}, \mathbf{s} - \mathbf{r}) \quad (34)$$

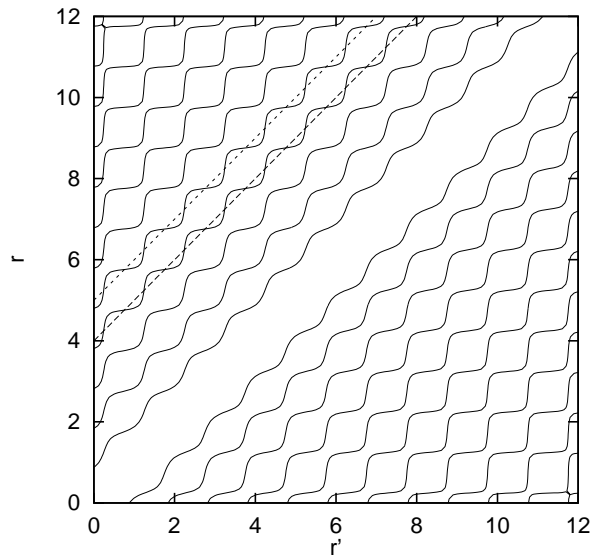
The numerically observed nodal structure of the density matrix can be motivated in a very similar way. Because of the orthogonality of the Wannier functions we have

$$0 = \int_{space} W_n^*(\mathbf{r}') W_n(\mathbf{r}' - \mathbf{R}) d\mathbf{r}' \quad (35)$$

for any non-zero lattice vector  $\mathbf{R}$ . Doing the same sequence of transformation as in Equation (33) one obtains

$$0 = F(s(\mathbf{R}), s(\mathbf{R}) - \mathbf{R}) \quad (36)$$

So there has to be one node in each cell. The numerically calculated nodal structure for a 1-dimensional model insulator is shown in Figure 2.



**Fig. 2:** The nodal structure of the density matrix for a 1-dimensional model insulator with a bandwidth of 4 a.u. and a band gap of 2 a.u.. The length of the primitive cell is 1. The nodes predicted by Equation (36) are at the intersections with diagonal lines, two of which are shown by the dashed lines.

The next step is to examine in a more quantitative way which factors determine the rate of this exponential decay for an insulator with a band gap  $\epsilon_{gap}$  at zero temperature.

Cloizeaux [10] proved the exponential decay behavior of the zero temperature density matrix, which is a projection operator. Considering the extension of the band energy  $\epsilon_n(\mathbf{k})$  into the complex  $\mathbf{k}$  plane he found that the minimal distance of the branch points of  $\epsilon_n(\mathbf{k})$  from the real axis determines the decay behavior. For the Wannier functions, which are closely related to the density matrix by Equation (23), Kohn [11] proved the same decay behavior in the case of a one-dimensional model crystal. In a later publication Kohn [12] claims that this distance to the real  $\mathbf{k}$  axis should be related to the square root of the gap. Even though he did not present a derivation of this result, it was widely accepted to be generally valid. Ismail-Beigi and Arias [7] have however shown that Kohn's claim is not generally valid. They demonstrated that in the Tight Binding limit the square root behavior can be found under certain circumstances, but that different behaviors can be found as well. In the weak binding limit, where the band-structure can be obtained by perturbation theory from the band structure of the free electron gas, they showed that the dependence is actually linear.

$$F(\mathbf{r}, \mathbf{r}') \propto \exp(-\gamma|\mathbf{r} - \mathbf{r}'|) \quad \text{where } \gamma = c \epsilon_{gap} a \quad (37)$$

The lattice constant is denoted by  $a$ , and  $c$  is an unknown constant of the order of 1. The dependence of the decay rate on the size of the band gap is a rather surprising relation. After all it follows from Equation (26) that only the properties of the occupied bands enter into the calculation of the density matrix, whereas the size of the gap is not directly related to the occupied states. In the following we will give an intuitive explanation of the factors determining the decay rate. This explanation will again be based on Equation (30) relating the bandstructure to the decay properties of the density matrix. As is known from complex analysis, the distance of the singularities from the real axis is comparable to the length over which one has very strong variations along the real axis of a complex function. Now, the long range decay properties of a Fourier transform are exactly determined by the length  $\Delta k$  of such a region of strongest variation. One thus regains Cloizeaux's result that the decay rate is proportional to the distance of singularities from the real axis. Let us now explain the behavior found in the weak binding limit by Ismail-Beigi and Arias. In the weak binding limit the effective mass establishes the connection between the gap and the important features of the occupied bands. The effective mass for the  $n$ -th band at the point  $\mathbf{k}_0$  is defined as

$$\frac{1}{m} = 1 + \frac{2}{3} \sum_{m \neq n} \frac{|\int \Psi_{n,\mathbf{k}_0}^*(\mathbf{r}) \nabla \Psi_{m,\mathbf{k}_0}(\mathbf{r}) d\mathbf{r}|^2}{\epsilon_n(\mathbf{k}_0) - \epsilon_m(\mathbf{k}_0)} \quad (38)$$

Since we are only interested in order of magnitudes, we have here averaged over the diagonal elements of the effective mass tensor in order to obtain a effective mass which is a scalar quantity. In the case of the weak binding limit, a gap will open up at the boundaries of the Brioullin zone and this gap will be small. The effective mass is therefore small and proportional to  $a^2 \epsilon_{gap}$ , where we have assumed that the dipole matrix elements  $\int \Psi_{n,\mathbf{k}_0}^*(\mathbf{r}) \nabla \Psi_{i,\mathbf{k}_0}(\mathbf{r}) d\mathbf{r}$  are on the order of  $\frac{1}{a}$ . The band-structure near the boundaries of the Brioullin zone is then given by

$$\frac{1}{2m} (\Delta k)^2 \propto \frac{1}{a^2 \epsilon_{gap}} (\Delta k)^2 \quad (39)$$

where  $\Delta k$  is the distance from the boundary, neglecting directional effects. Since the effective mass is small, the curvature of the band-structure is large in this region. Hence this region is just the region with the strongest variation. As is well known the perturbation theory arguments leading to Equation (39) are valid within an energy range of the order of  $\epsilon_{gap}$ . It then follows from Equation (39) that the corresponding range of  $\Delta k$  is  $\epsilon_{gap} a$ , confirming the linear decay of the density matrix with respect to the size of the gap, i.e.  $\gamma = c \epsilon_{gap} a$ .

Let us next show how a square root like behavior  $\gamma = c \sqrt{\epsilon_{gap}}$  can arise for real crystals with a big gap. In this case the effective mass is of the order of one at all stationary points  $\mathbf{k}_0$  in the Brioullin zone. Assuming that it is then of the order of one over the whole Brioullin zone, the region of largest variation is just the Brillouin zone itself. The decay constant is therefore simply related to the lattice constant  $a$ .

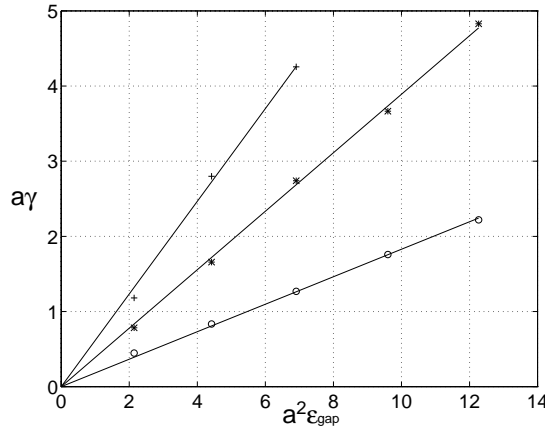
$$\gamma = c \frac{1}{a} \quad (40)$$

In order to get the square root dependence of the decay constant  $\gamma$ , one has to assume that

$$\epsilon_{gap} = C_{gap} \frac{1}{a^2} \quad (41)$$

where  $C_{gap}$  is a constant which is not or only weakly dependent on the material. Such a behavior has indeed been observed for certain classes of materials, where the tight binding limit is the most appropriate one, such as ionic crystals but with a non-negligible variation of  $C_{gap}$  across different materials. A square root behavior of  $\gamma$  can therefore be expected if one varies the lattice constant for a certain material, but the decay constants for different materials that happen to have the same gap are not necessarily comparable.

In practice the distinction between the Tight Binding and weak binding case may not always be clear. Unless the region of strongest variation is really a very small fraction of the whole Brioullin zone, all the prefactors which were neglected in these considerations might be important enough to blur out differences. The importance of these prefactors can also be seen from the fairly strong directional dependence of the decay rate. Ismail-Beigi and Arias [7] found such a strong directional dependence in numerical tests to confirm the linear dependence of the decay constant on the size of the gap (Figure 3). Stephan *et al.* [13] found the same behavior during Tight Binding studies of carbon. So a statement in an old paper by Kohn [2]), namely that the decay length of the Wannier functions is of the order of the interatomic spacing, is for practical purposes probably in many cases the best available characterization of localization.



**Fig. 3:** The dependence of the decay constant  $\gamma$  on the gap. Plotted are the two dimensionless quantities  $a\gamma$  versus  $a^2\epsilon_{gap}$ . The variation of the gap was obtained for a three-dimensional cubic model crystal by varying the strength of the potential. Circles refer to the  $[100]$ , stars to the  $[110]$  and pluses to the  $[111]$  direction. This figure is reproduced with kind permission of the authors from Ismail-Beigi and Arias (1998)

All the above arguments apply to simple and mainly periodic materials. Advanced electronic structure calculations however frequently study materials which are not in this class. The localization properties of such materials have not yet been studied systematically and so there is some incertitude about which orbitals are localized and to what extent (Kohn 1995). If the localization properties are unknown one should better not impose any localization constraints. In this case some of the discussed  $O(N)$  techniques still give a quadratic scaling, which also allows us to gain computational efficiency compared to the traditional cubically scaling algorithms.

### 3 Basic strategies for $O(N)$ scaling

Most  $O(N)$  algorithms are built around the density matrix or its representation in terms of Wannier functions and take advantage of its decay properties. To obtain linear scaling one has to

cut off the exponentially decaying quantities when they are small enough. This introduces the concept of a localization region. Only inside this localization region the quantity is calculated, outside it is assumed to vanish. For simplicity the localization region is usually taken to be a sphere, even though the optimal shape might be different [13]. In the Tight Binding context the boundary of the localization region can either be defined by a geometric distance criterion or in terms of the number of "hops", i.e. the number of steps one has to do along bonds connecting neighboring atoms to reach this boundary [14]. Different localization regions generally have significant overlaps. The localization regions thus do not form a partition of the computational volume and one atom in general belongs to several localization regions.

In a numerical calculation the density operator  $F(r, r')$  is discretized with respect to a basis. The basis set has to be chosen such that the matrix elements  $F_{i,j}$  reflect the decay properties of the operator  $F(r, r')$ . This will obviously only be the case if the basis set consists of localized functions, such as atom centered Gaussian type basis functions. Sets of orthonormal basis functions usually facilitate the calculations. Unfortunately all currently used localized basis sets are non-orthogonal. In the context of the orthogonal Tight Binding scheme one just assumes the existence of a basis set which is both atom centered and orthogonal. Since only the parameterized Hamiltonian matrix elements enter in the calculation, there is no need to explicitly ever construct such a basis set. In the following sections, we will follow this practice and assume in all relevant parts that we are dealing with such a localized orthogonal basis set. Whenever we refer from now on to a localization region, we actually mean the subset of all basis functions which are contained within this spatial localization region.

Obviously the size of the localization region needed to obtain a certain accuracy depends on the decay properties of the density matrix as well as on the selected accuracy threshold. It also depends on the quantity one wants to study. Generally, the total energy as well as derived quantities such as the geometric equilibrium configurations are surprisingly insensitive to finite localization regions, because these quantities are not strongly influenced by the exponentially small tails which are cut off by the introduction of a localization region. This insensitivity also holds true, even though to a much lesser extent, for metals. As we have seen above the introduction of a finite temperature leads to an exponential decay of the density matrix which in turn justifies truncation. In a metal, the difference between the finite and the zero temperature total energy  $\Delta E$  is proportional to the square of the temperature,  $\Delta E \propto T^2$ , and thus rather small. There are however quantities which are very sensitive to finite localization regions. In the modern theory of polarization in solids [15], the polarization can be expressed in terms of the centers of the Wannier functions  $\int W(\mathbf{r})\mathbf{r}W(\mathbf{r})d\mathbf{r}$ . Using this formula one has a strong influence of the tails of the Wannier functions because they get strongly weighted by the factor of  $\mathbf{r}$  in the integral. Since the tails are much more influenced by the boundary of the localization region than the central part, this quantity is more sensitive to the size of the localization region. There are even quantities which are not at all directly accessible by a solution which is given in terms of density matrices or Wannier functions. The Fermi surface in a metal which can be calculated via the eigenvalues of the band structure  $\epsilon_n(\mathbf{k})$  is such an example.

It is clear that one can gain significant computational efficiency only if the extent of the system is larger than the size of the localization region. The cross over point depends therefore on the decay properties of the density matrix of the system. It however also depends on the dimensionality of the system. For a linear-chain molecule with a large band gap, it might be enough to have a localization region containing just two neighboring atoms on each side. So the localization region would just contain 5 atoms and for systems larger than 5 atoms one might potentially gain computational efficiency by using an  $O(N)$  method. If one has a 3 dimensional

system with a comparable gap, then a spherical localization region extending out to the second neighbors would contain some 60 atoms and the crossover point would already be much larger. For a system with a small gap such as silicon or for metallic systems the crossover point is even larger.

There are essentially six basic approaches to achieve linear scaling.

- The Fermi Operator Expansion (FOE) is based on Equation (19). In this approach one finds a computable functional form of  $F$  as a function of  $H$  to build up the density matrix. Two possible representations based on a Chebychev expansion and a rational expansion will be discussed.
- The Fermi Operator Projection (FOP) method is closely related to the FOE method. The computable form of  $F$  is however not used to construct the entire density matrix but to find the space spanned by the occupied states, i.e. the space corresponding to the eigenfunctions associated with the unit eigenvalues of the Density matrix at zero temperature. These eigenfunctions can be considered as Wannier functions in the generalized sense defined before.
- In the Divide and Conquer (DC) method [16] for the density matrix the relevant parts of the density matrix are patched together from pieces that were calculated for smaller subsystems.
- In the Density Matrix Minimization (DMM) approach, one finds the density matrix by a minimization of an energy expression based on the density matrix.
- In the Orbital Minimization approach (OM) [17], one finds a set of Wannier functions by minimization of an energy expression.
- The Optimal Basis Density Matrix Minimization scheme (OBDMM) contains aspects of both the OM and DMM methods. In addition to finding a density matrix with respect to the basis, one also finds an optimal basis by additional minimization steps. The number of basis functions has to be at least equal to the number of electrons in the system, but can be bigger as well.

A major difference between these methods is whether they calculate the full density matrix or only its representation in terms of Wannier functions. The later approach applies only to insulators while the former is in also applicable to systems with fractional occupation numbers (i.e.  $f(\epsilon_n)$  is not either 1 or 0) such as metals or systems at finite electronic temperature.

In the following four of these six approaches will be presented in detail. The FOE [18] is the most straightforward approach for the calculation of the density matrix. The basic idea in this approach is to find a representation of the matrix function (19) which can be evaluated on a computer. Several such representations are possible. We will discuss a Chebychev and a rational representation.

### 3.1 The Chebychev Fermi Operator Expansion

One of the most basic operations a computer can do are matrix times vector multiplications. The simplest representation of the density matrix would therefore be a polynomial representation

$$F \approx p(H) = c_0 I + c_1 H + c_2 H^2 + \dots + c_{n_{pl}} H^{n_{pl}}.$$

where  $I$  is the identity matrix. Unfortunately polynomials of high degree become numerically unstable. This instability can however be avoided by introducing a Chebychev polynomial representation, which is a widely used numerical method (Press *et al.*, 1986)

$$p(H) = \frac{c_0}{2}I + \sum_{j=1}^{n_{pl}} c_j T_j(H) . \quad (42)$$

Since the Chebychev polynomials are defined only within the interval  $[-1;1]$ , we will assume in the following that the eigenvalue spectrum of  $H$  falls within this interval. This can always be easily achieved by scaling and shifting of the original Hamiltonian. The Chebyshev matrix polynomials  $T_j(H)$  satisfy the recursion relations

$$T_0(H) = I \quad (43)$$

$$T_1(H) = H \quad (44)$$

$$T_{j+1}(H) = 2 H T_j(H) - T_{j-1}(H) . \quad (45)$$

The expansion coefficients of the Chebychev expansion can easily be determined. The eigenfunction representation (Equation (21)) of  $F$  is,

$$\langle \Psi_n | F | \Psi_m \rangle = f(\epsilon_n) \delta_{n,m} . \quad (46)$$

Evaluating the polynomial expansion in the same eigenfunction representation we obtain

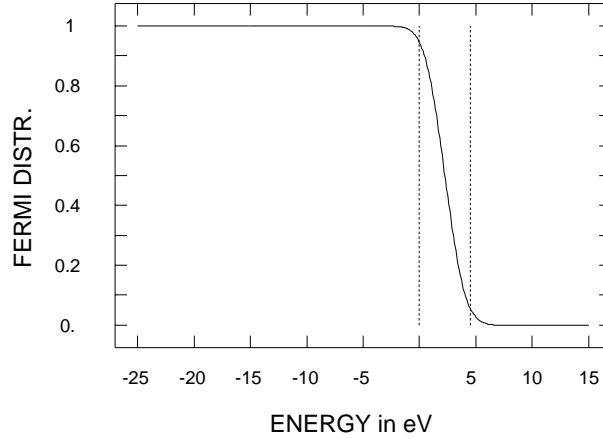
$$\langle \Psi_n | p(H) | \Psi_m \rangle = p(\epsilon_n) \delta_{n,m} , \quad (47)$$

where

$$p(\epsilon) = \frac{c_0}{2} + \sum_{j=1}^{n_{pl}} c_j T_j(\epsilon) . \quad (48)$$

Comparing Equation (46) and Equation (47) we see that the polynomial  $p(\epsilon)$  has to approximate the Fermi distribution in the energy interval  $[-1;1]$  where the scaled and shifted Hamiltonian has its eigenvalues. How to find the Chebychev expansion coefficients for a scalar function is described in standard textbooks on numerical analysis. Actually it is not necessary to take the exact Fermi distribution. In practically all situations one is interested in the limit of zero temperature. Hence any function which approaches a step functions in the limit of zero temperature can be used. In the case of simulations of insulators for instance it is advantageous to take the function  $f(\epsilon) = \frac{1}{2}(1 - \text{erf}(\frac{\epsilon - \mu}{\Delta\epsilon}))$  shown in Figure 4 since it decays faster to 0 respectively 1 away from the chemical potential. The term Fermi distribution will in the following always be used in this broader sense. The energy resolution  $\Delta\epsilon$  is chosen to be a certain fraction of the size of the gap. In the case of metals,  $\Delta\epsilon$  is chosen by considerations of numerical convenience. Large values of  $\Delta\epsilon$  will give lower accuracy results. However as pointed out before, the convergence of the total energy with respect to  $\Delta\epsilon$  is quadratic and so highly accurate total energies can be obtained with rather high values of  $\Delta\epsilon$ . Small values of  $\Delta\epsilon$  make the calculation numerically expensive. The detailed scaling behavior of the numerical effort in the limit of vanishing gaps is analyzed in section 4, where it is found that actually the increase in the size of the localization region is the limiting factor in all methods.

Even if one wants to study electronic properties in the limit of zero electronic temperature it is important that one nevertheless uses a finite temperature Fermi distribution for the Chebychev



**Fig. 4:** The Fermi distribution as obtained by a Chebychev fit of degree 40 in the case of a diamond structure. The bandgap is in between the two vertical lines.

fit. Using the zero temperature step function introduces so-called Gibbs oscillations in the fit and spoils the Chebychev fit over the whole interval.

How to eliminate these Gibbs oscillations in the zero temperature case by the so called kernel polynomial method [14] can be used as a starting point for an alternative derivation of the FOE method. The basic idea is to expand a delta function as a polynomial using damping factors to suppress large oscillations. This representation of an approximate delta function can then be integrated to obtain a smooth representation of the Fermi distribution. Used this way the kernel polynomial method is thus just another way to derive the expansion coefficients for the Chebychev expansion [19]. In addition the kernel polynomial method can also be used to smear out the density of states rather than the zero temperature Fermi distribution resulting in a method with practically identical computational requirements but some slightly different properties. One useful property is that the smeared density of states energy is an approximate lower bound to the energy, whereas the smeared Fermi energy is an approximate upper bound [14].

Coming back to the original motivation for a polynomial representation, let us now show how the density matrix can be constructed using only matrix times vector multiplications. Let us denote by  $t_l^j$  the  $l$ -th column of the Chebychev matrix  $T_j$ . Now each column of these Chebychev matrices satisfies the same recursion relations

$$\begin{aligned} |t_l^0\rangle &= |e_l\rangle \\ |t_l^1\rangle &= H|e_l\rangle \\ |t_l^{j+1}\rangle &= 2H|t_l^j\rangle - |t_l^{j-1}\rangle. \end{aligned} \tag{49}$$

where  $e_l$  is a unit vector that has zeroes everywhere except at the  $l$ -th entry. So Equation (49) demonstrates that we indeed need only matrix vector multiplications. Once we have generated the  $l$ -th columns of all the Chebychev matrices, we can obtain the  $l$ -th column  $f_l$  of the density matrix just by forming linear combinations

$$|f_l\rangle = \frac{c_0}{2}|t_l^0\rangle + \sum_{j=1}^{n_{pl}} c_j |t_l^j\rangle. \tag{50}$$

As we have described the method so far it has a quadratic scaling instead of the linear scaling we finally want to achieve. If we have  $M_b$  basis functions, the density matrix is a  $M_b \times M_b$  matrix



and we have to calculate  $M_b$  full columns. For the calculation of each column, we have to do  $n_{pl}$  matrix times vector multiplications, each of which costs  $M_b n_H$  operations assuming the matrix  $H$  is a sparse matrix with  $n_H$  off-diagonal elements per row/column. So the total computational cost is  $M_b^2 n_{pl} n_H$ . The degree of the polynomial  $n_{pl}$  and the width  $n_H$  of the Hamiltonian are independent of the size of the system, whereas  $M_b$  is proportional to the number of atoms in the system. The overall scaling with respect to the number of atoms is therefore quadratic.

In order to do the correct shifting and scaling of the original Hamiltonian to map its eigenvalue spectrum on the interval  $[-1:1]$  we have to know its lowest and highest eigenvalues  $\epsilon_{min}$  and  $\epsilon_{max}$ . In addition we have to know the chemical potential  $\mu$ . There are auxiliary matrix functions of  $H$  that can help us to determine these quantities. These functions of  $H$  can be built up in the same way as the density matrix. Since the recursive build up of the Chebychev matrices is the most costly part, the additional cost for evaluating other functions is negligible. To determine whether we have a vanishing density of states beyond an energy  $\epsilon_{up}$  we can for instance construct a Chebychev fit  $p_{up}(\epsilon)$  to a function which is zero (to within a certain tolerance) for energies below  $\epsilon_{up}$ , but blows up for energies larger than  $\epsilon_{up}$ . If  $Tr[p_{up}(H)]$  does not vanish we have a non-vanishing density of states beyond  $\epsilon_{up}$ . A similar procedure can be applied to determine a lower bound for the density of states. The determination of the chemical potential in an insulator can be done along the same lines as well (Bates and Scuseria 1998). Without any significant extra cost one can build up several Fermi distributions with different chemical potentials until one finds the correct chemical potential leading to charge neutrality. In a metallic system the search for the chemical potential can be accelerated since it is possible to predict with high accuracy how the number of electrons changes in response to a change in the chemical potential. From Equation (13) it follows

$$\frac{\partial N_{el}}{\partial \mu} = -Tr[p'(H)] , \quad (51)$$

where  $p'$  is the derivative of the Chebychev polynomial  $p$  that approximates the Fermi distribution. The Chebychev expansion coefficients of  $p'$  can be calculated from the coefficients for  $p$  (Press *et al.*, 1986). Using the finite difference approximation of Equation (51),

$$\Delta \mu = \frac{\Delta N_{el}}{Tr[p'(H)]} , \quad (52)$$

it is possible to find the correction  $\Delta \mu$  to the chemical potential which will nearly exactly eliminate an excess of  $\Delta N_{el}$  electrons due to an incorrect initial chemical potential. The correct chemical potential in a metallic system can thus be found with very high accuracy with a few iterations.

The desired linear scaling can be obtained by introducing a localization region for each column, outside of which the elements are negligibly small. For the  $k$ -th column, this localization region will be centered on the  $k$ -th basis function. If we use atom centered basis functions, then the localization region will consequently be centered on the atom to which this  $k$ -th basis function belongs. We have then to calculate only that part of each column which corresponds to this localization region. This means that we can use a truncated Hamiltonian  $H(k)$  which retains only the matrix elements corresponding to the basis functions contained within the localization region  $k$ . Denoting the number of basis functions in this region by  $M_{loc}$  (which might actually depend on the localization region  $k$  being considered), the overall computational cost is then  $M_b M_{loc} n_{pl} n_H$  and thus scales linearly. Let us stress, that the size of the localization region is independent of the degree of the polynomial. If one uses for instance a polynomial of degree

$n_{pl} = 50$ , the recursion in Equation (49) will extend over the 50 nearest neighbor shells without localization constraint for a Hamiltonian coupling only nearest neighbors. The localization region however is typically much smaller comprising just a few nearest neighbor shells. Imposing a localization region introduces some subtleties. For instance the eigenvalues of the truncated density matrix are not anymore exactly given by  $p(\epsilon_n)$  and  $F$  is not any more strictly symmetric. More importantly, strictly speaking we can no longer use the Trace notation, since we use different local Hamiltonians  $H(k)$  to build up the different columns of the density matrix. The band-structure energy  $E_{BS}$  has now to be written as

$$E_{BS} = \sum_k \sum_j [p(H(k))]_{k,j} [H(k)]_{j,k} . \quad (53)$$

Another important quantity are the forces. The force acting on the  $\alpha$ -th atom at position  $R_\alpha$  is obtained by differentiating the total energy with respect to these positions. The total energy consists of the band structure part and possibly other contributions. We will only discuss the non-trivial part of the force arising from the differentiation of the band structure energy  $E_{BS}$ . For simplicity let us assume that we have a simple polynomial expansion and not a Chebychev expansion. Let us also assume that we calculate the full density matrix, i.e. that we do not truncate  $H$  by introducing a localization region. We then obtain

$$\frac{dE_{BS}}{dR_\alpha} = \frac{d}{dR_\alpha} \text{Tr} \left[ H \sum_\nu c_\nu H^\nu \right] = \sum_\nu c_\nu \text{Tr} \left[ \frac{\partial H^{\nu+1}}{\partial R_\alpha} \right] . \quad (54)$$

Let us consider for instance the term for which  $\nu = 2$

$$\frac{d\text{Tr}[H^3]}{dR_\alpha} = \text{Tr} \left[ HH \frac{\partial H}{\partial R_\alpha} \right] + \text{Tr} \left[ H \frac{\partial H}{\partial R_\alpha} H \right] + \text{Tr} \left[ \frac{\partial H}{\partial R_\alpha} HH \right] = 3\text{Tr} \left[ HH \frac{\partial H}{\partial R_\alpha} \right] , \quad (55)$$

where we used that  $\text{Tr}[AB] = \text{Tr}[BA]$ . The final result for the force, which also holds in the case of a Chebychev expansion, is thus

$$\frac{dE_{BS}}{dR_\alpha} = \text{Tr} \left[ (p(H) + Hp'(H)) \frac{\partial H}{\partial R_\alpha} \right] . \quad (56)$$

In the case of an insulator, the second term in the brackets  $Hp'(H)$  is very small compared to the first term  $p(H)$  at small but finite temperatures and it vanishes in the limit of zero temperature. The reason for this is that the eigenvalues of the matrix  $p'(H)$  are  $p'(\epsilon_n)$ . Since at zero temperature  $p'(\epsilon)$  is nonzero only at the chemical potential which is in the middle of the gap, all eigenvalues are zero and the matrix is identically zero. Nevertheless it is recommendable to retain this term in numerical calculations because it leads to forces consistent with the total energy.

In the case where we calculate only part of the density matrix, i.e. where we have a truncated Hamiltonian  $H(k)$  going with the energy expression (53) we cannot use the properties of the trace to simplify the force expression as we did in Equation (55). The equation corresponding to Equation (55) therefore reads

$$\begin{aligned}
& \sum_{k,j1,j2,k} [H(k)]_{k,j1} [H(k)]_{j1,j2} \left( \frac{\partial H(k)}{\partial R_\alpha} \right)_{j2,k} + \\
& [H(k)]_{k,j1} \left( \frac{\partial H(k)}{\partial R_\alpha} \right)_{j1,j2} [H(k)]_{j2,k} + \\
& \left( \frac{\partial H(k)}{\partial R_\alpha} \right)_{k,j1} [H(k)]_{j1,j2} [H(k)]_{j2,k} .
\end{aligned} \tag{57}$$

Similar results hold for all the other terms with different values of  $\nu$ . In the case of a Chebychev expansion the situation is completely analogous, just the formulas are more complicated. The force formula has been worked out in this case by Voter *et al.* (1996) and is given by

$$\frac{dT_j(H)}{dR_\alpha} = \frac{dT_{j-2}(H)}{dR_\alpha} + \sum_{i=0}^{j-1} (1+k_i)(1+k_{j-1-i}) T_i(H) \frac{\partial H}{\partial R_\alpha} T_{j-1-i}(H) , \tag{58}$$

where  $k_j = 0$  if  $j \leq 0$  and  $k_j = 1$  otherwise. In the typical Tight Binding context  $\frac{\partial H}{\partial R_\alpha}$  is a very sparse matrix. If it contains  $n_D$  non-zero elements, we need of the order of  $n_{pl}^2 n_D M_b$  operations to evaluate all the forces according to Equation (58). The error incurred by using the approximate formula region is large enough. Since the approximate formula can be evaluated with order  $n_{pl} n_D M_b$  operations, it might actually be preferable to do so. In a molecular dynamics simulation, the largest deviations in the conservation of the total energy come from events where atoms enter or leave localization regions and this kind of error is not taken into account by either force formula.

All the above force formulas were derived for the case where we have a constant chemical potential and where the polynomial representing the Fermi distribution does thus not change. Frequently one wants however to do simulations for a fixed number of electrons rather than for a fixed chemical potential. In this case one has to readjust the chemical potential for each new atomic configuration. The chemical potential is thus a function of all the atomic positions  $\mu = \mu(R_\alpha)$ , but the explicit functional form of this dependence is not known. The force formula can however also be adapted to this case [29]. Ignoring the above warnings and using again trace notation for simplicity we have

$$E_{BS} = Tr[H p(H - \mu I)] \tag{59}$$

$$N_{el} = Tr[p(H - \mu I)] \tag{60}$$

and consequently

$$\frac{dE_{BS}}{dR_\alpha} = Tr \left[ (H p' + p) \frac{\partial H}{\partial R_\alpha} \right] - Tr[(H p')] \frac{\partial \mu}{\partial R_\alpha} \tag{61}$$

$$\frac{dN_{el}}{dR_\alpha} = Tr \left[ p' \frac{\partial H}{\partial R_\alpha} \right] - Tr[p'] \frac{\partial \mu}{\partial R_\alpha} . \tag{62}$$

Since  $\frac{dN_{el}}{dR_\alpha}$  has to be equal to zero, we can solve Equation (62) for  $\frac{\partial \mu}{\partial R_\alpha}$  and insert it into equation (61) to obtain the force under the constraint of a constant number of electrons.

Let us finally derive a force formula for the case where a local charge neutrality condition is enforced [19] The motivation for this approach is that in non-self-consistent Tight Binding calculations one frequently finds an unphysically large transfer of charge between atoms. In a

self-consistent calculation the electrostatic potential, built up by a charge transfer, is counteracting a further charge flow and thus limits charge transfer to reasonably small values. Some Tight Binding schemes enforce a so-called local charge neutrality condition requiring that the total charge associated with an atom in a molecule or solid be equal to the charge of the isolated atom. This is done by determining a potential offset  $u_\alpha$  for each atom  $\alpha$  in the system which will ensure this neutrality. The total Hamiltonian  $H$  of the system is then given by  $H_0 + U$  where  $H_0$  is the Hamiltonian without any potential bias and  $U$  a diagonal matrix containing the atomic potential offsets  $u_\alpha$ . The band structure energy is given by

$$E_{BS} = \text{Tr}[(H_0 + U) p(H_0 + U)] - \sum_{\alpha} Q_{\alpha} u_{\alpha} , \quad (63)$$

where the term containing the atomic valence charges  $Q_{\alpha}$  has been subtracted to make the expression invariant under the application of a uniform potential bias to all atoms in the system. Expressed in terms of the density matrix the local charge neutrality condition becomes

$$\sum_l p(H)_{\alpha,l;\alpha,l} = Q_{\alpha} . \quad (64)$$

In Equation (64) we have labeled the basis functions by a composite index where  $\alpha$  indicates on which atom the basis function is centered and where  $l$  describes the character of the atom centered basis function. If we have carbon atoms, for which  $Q_{\alpha} = 4$ ,  $l$  would for instance denote the 4 orbitals  $2s$ ,  $2px$ ,  $2py$ ,  $2pz$ . Using Equation (64), Equation (63) then simplifies to

$$E_{BS} = \text{Tr}[H_0 p(H_0 + U)] . \quad (65)$$

Taking the derivative we get

$$\frac{dE_{BS}}{dR_{\alpha}} = \sum_{\beta} \frac{\partial E_{BS}}{\partial u_{\beta}} \frac{\partial u_{\beta}}{\partial R_{\alpha}} + \frac{\partial E_{BS}}{\partial R_{\alpha}} , \quad (66)$$

where

$$\frac{\partial E_{BS}}{\partial u_{\beta}} = \text{Tr} \left[ H_0 p'(H) \frac{\partial H}{\partial u_{\beta}} \right] . \quad (67)$$

As discussed above, the matrix  $p'(H)$  is close to zero in an insulator at sufficiently low temperature and can often be neglected. The forces are therefore approximately given by

$$\frac{dE_{BS}}{dR_{\alpha}} = \text{Tr} \left[ H_0 p(H) \frac{\partial H}{\partial R_{\alpha}} \right] . \quad (68)$$

It has to be pointed out that to get sufficiently high accuracy the degree of the polynomial has to be higher than in the case of Tight Binding case without local charge neutrality.

The degree  $n_{pl}$  of the polynomial needed to represent the Fermi distribution is proportional to

$$n_{pl} \propto \frac{\epsilon_{max} - \epsilon_{min}}{\Delta\epsilon} . \quad (69)$$

This follows from the fact, that the  $n$ th order Chebychev polynomial has  $n$  roots and so a resolution that is roughly proportional to  $1/n$ . For the usual Tight Binding Hamiltonians the ratio in Equation (69) is not very large and for silicon and carbon systems without gap states polynomials of degree 50 are sufficient. In contexts other than Tight Binding this ratio can however be fairly large and polynomial representation would become very inefficient.

### 3.2 The Rational Fermi Operator Expansion

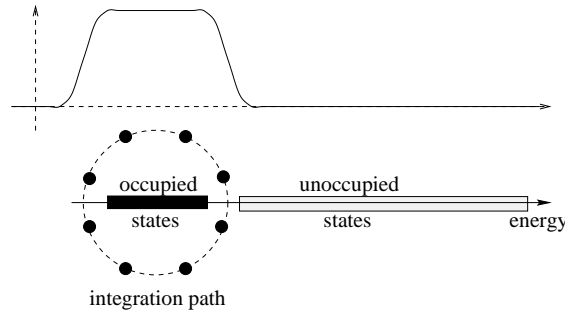
A rational representation of the density matrix [20] is in this case more efficient

$$F = \sum_{\nu} \frac{w_{\nu}}{H - z_{\nu}} . \quad (70)$$

As is well known, the function  $f(\epsilon)$  given by

$$f(\epsilon) = \frac{1}{2\pi i} \oint \frac{dz}{\epsilon - z} \quad (71)$$

is equal to 1 if  $\epsilon$  is within the volume encircled by the contour integration path and zero otherwise. If the integration path contains the occupied states as shown in Figure (5) it can therefore be used as a zero temperature Fermi distribution. Actually, as already mentioned above, it is usually not necessary to have the exact Fermi distribution. The electronic temperature is just determined by the slope (and possibly some higher derivatives) of the distribution at the Fermi energy. We will also refer to such generalized distributions as Fermi distributions. A distribution of this type can be obtained by discretizing the zero temperature contour integral from Equation (71) as shown in Figure 5.



**Fig. 5:** A discretization of the contour integral in the complex energy plane of Equation (71). The resulting Fermi distribution is shown on top.

In principle any other set of  $z_{\nu}$ 's and  $w_{\nu}$ 's can be used as long as it satisfies

$$f(\epsilon) \approx \sum_{\nu=1}^{n_{pd}} \frac{w_{\nu}}{\epsilon - z_{\nu}} , \quad (72)$$

where  $n_{pd}$  is the degree of the rational approximation. How can we now evaluate Equation (70) on a computer? Denoting  $\frac{I}{H - z_{\nu}}$  by  $F_{\nu}$  we have

$$(H - z_{\nu})F_{\nu} = I \quad (73)$$

$$F = \sum_{\nu} w_{\nu} F_{\nu} . \quad (74)$$

So we have first to invert all the matrices  $H - z_{\nu}$  and then to form linear combinations of them. The inversion is equivalent to the solution of  $M_b$  linear systems of equations. This can be effectuated using iterative techniques so that in the end everything can again be done by matrix times vector multiplications. A rational approximation can represent the sharp variation near the chemical potential of a low temperature Fermi distribution in a more efficient way than

a Chebychev approximation. Whereas in the Chebychev case the degree of the polynomial is given by Equation (69) the degree of the rational approximation  $n_{pd}$  is given by

$$n_{pd} \propto \frac{\mu - \epsilon_{min}}{\Delta\epsilon}. \quad (75)$$

This  $n_{pd}$  in contrast to  $n_{pl}$  does not depend on the largest eigenvalue  $\epsilon_{max}$ . Once  $n_{pd}$  is of the order of magnitude given by Equation (75) one has exponential convergence to the zero temperature Fermi distribution. In the case where the integration points and weights are obtained by discretizing the contour integral of Figure 5 this exponential behavior is immediately comprehensible since an equally spaced integration scheme gives exponential convergence for periodic functions. Since  $n_{pd}$  is usually reasonably small, the success of the method will hinge upon whether it is possible to solve the linear system of equations associated with each integration point with a small number of iterations. The number of iterations in an iterative method such as a conjugate gradient scheme is related to whether it is possible to find a good preconditioning scheme. In the case of plane wave calculations a good preconditioner can be obtained from the diagonal elements and of the order of 10 iterations are required. In other schemes using Gaussians for instance it is not quite clear whether good preconditioners can be found. When the Hamiltonian depends on the atomic positions  $R_\alpha$ , Equations (73) and (74) can be differentiated to obtain the derivative  $\frac{dF}{dR_\alpha}$ , which is needed for the calculation of the forces.

### 3.3 The Fermi Operator Projection Method

The FOE method is used to calculate the full density matrix. This can be inefficient if the number of basis functions per atom is very large. As was mentioned before, the density matrix at zero temperature does not have full rank. In the case of an insulator it can be constructed from  $N_{el}$  Wannier functions (23). If one has a numerical representation of the zero temperature density operator, which is actually a projection operator, that eliminates all components belonging to eigenvalues above the Fermi level, one can apply it to a set of trial Wannier functions  $\tilde{V}_n$ ,  $n = 1, \dots, N_{el}$  to generate a set of orbitals which span the space of the Wannier functions. The numerical representation of the density operator can again either be a Chebychev or rational one. We will first discuss the rational case [20].

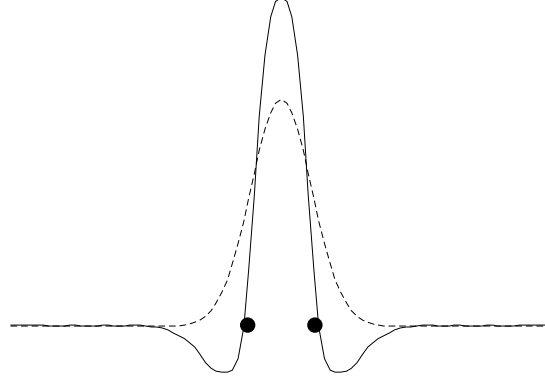
To do the projection with a rational representation, a system of equations analogous to (73) and (74) has to be solved for each trial Wannier function  $\tilde{V}_n$  and at each integration point  $\nu$

$$(H - z_\nu)\tilde{W}_{n,\nu} = \tilde{V}_n \quad (76)$$

$$\tilde{W}_n = \sum_\nu w_\nu \tilde{W}_{n,\nu}. \quad (77)$$

Thus the saving comes from the fact that one has to solve this system of equations (76) just for  $N_{el}$  right hand sides, whereas one has  $M_b$  right hand sides in Equation (73). Obviously the solution of the Equation (76) has to be done not within the whole computational volume but only within the localization region to obtain linear scaling. The functions  $\tilde{W}_n$  will now span our subspace unless one of our trial functions  $\tilde{V}_n$  was chosen in such a way that it has zero overlap with the space of the occupied orbitals, which is highly unlikely. To obtain a set of valid Wannier functions  $W_n$  one has still to orthogonalize the orbitals  $\tilde{W}_n$ . Since the  $W_n$ 's are localized the overlap matrix is a sparse matrix and can be calculated with linear scaling. In the typical Density Functional context, the inversion of this matrix is a rather small part, even if it

is done with cubic scaling. In a Tight Binding context it is much more important and a linear scaling method has been devised by Stephan [13] for the inversion. The construction of the Wannier functions by projection according to Equations (76) and (77) is illustrated in Figure 6 in the case of a silicon crystal. In this case one knows that the Wannier functions are bond centered and it is therefore natural to choose a set of bond centered functions as an initial guess. In this example we took simple Gaussians. As shown in Figure 6, the projection modifies the details of the Gaussian but does not significantly change its localization properties.



**Fig. 6:** The effect of applying the density operator, which is a projection operator in the eigenvalue space, to a Gaussian (dashed line) centered in the middle of a bond between two silicon atoms denoted by discs. The resulting function  $\tilde{W}$  is shown by the solid line. The orthogonal Wannier function  $W$  obtained by symmetric orthogonalization is practically indistinguishable from  $\tilde{W}$  on this scale. The calculation was done using Density Functional Theory with pseudopotentials

### 3.4 The Density Matrix Minimization approach

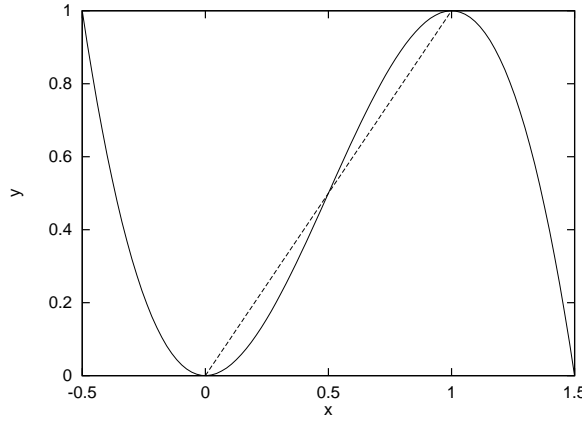
The DMM approach of Li, Nunes and Vanderbilt [21] is another approach where the full density matrix is constructed. In contrast to the FOE method one obtains the density matrix  $F$  in the limit of zero temperature, so no adjustable temperature parameter enters the calculation. The density matrix is obtained by minimizing the following functional for the grand potential  $\Omega$  with respect to  $F$

$$\Omega = \text{Tr}[(3F^2 - 2F^3)(H - \mu I)] . \quad (78)$$

There is no constraint imposed during the minimization so all the matrix elements of  $F$  are independent degrees of freedom. Nevertheless the final density matrix will obey the correct constraint of being a projector if no localization constraints are imposed. This is related to the fact that the matrix  $3F^2 - 2F^3$  is a purified version of  $F$  as can be seen from Figure 7. If  $F$  has eigenvalues close to zero or one then the purified matrix will have eigenvalues that are even closer to the same values. It is also clear from Figure 7 that the eigenvalues of the purified matrix are contained in the interval  $[0;1]$  as long as the eigenvalues of  $F$  are in the interval  $[-1/2 ; 3/2]$ .

The gradient of  $\Omega$  as given by Equation (78) with respect to  $F$  is itself a matrix and it is given by

$$\frac{\partial \Omega}{\partial F} = 3(F H' + H' F) - 2(F^2 H' + F H' F + H' F^2) , \quad (79)$$



**Fig. 7:** The McWeeny (1960) purification function  $3x^2 - 2x^3$

where  $H' = (H - \mu I)$ . In order to verify that Equation (78) defines a valid functional we have to show two things. First, that the grand potential expression (78) gives the correct result if we insert the exact density matrix  $F$ , and second, that the gradient (79) vanishes in this case. From Equation (24) we see that the exact  $F$  is a projection operator, i.e. that  $F^2 = F$ . Therefore  $(3F^2 - 2F^3) = F$  and the grand potential expression (78) agrees indeed with the correct result (11). Using in addition the fact that  $H'$  and the exact  $F$  commute (as follows from Equations (20), (21)) it is also evident that the gradient in Equation (79) vanishes. The gradient vanishes however not only for the ground state density matrix  $F$  but also for any excited state density matrix. In order to exclude the possibility of local minima, we have to verify, that these stationary points are no minima. This can easily be done (D. Vanderbilt, private communication) using the fact that the functional is a cubic polynomial with respect to all its degrees of freedom. Let us suppose that there are two minima. Inspecting the functional along the line connecting these two minima we would obviously again find these two minima, which is a contradiction because a cubic polynomial cannot have two minima. Thus we have proved by contradiction that the DMM functional has only one single minimum.

There is a second thing which is worrying at first sight with this functional. If the density matrix for an insulator at zero temperature is of the correct form (i.e. if the occupation numbers  $n_l$  are integers) the gradient (79) will vanish independently of the value of the chemical potential. This ambiguity however disappears as soon as one has fractional occupation numbers. Let us consider an approximate density matrix of the form

$$F = \sum_l n_l |\Psi_l\rangle \langle \Psi_l|. \quad (80)$$

Then it is easy to see that

$$\Omega = \sum_l (\epsilon_l - \mu) (3n_l^2 - 2n_l^3) \quad (81)$$

$$\frac{\partial \Omega}{\partial F} = \sum_l 6(\epsilon_l - \mu) n_l (1 - n_l) |\Psi_l\rangle \langle \Psi_l|. \quad (82)$$

The polynomial of Equation (81) is the same as the one shown in Figure 7 and we see that components corresponding to eigenvalues larger than the chemical potential are damped until they vanish in the minimization process, whereas components corresponding to smaller eigenvalues



are amplified until they reach the value one. Thus the chemical potential will determine the number of electrons to be found in the system as it should. The above statements are actually only correct if all the  $n_l$ 's are contained in the interval  $[-1/2: 3/2]$ . If this is not the case then one can see from Figure 7, that there can be runaway solutions, where some  $n_l$  tend to  $\pm\infty$ . When we implemented the scheme we however never encountered in practice such a runaway case. Having convinced ourselves, that the functional defined in Equation (78) is well behaved, let us now estimate the number of iterations which are necessary to minimize it. As is well known, the error reduction per iteration step depends on the condition number  $\kappa$  which is the ratio of the largest curvature  $a_{max}$  to the smallest curvature  $a_{min}$  at the minimum. These curvatures could be determined exactly by calculating the Hessian matrix at the minimum. Let us instead only derive an estimate of these curvatures by calculating the curvature along some representative directions. To do this let us now consider a ground state density matrix where some fraction  $x$  of an excited state is mixed in

$$F(x) = \sum_{n=1}^{N_{el}} \Psi_n^*(\mathbf{r}) \Psi_n(\mathbf{r}) - x \Psi_I^*(\mathbf{r}) \Psi_I(\mathbf{r}) + x \Psi_J^*(\mathbf{r}) \Psi_J(\mathbf{r}) . \quad (83)$$

The index  $I$  is a member of the  $N_{el}$  eigenstates below  $\mu$  and the index  $J$  refers to a state above  $\mu$ . The expectation value of the OM functional for this density matrix is given by

$$\begin{aligned} \Omega(x) &= Tr[(3F(x)^2 - 2F(x)^3)(H - \mu I)] \\ &= \sum_{n=1}^{N_{el}} \epsilon_n + (3x^2 - 2x^3) (\epsilon_J - \epsilon_I) \end{aligned} \quad (84)$$

and its second derivative by

$$\left. \frac{\partial^2 \Omega(x)}{\partial x^2} \right|_{x=0} = 6(\epsilon_J - \epsilon_I) . \quad (85)$$

The largest curvature will roughly be  $\epsilon_{max} - \epsilon_{min}$  and the smallest curvature of the order of the HOMO-LUMO separation  $\epsilon_{gap} = \epsilon_{N_{el}+1} - \epsilon_{N_{el}}$ . The condition number is thus given by

$$\kappa = \frac{a_{max}}{a_{min}} \approx \frac{\epsilon_{max} - \epsilon_{min}}{\epsilon_{gap}} . \quad (86)$$

In the conjugate gradient method, which is the most efficient method to minimize the DMM functional, the error  $e_k$  decreases as follows (Saad)

$$e_k \propto \left( \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^k . \quad (87)$$

The error  $e_k$  is defined in this context as the length of the vector which is the difference between the exact and approximate solution at the  $k$ -th iteration step. Under realistic conditions  $\kappa$  is large and the number of iterations  $n_{it}$  to achieve a certain accuracy is therefore proportional to

$$n_{it} \propto \sqrt{\kappa} = \sqrt{\frac{\epsilon_{max} - \epsilon_{min}}{\epsilon_{gap}}} . \quad (88)$$

This is an important result since it indicates that in an insulator the number of iterations is independent of system size. This result is also confirmed by numerical tests.

The use of a conjugate gradient scheme requires line minimizations along these conjugate directions. For arbitrary functional forms this has to be done by numerical techniques such as bisection. In the case of the DMM functional we have however a cubic form along each direction. The four coefficients determining the cubic form can be calculated with four evaluations of the functional. Once these 4 coefficients are known the minimum along this direction can easily be found.

The forces on the atoms are given by

$$\frac{d\Omega}{dR_\alpha} = \frac{\partial\Omega}{\partial F} \frac{\partial F}{\partial R_\alpha} + \frac{\partial\Omega}{\partial H} \frac{\partial H}{\partial R_\alpha}. \quad (89)$$

Since the method is variational,  $\frac{\partial\Omega}{\partial F}$  vanishes at the solution and the force formula simplifies to

$$\frac{d\Omega}{dR_\alpha} = \frac{\partial\Omega}{\partial H} \frac{\partial H}{\partial R_\alpha} = \text{Tr} \left[ (3F^2 - 2F^3) \frac{\partial H}{\partial R_\alpha} \right] \quad (90)$$

which can easily be evaluated.

The introduction of a localization region leads again to some subtleties. Whereas in the unconstrained case the eigenvalues of the final density matrix  $F$  will all be either zero or one, this is not any more the case when a localization region is introduced. So the truncated  $F$  is not any more a projection matrix but it is given by

$$F = \sum_{m=1}^{M_b} n_m \Psi_m^*(\mathbf{r}) \Psi_m(\mathbf{r}), \quad (91)$$

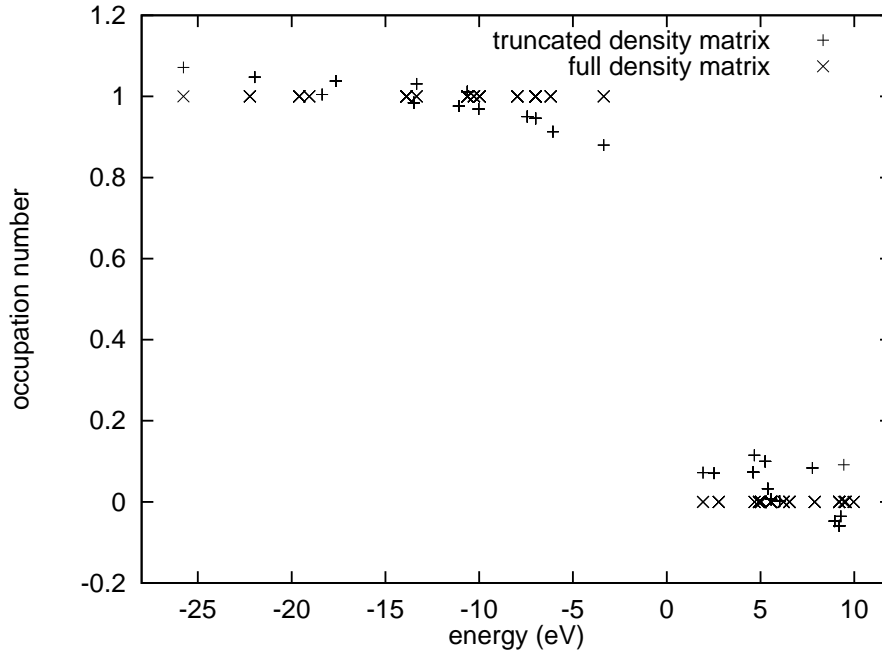
where now  $\Psi_m$  are the eigenfunctions of the truncated  $F$  and the occupation numbers  $n_m$  their eigenvalues. In a certain sense the localization constraint introduces a finite electronic temperature. This is actually not surprising after the discussion of the relation between the temperature and the localization properties in section 2. Figure 8 shows the energy expectation values of the eigenvectors of  $F$  versus the occupation numbers, for the case of a crystalline Si cell of 64 atoms, where the localization region extends up to the second nearest neighbors. As one sees, the energy expectation values  $\langle \Psi_m | H | \Psi_m \rangle$  of the eigenvectors of  $F$  are very close to the exact eigenvalues of  $H$ .

This close correspondence of the eigenvectors of  $F$  to the eigenvectors of  $H$  explains why the number of iterations needed to find the minimum does not increase as one introduces localization constraints. Equation (85) remains approximately valid if the occupation numbers for the occupied states are close to 1 and if the occupation numbers for the unoccupied states are very small as well as if the energy expectation values  $\langle \Psi_m | H | \Psi_m \rangle$  are close to the exact eigenvalues of the Hamiltonian. These conditions are fulfilled as discussed above. Hence the condition number for the minimization process does not change appreciably in the truncated case.

All the arguments used to prove the absence of local minima remain valid in the truncated case as well. The force formula Equation (90) remains equally valid.

### 3.5 The Optimal Basis Density Matrix Minimization method

Despite its many advantages in the Tight Binding context, the DMM method has the big disadvantage that it is very inefficient if one needs very large basis sets (i.e. many basis functions per atom). Large basis sets are typically required in grid based Density Functional calculations. In



**Fig. 8:** An analysis of the eigenvectors of the full and truncated density matrix. In the case of the full density matrix the eigenvectors were chosen to be simultaneously eigenvectors of both  $F$  and  $H$ , and the eigenvalues with respect to  $F$  (occupation numbers) are plotted versus the eigenvalues with respect to  $H$ . In the case of the truncated density matrix, the eigenvectors cannot anymore simultaneously diagonalize  $F$  and  $H$ . Therefore the eigenvalues with respect to  $F$  are plotted versus their energy expectation values with respect to  $H$ . Note that in the energy expression (78) the purified density matrix  $3F^2 - 2F^3$  enters instead of  $F$ . The occupation numbers of the purified version are closer to zero or one.

this case it just becomes impossible to calculate and store the full density matrix in the DMM method even though it is a sparse matrix. From this point of view the Wannier function based methods are advantageous since they do not require the full density matrix. The basic idea of the OBDMM method as put forward by Hierse and Stechel [22] and Hernandez and Gillan [23] is now to contract first the fundamental basis functions into a small number of new basis functions and then to set up the Hamiltonian and overlap matrix in this new small basis. A generalized version of the DMM method which can be applied to the non-orthogonal context is then used to solve the electronic structure problem in this basis. The essential point is that one tries to do the contraction in an optimal way. This is done by minimizing the total energy also with respect to the degrees of freedom determining the contracted basis functions  $\Psi_n$ . Formulated mathematically the density matrix is given by

$$F(\mathbf{r}, \mathbf{r}') = \sum_{i,j} \Psi_i^*(\mathbf{r}) K_{i,j} \Psi_j(\mathbf{r}') . \quad (92)$$

The matrix  $K$  is a purified version of the the density matrix within the contracted basis  $L$  and it is given by

$$K = 3LOL - 2LOLOL , \quad (93)$$

where  $O$  is the overlap matrix among the contracted orbitals. The main difference between the formulation of Hierse and Stechel and of Hernandez and Gillan is that in the first formulation the number of contracted basis functions  $\Psi_i$  is equal to the number of electrons, whereas in

the second approach it can be larger. In the formulation of Hernandez and Gillan the basis set can for instance be chosen to have the size of a minimal basis set. The difference to standard minimal basis sets from quantum chemistry is that it is optimally adapted to its chemical environment since the contraction coefficients are not predetermined but found variationally. In practice the full density matrix is found by a double loop minimization procedure. In the inner loop one has the ordinary DMM procedure to find the density matrix for a given contracted basis set. In the outer loop one searches for the optimally contracted basis functions  $\Psi_i$  for fixed  $L$ .

Unfortunately the minimization of the contracted basis functions  $\Psi_i$  is ill conditioned [24] and the number of iterations is therefore at present very large. As already explained before ill-conditioning occurs if the curvatures in the minimum along different directions are widely different. Three causes for the ill-conditioning exist in the OBDMM method

- Length scale ill-conditioning:

This problem is actually not related to the OBDMM algorithm itself but to the (uncontracted) basis functions which are taken to be so-called "Blip" functions in the present implementation. This kind of problem can be found in all iterative electronic structure algorithms if grid based basis functions such as finite elements are used. Its origin is easy to understand. Let us imagine that we are searching for the lowest state of jellium using a localized basis set associated with an equally spaced grid. By symmetry the solution is a constant vector, i.e. all basis functions have the same amplitude in the solution vector. Let us now assume that we explore the energy surface around the minimum along several directions. Let us first "go" into a direction where we add components in such a way that the sign of the amplitude of each neighboring basis function changes. This corresponds to a high frequency plane wave and since the kinetic energy of such a plane wave is big, the total energy will rapidly increase if we add a such a contribution to our solution vector. If on the other hand we add contributions that correspond to low frequency plane waves the energy will increase much more slowly. Since in grid based methods the basis functions are usually narrow and since one can thus construct high frequency functions the condition number can be very bad. As one can suspect from the above explanation the different curvatures can be estimated by doing a Fourier analysis. With this information one can then use preconditioning techniques to cure the length scale ill-conditioning problem. Such a scheme has been proposed [25].

- Superposition ill-conditioning:

This ill-conditioning problem is essentially identical to the ill-conditioning problem of the OM functional. If we have  $N_{el}$  contracted basis functions and no localization constraints the total energy is invariant with respect to unitary transformations of these functions. The introduction of a localization constraint destroys this invariance but there is an approximate invariance left which manifests itself in very small curvatures in the minimum along certain directions.

- Redundancy ill-conditioning:

This problem can only be found in the formulation of Hernandez and Gillan, where the number of contracted basis functions is larger than the number of electrons. In this case one spans a space that contains not only the occupied orbitals but also some unoccupied. As was shown before in the context of the DMM functional introducing a localization constraint will not assign zero occupation numbers, but only very small occupa-

tion numbers to components corresponding to the unoccupied states in the unconstrained case. Since these components corresponding to the unoccupied states have now very little weight they have little influence on the total energy and one has again certain directions where the total energy changes very slowly resulting in very small curvatures.

Another open question is whether the OBDMM has local minima. The functional is a 6-th order polynomial with respect to the expansion coefficients of the contracted basis functions as can be seen from Equation (92) and (93). The two overlap matrices in Equation (93) give each a quadratic term, the two contracted orbitals in Equation (92) a linear term. Minimization with respect to the contracted basis functions should therefore exhibit local minima. Local minima have however not been reported with this method so far. Perhaps the following DMM minimization step which is free of local minima saves the method from overall local minima.

## 4 Comparison of the basic methods

It is certainly not possible to claim that a specific method is the best for all applications. Nevertheless the methods presented so far differ in many respects and one can therefore clearly judge under which limiting circumstances certain methods will fail or perform well. In the following the methods presented so far will therefore be compared under several important aspects. The comparison will be done in two categories. The first category applies to electronic structure methods requiring a small number of degrees of freedom per atom. The Tight Binding method belongs to the first category requiring a few basis functions per atom (or just a few degrees of freedom in the case of semiempirical Tight Binding). But we will also include the standard quantum chemistry methods into this first category, where one typically needs from a few up to a few dozen Gaussian type basis functions per atom. The second category contains methods which are grid based such as finite difference schemes, or where the basis functions can be associated with grid points such as in finite element basis functions or blip basis functions. In these methods one has typically many hundred degrees of freedom per atom. Even though the density matrix is a sparse matrix,  $O(N)$  methods which calculate the full density matrix can not be applied to the second category of electronic structure methods. The memory requirements alone are already prohibitive. As pointed out before we can expect that the localization region in a 3-dimensional structure comprises on the order of 100 atoms. The density matrix will exhibit significant sparseness only for larger system. Assuming that we just have 100 basis functions per atom the storage of the density matrix would require about 1 Gigabyte of memory which is the upper limit of current workstations. The comparison in the large basis set class will therefore comprise only the methods which are Wannier function based namely FOP, OM and OBDMM. The comparison in the small basis set class will comprise FOE, DC, DMM and OM, excluding two methods which are explicitly targeted at large basis sets, namely FOP and OBDMM.

### 4.1 Small basis sets

The comparison of the methods applicable to small basis sets is based on the following criteria:

- Scaling with respect to the size of the localization region:  
The size of the localization region is taken as the number of atoms contained within it.

Only the FOE method has a linear scaling with respect to the size of the localization region. As one increases the size of the localization region the nonzero part of each column of the Chebychev matrices increases linearly implying also a linear increase in the basic matrix time vector multiplication part. In the DMM method the CPU time increases quadratically since the numerical effort for the basic matrix times matrix multiplications grows quadratically with respect to the number of off-diagonal elements of the matrix. From the comparison of the scaling behavior of all these methods one can thus conclude, that the FOE method will clearly perform best if large localization regions are needed. The FOE method is thus also the only method which can be faster than traditional cubically scaling algorithms if no localization constraints are imposed. In this case its overall scaling behavior is quadratic whereas all other methods have a cubic scaling with a prefactor which is significantly larger than the one for exact diagonalization.

- Scaling with respect to the accuracy:

A detailed comparison of the polynomial FOE method and the DMM method under this aspect has recently been given by Baer and Head-Gordon [26] for systems of different dimensionality. Their analysis is based on the assumption that the decay constant  $\gamma$  is given by the tight binding limit of Equation (37). They conclude, that in the one dimensional case the DMM has the best asymptotic behavior, but its prefactor is much larger than the one of the FOE method, so that the FOE method is more efficient in the relevant accuracy regime. In the two dimensional case they have the same asymptotic behavior, but the FOE method has again a much smaller prefactor. In the most relevant three dimensional case the FOE method has both the best asymptotic behavior and prefactor. These results are plausible after the preceding discussion of the scaling with respect to increasing localization region size. When one wants to improve the accuracy the most important factor is the enlargement of the localization region. It is also clear that in higher dimensions the number of atoms within the localization region grows faster than in lower dimensions and that the scaling with respect to the number of atoms will thus become the decisive factor in 3 dimensions. In lower dimensions the number of iterations has higher relative importance, favoring the DMM method. A comparison of the FOE and DMM method applied to quasi two-dimensional huge tight binding fullerenes by Bates and Scuseria [27] is also in agreement with the above statements. They found that the FOE and DMM methods give nearly the same performance with a small advantage for the FOE method.

- Scaling with respect to the size of the gap:

In the FOE method the degree  $n_{pl}$  of the Chebychev polynomial increases linearly with decreasing gap (Equation (69)). At the same time the density matrix decays more slowly. It follows from Equation (37) that the linear extension of the localization region grows as  $\epsilon_{gap}^{-1}$  in the applicable weak binding limit. The volume of the localization region and the number of atoms contained in it grow consequently as  $\epsilon_{gap}^{-3}$ . Taking into account the number of iterations (Equation (69)), the total numerical effort increases thus as  $\epsilon_{gap}^{-4}$ . In the DMM method the number of iterations also increases with decreasing gap but more slowly namely like  $\epsilon_{gap}^{-1/2}$  as follows from Equation (88). Taking into account the above discussion of the scaling properties of the DMM method with increasing localization region we obtain the overall scaling of  $\epsilon_{gap}^{-13/2}$ , which is higher than the scaling behavior of the FOE method. So in contrast to what one might first think the FOE method performs best in this limit. In three-dimensional metallic systems, the FOE method is thus to be

expected to be the only method which will work efficiently at good accuracies.

- Finding a first initial guess:  
No initial guess is required in the FOE method (except perhaps for the potential in a selfconsistent calculation). In the DMM method an extremely simple and efficient input guess for the density matrix is just a diagonal matrix that sums up to the correct number of electrons.
- Cross over point for standard Tight Binding systems:  
The FOE method has the lowest reported cross over point for the standard carbon test-system in the crystalline diamond structure. For the FOE method it is around 20 atoms [20] and for the DMM it is estimated [21] to be around 90 atoms.
- Influence of the range of a sparse Hamiltonian matrix on the performance:  
In the FOE method the numerical effort increases strictly linearly with respect to the number of nonzero elements per column which depends cubically on the range of the Hamiltonian matrix. In the case of the DMM method it can be shown [21] that one has to calculate intermediate product matrices only up to a range which is the sum of the range of the density matrix and the Hamiltonian matrix. As long as the range of the Hamiltonian is small compared to the range of the density matrix the number of operation increases therefore only very weakly with respect to an increasing Hamiltonian range. The DMM method therefore outperforms the FOE method under such circumstances [28]. Hamiltonian matrices of relatively large range are found in the context of Density Functional calculations using Gaussian basis sets. For Tight Binding calculations, in contrast, the range of the Hamiltonian is usually small.
- Scaling with respect to the size of the basis set:  
Let us now consider the case, where the number of atoms as well as all other relevant quantities, such as the size of the localization region, are fixed and where we only increase the number of basis functions per atom  $m_b$ . Both the number of columns  $n$  and the number of off-diagonal elements per column  $m$  of the density matrix will then increase linearly with respect to  $m_b$ . We will also assume that the Hamiltonian is a sparse matrix with  $m_b$  off diagonal elements per column. In the DMM method the numerical effort will consequently grow cubically with respect to  $m_b$ , since the number of operations needed for the multiplication of two sparse matrices of linear dimension  $n$  with  $m$  off-diagonal elements per column is proportional to  $n m^2$ . The FOE method scales cubically, since three factors are increasing. The number of columns in the density matrix, the number of coefficients in each column and the number of off-diagonal elements of the Hamiltonian matrix. In addition to the arguments showing the unrealistically large memory requirements of these methods when used with large basis sets, we thus also find a cubic scaling which prohibits the use of these algorithms in this context.
- Memory requirements:  
The DMM method requires the storage of the whole sparse density matrix. If one takes advantage of the fact that the matrix is symmetric storage can actually be cut into half. In the method only the subparts respectively the columns of the density matrix which are consecutively calculated need to be stored. The storage requirements are therefore greatly reduced compared to the DMM method, namely by a factor of roughly  $N_{el}$ .

- Quality of forces:

In the case of the variational DMM method the force formula is particularly simple (Equation 90) since only the Hellman Feynman term survives. It has to be stressed that this formula is however only exact if one has succeeded in reducing the gradient with respect to all variational quantities really to zero. If, in a simulation, the gradient is not reduced to zero within the required precision because too many iterations would be required, errors will creep into the calculated forces, making them inconsistent with the total energy. From this point of view the situation is easier in the FOE method. Since the FOE method is not an iterative method (in the sense that one iterates until a certain accuracy criterion is met), the force formula of Voter (Equation (58)) will always give forces consistent with the total energy.

Consistent forces are a prerequisite for the conservation of the total energy in Molecular Dynamics simulations. Even with consistent forces there are however other factors which can cause deviations from perfect total energy conservation in Molecular Dynamics simulations such as finite time steps and events where atoms enter or leave the localization region.

In summary, we see that the performance depends critically on many parameters which can change from one application to another. Performance superiority claims based on test runs of a particular system have therefore to be taken with caution.

## 4.2 Large basis sets

Whereas the methods which are mainly applicable in the context of small basis sets showed important differences under the various comparison criteria, the behavior of the FOP, OM and OBDMM methods are quite similar under most of these criteria. The comparison of the methods which are applicable to large basis sets will therefore be based only on a smaller set of important criteria.

- Scaling with respect to the size of the basis set:

As pointed out before the methods compared in this sections all have a reasonable scaling with respect to the size of the basis set allowing thus their use in the context of very large basis sets. In contrast to the discussion of the same point within the small basis set framework, the number of nonzero matrix elements of the Hamiltonian is typically independent of the resolution of the grid, i.e. of the number of basis functions. The most important part of the FOP and OBDMM algorithms, the application of the Hamiltonian matrix to a wave vector scales therefore linearly. At the same time all these algorithms require at some stage the calculation of an overlap matrix among the occupied orbitals. This part scales quadratically as discussed before.

- Finding a first initial guess:

As discussed in the comparison part dealing with small basis sets, it can be difficult to find an initial guess for Wannier function based methods. This problem does not exist in the OBDMM method if the number of orbitals is larger than the number of electrons. In this case the orbitals are just basis functions and by analogy with the usual tight binding or LCAO basis sets it should always be possible to generate a physically motivated initial guess for these orbitals.



- Required number of iterations:  
As mentioned both the OBDMM method suffers from ill-conditioning problems and requires therefore frequently an excessive number of iterations for the iterative minimization. No such ill-conditioning exists for the FOP method.
- Cases where the methods become highly inefficient:  
None of the 3 methods have ever been applied to metallic systems, and they are all expected to fail in this case.

## References

- [1] V. Heine, Solid State Phys. **35**, 1 (1980)
- [2] W. Kohn, Phys. Rev. **133**, A 171 (1964)
- [3] W. Kohn, Phys. Rev. Lett. **76**, 3168 (1996)
- [4] E. Blount, in *Solid State Physics* vol. 13, page 305, edited by F. Seitz and C. Turnbull, 1962, (Academic Press, New York)
- [5] N. Marzari, and D. Vanderbilt, Phys. Rev. B **56**, 12847 (1997)
- [6] S. Ismail-Beigi, and T. Arias, Phys. Rev. B **57**, 11923 (1998)
- [7] S. Ismail-Beigi, and T. Arias, Phys. Rev. Lett. **82**, 2127 (1999)
- [8] S. Goedecker, Phys. Rev. B **58**, 3501 (1998a)
- [9] P. Maslen, C. Ochsenfeld, C. White, M. Lee and M. Head-Gordon, J. Phys. Chem. **102**, 2215 (1998)
- [10] J. des Cloizeaux, Phys. Rev. **135**, A685 and A698 (1964)
- [11] W. Kohn, Phys. Rev. **115**, 809 (1959)
- [12] W. Kohn, Chem. Phys. Lett. **208**, 167 (1993)
- [13] U. Stephan and D. Drabold, Phys. Rev. B **57**, 6391 (1998)
- [14] A. Voter, J. Kress and R. Silver, Phys. Rev. B **53**, 12733 (1996)
- [15] R. King-Smith, and D. Vanderbilt, Phys. Rev. B **47**, 1651 (1993)
- [16] W. Yang, Phys. Rev. Lett. **66**, 1438 (1991) ; W. Yang, J. Chem Phys. **94**, 1208 (1991)
- [17] F. Mauri, and G. Galli, Phys. Rev. B **50**, 4316 (1994) ; P. Ordejon, D. Drabold, M. Grumbach and R. Martin, Phys. Rev. B **48**, 14646 (1993)
- [18] S. Goedecker, L. Colombo, Phys. Rev. Lett. **73**, 122 (1994a) ; S. Goedecker, M. Teter, Phys. Rev. B **51**, 9455 (1995)
- [19] J. Kress and A. Voter, Phys. Rev. B **52**, 8766 (1995)

- [20] S. Goedecker, J. of Comp. Phys. **118**, 261 (1995)
- [21] X.-P. Li, W. Nunes and D. Vanderbilt, Phys. Rev. B **47**, 10891 (1993)
- [22] W. Hierse, and E. Stechel, Phys. Rev. B **50**, 17811 (1994)
- [23] E. Hernandez, and M. Gillan, Phys. Rev. B **51**, 10157 (1995)
- [24] M. Gillan, D. Bowler, C. Goringe and E. Hernández, 1998, in Proceedings of the Symposium on Complex Liquids, 10-12 November 1997, Nagoya, Japan, edited by T. Fujiwara and M. Doi (World Scientific, 1998) page
- [25] D. Bowler and M. Gillan, Comp. Phys. Comm. to be published (1998)
- [26] R. Baer and M. Head-Gordon, Phys. Rev. Lett. **79**, 3962 (1997a) ; Baer, R., and M. Head-Gordon, J. Phys. Chem. **107**, 10003 (1997b)
- [27] K. R. Bates, A. D. Daniels, and G. E. Scuseria, J. Chem. Phys. **109**, 3308 (1998a)
- [28] A. Daniels, and G. E. Scuseria, J. Chem. Phys. **110**, 1321 (1998b)
- [29] B. Roberts, Roberts, K. Johnson, W. Luo and P. Clancy, Chemical Engineering Journal, 1999