Publication Series of the John von Neumann Institute for Computing (NIC)
NIC Series Volume 29

John von Neumann Institute for Computing (NIC)

Mathilde Romberg (Editor)

OpenMolGRID – Open Computing Grid for Molecular Science and Engineering

Final Report

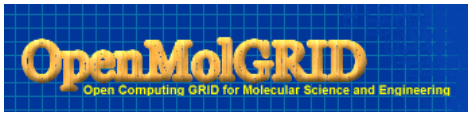
NIC Series Volume 29 ISBN 3-00-016007-8

Central Institute for Applied Mathematics

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

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

NIC Series Volume 29 ISBN 3-00-016007-8



Open Computing Grid for Molecular Science and Engineering

Final Report

of the IST Project OpenMolGRID

Grant Number: IST-2001-37238

Duration: September 2002 - February 2005

The project on which this report is based was funded in part by the European Commission under grant IST-2001-37238. The authors who are listed in appendix 10.3 are responsible for their contributions.

Editor: Mathilde Romberg
Forschungszentrum Jülich GmbH
Zentralinstitut für Angewandte Mathematik
52425 Jülich
Germany

Distribution: public

Internet: www.openmolgrid.org

TABLE OF CONTENTS

1. Executive Summary

Molecular modelling, molecular engineering and drug discovery, as closely related fields, provided the set of real life applications central to the OpenMolGRID project. OpenMolGRID was conceived to exploit the power of Grid Computing to shorten the time to solution for drug discovery, specifically the identification of promising new compounds as potential drug candidates.

The integration of data sources, methods from computational molecular engineering, and knowledge from chemists, toxicologists, pharmacists, and computer scientists has been a major challenge. The integration process required the building of a common understanding and language between the disciplines to aid the production of automated workflows that could be mapped onto Grid resources governed by the UNICORE middleware. This multidisciplinary approach resulted in a set of software components, data and applications integrated into the UNICORE system, which are successfully used on the OpenMolGRID testbed set up among the project partners' sites.

The project demonstrated that the automation of the molecular design pipeline has promising advantages with respect to the manual, stepwise approach. It is much faster and more reliable, offering new tools. This will broaden horizons and open up new challenges related to the larger sets of potential promising compounds.

2. Goals of the Project

Molecular engineering is the task of designing molecular compounds and materials with predefined target properties. The challenge in the industrial application of molecular engineering is to design compounds that up to the present have not been discovered for the intended purpose and can be patented. The design of molecular compounds relies on the knowledge that the properties of molecular compounds are determined by the properties of the molecular fragments and their interaction ([1]). Molecular modelling makes use of this fact by building candidates for chemical compounds with predetermined target properties from appropriate fragments according to established rules. For all generated candidates the target properties are estimated by the quantitative structure-property relationship / quantitative structure-activity relationship (QSPR/QSAR, [2], [3], [4]), or quantum-chemical modelling. Finally, candidates that match the predefined target property are selected for laboratory tests.

QSPR/QSAR relies on the observation that molecular compounds with similar structure have similar properties. For each specific target property a set of molecules is needed for which the experimental property is known. This requires searching globally distributed information resources for appropriate data. For the purpose of exploring molecular similarity, descriptors are calculated from the molecular structure. Thousands of molecular descriptors have been proposed and are used to characterise molecular structures with respect to different properties. Their calculation puts high demands on computer resources and requires high-performance computing. For the available set of compounds with the appropriate target property a model for QSPR/QSAR is developed. This involves finding the most suitable theoretical method and set of descriptors. Finally, the developed model is used to predict the properties for the new molecular compounds.

The main objective of the proposed OpenMolGRID project was to provide a unified and extensible information-rich environment for solving molecular design/engineering tasks relevant to chemistry, pharmacy and life sciences. This was to be achieved by extending the currently used local approach where everything is processed on a local resource to the global dimension by building the OpenMolGRID environment on top of the Grid infrastructure provided by UNICORE (Uniform Interface to Computing Resources, [5]). The planned system was to realise seamless integration of existing, widely accepted, relevant computing tools and data sources ([6]). The proposed system targeted both academic and commercial end-users (especially the chemical and pharmaceutical industries).

The OpenMolGRID system was to comprise a set of application-oriented tools that are built on core Grid services and functions provided by the UNICORE infrastructure ([7]). The specific objectives of the project were as follows:

1. To develop tools that permit end-users to securely and seamlessly access, integrate, and use globally distributed information resources and systems relevant to molecular engineering.

2. To develop tools that permit end-users to securely and seamlessly access, integrate, use, and schedule globally distributed computational methods and tools used for molecular engineering.

3. To provide a realistic testbed and reference application for similar Grid projects in life science and beyond.

4. To promote the use and evolution of both the UNICORE and the OpenMolGRID environment for scientific and industrial end-users.

5. To provide foundations and design principles for developing and constructing next-generation molecular engineering systems.

These initial project goals stayed valid throughout the project but were refined over the time. The project's initial idea for achieving these objectives by focusing on building a powerful toxicity prediction model based on 30,000 newly synthesised and analysed compounds was redirected to put more emphasis on the automation of the drug discovery pipeline. Data management and Grid integration became more important in the light of building a flexible system, which can really serve as the basis for next generation molecular engineering systems (objective 5). Therefore, after the first project year, the work plan was updated to include the grid enabling of the data warehouse ([8]) processes, which are the prominent data management components responsible for harvesting and transforming data from globally distributed information resources (objective 1). The support for classes of applications by OpenMolGRID on top of UNICORE, as well as the integration of specific applications as representatives of specific classes, was the basis for achieving objective 2. Within the testbed, which was established and evolved over the course of the project, the priority for the compute resources to be integrated changed as the project progressed. The existence of a specific application on a computing platform was considered more important for inclusion into the testbed than the computing power of a candidate system. As all application software selected for integration were non-parallelised code and implemented so far for Linux workstations and personal computers only, it did not make any sense to integrate supercomputer systems such as the IBM p690 cluster or Cray T3E.

The technical goal of the OpenMolGRID project can be summarised as ‘automate and speed-up the drug discovery pipeline using Grid technology’. This has been successfully achieved as the remainder of the report will discuss in more detail. Section 3 elaborates on the main results, Section 4 deals with its contribution to standardization, Sections 5 and 6 go into detail about result dissemination and exploitation, while Section 7 deals with project management issues and Section 9 details on the lessons learnt from the project.

3. Main Results

OpenMolGRID achieved an automation of the drug discovery pipeline as laid out in the following figure:

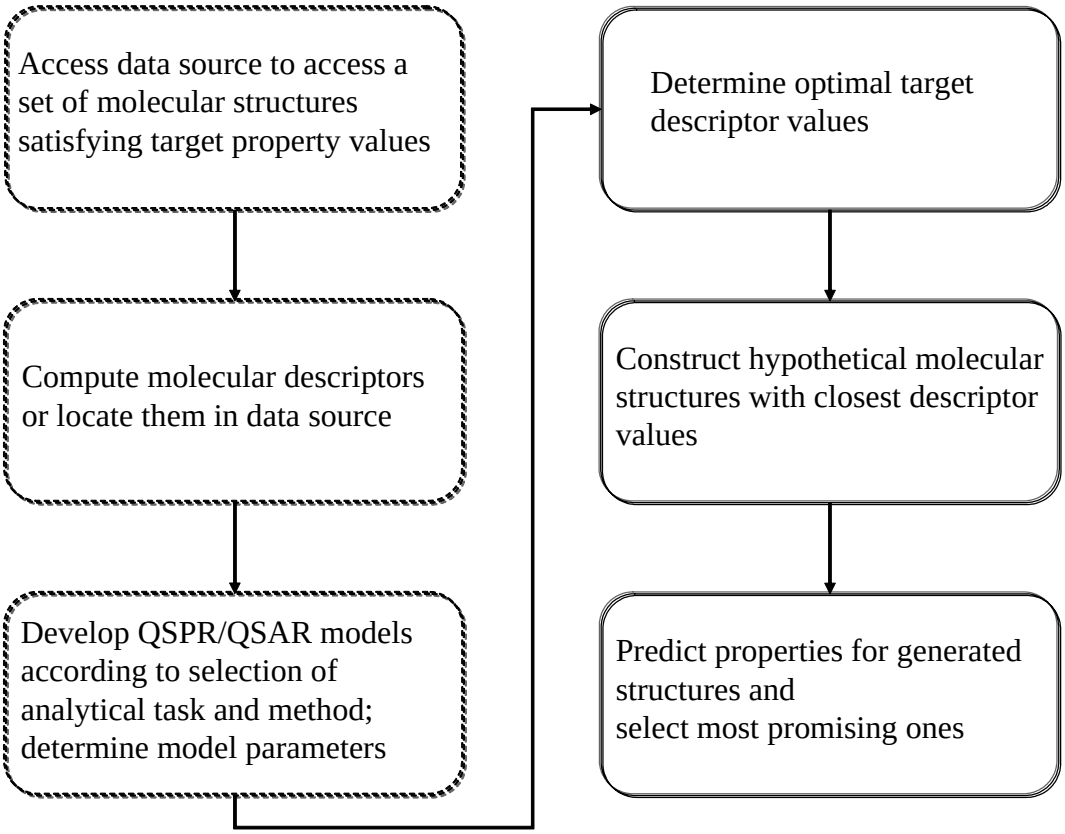


Figure 1: Flowchart of the drug discovery pipeline

The flow chart in Figure 1 shows two major sub-processes of the pipeline: the data mining process (doted line boxes) and the molecular engineering process (continuous line boxes). The whole pipeline relies heavily on high-quality data sources. Therefore a data warehouse was developed which harvests relevant public and private data bases and transforms the data as necessary.

The general approach for automating the drug discovery pipeline on the basis of the UNICORE Grid infrastructure was to add an additional abstraction layer between applications and data sources on the server side and the Client access to these resources as shown in Figure 2.

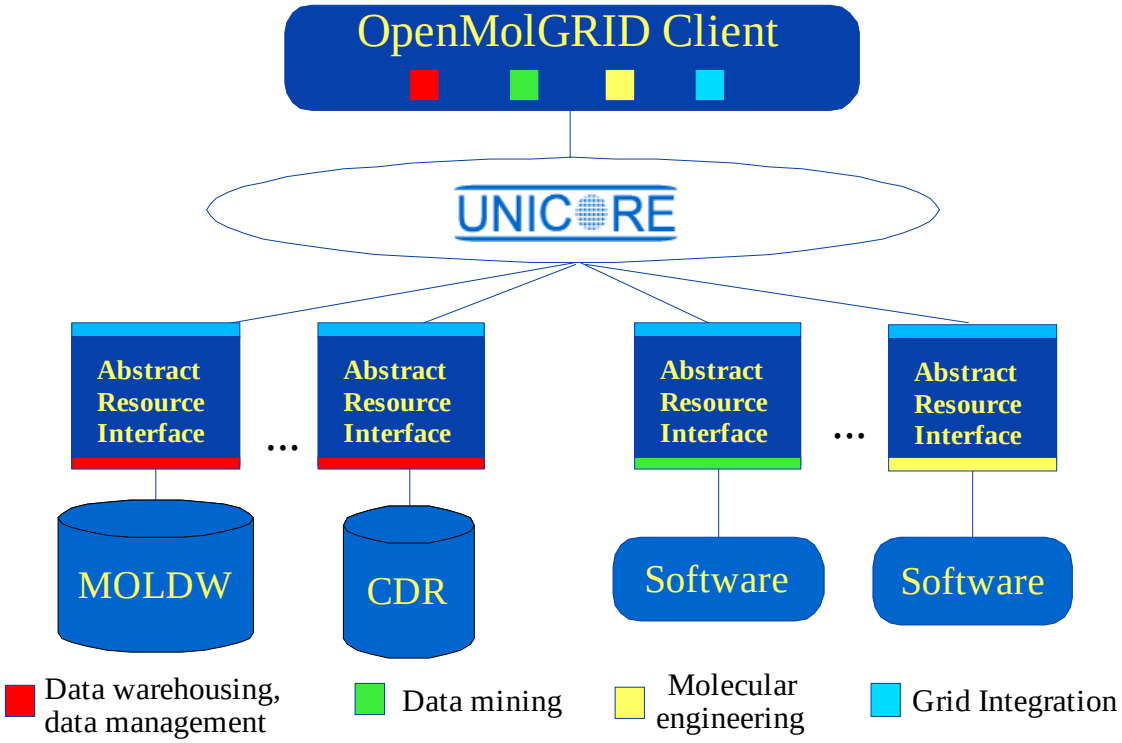


Figure 2: OpenMolGRID overall architecture

With this approach the flexible integration of data sources and application software modules was achieved. It built the basis for the automated workflow support which was realised as Client Plugins. The MetaPlugin (shown in Figure 3, "Add Workflow") together with the Resource Information Provider Plugin (see Figure 4) enable the user to generate a complex UNICORE job from an XML workflow description. The Plugins take care of resource allocation, addition of auxiliary tasks such as data transfers, data format conversions, and distribution of data parallel tasks onto the available systems thereby facilitating the user's task and shortening the time to solution.

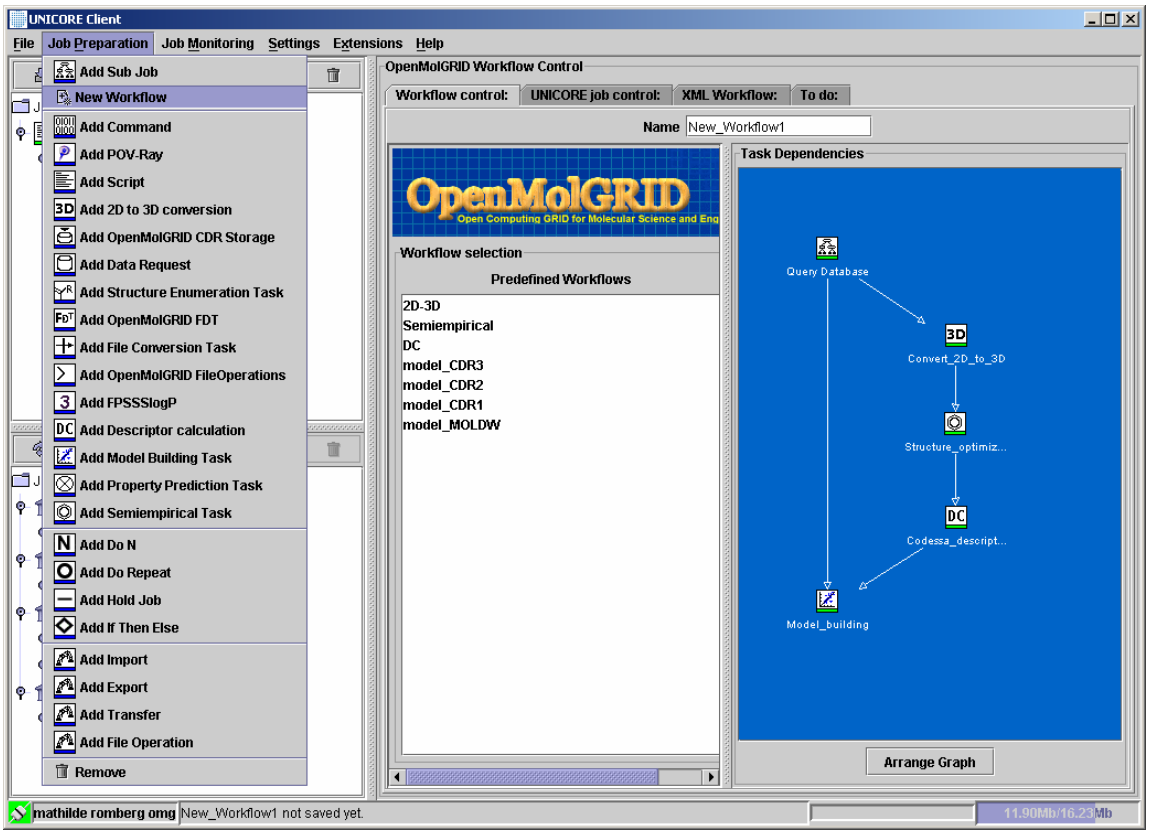


Figure 3: UNICORE Client with OpenMolGRID extensions

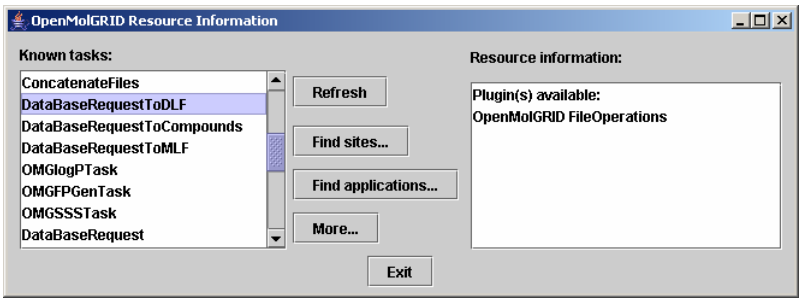


Figure 4: Resource Information Provider Extension Plugin

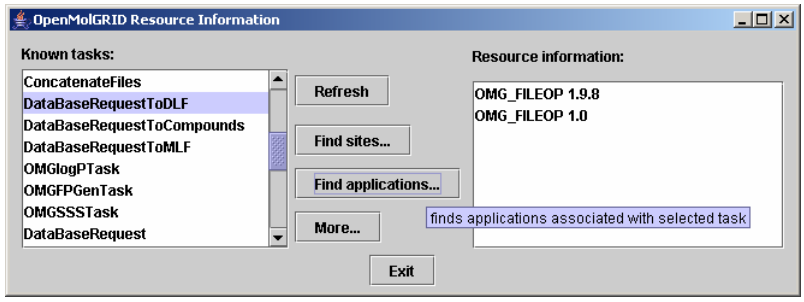


Figure 5: Resource Information Provider's application information

Figure 3 also shows the list of developed Client Plugins in the Job Preparation list on the left: from 'Add 2D to 3D Conversion' to 'Add Semiempirical Task'. All of these correspond to one or more software modules and their abstract interfaces on the server side. Figure 5 shows an example where two versions of an application are available for a task.

On the basis of this system QSAR/QSPR models were developed and predictions of properties of newly generated compounds were performed. The developed models were compared with manually built models, the outcome of which illustrated that the OpenMolGRID system achieved *at least* the same quality of results. The remainder of this Section presents the project results in detail.

3.1. Data Management

3.1.1 Data Warehouse

Predictive QSAR/QSPR modelling requires the handling and management of chemical structure and property data and data relating to molecular descriptors. This data is often not readily available and must be retrieved from public data repositories. Furthermore, the data must be integrated and formatted so that it is amenable to data mining methods such as linear regression methods, artificial neural networks, and decision tree algorithms. Data warehousing (see [10]) is often employed to provide the data integration and formatting functionality needed by data mining applications. A data warehouse integrates, cleanses, normalizes, and consolidates data from different sources and maps them onto "ready-to-use" data structures (e.g. by de-normalizing relational database tables). A key component of the OpenMolGRID system is to provide a Grid-enabled data warehouse for molecular engineering environments.

The main purpose of the OpenMolGRID data warehouse is to provide integrated and consolidated data originating from selected public data resources relevant to

molecular engineering. Currently, the following data resources have been integrated:

- National Toxicology Program database provides information about potentially toxic chemicals to health regulatory and research agencies, the scientific and medical communities, and the public NTP ([11]),
- ECOTOX (ECOTOXicology) databases Aquire and Terretox, which provide chemical toxicity information for aquatic and terrestrial life respectively([12]),
- MDR (Multi-Drug Resistance) Data Set, provided by ComGenex, and
- GPCR (G-Protein Coupled Receptor) Data Set, provided by ComGenex.

The databases integrated in the OpenMolGRID data warehouse are harvested from public web sites and are mapped into the warehouse and its physical repository. Currently, access to the warehouse is restricted to members of the OpenMolGRID consortium, involving partners from Germany, Hungary, Estonia, Italy, and the UK. A single physical repository resides on a server in the UK. A more detailed view of the OpenMolGRID data warehouse and its relation to the web and other OpenMolGRID components is depicted in Figure 6.

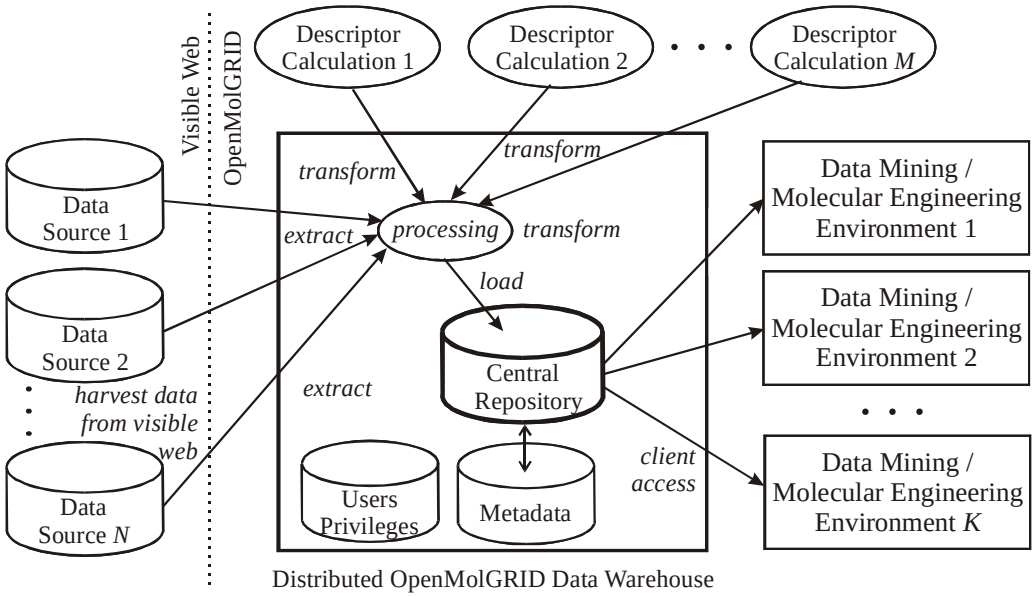


Figure 6: OpenMolGRID data warehouse and related components

The warehouse processes follow the typical *extract*, *transform*, and *load* scheme (also known as ETL). The entire process is performed periodically, thus the warehouse is able to reflect changes (i.e., updates) in the underlying databases.

The extract component accesses the underlying molecular compound data from public web sites and transfers the database as single or multiple files (depending on the database) to the data warehouse. Access to databases is performed over HTTP, HTTPS and FTP protocols. Each database has its own implementation specific format e.g. self-extracting compressed data archives, well-defined XML formats, consisting of single or multiple files ranging from a few kilobytes in size to several tens of megabytes. Once within the data warehouse environment, the data is extracted from the files and mapped into the data transformation environment.

Key functions of the data transformation environment are to de-normalize the data from relational databases, cleanse the data (remove inconsistent entries), enrich and standardise the data based on the requirements of the molecular engineering environments. For example, new fields are computed to facilitate different types of analysis or as “convenience” fields. The log-inverse of the measured dosage of a chemical’s toxicity, for instance, is often more useful for certain calculations and models than the toxicity value itself. By providing this value within the warehouse’s data structures, the user does not have to perform this calculation and can focus on the more intricate aspects of the modelling task at hand. Data normalization may involve, for example, missing value imputation, mean centring, or alignment to canonical units. In addition, complex data transformations (descriptor calculations) were integrated into the warehouse (see section 3.1.4).

The transformed data is then loaded into the data warehouse’s physical data storage, which is realised as relational database management system in OpenMolGRID. After a detailed comparison of open source solutions, the PostgreSQL platform was selected. Client access to data in the OpenMolGRID data warehouse is enabled via a generic database access tool developed by the OpenMolGRID project. Inputs and outputs are encapsulated in an OpenMolGRID-specific XML syntax and data are easily identifiable due to being associated with generic data types defined especially for OpenMolGRID’s data needs. These data types are used throughout all applications in the OpenMolGRID system. In effect, data is abstracted and translated to the particular data format required at a target site (in accordance with UNICORE’s approach to resource abstraction). The database access tool is a command-line tool, but an intuitive GUI has also been developed to formulate queries by using data entry forms. Data is transported using UNICORE’s file transfer mechanism.

What is interesting, and to some extent novel, from a Grid perspective are the following aspects of the OpenMolGRID data warehouse:

- First, the realisation of a complex workflow that requires the interoperation of various data *and* computing systems across several European countries. This is made possible by the components that have been developed on top of the UNICORE infrastructure as part of the project: The components include the Database Access and Database Input Tools, Abstract Resource Interfaces and Client plugins for a set of software modules, the MetaPlugin and Resource Information Provider plugin for automated workflow support, and the Command Line Client.

- Second, the integration of physically distributed, complex data transformation procedures as part of the data warehouse's transformation environment. Specialised software is required to perform these calculations and typically they are expensive to compute, especially if there are a large number of chemicals and several representations of the same chemical. Clearly, when the data warehouse is updated, it will only update an entry and re-compute descriptors if the entry in the underlying database has been modified, avoiding needless computation. The OpenMolGRID data warehouse effectively "caches" computations (i.e. stores the results of computations) and is thus facilitating more efficient data mining downstream, as it removes the burden from data miners and molecular engineers to carry out the required integration and transformations.

Within OpenMolGRID, we have developed a highly generic data transformation environment, using a flexible XML approach. The overall process can be broken as follows:

```
Input:      XML script
Taxonomy:   Allows for substitution of values.
Unit:       Provides functionality to standardize units within
            the OpenMolGRID data warehouse.
Calculation:    This is a generic engine for calculating and
transforming values.
Output:         XML script
```

In many data resources there are inconsistencies in the way the same data (types) are represented in different records (the idea of a record varies from source to source). For example, supposing we have decided that the standardised data unit for a particular dosage field is milligrams per kilogram (mg/kg), there may be variations in the way a source represents this. Some records may contain Mg/Kg (or some other variation) and thereby causing inconsistencies with the standard realised in the OpenMolGRID warehouse. The Taxonomy step in the process flow

described above enables any number of substitutions to be defined to ensure that consistency is maintained within, and between, data resources entering the warehouse.

Characteristic of many data sources is the idea that each data field contains a value from a set of allowable values. This can be problematic in cases where a number of resources are being integrated into a data warehouse. A dosage field can have several measurement units associated with it, e.g. g/kg, mg/kg, or µg/kg. As this data is often used in data mining and molecular engineering, there is a need to align these units. In the absence of a data warehouse, this must be done manually, but in the OpenMolGRID data warehouse, we require an automated mechanism. The mechanism developed for the OpenMolGRID data warehouse revolves around the concept of a *canonical unit* or *unit primitive*. Measurement units can be broken down into several categories, e.g. length, weight, distance. Each of these categories has an associated base unit, the unit primitive, e.g. kilograms for weight. To convert between various forms of this category, mathematical calculations are performed, e.g. to convert from grams to kilograms, the grams value is scaled with a division by 1000. By defining the scaling factors (which can be more complex mathematical formulations) between various forms of the same measurement category, in both directions, it is possible to dynamically convert one unit to another.

Data transformation environments are usually characterised by more complex data transformations than simple unit standardisations. Within the OpenMolGRID data warehouse there are several complex transformations that can be performed. For example, based on a particular dosage it is possible to calculate the log-inverse of this dosage. To do this we encapsulate the processing logic in a rule, as shown below.

```
<RULE name="Calc Log Inverse“ link=“Log Inverse">
        $Log Inverse$=Math.log(1/$Mol Dose$);
</RULE>
```

Given that any piece of Java code (that will compile) can be embedded within the rule, effectively any complex transformation can be enabled via this mechanism. Notably this mechanism can be used to call Grid interfaces.

3.1.2 Custom Data Repository

While the OpenMolGRID data warehouse provided the read-only data source with the well prepared data as described above another database was needed for storage and retrieval of results generated during the use of the OpenMolGRID system

(e.g. generated structures, predicted properties, QSAR models, etc.). Therefore the *Custom Data Repository* (CDR) was developed.

The CDR is based on a relational database management system and has been designed to support the complete molecular engineering process realized in the OpenMolGRID system. It is capable of handling all data generated during the normal use of the OpenMolGRID system (molecules, descriptors, models, experimental property values, predicted property values, etc.). The database part of CDR is implemented in PostgreSQL considering all features of this database management system.

During the course of the project, the CDR database has been populated with compound structures and related experimental data values (100 Multi-Drug Resistance, 100 G-Protein Coupled Receptor Activity and 30000 *in vitro* human fibroblast cytotoxicity), as well as several thousands of molecular structures generated using the OpenMolGRID system.

For data retrieval, the Data Base Access Tool component (DBAT_CDR) was used. DBAT_CDR is based on the generic version of the Database Access Tool (DBAT) developed for SQL databases, but minor changes to make it CDR specific have been applied.

For the import of data, the CDR Data Base Import Tool (DBIT_CDR) was designed and implemented. It is responsible for inserting data generated by OpenMolGRID modules into the CDR database, as well as to import user specific (e.g. experimental) data. The generalised architecture is shown in Figure 7.

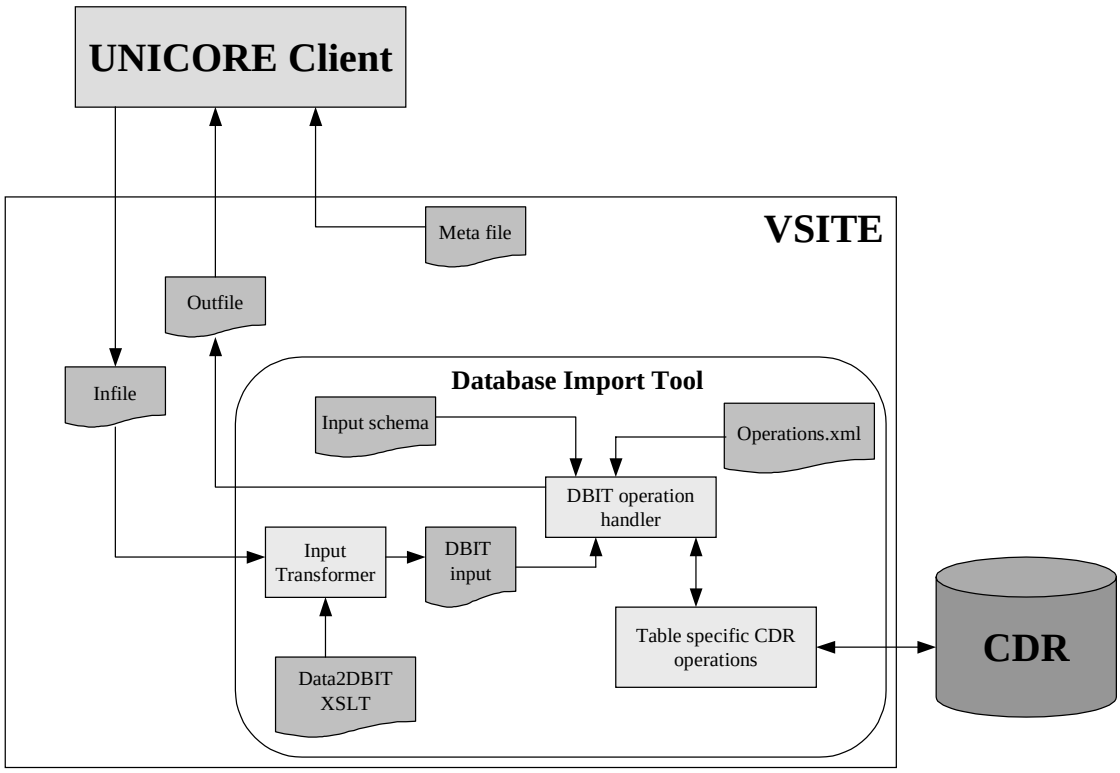


Figure 7: DBIT_CDR architecture

DBIT_CDR has the following components:

- The Input Transformer is responsible for processing incoming XML files from UNICORE modules such as descriptor calculation, model building, 2D-to-3D structure conversion, SDFile upload, etc. This component contains the business logic of DBIT_CDR mainly in XSLT and STX form. For each XML input schema, an XSLT/STX file has been created. For large input XML files the XSLT/STX transformers have been replaced with dedicated transformers written in Java.
- DBIT Operation Handler processes incoming XML files according to the operations.xml file containing all implemented table operations.
- CDR Operation component contains the table specific insert functions with necessary data existence checking.

3.1.3 Substructure Search

The data warehousing and data storage concept in the OpenMolGRID system includes substructure search capabilities. This function is important for identifying

the best subset of data (chemical compounds) to be used for further analysis and is fundamental in chemical and related communities. The substructure search is essentially a two-part process and results in the need for two different queries. The first query aims to select a subset of structures that may contain the substructure. The selection procedure is carried out using a fingerprinting approach, which significantly accelerates the search. Fingerprints of the structures are matched against the fingerprint of the substructure and those structures that cannot possibly match are removed from the set of matches. The second query is performed within the matching subset to select structures that actually contain the full chemical substructure. The comparison in the second step is computationally expensive. Therefore, the first step reduces the input set for the second query. To improve the performance of substructure search the data warehouse adopted the "fingerprint" approach outlined above. This required the fingerprint data to be stored in the warehouse for chemical structures. Special fingerprint generation and substructure search programs were invoked remotely via OpenMolGRID. This process illustrates the ability to use other Grid resources in a distributed data warehouse solution.

3.1.4 Complex Data Transformations

The data warehouse also contains a set of molecular descriptors, which are derived from the molecular structure information of the stored compounds via computationally intensive descriptor calculation workflows. From a data warehouse perspective, these descriptor calculations are complex data transformations. Therefore, the most frequently used molecular descriptors are calculated for each molecular structure in the data warehouse. Beside the traditional molecular descriptor types, a physico-chemical parameter, the log*P* value (octanol-water partition coefficient) is also calculated for the compounds (using an adapted version of the PrologP software [13]).

Essentially, the descriptor calculation procedure amounts to virtualisation of parts of the data warehouse's data transformation processes. This virtualisation functionality is realised by the development of the Command Line Interface, which is detailed in section 3.4.3.

3.2. Model Development

The predictive model development with the help of data mining techniques is one of the main applications of the OpenMolGRID system. Several state of the art tools that are required for the development of quantitative structure-property/activity relationship (QSPR/QSAR) models have been adapted, as summarised in Section 3.2.1. The QSPR/QSAR models can be applied for the

estimation of various chemical properties and biological activities. The QSPR/QSAR model describes the modelled activity or property as a mathematical function of a molecular structure. The molecular structure is characterised in these models by molecular descriptors. The QSAR/QSPR models are particularly suitable for drug design, material design, molecular modelling, and chemical engineering problems.

The adaptation of QSPR/QSAR modelling tools for the Grid environment was especially challenging for several reasons. One of the main reasons is that the model development is multidisciplinary by its nature. It involves knowledge from such diverse fields as data mining and management, computational chemistry, and statistical analysis, not to mention the fact that this was the first attempt to bring these tools to a Grid environment. Sections from 3.2.2 to 3.2.5 describe experiences collected through different phases of the project, including specification, implementation, and testing.

3.2.1 Available applications

The OpenMolGRID system has been augmented with Grid adapters (implementations of the abstract interface) for several existing software packages that are required for carrying out tasks in QSAR/QSPR model development workflows.

- **2D to 3D conversion:** The MOLGEO ([18]) software has been adapted for the conversion of 2D structures to 3D representations. This is a common data pre-processing task in the QSPR/QSAR modelling, since the 2D representation is very convenient for the end-user to sketch molecular structures and most chemical databases have only 2D representations available. However, all quantum chemical and most molecular descriptor calculation programs require the 3D representation of molecular structures as an input.
- **Semi-empirical quantum chemical calculations:** The MOPAC (version 7, [19]) software has been adapted for the semi-empirical quantum chemical calculations. MOPAC is a general-purpose semi-empirical quantum mechanics package for the study of chemical properties and reactions in gas, solution or solid-state. The output from MOPAC calculations is used to calculate quantum-chemical descriptors (e.g. dipole moment, heat of formation, energy partitioning, reactivity indexes, etc.) for QSAR/QSPR model development.
- **Descriptor calculation:** The MDC (Molecular Descriptor Calculator) module from the CODESSA PRO ([20]) software has been adapted for the molecular descriptor calculation. In addition, the prediction engine of the PrologP software [13] has been adapted for the calculation of logP. Currently the

system incorporates a wide range (about 1000) of molecular descriptors, describing constitutional, topological, structural and electronic features of structures. The descriptor calculation task is applicable both to 2D and 3D structures, although 3D structures provide more information rich description of molecules.

- **Model development:** The MDA (Molecular Descriptor Analyser) module from the CODESSA PRO software has been adapted for the QSAR/QSPR model development. Multiple statistical methods are available for the development of predictive models, including Multilinear Regression Models (MLR) and Partial Least Squares (PLS). Several selection algorithms are available for descriptor selection in the effective search of the best (most informative) multi-parameter correlations in the large space of the natural descriptors. The prediction capability of the model is judged by statistical parameters calculated for the model, various cross-validation techniques, internal and external validation sets. Visualisation tools are available for plotting actual vs. predicted activities/properties and residuals.

3.2.2 Specification

The specification phase was extremely important, because use cases and user requirements were collected and communicated between all partners. This was a challenging task, because a common language had to be developed between chemists, toxicologists, and computer scientists. In addition, integration with different subsystems had to be considered (e.g. UNICORE middleware, data warehouse, custom data repository, and molecular engineering environment). Based on this information a general strategy for the OpenMolGRID system was designed.

From the very beginning of the specification process a need for the automated workflows was apparent. A typical workflow is described in Figure 8. Other requirements were flexibility and extendibility. For example, each step in this workflow could be performed with different software packages, some processing steps removed, or additional processing steps added. Therefore a service-oriented architecture was selected. Instead of adapting specific software packages, we designed high level interfaces for more specific application classes or tasks. As a result, the communication in the Grid layer can be performed without going into software specific details. Only one client plugin has to be developed for a task, while multiple software packages can be adapted for carrying out the same task. The workflow support architecture of the OpenMolGRID system is described in more details in Section 3.4.2.

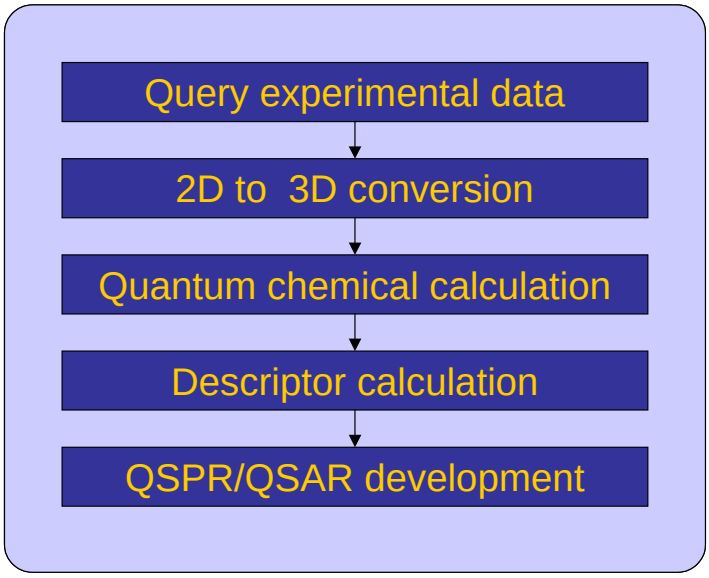


Figure 8: A typical model building workflow

Application neutral data formats were defined for relevant data types. Available data formats were analysed and reused when appropriate. For example, the chemical mark-up language (CML) is very versatile and perfectly suitable for representing molecular structures. In some cases (the output of semi-empirical calculation and predictive models), no appropriate format was readily available and it was not feasible to develop a new one from scratch. In these cases, more specific data formats were used. Details about application neutral data formats are available in [21].

3.2.3 Server side components

A typical model building software package does not consider the communication of Grid resources so it has to be adapted for the Grid later. The main requirement from the UNICORE infrastructure for adapting a software package is the possibility to use it from the command line. The UNICORE application is normally a thin wrapper that is called by the target site interface (TSI) to execute the software package. If this requirement is met, then it is very straightforward to adapt new applications for UNICORE, because no changes, or very minimal changes at worst, are required to software packages. This approach is particularly well suited for integrating legacy applications. Of course, this means that without access to the source code it is not possible to adapt proprietary applications that can be controlled only from an already existing GUI. Since all required software packages had a command line interface (CLI) available, integrating them to the UNICORE was straightforward. In a few cases minor modifications to the source

code was required compile it and get make it executable under Linux operating systems.

The main development effort was spent on the development of application neutral data formats for input and output files. The data was represented by XML mark-up and designed for easy transformation between different application classes. In most cases transformations are possible with simple XSL transformations (XSLT).

Based on these experiences, we can conclude that the integration of existing software packages to UNICORE infrastructure is very straightforward. If the application is required as a part of a workflow, then it is worthwhile to standardise the file formats, although for some applications this can be a major undertaking.

3.2.4 Client plugins

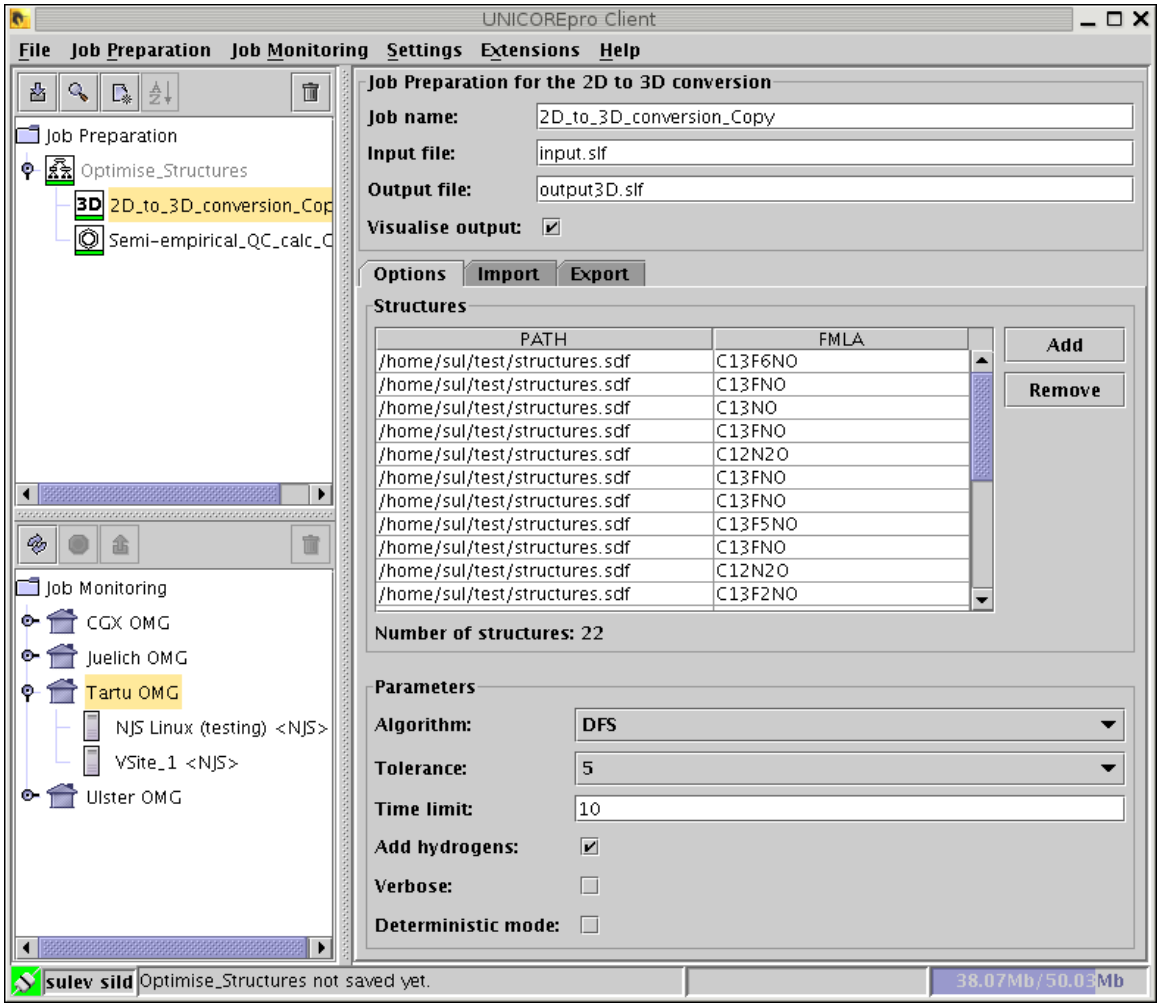


Figure 9: Input preparation for the 2D to 3D conversion task

For any kind of application a good user interface is essential to make the application accessible to the user. The strength of the UNICORE infrastructure is that it provides a basic framework for preparing client plugins for UNICORE applications. While creating a client plugin, we only had to provide the following functionality to the plugin:

- Implement IChainable interface for the workflow support.
- Implement GUI panels for the preparation of input data for a job. See an example screenshot for 2D to 3D conversion task in Figure 9.
- Implement GUI panels for the visualisation of results. See an example screenshot for the model building output in Figure 10.

- Convert input files provided by the end user to application neutral formats when they are in popular formats supported by major third party software.

All other Grid specific details (e.g. the submission of jobs, data transfer, job monitoring, etc) are carried out by the UNICORE infrastructure. At the same time it also is a weakness of UNICORE, because existing user interfaces cannot be reused unless they have been implemented in Java. This was the case in this project, since the GUI for CODESSA Pro software has been implemented in C++ and all the GUI panels for client plugins had to be re-implemented in Java. Fortunately, there were several high quality open source libraries (e.g. Chemistry Development Kit, Jmol, and JFreeChart) available that made this task much easier.

The design of the GUI in client plugins is important, because it is necessary to handle the added complexity from the Grid interaction and hide it from the user wherever possible. For instance the different formats of input data sets from different sources have to be dealt with properly as within OpenMolGRID the input data may come from different sources, such as a remote database, local or remote file systems, and output from some other process. Some users have criticized that this makes the system more complex and sometimes slower for very simple tasks.

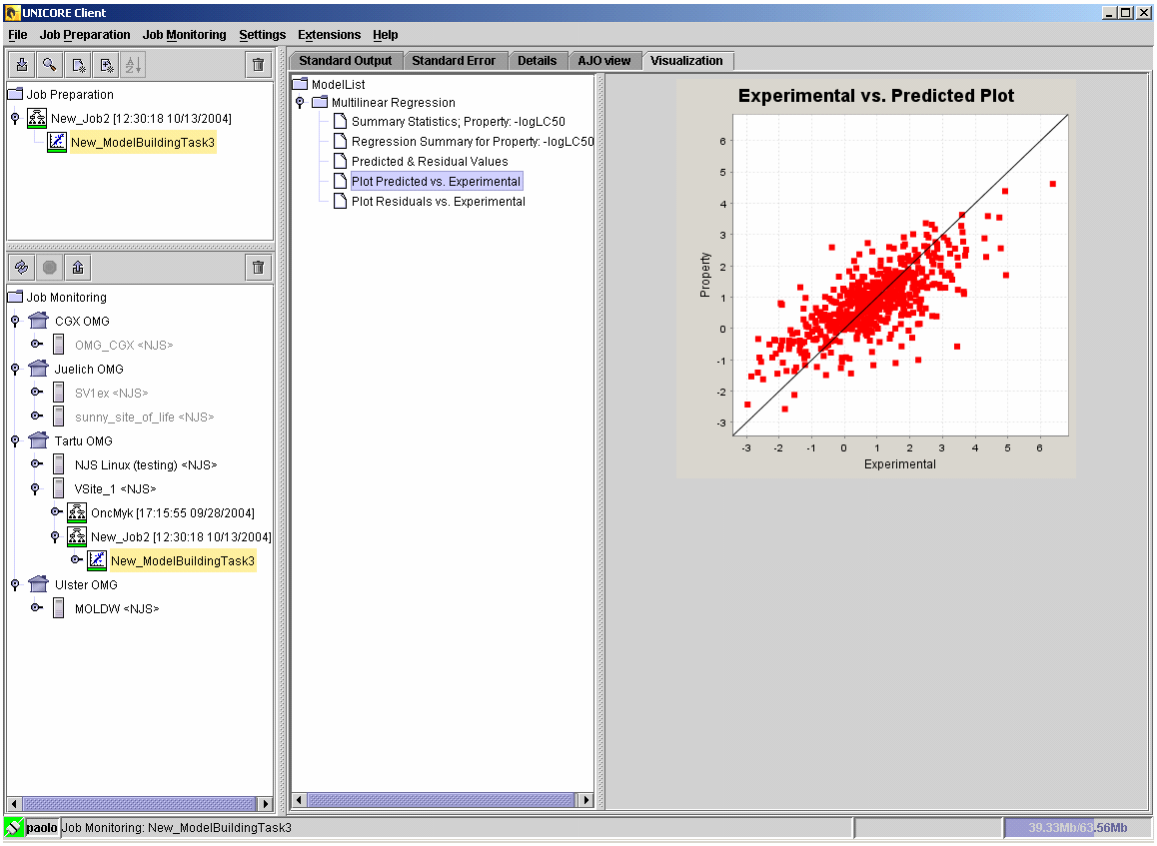


Figure 10: Visualisation panel for the model building output

3.2.5 Testing summary

The model building applications have been in production use since June 2004. They have undergone extensive testing. The applications have been used with different real life data sets and have been subject to integration testing under various workflows.

Initial tests with smaller data sets were successful. However, when testing started with larger data sets and more complex workflows, then some performance problems were experienced. In particular, the memory usage on the client side caused problems when large output files were visualised. As a response, rather minor changes were required to the relevant data structures to reduce the memory consumption. The server side components had some performance problems as well. Most of them were addressed by tuning configuration parameters (e.g. memory limits and number of concurrent connections). Overall, the OpenMolGRID system has been proved to be stable and reliable.

We have found that the UNICORE infrastructure combined with the OpenMolGRID architecture provides a solid foundation for various model-building applications. It's service-oriented architecture makes it easy to plug in different applications to automated workflows and eliminates the time consuming manual processing of intermediate results. In addition, significant speedups are achieved by the seamless distribution of data parallel tasks over the available Grid resources (see section 3.4.2).

3.3. Molecular Engineering

The application of predictive models has a major importance to molecular engineering tasks. Within the OpenMolGRID project we have developed Grid-enabled tools that can be used for predicting various chemical properties and biological activities. In addition we have designed and implemented algorithms for constructing new molecular structures and rapid methods for testing their properties.

This has been done in the same way as for the development of model development tools. This ensured that the developed prediction tools could be easily combined with different tools used in the QSPR/QSAR model development process. The following sections summarise the experiences that were acquired during the development of different prediction tools.

3.3.1 Property/Activity prediction

As described in Section 3.2.3, the main requirement for adapting existing applications is the possibility to use it through a command line interface. While the integration of model building applications was rather straightforward, it was not the case with the prediction software. The prediction task was planned to be implemented from the CODESSA Pro software, but in CODESSA Pro the prediction task is directly performed through the graphical user interface, so it was not possible do develop wrappers that can control this task through the command line interface. Therefore, it was necessary to use a different approach. Since we had access to the source code of the CODESSA Pro software; we could use the prediction module as a library and link it with the UNICORE application.

The client plugin offers a GUI panel for preparing the input for the prediction task. It is possible to select models for the prediction from the client machine or the model location on the target system. The predicted properties can be seen on the visualisation panel after the execution of the prediction task.

3.3.2 Molecular descriptor prediction

The rapid validation of candidate molecules is extremely important to select best molecules that match the target properties or activities. The traditional QSPR/QSAR models are good for this kind of prediction. Unfortunately, this approach depends on a time-consuming molecular descriptor calculation step, which limits this approach to smaller data sets. Alternatively, it is possible to use fragment descriptors to predict molecular descriptor values, which in turn can be used in traditional QSPR/QSAR predictions. This approach is suitable for huge data sets, because the time consuming molecular descriptor calculation step is avoided and the QSPR/QSAR predictions can be performed at a much higher rate.

The OpenMolGRID system offers a component for molecular descriptor prediction that is integrated with the structure generation engine. Structure generation algorithms use this component to eliminate unsuitable candidates on the fly.

3.3.3 Structure generation

We have developed software implementing various algorithms that use the fragment-based approach for generating new molecular structures. Based on the fragment structures, a combinatorial library is defined that contains one core structure with one or more substitution sites. The combinatorial library has rules for attaching available substituent fragments to the substitution site(s). Since the number of potential combinations can be huge, various structure generation algorithms (e.g. full enumeration, stochastic algorithms) can be used to generate molecular structures. The design of the system allows for easy integration of different structure enumeration algorithms into the OpenMolGRID system. Like for the model building applications, the structure generation application consists of a generic client plugin and UNICORE application wrapper.

The client plugin provides options for selecting an algorithm for the structure generation. Since the data representation of the combinatorial library is rather complex, a user-friendly tool (Figure 11) is available in the job preparation area for its definition and editing. In addition, it is possible to select predictive models for eliminating unsuitable candidates. After the execution of the structure enumeration task the generated structures (Figure 12) can be visualised.

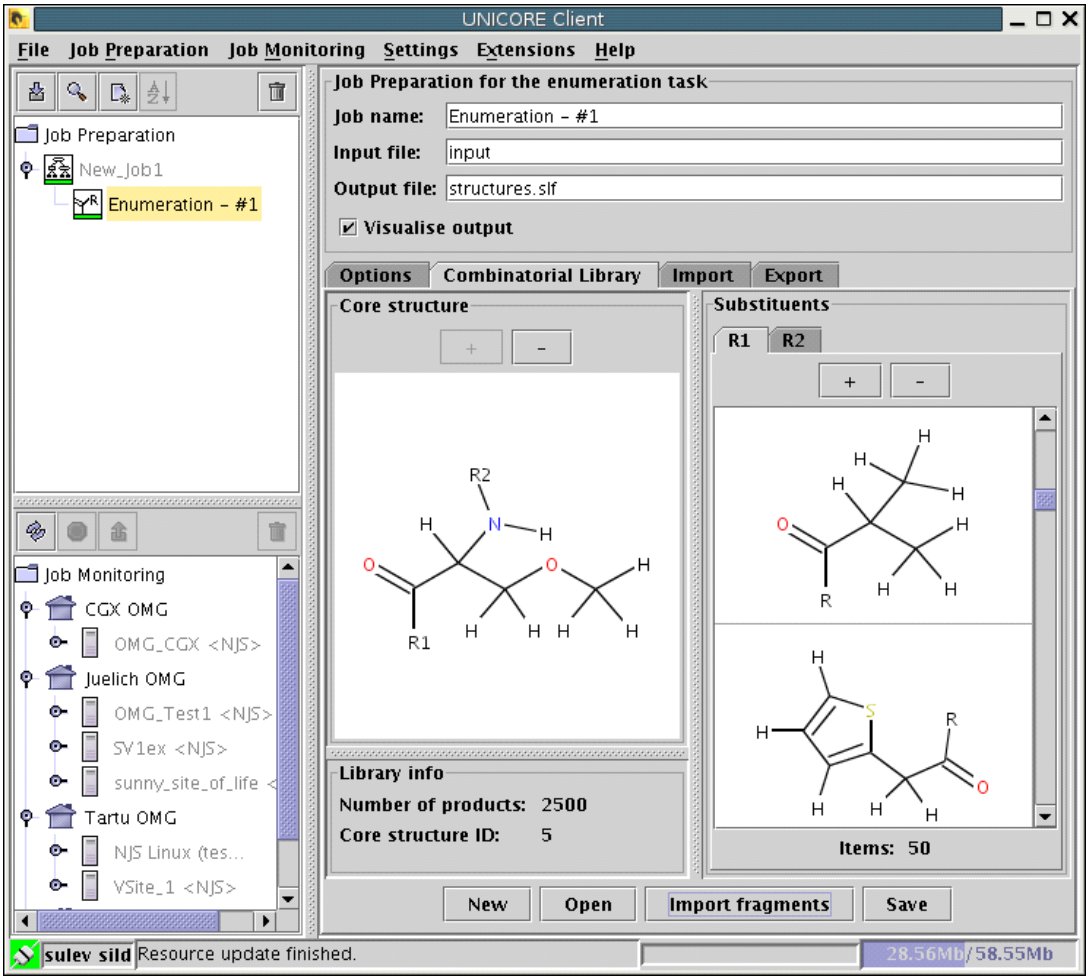


Figure 11: Input preparation for structure generation task

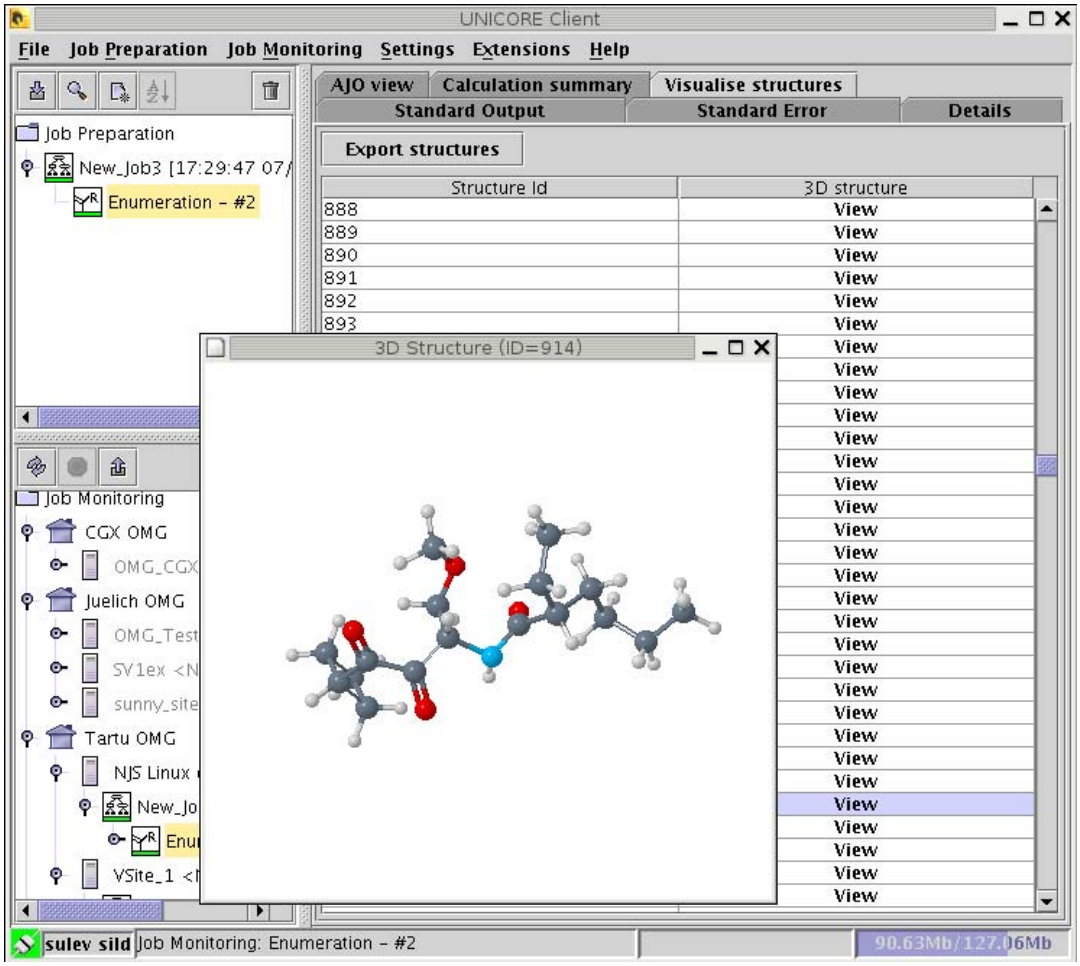


Figure 12: Output form the structure generation task

3.3.4 Fragment library

The fragment library contains structures that can be combined by structure generation algorithms to produce complete molecules. The fragment structures are divided into two classes, i.e. (1) generic core structures and (2) substituents, as illustrated in Figure 13. The core fragment has one or more open valence nodes, marked with * that are called substitution sites. Each substituent has only one open valence node (substitution site). The representation of fragment structures is virtually identical to the representation of molecular structures. The only difference is that one or more atoms in the molecule are defined as substitution sites.

Fragments are represented by their 2D and/or 3D structures and characterised by several attributes, such as fragment type, fragment code, and fragment descriptors.

For the calculation of fragment descriptors we have extended the descriptor calculation tools described in Section 3.2. It required some changes to the adapted software packages. There were no changes required to applications neutral data, because the chosen data formats are equally applicable to molecules and fragments.

The fragment library uses the Custom Data Repository for the storage of fragments. This approach allows exploiting the generic access mechanisms provided by the database access tool and using the DataBaseRequest Plugin as the user interface.

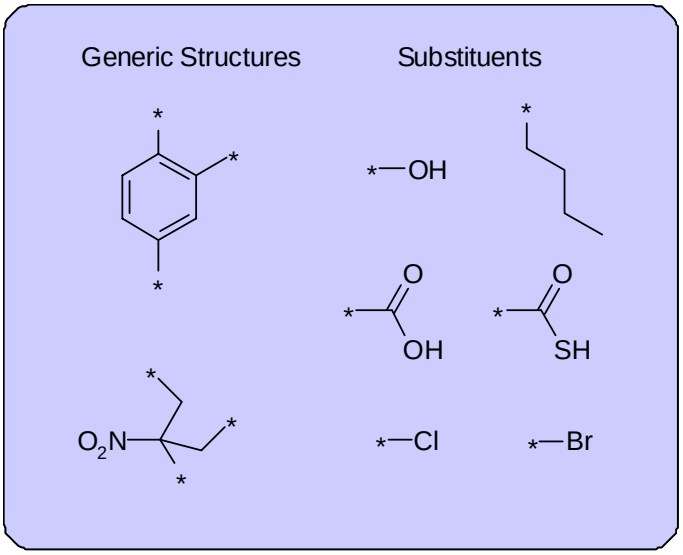


Figure 13: Fragment structures

3.4. UNICORE Extensions to Support Molecular Engineering

3.4.1 DataBase Access

A general database access mechanism through UNICORE has been designed, and implemented for the databases of interest within the project. Figure 14 illustrates the basic architecture. To achieve a seamless interface, a server-side wrapper application called Database access tool encapsulates the communication with the underlying database system. The output data are sent to the client in an XML format that was designed for easy automatic processing using for example XML transformation stylesheets.

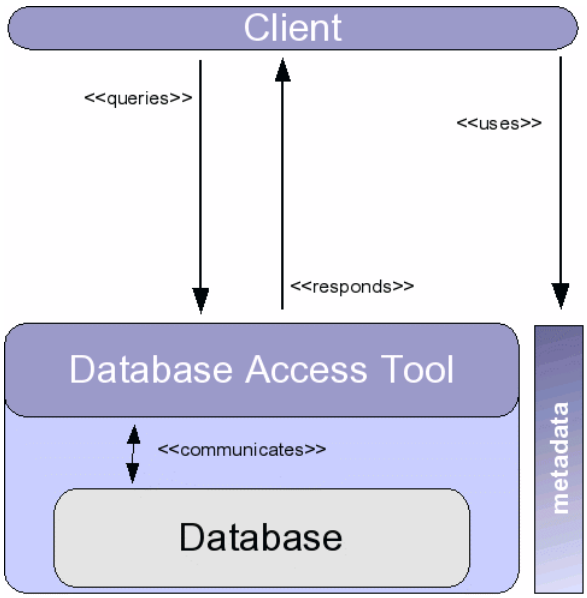


Figure 14: Database access architecture

Metadata are used to provide a description of the database's layout as well as semantic information about the database's content to the client. This architecture allows accessing all of the databases within the project with one client-side plugin. In this way, the UNICORE paradigm of seamless access to different resources has been followed.

The client-side plugin offers several GUI panels for preparing the query (Figure 15). For some fields that are modelled as free text input fields, such as for example the "target species" or "property name", a mechanism has been implemented to get the list of possible values from the database. From the user input, a database-specific query is generated, making use of the metadata to prepare correct SQL statements. If needed, the statement can be checked and edited before submission. The same editor can also be used for direct input of custom SQL statements, providing an interface that is very useful for users having experience in constructing SQL queries. The query results can be visualised in table form in the Client's job monitoring part (Figure 16). External viewers can be plugged in for visualisation of complex data such as chemical structures.

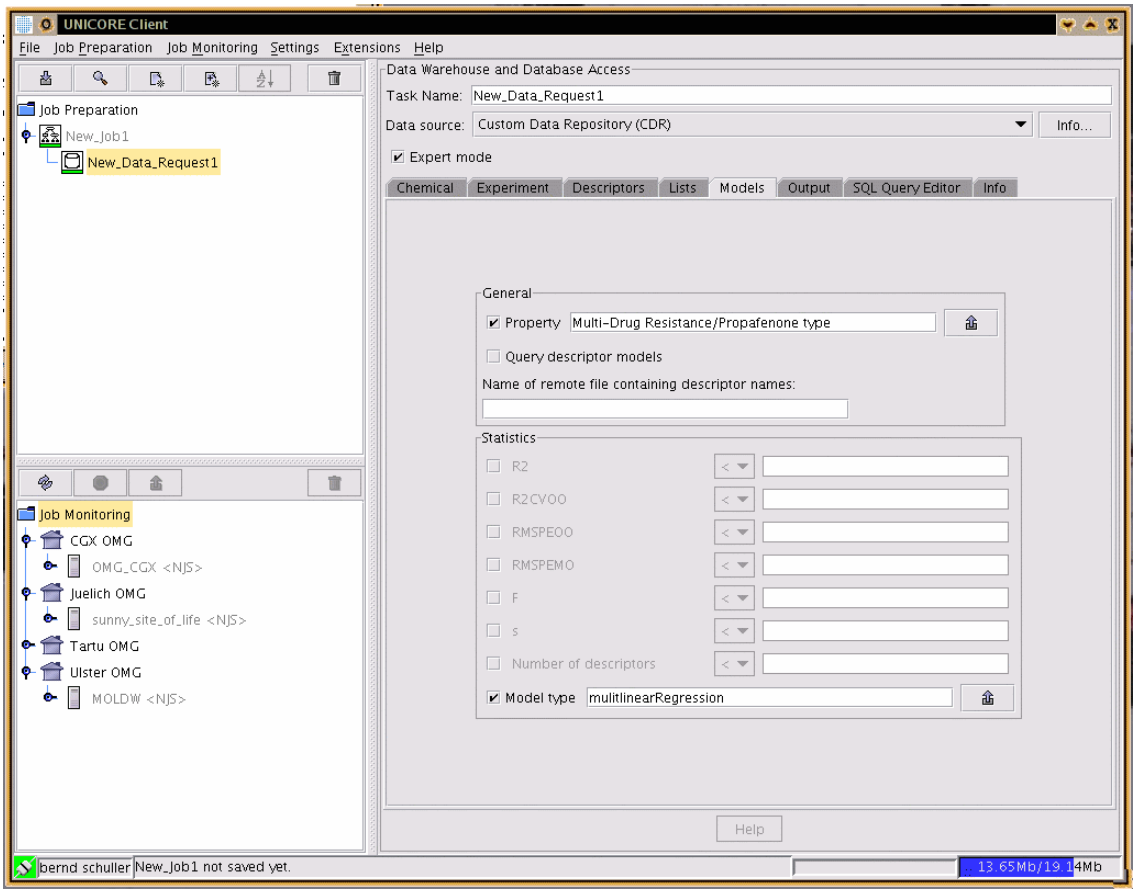


Figure 15: DataBaseRequest Plugin input panel

Within the project the following experiences were made

- Server-side related: The overall architecture had proven to be very well suited for the needs of the project, because it proved fairly easy to generate the diverse file formats needed for data mining. Initially, these conversions were done using XSLT. However, this proved to be not scalable enough. Eventually, the data conversions were done using STX (a streaming version of XSLT) and explicit conversion tools in more complicated cases.
- Graphical user interface related: While the GUI is very useful for preparing queries, and covers the most common usage scenarios within the project, often some manual fine tuning of the SQL query is necessary. Some of the end users felt that the client plugin was very complex, and not too easy to use. While this criticism can be understood, it is important to note that often requirements to the GUI would change, and unexpected usage scenarios had to be taken into account. Still, a more streamlined GUI more appropriate for the non-expert user would be a useful, for example in the form of a "query wizard". While

such an approach sacrifices some generality, it would definitely be useful for the "common user".

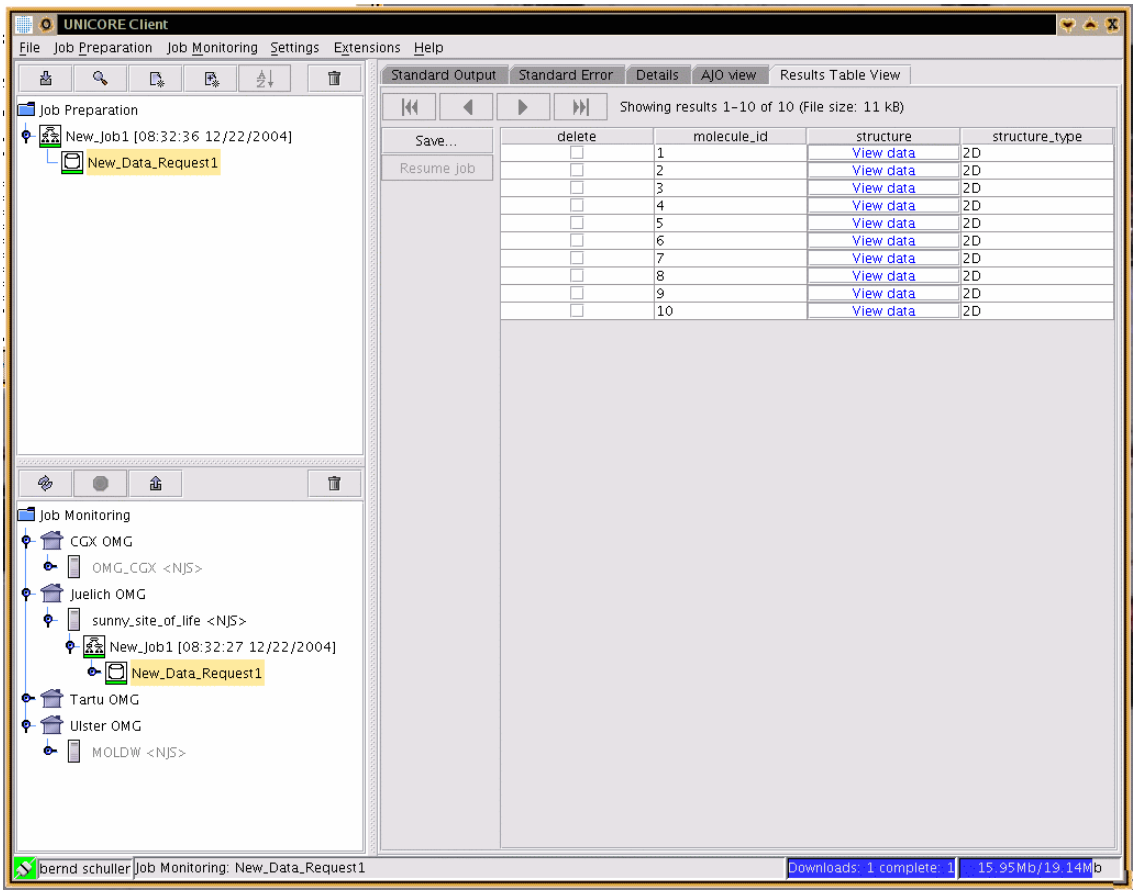


Figure 16: DataBaseRequest Plugin monitoring panel

3.4.2 Workflow Support

To support the complex molecular design and engineering applications within OpenMolGRID, the basic UNICORE middleware has been enhanced significantly to better support complex scientific workflows in a Grid environment. We have:

- simplified and automated the mapping of scientific workflows onto a set of Grid sites,
- automated the resource selection process,
- made better use of compute capacity in the Grid by splitting data-parallel tasks.

Since it has been a fundamental design principle to not modify the basic UNICORE software, if possible, these enhancements have been done using the

extension interfaces and mechanisms provided by UNICORE. The basic workflow support architecture is shown in Figure 17.

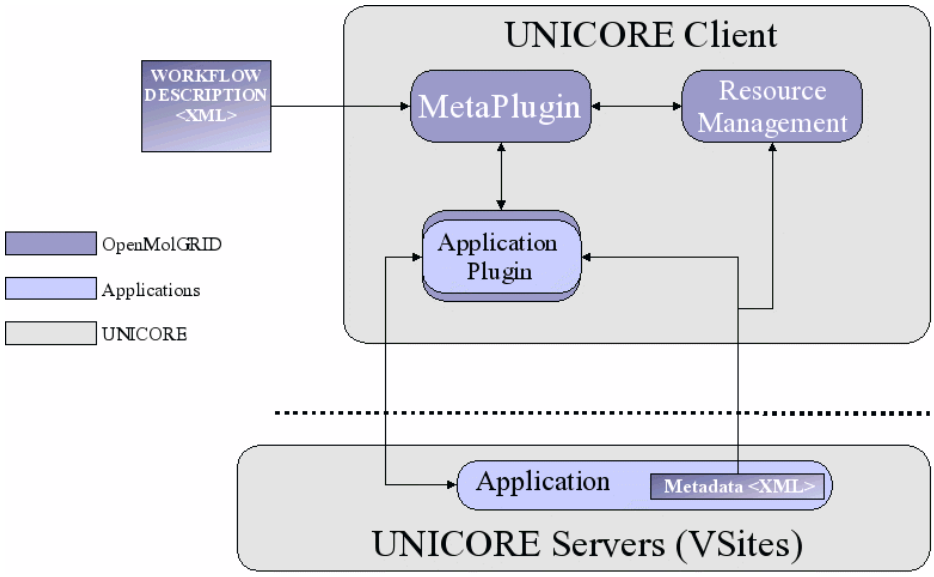


Figure 17: Basic workflow support architecture

Enhanced workflow support is achieved by three elements: application metadata, an extended client plugin interface and the MetaPlugin.

Each application supplies information about its capabilities and its input and output in a metadata file. Associated to each input and output file is a data type, comparable to a MIME type, that is used to map input/output for subsequent workflow steps. On the client side, each OpenMolGRID application plugin has an additional interface that supports setting filenames, adding exports, etc. Finally, a new component, the MetaPlugin (supported by a resource management or "service registry" component) deals with coordinating the individual plugins and building a UNICORE job from a workflow description.

Figure 18 shows a screenshot of the UNICORE client with a workflow loaded into the MetaPlugin. In addition to the basic functionality of setting up the job, the MetaPlugin can split data-parallel tasks in order to make the best use of the available Grid resources. Figure 19 shows a screenshot of a loaded workflow with one task split to four sites.

The MetaPlugin is highly modular, and has various interfaces that can be used for later enhancements, especially in the area of resource selection or brokering.

Within the project the following experiences were made

- Server-side related: The OpenMolGRID architecture has proven its worth and has been used successfully in the project. While it is generic and extensible enough to be used in other application contexts, it is not currently based on any standards beside UNICORE itself. However, since the architecture follows the "service oriented architecture" (SOA) paradigm, it can be easily adapted to other frameworks. In any Grid context, a service registry, execution of applications and file handling are core functionality, and this is all that is needed for OpenMolGRID.
- Graphical user interface related: The workflow support is built on top of the standard UNICORE client, thus the GUI is not optimal. What is missing most of all is a user friendly graphical workflow editor. Currently, workflows have to be either written from scratch or created by modifying existing workflows. While this is OK in a research setting, a commercial/production system should have a customised, streamlined GUI. The full details as provided by the standard UNICORE client should be kept hidden from the user.

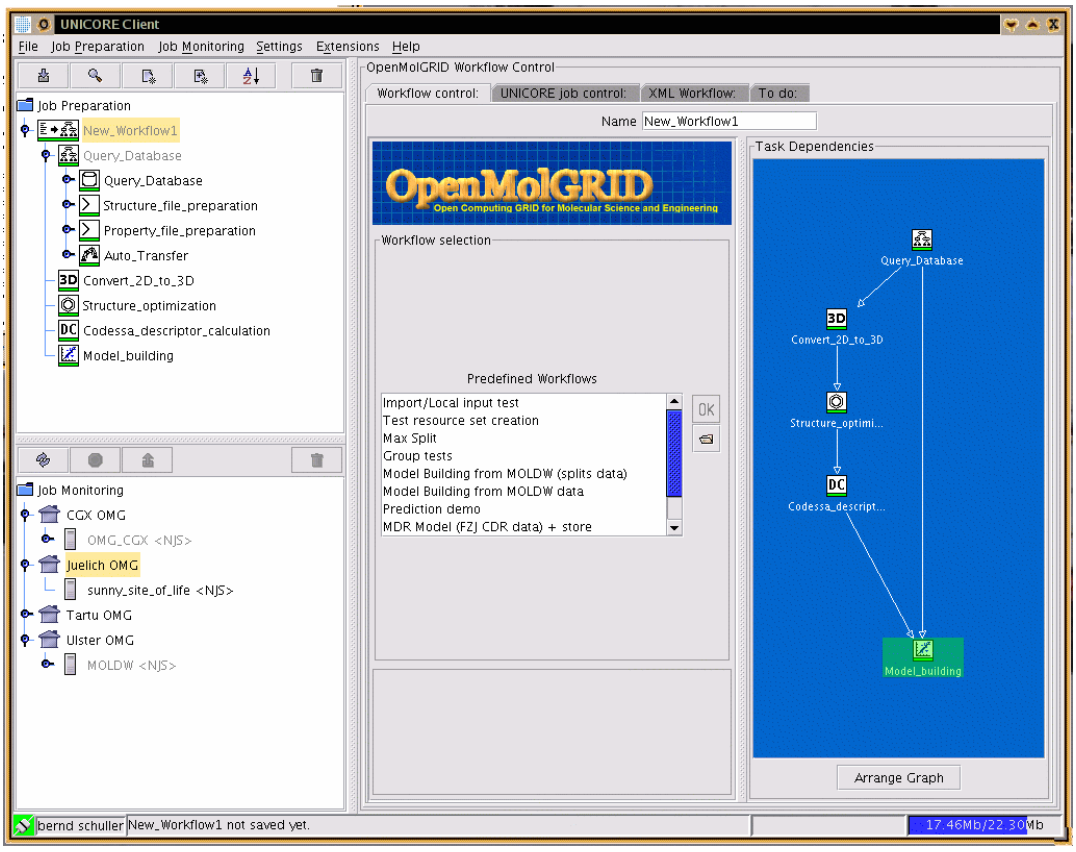


Figure 18: MetaPlugin input screen

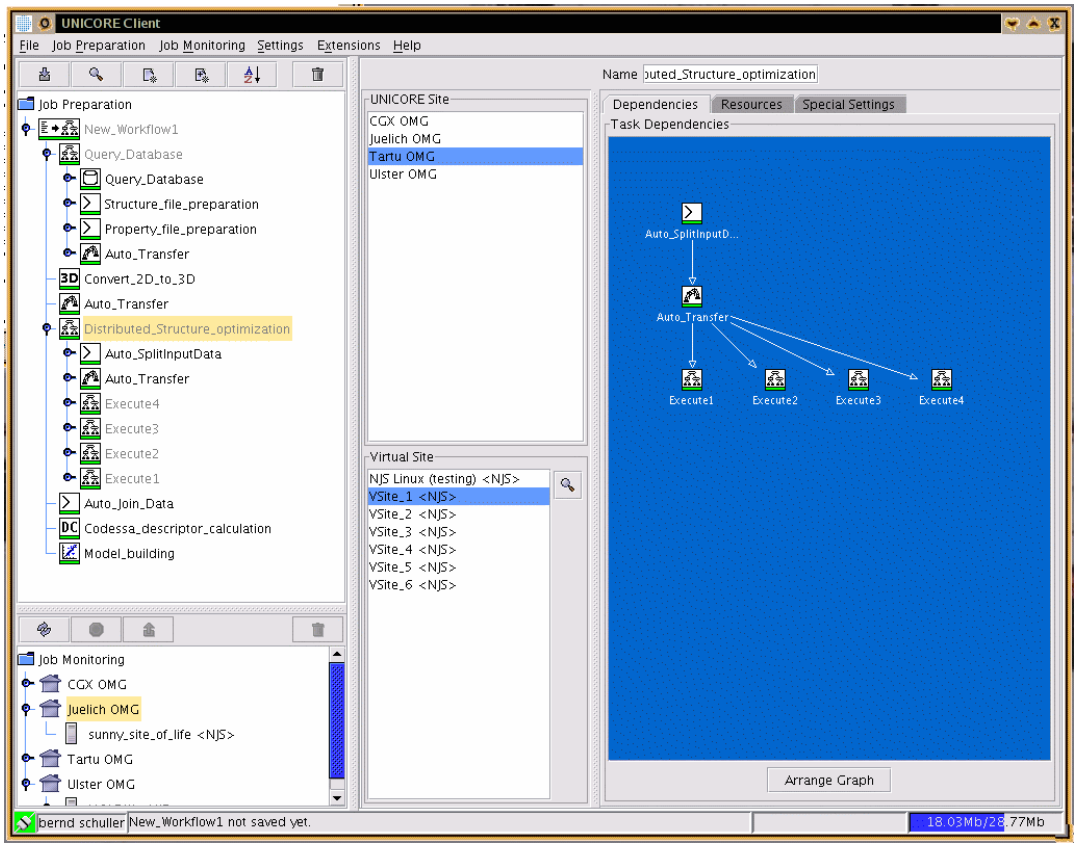


Figure 19: MetaPlugin input with one task being split

3.4.3 Command Line Interface

The usage of UNICORE is mainly based on user-interaction with the Graphical User Interface (GUI), allowing the user to generate jobs, submit and monitor them and retrieve results. This GUI-based interaction makes UNICORE unsuitable for usage within other programs, so a different interface had to be developed. The Command Line Interface (CLI) and its API (CLAPI) provide means to access UNICORE resources and features from the command prompt or from within programs.

One of the most important functions of the CLI is job generation. It can build abstract job objects from XML workflow descriptions and save them to files. The main advantage of this feature is that UNICORE jobs can be built automatically without user intervention, thus enabling even programs or scripts to generate jobs,

so far a highly interactive process. Furthermore, the CLI users - no matter if persons or processes - do not have to take care about server or local resources available at runtime. It is part of the CLI to dynamically collect and manage all required resources and to set the target systems for the generated job properly.

The CLAPI allows the usage of CLI features within a program acting as a UNICORE user. At start-up an instance of the CLAPI collects all required resources and caches them internally, allowing successive processing of workflows and handling of generated jobs without further, redundant data collection.

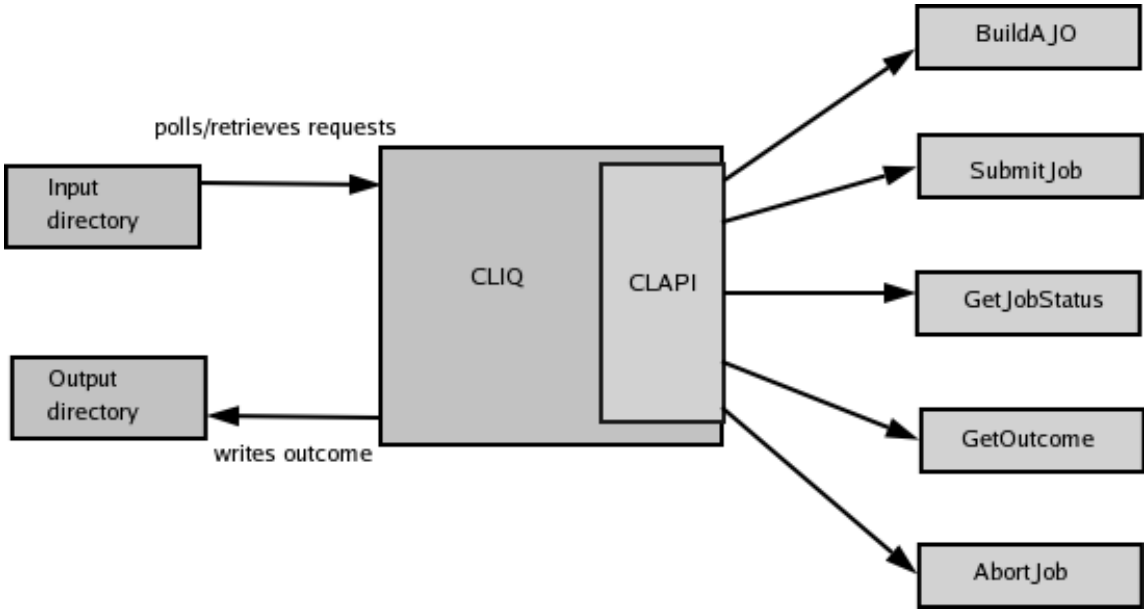


Figure 20: Interaction of Command Line API and Command Line Queuing component

Based on the CLAPI, a queuing mechanism, CLI Queue (CLIQ), has been developed (see Figure 20). It is a software component linking a software entity acting as a UNICORE user to the CLAPI. CLIQ offers a mechanism for subsequent processing and executing of multiple workflows, accounting of submitted jobs and returning their results to the calling software entity via CLAPI.

Within the OpenMolGRID-project the CLI, CLAPI and CLIQ are used by the Data Warehouse to perform several computations on the Grid during its data upload procedures.

3.4.4 Testbed

A testbed had been set up between the partners' sites which evolved during the course of the project: From one Linux PC hosting the UNICORE server components Gateway, NJS, and TSI per site to multiple target systems with the developed data sources and applications. A certification authority had been established in Juelich to provide the necessary X.509 certificates for users and servers.

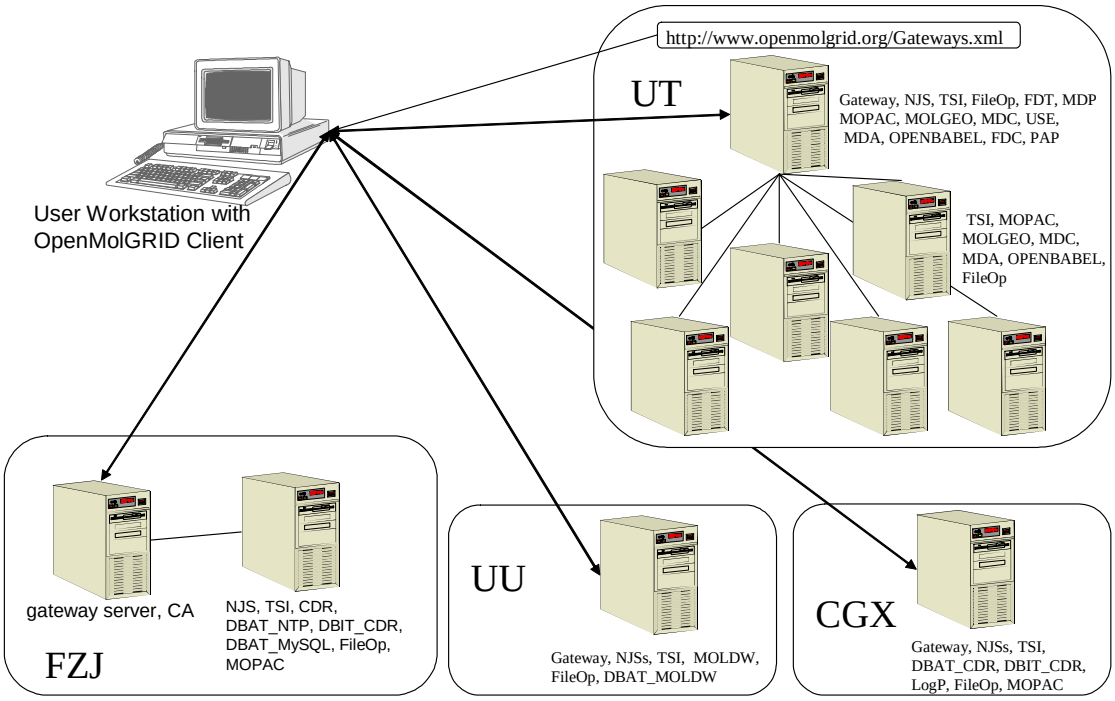


Figure 21: The final OpenMolGRID testbed

3.5. In Silico Testing, Real Life Model Development and Prediction

3.5.1 In Silico Testing

Under the frame of *in silico* testing, the system has been tested whether it is operational in Test Cases targeting the original goals of proposed use of the system. During this test phase, QSAR/QSPR models were built on *in vitro* measurements of Multi-drug resistance (MDR) and G-protein coupled receptor (GPCR) activity, as well as a model was built on a selected set of the *Colinus*

Virginianus (bobwhite quail) toxicity data collected from the public ECOTOX database.

The overall test results were very good; all major functionalities were operational. Only a few, minor errors had been discovered, and all of the errors had been corrected. The testing of the model building capabilities resulted in several QSAR models for the above three different biological activities. The models, especially on MDR activity show considerably high accuracy. The predictive power of the models, however, has not yet been tested in production use.

3.5.2 Model Building and Evaluation Results

The major goal of one of the tasks in the project was to build a QSPR model on a real-life data set (20,000 chemical structures) and to find the best QSAR/QSPR model for cytotoxicity. Measurement of cytotoxicity, preparation of the database upon the experimental data and finding the best QSAR/QSPR model for the measured cytotoxicity values have been accomplished. In order to achieve the goal, there was a code freeze of the system on December 1, 2004. The code freeze system was used for the model building procedure and the testing environment was not changed until the finish of this test phase.

All tests have been made from the user perspective. Additionally, the tests have been made using workflows allowing the repeatability of any test cases. Test results and model building procedures, including detailed test conditions were recorded. The system has been tested in many different aspects. The workflow control and automation have been widely tested. There have been workflows prepared for all tasks. The parameters for the preparation steps for model building, as well as for the model definition itself have been optimised. The experimental dataset has been investigated, the chemical structures and the corresponding descriptors have been analysed. Data and descriptor subsets have been selected to ease the model building procedure. Besides the IC_{50} values, an additional property value, the viability at 60 μMol concentration (which is available for all compounds) has been extracted from the experimental results.

The most important overall result of this test phase is that the model building procedure using the OpenMolGRID system works properly.

Several models have been developed on the experimental cytotoxicity values, and the results have been compared to non-linear models built outside of the system. Unfortunately, the statistical parameters of the models are not too good, but the models are still usable in the next test phase, prediction. In general, there is no significant difference between the results achieved by the linear and the non-liner

models. The model built on approximately 20,000 compounds enables us to test the predictive power of it, and use it in molecular engineering scenarios. Note that there is no any published QSAR model available on such a large dataset in the world so far.

The analysis of the experimental data set, the structures, and the calculated descriptors pointed out several problematic areas both with the investigated property and the QSAR techniques implemented into the OpenMolGRID system. These discoveries enable further improvement of the cytotoxicity models, as well as further development of the system in the future. The most promising possibility is the calculation of descriptors based on several optimised conformers of the compound.

The tests proved that the OpenMolGRID system is a real Grid system, because it could smoothly run complex workflows on distributed and heterogeneous resources. The results also showed that the efforts to decrease the time needed for heavy calculations were successful, proving that the distributed computing technology was properly applied.

An additional general result of the tests is that the OpenMolGRID system has a real open architecture, and the testbed is easily extendable. This enables a further increase in the calculation speed in the future, and establishes the possibility of further developing the system to a commercial product.

3.5.3 Prediction

The major goal of the prediction tasks in the project was to test the prediction power of the QSPR model(s) built on a real-life data set (10,000 chemical structures) under the frame of the previous test phase, model building. The models built on approximately 20,000 compounds have been used in molecular engineering scenarios.

Besides testing the molecular engineering process, a comprehensive investigation of the universal structure enumerator, and the property/activity prediction component, as well as the usage of the fragment descriptors and the corresponding descriptor models has also been accomplished. All tests have been made using workflows allowing the repeatability of any test cases.

4. Contributions to Standards

The general policy within OpenMolGRID has been to use UNICORE as-is, and build services on top of the basic software, using only the extension interfaces and features that UNICORE provides. On the server side, the basic UNICORE software is used as-is in OpenMolGRID. However, on the client side several contributions to the basic UNICORE client have been necessary to achieve efficient application integration and extended workflow support. Most notably, the application specific plugins that are used in OpenMolGRID use an additional interface that supports the automated workflow processing through the MetaPlugin. In addition, for efficient use of the MetaPlugin, the UNICORE client itself has been modified slightly. The most important of these changes have been already incorporated into the standard UNICORE client by the UNICORE developers. We expect that all of the changes made in OpenMolGRID to the UNICORE client will be incorporated in the basic software eventually, as they are generic, and thus of general interest. Another important component of general interest is the command line client for UNICORE that we have developed. It will be distributed through the UNICORE project at SourceForge, and thus will be available worldwide.

OpenMolGRID had a strong focus on support for high-level workflows and application integration. The unique OpenMolGRID approach to treating these problems will be carried forward in Framework 6 projects CoreGRID and NextGRID through the involvement of Forschungszentrum Juelich. Currently, the standardisation efforts of the Grid community in these areas are still in an early stage, and we are optimistic that results from and experience gained in the OpenMolGRID project will have a significant impact on next generation Grid software.

Another important issue is the interesting possibilities that the OpenMolGRID approach can offer to standardisation of QSAR models. These models are currently investigated as an approach to the EU legislation for chemicals, for instance within REACH (Registration, Evaluation and Authorisation of Chemicals, EC Initiative within the SMEAP 4C program). A fundamental issue for the acceptance of QSAR models is their reproducibility. Currently the calculation of 3D descriptors is deeply affected by the manual process of optimisation of the 3D structure. Different expert will achieve different conformations of the same compound. However, OpenMolGRID can offer an innovative tool to obtain more reproducible results.

5. Dissemination

The dissemination of results has been an important part in the OpenMolGRID project. Raising awareness of the OpenMolGRID project, disseminating results achieved during the project, and promoting take-up of Grid computing among end-users in different scientific fields both in academia and in industry has been in focus. We learned that Grid computing is a truly global undertaking. Projects in this field can prosper and gain credibility only if they are promoted and exposed internationally. We have performed the information dissemination to target audiences in Europe and worldwide by the following public channels:

- Articles in journals and conference proceedings;
- Presentations and demonstrations at key conferences regarding the field of high performance- and data-intensive computing;
- Press Releases;
- Through participation in the Global Grid Forum, in GRIDSTART, and in conferences or collaborations that emerged towards the end of the project;
- With an extensive Web presence. The information about the project is hosted on the OpenMolGRID Web site (http://www.openmolgrid.org/) and is maintained together by all partners. Additional Web presence as part of the commercial exploitation is performed through ComGenex' and its affiliates' web sites.

Within the lifetime of the project a lot of effort has been devoted towards dissemination. This work has resulted in a number of publications in journals and conference proceedings, and presentations and demonstrations in conferences, as summarised in the following table:

Papers in reviewed journals	1
Papers in conference proceedings	11
Oral presentations	37
Poster presentations	3
Exhibitions	3

Other publications	9
Diploma thesis	2

5.1. Collaborative links to other projects

OpenMolGRID has established collaborative links with the following EC funded and national projects. The particular projects are as follows:

- OpenMolGRID has a close relationship to the EC funded projects EUROGRID and GRIP. It uses the UNICORE developments performed in these projects. In addition, it collaborates with all projects in the GRIDSTART (http://www.gridstart.org) cluster of FP5 projects.
- DEMETRA, Development of Environmental Modules for Evaluation of Toxicity of pesticide Residues in Agriculture; QLK5-CT-2002-00691, 2003-; http://www.demetra-tox.net/. The aim of DEMETRA is to build up software to predict toxicity of pesticides, on the basis of the chemical structure. The software will be available for free of charge to end-users;
- FATEALLCHEM, Fate and toxicity of allelochemicals (natural plant toxins) in relation to the environment and consumers; QLRT-2000-01967. 2001-; http://www.fateallchem.dk/; The overall objective of FATEALLCHEM is to perform an environmental and human risk assessment of exploiting the allelopathic properties of wheat in modern farming and to develop a framework for future assessments of allelopathic crops;
- EASYRING, Environmental Agent Susceptibility Assessment utilizing existing and novel bio-markers as Rapid non-Invasive testing methods; QLRT-2001-02286, 2003-; http://www.easyring.org (not yet active); http://www.credocluster.info/assoc.html; EASYRING aims to develop and validate novel non-invasive methods for detecting known and new bio-markers of endocrine disrupters directly in the mucus of aquatic species;
- IMAGETOX, Intelligent Modelling Algorithms for the General Evaluation of Toxicity; HPRN-CT-1999-00015, 1999-; http://airlab0.elet.polimi.it/imagetox/; IMAGETOX is a Research Training Network aimed to train young researchers in computer methods that predict the toxic and the environmental properties of chemicals;
- Development of software to predict the behaviour and the toxicity of environmental pollutants. The project is funded by the Italian

Environmental Minister. The aim of the project is to develop computer models for the prediction of the behaviour and the toxicity of environmental pollutants, with special attention to pollutants found in Italy.

6. Exploitation

The OpenMolGRID technology and expertise have exploited by all partners throughout several exploitation channels and links with other EC funded and national projects. While the partners from research focus on exploitation in teaching, further research projects and standardisation efforts, and Open Source Software development the commercial partner, ComGenex, perform industrial exploitation as it is described below.

6.1. University of Tartu

Since the early development phases of the OpenMolGRID system, University of Tartu (UT) has been testing and using it for academic and scientific research activities. The resources from the OpenMolGRID testbed have been used by some graduate and undergraduate students with good success. Therefore, the Grid resources that were provided to the OpenMolGRID testbed will remain operational and will be maintained in future. In addition, it is possible to use OpenMolGRID components as a tool for teaching (e.g. lab sessions in molecular engineering, molecular design, etc.). During the implementation and integration of existing technologies, several scientific issues (the treatment of multiple conformations, the validation of the predictive power of QSAR models, quality of data sources, etc.) were experienced that need further research and development of technology. UT will make its best effort in future to address these issues and improve the existing OpenMolGRID system.

UT is involved in several national and European research projects, where the results and experiences obtained from the OpenMolGRID project are relevant. We are participating in the EC funded project NANOQUANT (Understanding Nanomaterials from the Quantum Perspective) started April 1, 2004. This project integrates basic and applied science by combining developments of theory and computational technology for the study of nanomaterials with their application to the design and characterization of such materials. NANOQUANT will provide new insight and better techniques for electronic-structure simulations of new materials. In addition, in the second half of 2004 we were involved in the submission of two EU FP6 proposals.

6.2. University of Ulster

We have developed UNICORE-enabled components for grid-based data warehousing. The components developed at the University of Ulster (UU) are made available as open source. Currently, our key exploitaiton stratgy is focused

on exploiting the knowledge gained from the OpenMolGRID project. In particular, we have already been successful in launching the DataMiningGrid STRP proposal/project under FP6 IST Grids and Complex Problem Solving Call. This project could be viewed as a result of our efforts in the field of grid comuting and in particular as a result coming out of the OpenMolGRID project. Together, both grid projects will position our research group in an excellent position to develop further proposals (national and European) both on grid technology and bioinformatics. Another result of the OpenMolGRID project is a national proposal entitled Data Mining Services – A Grid/Web Services Infrastructure for Data Mining in Modern e-Businesses and e-Organizations (submitted in September 2004). This proposal has been submitted to Invest Northern Ireland. Another sucessful indirect outcome of OpenMolGRID is a successful funding awarded to the Bioinformatics Research Group by the Northern Irish Department and Learning.

6.3. Mario NEGRI

Nowadays, the necessity to target regulatory policies in the area of environmental risk assessment and human health-care pushes governmental agencies toward the adoption of grid-based solutions. The reason for such a tendency is simple. The lack of standardised, reproducible, and flexible procedures have raised serious concerns about the reliability of current in-silico risk assessment. The absence of standardized protocols has stimulated the uncontrolled growth of in-silico predictive models to target risk assessment. Up to date, more than 20,000 quantitative structure-activity relationship (QSAR) models have been developed and listed, each adopting different combinations of human hand-feeding actions and computational resources. Use of such “hand-made” tools is not acceptable for regulatory purposes because of the lack of reproducibility and the narrow range of applicability. Commercially available software packages offer an alternative, highly questionable solution to this problem, owing to a restricted flexibility. Acceptance and implementation of automated grid-based solutions concretely overcomes such drawbacks, paving the way to high-quality standardized predictive tools. In this sense, OpenMolGRID represents the ideal tool for promoting the development and the diffusion of standardized protocols for toxicology prediction. Use of OpenMolGRID offers an automated and consistent way to generate predictive models, regardless the user’s background, knowledge, or preferences. The possibility to standardise computational predictive protocols will make use of QSAR very attractive for regulatory purposes. This goes into the direction stated by the European Centre for Ecotoxicology and Toxicology of Chemicals (ECETOC), which has posed several concerns on the appropriate selection, development, and use of QSARs. (see [22]). In this perspective, the

Mario Negri Institute is encouraging and spreading the use of OpenMolGRID throughout several EU projects. In order to concretely target standardization-related issues, OpenMolGRID will be used for:

1) Promoting the adoption of standardized protocols for QSAR model building. To be effective, standardized protocols require general acceptance from end-users, such as governmental institutions, pharma companies, and academies. OpenMolGRID is the gate through which such a compliance can be reached.
2) Devising the guidelines to be further used in developing standard protocols. The automated processing strategy currently offered by OpenMolGRID workflows provides useful indications on how standardization can be promoted and implemented within computational chemistry. In contrast to human users, whose choices might vary time-to-time, OpenMolGRID offers consistent and reproducible processing strategies. The final goal will be that of providing a standardized protocol for QSAR modelling to be used for regulatory purposes.

6.4. Forschungszentrum Jülich

The extensions to the UNICORE Grid Infrastructure developed in OpenMolGRID, the abstract resource interface layer, the integration of database access, the automated workflow support, the resource selection mechanism, and the command line interface and client, will be inputs to FP6 IST projects DEISA and UniGrids and into the German project VIOLA. In addition, they will be used in FZJ's UNICORE production environment for its supercomputer users. All components developed by FZJ are distributed as Open Source under BSD license together with the UNICORE software on SourceForge (http://unicore.sourceforge.net). Figure 22 and Figure 23 show screenshots from the UNICORE project at SourceForge maintained by FZJ.

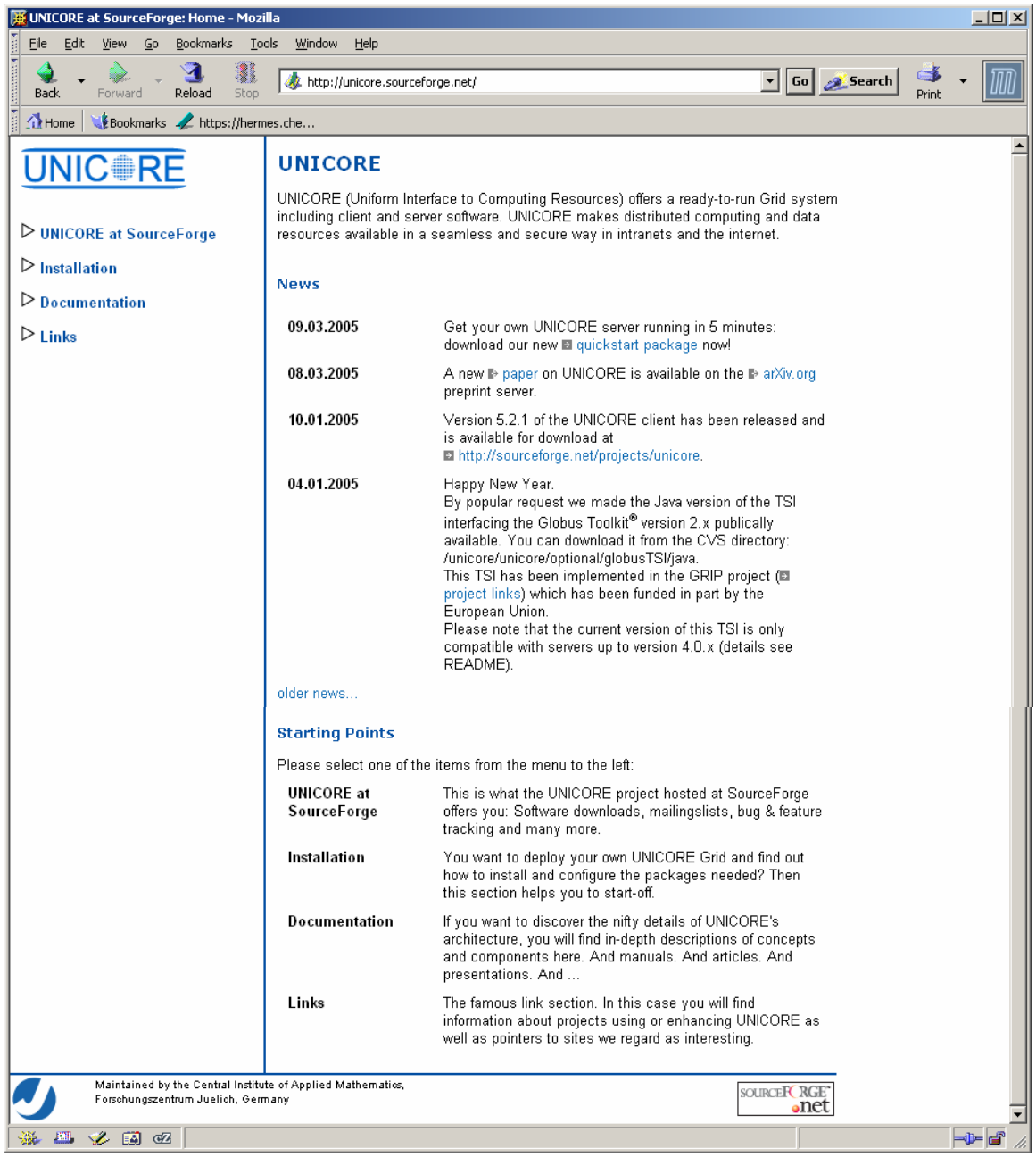


Figure 22: UNICORE Introduction at SourceForge

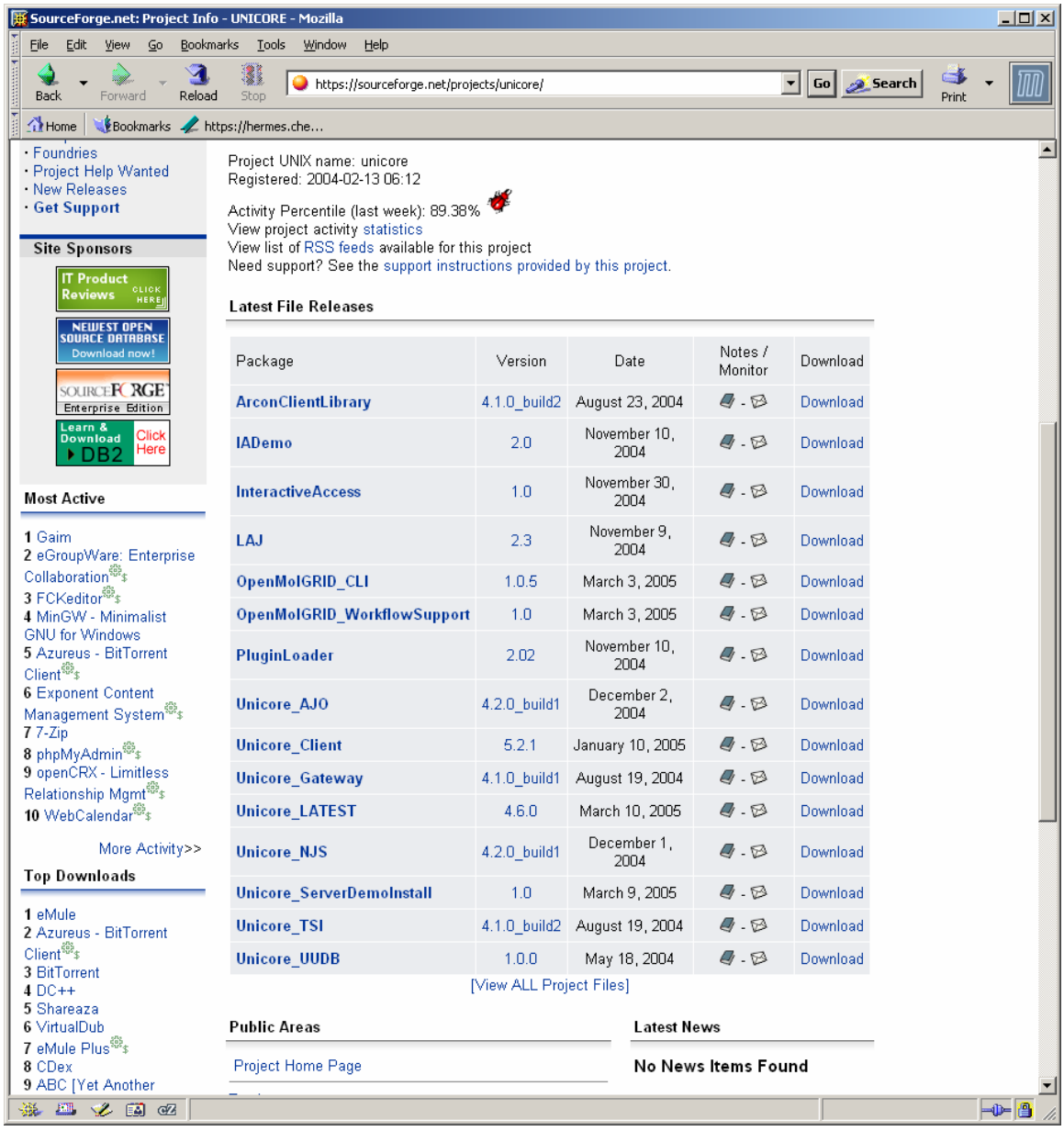


Figure 23: OpenMolGRID packages for download

6.5. ComGenex

CGX is actively serving an estimated 90%+ of the leading pharmaceutical companies worldwide. With offices in the US and Europe as well as representatives in Japan it is able to provide professional contact to our 200+ clients in the pharma and biotech industry. CGX' client network is looking back to more than 10 years of experience and includes CGX' well established and

maintained communication routes to its customers throughout major sales and marketing contacts.

ComGenex (CGX) has made strong efforts to actively disseminate the results and ongoing activities of the OpenMolGRID project to their industrial partners with respect to the commercial exploitation of the project. The aims of such activities were to

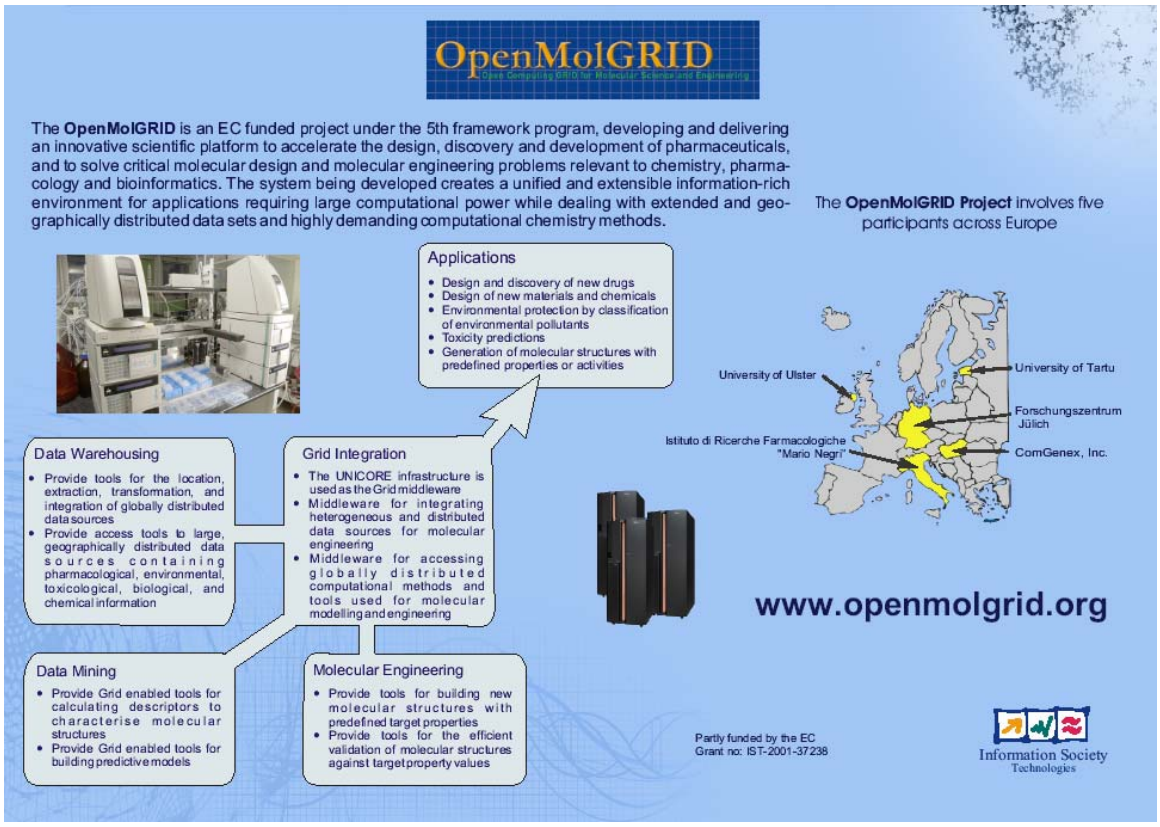
i. create awareness of the project;
ii. assist the clients to gain early insight of the project in the development stage;
iii. stimulate clients to provide opinions and recommendations useable for the project to deliver a better product
iv. find industrial partners who have demand of using and testing the system in its development phase and
v. explore exploitation possibilities tailored to the best needs of the industrial users.

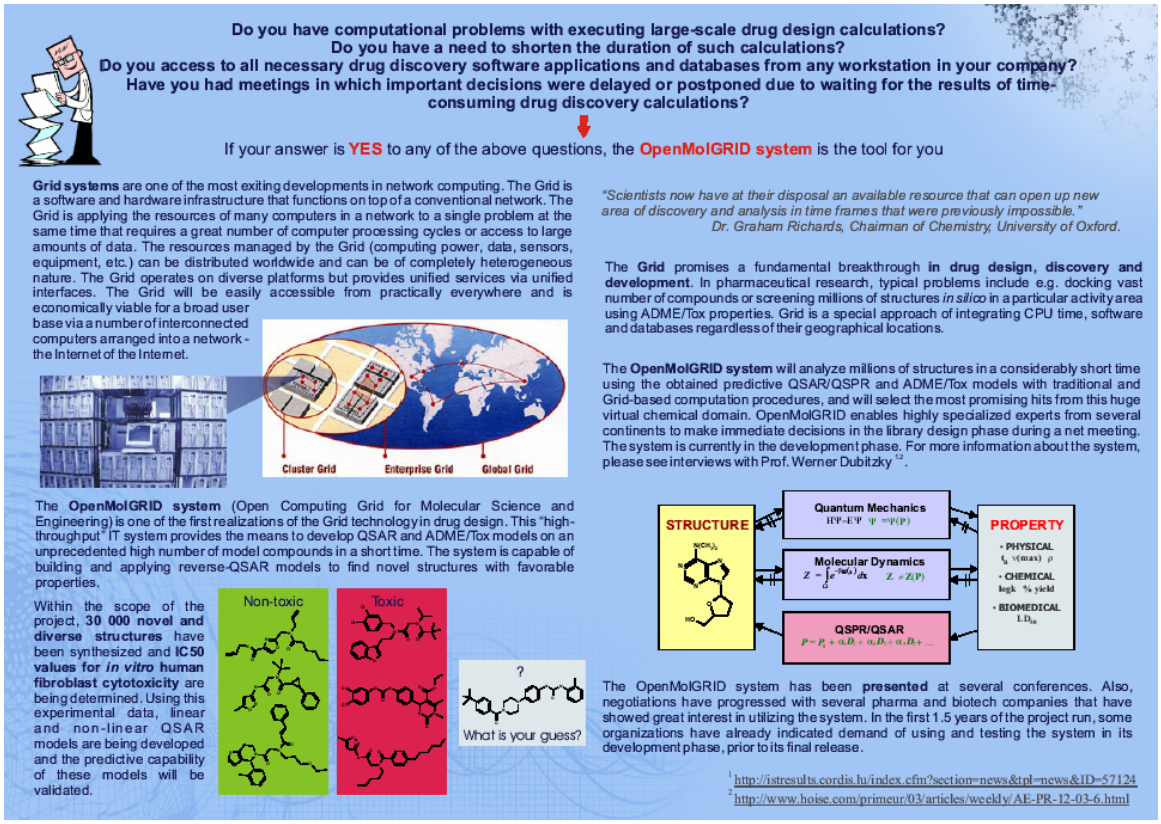
CGX has introduced the OpenMolGRID system to its collaborative partners including distinguished partners, research collaborators and contracted partners for long-term investigations in the scientific field of biotechnology and drug design, discovery and development and related fields.

CGX has made several road shows in Japan, the USA, Korea and quick client visits. CGX and its representatives routinely visited their collaborative partners and clients of their client network. A number of Japanese, Korean, American and European companies have been informed about the OpenMolGRID project, which raised great interest. During the road shows and visits, clients received introductory information (presentations and discussions) related to the system and system demonstrations. Partners were kept up-to-date with information regarding the developments in the project.

CGX organized net meetings with its industrial partners. CGX' scientific professionals presented information and results of the ongoing project via computer and telephone conference aided slide shows. CGX has designed, printed, and distributed the marketing materials including brochures and feedback forms. The brochures and feedback forms have been incorporated into CGX' official *Scientific Newsletter Spring Issue*, which have been printed and distributed in 1000 copies to CGX' partners. Such marketing materials have been burned onto the compact disks of ComGenex' Chemistry Novelties. CGX' sales and marketing group regularly mails out such CDs to most of its clients bimonthly. The CD comprises sales and marketing information including updated information on the CGX' Compound Repository Stock, animated company information, the latest

Press Releases, and scientific updates including the latest Scientific Newsletter, scientific publication request form, as well as technical information. The first release of the OpenMolGRID materials on the CDs was back in February 2004.





CGX is actively participating at Scientific Conferences and Trade Shows (e.g. over 60 conferences have attended in 2004), presenting and exhibiting its innovations. CGX has been continuously presenting the latest news and information on the OpenMolGRID. During the project, CGX provided scientific details regarding the development of the system in relevant scientific presentations and posters at European (including UK, Germany, France, Switzerland, Belgium, Denmark, Poland, Austria, Russia and Hungary), American, Australian and Asian (including Japan, Korea and India) conferences and tradeshows. More than 600 people were directly informed during such occasions.

As the result of CGX' above exploitation activities, client feedbacks accumulated clearly show great interests in the OpenMolGRID system. Several organisations have been found to be interested to test the OpenMolGRID system or requested to be kept further informed. One of the experts working in the drug discovery software market expressed his enthusiasm about the OpenMolGRID system after a net meeting session. His words were as follows:

> *"I was impressed by the demonstration. In my point of view, OpenMolGRID addresses two new opportunities in QSAR (Quantitative Structure Activity Relationship) studies. Many*

scientists have tried to work with 3D descriptors, and the success was very limited. This is easy to understand because we calculate 3D structures in the gas phase and in reality the structures is in liquid (water) and probably interacts always in some ways with its surrounding. This makes the prediction of 3D structures near active sites in biological systems very difficult, but not impossible.

1. *OpenMolGRID could provide enough computer power to calculate more realistic 3D structures.*
2. *OpenMolGRID makes large-scale QSAR predictions possible. "*

6.6. Joint Exploitation Efforts

Joint exploitation activities have been focused on implementing the recommendations of the *Individual Project Review Report of the Special Review on Project Results Exploitation and Dissemination Activities* of the European Commission. Efforts have been made in defining the suite of products and exploring potential applications areas in information technology (middleware) and in the chemical and pharmaceutical academic environment and industry. Based on an overall market analysis for the identified products, an exploitation strategy has developed. As basis for the actual exploitation of the products developed in the OpenMolGRID project an Intellectual Property Right and Licensing Document has been generated covering the interest of all parties. Web pages (at www.openmolgrid.com) focused towards the exploitation of the results of the project has developed including parts addressing the commercial exploitation of the identified products.

As a continuation of the project, the OpenMolGRID Steering Committee has considered particular steps for future exploitation: the OpenMolGRID testbed will be available until end of 2005 for project partners and for selected beta-test partners as well as the OpenMolGRID web site will continue.

The Steering Committee has recognised the importance of IPR protection of several OpenMolGRID items including the name of "OpenMolGRID". The "OpenMolGRID" name as a trademark has been registered after the project duration within all countries of the European Union. The Internet domain name of www.openmolgrid.org has been registered at the beginning of the project. More variations (.com and .net) were registered at the end of the project.

Future joint exploitation would entail some new legal entity or consortium to be formed by all or a subset of the current project partners. In order to seamlessly ensure the future commercial exploitation of the project results, the industrial partner, CGX has advised to the project partners to establish a joint organization. The partners of the new organization shall prepare a viable business plan including exploitation strategy.

7. Project Structure and Management

The project established an efficient structure with clear roles and responsibilities for the diverse duties and a well-defined decision and conflict resolution structure. Each topic covered in the project was in the responsibility of one person who had to interface with the other topics. A technical coordinator was implemented to survey the technical progress of the project and to insure that developments in the different topics remained synchronised.

The project structure had been build according to the topics, it consisted of the development fields Data Warehousing/Data Management (led by UU), Data Mining (led by UT), Molecular Engineering (led by UT), and Grid Integration (led by FZJ) and the topics Real-life Testing (led by CGX), Dissemination (led by UT), and Project Management (led by UT[1] / FZJ[2]).

The internal communication within the project made extensive use of e-mail and BSCW (Basic Support for Collaborative Work) mainly used as document and software repository, which was installed and managed at UT. Regular project meetings were held at partner sites approximately every four months. It turned out that especially in the starting phase of the project it was absolutely necessary to meet more often to find a common language within the multidisciplinary group and to define the interfaces. In the last phase of the project, meetings were held as integration sessions, which helped to identify and solve software problems efficiently.

During the course of the project, a major revision of the Description of Work has been completed. After the first project year, the focus of the project had been relocated from building a toxicology prediction model based on 30,000 newly generated and analysed molecules to the Grid integration and the Grid-enabling of the data warehouse. With this shift the development of a Command Line Client for UNICORE was introduced, which was used by the data warehouse to run the calculation of prominent molecular descriptors on the Grid infrastructure.

[1] During the first half of the project

[2] During the second half of the project

8. Summary and Lessons Learnt

The project OpenMolGRID developed a solution for the automation, speed-up, and integration of the drug-discovery pipeline using well established Grid middleware, namely UNICORE. The drug-design pipeline consists of three major sequenced parts: data preparation (data warehousing), data mining, and molecular engineering.

The project developed a data warehouse (MOLDW), which consists of a database and a set of harvesting and data transformation tools. Public data sources containing relevant data were harvested and the data was transformed into a unified format. This allowed for the subsequent steps to access the relevant data in a uniform way at one source, namely MOLDW. Within the transformation process some calculations were done using the newly developed Command Line Client to provide the users with data most of them would calculate anyway.

For data mining all necessary applications were made available as single software modules: 2D to 3D structure conversion, structure optimisation (semi-empirical calculations), descriptor calculation, and model building. These applications were augmented with an application wrapper providing a standardised interface for the corresponding class of applications, metadata describing input and output formats and its function, and Plugins for the UNICORE Client.

Applications for the molecular engineering part had partly been developed from scratch like the universal structure enumerator which generates new structures on the basis of one core structure and a number of fragments, or the filtering module which selects a subset of the generated structures according to predefined rules. Like for the data mining applications the molecular engineering applications were augmented with wrapper, metadata, and Plugin.

The integration of these pieces of software described above to make up the unified OpenMolGRID system was achieved by the definition of an abstract resource interface for data sources as well as applications and by the implementation of automated workflow support (MetaPlugin). We achieved the automated generation of a UNICORE job with all auxiliary tasks and resource assignments from a brief workflow description giving the major tasks and their dependencies only. The system also allows the automated splitting of data-parallel tasks to be executed on different target systems.

The developed system proved to be fully functional and reliable during the real-life and reliability tests. The key questions dealt with in the testing phase were:

- Does the quality of a prediction model developed within the OpenMolGRID system differs from the quality of a model developed manually step-by step?
- Is the system capable of speeding up the drug-design pipeline to support the decision making process?
- Is the system capable to generate toxicity prediction models based on 30,000 newly created and analysed molecules?

All these questions received positive answers but with a grain of salt as it turned out that some questions need answers from the QSAR/QSPR community and the selected middleware encloses performance issues that have to be addressed by the UNICORE developers. As we used prototype software developed in other research projects we reported the issues but had no direct influence on the time frame of bug fixing or the implementation of software improvements. This is something one has to be aware of in such a project and find a way to work around it. The project partners' expectations of the quality of prototype software compared to commercially supported production software differed a lot in the beginning and in some cases made it difficult to use workarounds.

Besides the IT-related challenges the project faced challenges from chemistry and QSAR/QSPR:

- There is no established standard for globally unique identifiers for descriptors yet,
- PMML (predictive model markup language) for models is not sufficient as it cannot be used to describe PLS (partial least square), or PCR (principal components regression) models - extension of PMML would be a solution,
- Handling of necessary legacy data together with models – storage of models is an open question,
- Handling of multiple data (e.g. multiple conformations of a structure) including storage and selection and processing.

These issues have partly been solved for the project but more general solutions will be needed.

The biggest challenge the project faced was it's multidisciplinarity: Finding a common language between chemists, pharmacists, toxicologists, and computer scientist from bioinformatics and Grid computing for developing a uniform IT system providing solutions for chemists/pharmacists. Therefore it was extremely important to start with a kick-off meeting, which clarified roles and duties and laid the basis for good communication between the partners. This included the

selection and introduction (tutorial) of the collaboration tools (BSCW, e-mail) and an agreement on a project discipline.

It turned out that it is very important to have intensive technical face-to-face meetings in the beginning of the project to support collaboration and an overall approach understood by everyone. In addition the good technical coordination plays key role for success.

While in the beginning of the project potential users of the system could hardly imagine that the drug-discovery process could work properly when executed automatically without manual intervention this changed after the first data mining workflows were run. The results had good quality and the broadened the view for the potential of this solution. The automation of model development from data query via structure optimization, descriptor calculation to model building gives the opportunity to standardise the process and leads to reproducible results. The project partners are very confident that further applications of the OpenMolGRID results will turn up.

9. References

This is the reference to external literature, the project dissemination results are given in Appendix 10.2.

[1] F. Darvas, I. Szabó, Gy. Dormán, “High-Throughput Combinatorial Chemistry Combined with Predictive Tools: Application in Early Metabolism/Retrometabolism Studies”, in 6th European Congress of Pharmaceutical Sciences, Budapest, Hungary, September 16-19, 2000

[2] M. Karelson, G.H.F. Diercksen, “Models for Simulating Molecular Properties in Condensed Systems”, in "Problem Solving in Computational Molecular Science: Molecules in Different Environments", S. Wilson and G.H.F. Diercksen (Eds.), Kluwer Academic Publ., Dordrecht, 1997, pp 215-248.

[3] M. Karelson, U. Maran, Y. Wang, A.R. Katritzky, “QSPR and QSAR Models Derived with CODESSA Multipurpose Statistical Analysis Software”, AAAI Tech. Report, SS-99-01, pp 12-23 (1999).

[4] A.R. Katritzky, R. Petrukhin, D. Tatham, S. Basak, E. Benfenati, M. Karelson, U. Maran, “Interpretation of quantitative structure-property and -activity relationships”, in Journal of Chemical Information and Computer Sciences, Volume 41, Issue 3, pp 679-685, 2001.

[5] D. Erwin, ed., “UNICORE Plus Final Report – Uniform Interface to Computing Resources”, UNICORE Forum e.V. 2003.

[6] M. Atkinson, P. Kunszt, I. Narang, N. W. Paton, D. Pearson, P. Watson, “Database Access and Integration”, Draft chapter commissioned for the second edition of *The Grid*, Foster & Kesselman, April 2003, http://www.ogsadai.org.uk/docs/OtherDocs/DAIChapterFinalV1.4-1April03.pdf

[7] J. Pytlinski, L. Skorwider, V. Huber, P. Bala, „UNICORE - A uniform platform for chemistry on the Grid“; Journal of Computational Methods in Science and Engineering, 2 (2002), pp 369–376.

[8] F. Azuaje, W. Dubitzky, N. Black, K. Adamson, “Discovering Relevance Knowledge in Data: A Growing Cell Structures Approach”, in IEEE Transactions On Systems, Man And Cybernetics. Part B: Cybernetics, Vol. 30, No 3, pp448-460, 2000.

[9] D. Berrar, W. Dubitzky, S. Solinas-Toldo, S. Bulashevska, M. Granzow, C. Conrad., J. Kalla, P. Lichter, R. Eils, "Design and Implementation of a Database System for Comparative Genomic Hybridization Analysis", in IEEE Engineering in Medicine and Biology, Vol. 20, Number 4, pp75-83, July/August, 2001.

[10] L. Moss, A. Adelman., Data Warehousing Methodology, *Journal of Data Warehousing*, 5: 23-31, 2000.

[11] NTP, National Toxicology Program server at http://ntp-server.niehs.nih.gov/

[12] ECOTOX at http://www.epa.gov/ecotox/

[13] The PrologP software is a trademark of CompuDrug, Inc., http://www.compudrug.com

[14] I. Foster, C, Kesselman, J. M. Nick S. Tuecke, “The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration“, http://www.globus.org/research/papers/ogsa.pdf

[15] G. Gini, M. Lorenzini, E. Benfenati, P. Grasso, M. Bruschi, “Predictive Carcinogenicity: a Model for Aromatic Compounds, with Nitrogen-containing Substituents, Based on Molecular Descriptors Using an Artificial Neural Network”, J. Chem. Inf. Comp. Sci., 39, pp 1076-1080, 1999.

[16] U. Maran, S. Sild, “QSAR Modelling of Mutagenicity on Non-congeneric Sets of Organic Compounds”, In Artificial Intelligence Methods and Tools for Systems Biology, Dubitzky, W.; Azuaje, F. (Eds.), Kluwer Academic Publishers, Boston/Dordrecht/London Copyright 2004, pp 19-36.

[17] P.Hliva, “Specification, Design, and Implmentation of the Custom Data Repository”OpenMolGRID Project Deliverable D1.5

[18] E.V. Gordeeva, A.R. Katritzky, V.V. Shcherbukhin, N.S. Zefirov, "Rapid conversion of molecular graphs to three-dimensional representation using the MOLGEO program", J. Chem. Inf. Comput. Sci., 33, pp 102-111, 1993.

[19] J.J. Stewart, "MOPAC: a semiempirical molecular orbital program", J. Comput. Aid. Mol. Des., 4, pp 1-45, 1990.

[20] Codessa Pro, http://www.codessa-pro.com/

[21] S. Sild, A. Lomaka, “Specification of software modules for descriptor calculation and model development and their GRID interface components”, OpenMolGRID Project Deliverable D2.1
(http://www.openmolgrid.org/downloads/D2.1.pdf)

[22] Feijtel, Comber. “QSARs need to be consistently checked and updated. Transparency is crucial to the confidence of the user of such QSARs and for checking that QSARs evolve with new data”,
in QSARs for pollution prevention, toxicity screening, risk assessment, and Web applications. Edited by John D. Walker, Ph.D., M.P.H. Office of Pollution Prevention and Toxics; U.S. Environmental Protection Agency, Washington DC, USA, 1997

10. Appendices

10.1. List of OpenMolGRID Project Partners

- University of Tartu, Estonia (coordinator 09/02-11/03)
  http://www.ut.ee/
- University of Ulster, Northern Ireland
  http://www.ulst.ac.uk/
- Istituto di Ricerche Farmacologiche "Mario Negri", Italy
  http://www.marionegri.it/
- Forschungszentrum Jülich GmbH, Germany (coordinator)
  http://www.fz-juelich.de/
- ComGenex, Inc., Hungary
  http://www.comgenex.hu
- OpenMolCONSULTING, Germany (subcontractor)
- Politecnico di Milano, Italy (subcontractor)
  http://www.polimi.it

10.2. List of Publications and presentations

10.2.1 Papers in Reviewed Journals

- **Grid-enabled Data Warehousing for Molecular Engineering**
  W. Dubitzky, D. McCourt, M. Galushka, M. Romberg, B. Schuller
  Special Issue on High-performance and Parallel Bio-computing in *Parallel Computing*, Volume 30, Issues 9-10, September-October 2004, pp1019-1035, 2004

10.2.2 Papers in Conference Proceedings

- **Ex Silico ADME/Tox Approaches for Drug Discovery**
  F. Darvas
  Proceedings of the Conference on New Models for Faster, More Effective ADME, London, UK, February 18-19, 2003

- **Towards an Intelligent Data Type for Toxicity**
  D. McCourt, J. Lopez, E. Benfenati., P. Mazzatorta, M. Romberg, B. Schuller, W. Dubitzky
  Proceedings of the International Conference on Artificial Intelligence, Las Vegas, USA, 2003, 328-334

- **Can the Grid Help to Solve the Data Integration Problems in Molecular Biology?**
  B**.** Sturgeon, D. McCourt, J. Cowper, F. Palmer, S. McClean, W. Dubitzky
  Proceedings of the 3rd IEEE/ACM International Symposium on Cluster Computing and the Grid, Biogrid, Tokyo, Japan, 2003, 588-593

- **The OpenMolGRID Data Warehouse, MOLDW**
  D. McCourt, W. Jing, W.Dubitzky
  Proceedings of AXGrids 2004, January 28-30, 2004; Nicosia Cyprus
  http://grid.ucy.ac.cy/axgrids04/AxGrids/papers/E00-1806053365.pdf

- **Support for Classes of Applications on the Grid**
  M.Romberg, B.Schuller
  Proceedings of AXGrids 2004, January 28-30, 2004; Nicosia Cyprus
  http://grid.ucy.ac.cy/axgrids04/AxGrids/papers/E00-640437544.pdf

- **OpenMolGRID, a GRID based system for solving large-scale drug design problems**;

F. Darvas, A. Papp, I. Bágyi, G. Ambrus-Aikelin, L. Urge;
Lecture Notes in Computer Science, Springer-Verlag GmbH, ISSN: 0302-9743, Volume 3165 / 2004, Title: Grid Computing: Second European AcrossGrids Conference, AxGrids 2004, Nicosia, Cyprus, January 28-30, 2004, Editors: Marios D. Dikaiakos, ISBN: 3-540-22888-8, p. 69ff

- **Scientific Computing with UNICORE**
  D. Breuer, D. Erwin, D. Mallmann, R. Menday, M. Romberg, V. Sander, B. Schuller, Ph. Wieder
  Proceedings of the NIC Symposium 2004: 17. - 18. Feburary 2004, Forschungszentrum Jülich; ed.: D.Wolf, G. Münster, M. Kremer. - Jülich, Forschungszentrum, 2004. - (NIC series 20). - 3-00-012372-5. - pp. 429 - 440

- **OpenMolGRID: Molecular Science and Engineering in a Grid Context**
  P.Mazzatorta, E.Benfenati, B.Schuller, M.Romberg, D.McCourt, W.Dubitzky, S.Sild, M.Karelson, A.Papp, I.Bágyi, F.Darvas
  Proceedings of the PDPTA2004; June 21-24, 2004; Las Vegas, Nevada, USA

- **OpenMolGRID: QSAR/QSPR applications in Grid environment**
  S. Sild, A. Lomaka, U. Maran
  Proceedings of 4th Cracow Grid Workshop 2004, Cracow, December 12-15, 2004

- **OpenMolGRID: Using Automated Workflows in GRID Computing Environment**
  S. Sild, U. Maran, M. Romberg, B. Schuller, E. Benfenati,
  in P.M.A. Sloot, A.G. Hoekstra, T. Priol, A. Reinefeld, M. Bubak (eds),
  Advances in Grid Computing - EGC 2005, European Grid Conference, Amsterdam, The Netherlands, February 14-16, 2005, Revised Selected Papers, June 2005, pp.464-473

- **Application driven Grid developments in the OpenMolGRID project**
  B. Schuller, M. Romberg, L. Kirtchakova
  in P.M.A. Sloot, A.G. Hoekstra, T. Priol, A. Reinefeld, M. Bubak (eds),
  Advances in Grid Computing - EGC 2005, European Grid Conference, Amsterdam, The Netherlands, February 14-16, 2005, Revised Selected Papers, June 2005, pp.23-29

- **UNICORE - from project results to production grids**;
  A. Streit, D. Erwin, D. Mallmann, R. Menday, Th. Lippert, M. Rambadt, M. Riedel, M. Romberg, B. Schuller, Ph. Wieder
  Proceedings of the Cetraro Workshop HPC 2004, to be published in 2005 by

Elsevier in the series "Grid Computing: New Frontiers of High Performance Computing"

- **Ein Command Line Interface als Zugang zu Grid-Ressourcen**;
  L. Kirtchakova
  accepted by PARS-Workshop (organized by Fachgruppe , Parallel-Algorithmen, -Rechnerstrukturen und –Systemsoftware (PARS)' der Gesellschaft für Informatik e. V., Informationstechnische Gesellschaft im VDE), in German

10.2.3 Presentations

- November 27-28, 2002; Bonn, Germany; Symposium on Grid Computing;
  GRID Computing in Large Scale Molecular Engineering; M. Karelson

- March 25, 2003; Brussels, Belgium; Grids Information Day;
  The OpenMolGRID Project; M. Karelson

- June 4, 2003; Tartu, Estonia; IMAGETOX Seminar;
  Overview of the OpenMolGRID Project; S. Sild

- June 17 2003; Brussels, Belgium; First Project Review;
  - **Overview of OpenMolGRID**; M. Karelson;
  - **Grid Data Warehousing of Molecular Structure-Property (-Activity) Information**
    W. Dubitzky;
  - **Progress and Future Activities In Molecular Descriptor Generation and QSPR Model Building on the Grid & Computational Molecular Engineering of New Compounds and Materials**;.
    S. Sild;
  - **Progress and Future Activities of Grid Integration**; M. Romberg;
  - **Test Application of the OpenMolGRID System for Chemical and Pharmaceutical Predictions;** A. Papp;
  - **Information Dissemination**;G.H.F. Diercksen;
  - **Project Management**; G.H.F. Diercksen

- June 18-19, 2003; Brussels, Belgium; 1st Grid Concertation Meeting
  - Contribution to Working Group **Grid Applications & Grid Benefits**; E.Benfenati;

- Contribution to Working Group **Complex Problem Solving Aspects**; W. Dubitzky;
- Contribution to Working Group **Data Management**; A.Papp;
- Contribution to Working Group **Grid Architectures, Middleware, Interoperability, Scheduling & Resource Discovery**: M.Romberg

- September 7-11, 2003; New York, USA; 226th American Chemical Society National Meeting; **OpenMolGRID, a GRID based system for solving large-scale drug design problems**;
  L. Urge, A. Papp, I. Bágyi, G. Ambrus-Aikelin, F. Darvas

- September 17-19, 2003, Thessaloniki, Greece; Computational Methods in Toxicology and Pharmacology Integrating Internet Resources (CMTPI-2003); **OpenMolGRID, a GRID based system for solving large-scale drug design problems**;
  G.Dorman, A. Papp, L. Urge, I. Bágyi, G. Ambrus-Aikelin, F. Darvas

- October 2-4, 2003; Milan, Italy; IST2003 Exhibition "The Opportunities Ahead";
  OpenMolGRID: Large Scale Molecular Design in Grid;
  S. Sild, A. Papp, G. Pocze

- October 29, 2003; Cracow, Poland; EUROGRID Workshop at 3rd Cracow Grid Workshop;
  OpenMolGRID: Complex Problem Solving in Molecular Design;
  M. Romberg
  http://www.cyfronet.krakow.pl/cgw03/abstracts.html#romberg

- November 10, 2003; Sankt Augustin, Germany;
  Parallelverarbeitungskolloquium;
  OpenMolGRID: Complex Problem Solving in Molecular Design;
  B. Schuller
  http://www.fz-juelich.de/zam/pkoll/Nov2003.html

- January 28-30, 2004; Nicosia Cyprus; AxGrids2004;
  OpenMolGRID, a GRID based system for solving large-scale drug design problems;
  F.Darvas, A. Papp, I. Bágyi, G. Ambrus-Aikelin, L. Ürge

- March 28-April 1, 2004; Anaheim, CA, USA; ACS National Meeting;
  OpenMolGRID, a Grid-based large-scale drug design system;
  L. Ürge, Á. Papp, I. Bágyi, G. Ambrus, and F. Darvas

- May 6, 2004; Ljubljana, Slovenia; International Workshop on Grids for Complex Problem Solving;
  UNICORE and OpenMolGRID (demo); D. McCourt

- May 9-13, 2004; Liverpool, England; The 11th International Workshop on Quantitative Structure-Activity Relationships in Environmental Sciences (QSAR 2004);
  Large Scale Molecular Design and Engineering in Grid;
  U. Maran, S. Sild, A. Lomaka, and M.Karelson

- May 18, 2004; Brussels, Belgium; Project Review;
  - **Overview of OpenMolGRID**; M. Romberg
  - **User Demonstration of the Current OpenMolGRID System**; P. Mazzatorta
  - **Achievements of WP1: Grid Data Warehousing of Molecular Structure – Property (Activity) Information**; D. McCourt
  - **Achievements of WP2: Molecular Descriptor Generation and QSPR Model Building on the Grid**; S. Sild
  - **Achievements od WP3: Computational Molecular Engineering of New Compounds and Materials**; S. Sild
  - **Achievements of WP4: Grid Integration**; B. Schuller
  - **Achievements of WP5: Test Application of the OpenMolGRID System for Chemical and Pharmaceutical Predictions;** A. Papp
  - **Activities of WP6: Dissemination and Exploitation**; G.H.F. Diercksen
  - **OpenMolGRID Project Management**; M. Romberg

- June 04, 2004; Strobl, Austria; AURORA Meeting;
  - **UNICORE Tutorial**; M. Romberg
  - **Applications and UNICORE: OpenMolGRID**; M. Romberg

- June 14, 2004; Brussels, Belgium; Special Review on Project Results Exploitation and Dissemination Activities;
  Exploitation Plans and Activities OpenMolGRID Project, IST-2001-37238 ; I. Bagyi

- June 21-24, 2004; Las Vegas, USA; 2004 International Conference on Parallel and Distributed Processing Techniques and application as part of the 2004 International Multiconference on Computer Science and Computer Engineering;
  OpenMolGRID: Molecular Science and Engineering in a Grid Context; D. McCourt

- June 23-25, 2004, Heidelberg, Germany; International Supercomputer Conference and Exhibition (ISC 2004)
  Demonstration of the OpenMolGRID system; M. Romberg

- July 12, 2004, Münster, Germany; Regional group of the German Society for Informatics (GI)
  Grid Computing; M. Romberg

- July 16, 2004; Jülich, Germany; Visit of participants of the Japanese NAREGI project at FZ Jülich;
  - **Overview of the OpenMolGRID Project**; M. Romberg
  - **OpenMolGRID's Workflow and Resource Management solution**; B. Schuller
  - **A Command Line Interface for UNICORE**; L. Kirtchakova

- September 21, 2004; Brussels; Global Grid Forum 12;
  UNICORE Deployment - Experiences from Testbeds and Production; M. Romberg

- October 15, 2004; Tartu, Estonia; Estonian Grid Seminars
  Invited lecture and OpenMolGRID demonstration: **Keemialalane tarkvara Griidis** (translation: Chemistry Related software in the Grid), S. Sild

- October 27-29, 2004; Vienna; eChallenges 2004
  OpenMolGRID demonstrations; P. Mazzatorta, A. Papp. M. Romberg, S. Sild

- November 2, 2004; Milan; Data, Algorithms and Results in QSAR (DARC2004);
  OpenMolGRID Data Management; B. Schuller

- November 3, 2004; Milan; Data, Algorithms and Results in QSAR (DARC2004)
  QSAR Modelling in OpenMolGRID; S. Sild, U. Maran

- November 15, 2004; Jülich; Visitor from Japanese NAREGI project;
  OpenMolGRID; M. Romberg

- November 18, 2004; Jülich; Visitor from University Linz, Austrian Grid Project;
  UNICORE and OpenMolGRID demonstration; M. Romberg

- December 13, 2004; 4th Cracow Grid Workshop 2004, CGW'04;
  OpenMolGRID: QSAR/QSPR applications in Grid environment; S. Sild

- December 15, 2004; ZAM internal Seminar, Jülich;
  Ein Command-Line-Interface als Zugang zu Grid-Ressourcen; L. Kirtchkova

- December 16,.2004; Jülich; Jahresabschluss-Kolloquium des ZAM;
  OpenMolGRID: Suche nach der Nadel im Heuhaufen; B. Schuller

- January 25, 2005; Jülich; Final Project Review;
  - **The OpenMolGRID system and its application**; M. Romberg
  - **Exploitation of Project Results**; I. Bagy
  - **Demonstration of the basic OpenMolGRID functionality**; P. Mazzatorta
  - **Demonstration of the Data Warehouse transformation process using Grid**; L. Kirtchakova
  - **Demonstration of Molecular Engineering**; S. Sild

- February 14, 2005; Amsterdam; European Grid Conference EGC2005;
  OpenMolGRID: Using Automated Workflows in GRID Computing Environment; U.Maran

- February 15, 2005; Amsterdam; European Grid Conference EGC2005;
  Application driven Grid developments in the OpenMolGRID project;
  B. Schuller

- February 15, 2005; Amsterdam; European Grid Conference EGC2005, Special Events Track;
  OpenMolGRID Demonstration; M.Romberg, B.Schuller, S.Sild, U.Maran

- March 9, 2005; Stuttgart; 8th HLRS Metacomputing and Grid Workshop;
  UNICORE's Evolution Towards a Service Oriented Architecture; M.Romberg

- March 11, 2005; Heidelberg; Grid Symposium;
  UNICORE; M.Romberg

- March 11, 2005; Heidelberg; Grid Symposium;
  Molekül-Design im Grid: Das Projekt OpenMolGRID; M.Romberg

10.2.4 Poster Presentations

- August 10-15, 2003; Drug Discovery Technology, Boston, MA, USA;
  OpenMolGRID, A GRID based system for solving large-scale drug design problems;
  A. Papp., I. Bágyi, G. Ambrus-Aikelin, K. Frobel, L. Ürge, F. Darvas

- January 28-30, 2004; Nicosia Cyprus; AxGrids2004;
  Support for Classes of Applications on the Grid; M.Romberg, B.Schuller

- January 28-30, 2004; Nicosia Cyprus; AxGrids2004;
  The OpenMolGRID Data Warehouse, MOLDW; D.McCourt, W.Jing, W.Dubitzky

10.2.5 Other

- **Article** "Project OpenMolGRID has started" (in German)
  M. Romberg
  ZAMaktuell, November 2002
  http://www.fz-juelich.de/zam/docs/za/2002/za-110.html

- **Presentation on Radio** "The OpenMolGRID Project"
  M. Karelson
  Tervis program, Estonian Radio 4, March 4, 2003

- **Presentation** of the OpenMolGRID project as part of the COST Action 282 Annual Report
  W. Dubitzky
  COST Telecommunication, Information Science and Technology (TIST) Technical Committee (TC) Action Chairs Meeting in Dubrovnik, June 5, 2003

- **Presentation** of the OpenMolGRID project as part of the Grid activities at FZJ
  M. Romberg
  International Supercomputer Conference (ISC 2003) and Exhibition, Heidelberg, Germany, June 25-27, 2003

- **Article** “OpenMolGRID: Application-driven Development of Grid Tools & Services”
  B. Schuller, D. McCourt
  GRIDSTART Newsletter, Issue 6, April 2004; pp. 12-13
  http://www.gridstart.org/download/GRIDSTARTNewsletterApr2004.pdf

- **Interview** “Maailma farmaatsiatööstus saab Tartust topeltkiirenduse (“Tartu is Contributing to World's Pharmaceutical Industry”);
  M. Karelson;
  "Postimees" (one of the major newspapers in Estonia), March 29, 2004
  http://www.postimees.ee/290304/tartu_postimees/130183.php

- **Tutorial** “Unicore: Advanced User Support and Interoperability” as part of a UNICORE tutorial;
  Ph. Wieder (FZJ);
  11th Meeting of the Global Grid Forum in Hawaii on June 9, 2004
  http://unicore.sourceforge.net/docs/ggf11_tutorial_extensions.pdf

- **Article** “OpenMolGRID“
  ComGenex Scientific Newsletter, Spring-Summer 2004, pp.13-16
  http://www.comgenex.com/pdf/Newsletter_2004_SpringSummer.pdf

- **Article** “OpenMolGrid automates 3D molecular structure search using Unicore”
  EnterTheGrid – Primeur Weekly, July 5, 2004
  http://www.hoise.com/primeur/04/articles/weekly/AE-PR-08-04-16.html

- **Flyer** “Speed-up, automatise, and standardise Drug Design using Grid Technology”
  http://www.openmolgrid.org/downloads/flyer.pdf , September 2004

- **Article** “OpenMolGRID erfolgreich abgeschlossen” (in German)
  ZAM aktuell Nr. 132, March 2005
  (http://www.fz-juelich.de/zam/docs/za/2005/za-132#openmolgrid)

10.2.6 Theses

- **Ein Command-Line-Interface als Zugang zu Grid-Ressourcen** (in German)
  L. Kirtchakova
  Diploma Thesis presented at Aachen University of Applied Science, Department Applied Science and Technology, Technomathematics; December 2004

- **QSPR/QSAR modelling of HIV-1 protease inhibitors**
  K. Takkis.
  Bachelor Thesis at the University of Tartu, July 2004

10.3. List of Authors of the Report

1	Executive Summary	M.Romberg
2	Goals of the Project	M.Romberg
3.1.1	Data Warehouse	W.Dubitzky
3.1.2	Custom Data Repository	A.Papp
3.1.3	Substructure Search	A.Papp
3.1.4	Complex Data Transformations	A.Papp
3.2	Model Development	S.Sild
3.3	Molecular Engineering	S.Sild
3.4.1	DataBase Access	B.Schuller
3.4.2	Workflow Support	B.Schuller
3.4.3	Command Line Interface	L.Kirtchakova
3.4.4	Testbed	M.Romberg
3.5	In Silico Testing, Real Life Model Development and Prediction	A.Papp, P.Mazzatorta
4	Contributions to Standards	B.Schuller, E.Benfenati
5	Dissemination	S.Sild
6	Exploitation	S.Sild, I.Bagyi
7	Project Structure and Management	M.Romberg
8	Summary and Lessons Learnt	M.Romberg

10.4. List of Figures

10.5. Glossary of Terms

BSCW	Basic Support for Cooperative Work
CA	Certification Authority
CGX	ComGenex, Budapest
CLI	Command Line Interface
DB	Database
DBAT	Database Access Tool
DW	Data Warehouse
FZJ	Forschungszentrum Jülich GmbH
GRIP	EU Grid Interoperability project
GUI	Graphical User Interface
MDA	Molecular Descriptor Analyser
MDC	Molecular Descriptor Calculation
MOLDW	OpenMolGRID Data Warehouse
NJS	Network Job Supervisor
OGSA	Open Grid Services Architecture
OGSI	Open Grid Services Infrastructure
OLS	Ordinary Least Square
PCR	Principal Components Regression
PKI	Pubic Key Infrastructure
PLS	Partial Least Square
RA	Registration Authority
QSAR	Quantitative Structure-Activity Relationship
QSPR	Quantitative Structure-Property Relationship
SQL	Structured Query Language
TSI	Target System Interface
UNICORE	Uniform Interface to Computer Resources
Usite	UNICORE site
UT	University of Tartu

UU	University of Ulster
Vsite	UNICORE target system at a Usite
XML	Extensible Markup Language
XSL	Extensible Stylesheet Language
XSLT	Extensible Transformations

Already published:

Modern Methods and Algorithms of Quantum Chemistry - Proceedings
Johannes Grotendorst (Editor)
Winter School, 21 - 25 February 2000, Forschungszentrum Jülich
NIC Series Volume 1
ISBN 3-00-005618-1, February 2000, 562 pages
out of print

Modern Methods and Algorithms of Quantum Chemistry - Poster Presentations
Johannes Grotendorst (Editor)
Winter School, 21 - 25 February 2000, Forschungszentrum Jülich
NIC Series Volume 2
ISBN 3-00-005746-3, February 2000, 77 pages
out of print

Modern Methods and Algorithms of Quantum Chemistry - Proceedings, Second Edition
Johannes Grotendorst (Editor)
Winter School, 21 - 25 February 2000, Forschungszentrum Jülich
NIC Series Volume 3
ISBN 3-00-005834-6, December 2000, 638 pages

Nichtlineare Analyse raum-zeitlicher Aspekte der hirnelektrischen Aktivität von Epilepsiepatienten
Jochen Arnold
NIC Series Volume 4
ISBN 3-00-006221-1, September 2000, 120 pages

Elektron-Elektron-Wechselwirkung in Halbleitern: Von hochkorrelierten kohärenten Anfangszuständen zu inkohärentem Transport
Reinhold Lövenich
NIC Series Volume 5
ISBN 3-00-006329-3, August 2000, 146 pages

Erkennung von Nichtlinearitäten und wechselseitigen Abhängigkeiten in Zeitreihen
Andreas Schmitz
NIC Series Volume 6
ISBN 3-00-007871-1, May 2001, 142 pages

Multiparadigm Programming with Object-Oriented Languages - Proceedings
Kei Davis, Yannis Smaragdakis, Jörg Striegnitz (Editors)
Workshop MPOOL, 18 May 2001, Budapest
NIC Series Volume 7
ISBN 3-00-007968-8, June 2001, 160 pages

Europhysics Conference on Computational Physics - Book of Abstracts
Friedel Hossfeld, Kurt Binder (Editors)
Conference, 5 - 8 September 2001, Aachen
NIC Series Volume 8
ISBN 3-00-008236-0, September 2001, 500 pages

NIC Symposium 2001 - Proceedings
Horst Rollnik, Dietrich Wolf (Editors)
Symposium, 5 - 6 December 2001, Forschungszentrum Jülich
NIC Series Volume 9
ISBN 3-00-009055-X, May 2002, 514 pages

Quantum Simulations of Complex Many-Body Systems: From Theory to Algorithms - Lecture Notes
Johannes Grotendorst, Dominik Marx, Alejandro Muramatsu (Editors)
Winter School, 25 February - 1 March 2002, Rolduc Conference Centre, Kerkrade, The Netherlands
NIC Series Volume 10
ISBN 3-00-009057-6, February 2002, 548 pages

Quantum Simulations of Complex Many-Body Systems: From Theory to Algorithms- Poster Presentations
Johannes Grotendorst, Dominik Marx, Alejandro Muramatsu (Editors)
Winter School, 25 February - 1 March 2002, Rolduc Conference Centre, Kerkrade, The Netherlands
NIC Series Volume 11
ISBN 3-00-009058-4, February 2002, 194 pages

Strongly Disordered Quantum Spin Systems in Low Dimensions: Numerical Study of Spin Chains, Spin Ladders and Two-Dimensional Systems
Yu-cheng Lin
NIC Series Volume 12
ISBN 3-00-009056-8, May 2002, 146 pages

Multiparadigm Programming with Object-Oriented Languages - Proceedings
Jörg Striegnitz, Kei Davis, Yannis Smaragdakis (Editors)
Workshop MPOOL 2002, 11 June 2002, Malaga
NIC Series Volume 13
ISBN 3-00-009099-1, June 2002, 132 pages

Quantum Simulations of Complex Many-Body Systems: From Theory to Algorithms - Audio-Visual Lecture Notes
Johannes Grotendorst, Dominik Marx, Alejandro Muramatsu (Editors)
Winter School, 25 February - 1 March 2002, Rolduc Conference Centre, Kerkrade, The Netherlands
NIC Series Volume 14
ISBN 3-00-010000-8, November 2002, DVD

Numerical Methods for Limit and Shakedown Analysis
Manfred Staat, Michael Heitzer (Eds.)
NIC Series Volume 15
ISBN 3-00-010001-6, February 2003, 306 pages

Design and Evaluation of a Bandwidth Broker that Provides Network Quality of Service for Grid Applications
Volker Sander
NIC Series Volume 16
ISBN 3-00-010002-4, February 2003, 208 pages

Automatic Performance Analysis on Parallel Computers with SMP Nodes
Felix Wolf
NIC Series Volume 17
ISBN 3-00-010003-2, February 2003, 168 pages

Haptisches Rendern zum Einpassen von hochaufgelösten Molekülstrukturdaten in niedrigaufgelöste Elektronenmikroskopie-Dichteverteilungen
Stefan Birmanns
NIC Series Volume 18
ISBN 3-00-010004-0, September 2003, 178 pages

Auswirkungen der Virtualisierung auf den IT-Betrieb
Wolfgang Gürich (Editor)
GI Conference, 4 - 5 November 2003, Forschungszentrum Jülich
NIC Series Volume 19
ISBN 3-00-009100-9, October 2003, 126 pages

NIC Symposium 2004
Dietrich Wolf, Gernot Münster, Manfred Kremer (Editors)
Symposium, 17 - 18 February 2004, Forschungszentrum Jülich
NIC Series Volume 20
ISBN 3-00-012372-5, February 2004, 482 pages

Measuring Synchronization in Model Systems and Electroencephalographic Time Series from Epilepsy Patients
Thomas Kreutz
NIC Series Volume 21
ISBN 3-00-012373-3, February 2004, 138 pages

Computational Soft Matter: From Synthetic Polymers to Proteins - Poster Abstracts
Norbert Attig, Kurt Binder, Helmut Grubmüller, Kurt Kremer (Editors)
Winter School, 29 February - 6 March 2004, Gustav-Stresemann-Institut Bonn
NIC Series Volume 22
ISBN 3-00-012374-1, February 2004, 120 pages

Computational Soft Matter: From Synthetic Polymers to Proteins - Lecture Notes
Norbert Attig, Kurt Binder, Helmut Grubmüller, Kurt Kremer (Editors)
Winter School, 29 February - 6 March 2004, Gustav-Stresemann-Institut Bonn
NIC Series Volume 23
ISBN 3-00-012641-4, February 2004, 440 pages

Synchronization and Interdependence Measures and their Applications to the Electroencephalogram of Epilepsy Patients and Clustering of Data
Alexander Kraskov
NIC Series Volume 24
ISBN 3-00-013619-3, May 2004, 106 pages

High Performance Computing in Chemistry
Johannes Grotendorst (Editor)
Report of the Joint Research Project:
High Performance Computing in Chemistry - HPC-Chem
NIC Series Volume 25
ISBN 3-00-013618-5, December 2004, 160 pages

Zerlegung von Signalen in unabhängige Komponenten: Ein informationstheoretischer Zugang
Harald Stögbauer
NIC Series Volume 26
ISBN 3-00-013620-7, April 2005, 110 pages

Integration von Programmiersprachen durch strukturelle Typanalyse und partielle Auswertung
Jörg Striegnitz
NIC Series Volume 28
ISBN 3-00-016006-X, May 2005, 306 pages

All volumes are available online at http://www.fz-juelich.de/nic-series/.