

# OpenMolGRID: Molecular Science and Engineering in a Grid Context

P. Mazzatorta<sup>1</sup>, E. Benfenati<sup>1</sup>, B. Schuller<sup>2</sup>, M. Romberg<sup>2</sup>, D. McCourt<sup>3</sup>, W. Dubitzky<sup>3</sup>, S. Sild<sup>4</sup>, M. Karelson<sup>4</sup>, A. Papp<sup>5</sup>, I. Bágyi<sup>5</sup>, and F. Darvas<sup>5</sup>

<sup>1</sup> Istituto di Ricerche Farmacologiche "Mario Negri" Milano,  
Via Eritrea, 62, 20157 Milano, Italy;

<sup>2</sup> ZAM, Forschungszentrum Jülich, 52425 Jülich, Germany;

<sup>3</sup> University of Ulster, Cromore Road, Coleraine BT521SA, Northern Ireland;

<sup>4</sup> Department of Chemistry, University of Tartu, Jakobi Str 2, 51014 Tartu, Estonia;

<sup>5</sup> ComGenex, Inc., Bem rkp. 33-34, H-1027 Budapest, Hungary

## Abstract

*Modern approaches to chemistry and pharmacology deal with large-scale molecular design problems. The molecular design is essentially based on data warehousing and data mining. Data warehousing techniques are needed to collect relevant data from distributed and heterogeneous databases. Data mining techniques are used to build predictive quantitative structure-property and activity relationship models. Increasingly, the computational resources (databases and analysis and modeling programs) needed for molecular engineering are geographically distributed. Grid technology offers a framework to build distributed molecular engineering systems. The OpenMolGRID project is focused on integrating existing software modules into a Grid infrastructure, and on making a solid foundation for next step molecular engineering tools. This involves the design of a seamless and unified user interface and the provision of adapters to make software Grid-aware. The system will be used to develop prototype applications for the generation of molecular structures with given chemical properties or biological activities. It will be intensively tested by academic and industrial users on real applications (multi-drug resistance, G-protein-coupled receptors activity, and toxicity).*

**Keywords:** Molecular Engineering, Grid, Data Warehousing, Data Mining, QSAR/QSPR, OpenMolGRID

## 1. Introduction

The use of *quantitative structure-property/activity relationship* (QSPR/QSAR) methods responds to the need of pharmaceutical and chemical companies for the *in-silico* screening of millions of new potential drugs or chemicals and to the need of research institutes and regulatory bodies for having fast, accurate and reliable models to understand and predict the consequence of chemicals to human health, wildlife, and the environment. The basic concepts together with new approaches of QSPR/QSAR have been reviewed several times [1]-[6]. In a recent paper [7], Schultz and Cronin identified the essential and desirable characteristics of ecotoxicity quantitative structure-activity relationships, which can be generally applied to every QSPR/QSAR method. According to them, the development of QSPR/QSARs should be based on:

1. a reliable dataset, which differ both in terms of potency and chemical structure;
2. a set of descriptors of superior quality, reproducible, of a number and type so as to be constant with the property being modeled, and when possible, allow for a mechanistic explanation;
3. a rigorous and appropriate statistical process; and
4. a strict validation procedure of the model.

For the purpose of general models, following these four rules, researchers easily end up with an extremely large dataset to be analyzed, and in spite of the potential capabilities of modern computers, the computational effort for such studies will be massive. The well-known axiom *time is money* applies also to QSPR/QSARs and the calculation procedure has to be reduced in terms of computation time. The new Grid technology seems to be adequate to such a challenge because of its ability to exploit distributed computational resources.

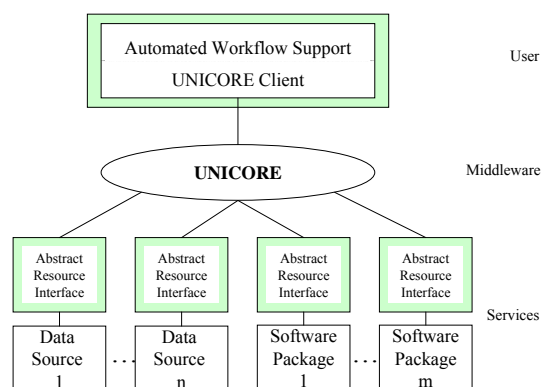
The main objective of the OpenMolGRID project is to provide a unified and extensible information-rich environment for solving molecular design/engineering tasks relevant to chemistry, pharmacy and life sciences. The OpenMolGRID system comprises a set of application-oriented tools that are built on core Grid services and functions provided by the UNICORE [8] infrastructure. The specific components of OpenMolGRID are discussed below. The general structure and technology of the project is described in Section 2. Section 3 outlines the data warehousing component, which is designed to integrate from disparate locations. The choice and the adaptation of existing QSPR and QSAR analysis software for the Grid environment is the subject of Section 4. The paper ends with conclusions and an outlook on further applications of the OpenMolGRID system.

## 2. OpenMolGRID Architecture

The OpenMolGRID project has several requirements in terms of the underlying Grid infrastructure: (1) Existing computational software packages need to be integrated, with particular emphasis on support for complex, multi-step workflows. (2) Computationally intensive tasks need to be executed in a distributed fashion to reduce turn-around times. (3) Access to heterogeneous data sources is needed, where the strict security requirements of the pharmaceutical industry need to be taken into account. (4) And above all, the user interface of the system has to be as user-friendly as possible, with most of the Grid-related complexity hidden from the user, while still providing all of the flexibility and power of the underlying Grid system for advanced users.

The UNICORE (UNified Interface to COmputing RESources) [8] Grid infrastructure was chosen as the foundation of the OpenMolGRID system. UNICORE can be briefly characterized as a vertically integrated Grid system, with an emphasis on seamless and secure access to Grid resources. It offers a powerful and easy-to-use graphical user interface, single sign-on, and

strong security through X.509 public key cryptography. The UNICORE plugin interface allows for the straightforward integration of new applications. OpenMolGRID uses the open interfaces provided by UNICORE to integrate novel applications such as databases and software packages. Figure 1 shows the general structure of the OpenMolGRID system.



**Figure 1. General OpenMolGRID architecture.**

OpenMolGRID extends UNICORE, adding significant new functionality in the areas of workflow support and resource management. On the client side, a new type of UNICORE plugin, called MetaPlugin, supports the user in dealing with complex workflows. It enables users to build UNICORE jobs from abstract workflow descriptions, where details such as file transfers, dependencies and resource allocation are taken care of automatically. Computationally intensive tasks can be run on multiple sites, if the input data can be split into smaller pieces and distributed. The resources needed for the job are identified and allocated automatically by the MetaPlugin.

On the server side, an abstraction layer, called Abstract Resource Interface, is used to access software resources in a generic fashion. Data sources are integrated in the same general way using an Abstract Resource Interface. An important task of the Abstract Resource Interface is to allow the use of standardized input/output formats, thus creating an abstraction layer around the underlying software package.

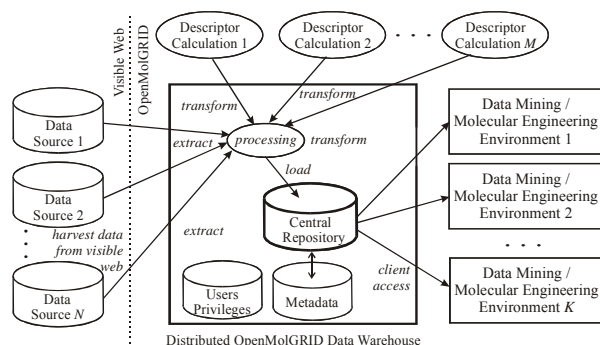
The extension of the UNICORE infrastructure that is required to provide the functionality offered by the OpenMolGRID system is performed in an application domain in an independent, flexible and extensible fashion by using an XML-based metadata layer.

All GUI plugins for new software packages developed within OpenMolGRID are also useable as standalone components, thus creating added value for

UNICORE, even without taking advantage of the full OpenMolGRID system.

### 3. Data warehousing

The molecular engineering process of the OpenMolGRID system is supported by data mining tools and systems. Data mining techniques, such as *multi-linear regression* (MLR), *principle component analysis* (PCA), *partial least squares* (PLS), and *artificial neural networks* (ANN), are used to build predictive QSPR/QSAR models [1]. Data warehousing is often employed as a prerequisite to data mining [9]. A data warehouse integrates, cleanses, normalizes, and consolidates data from different sources and maps them onto “ready-to-use” data structures (e.g. by denormalizing relational database tables). The main function of the *OpenMolGRID data warehouse* (MOLDW) is to harvest chemical-compound data from public resources and integrate and pre-process them for the data mining and molecular engineering process within OpenMolGRID. The diagram in Figure 2 illustrates the basic logical structure of the MOLDW and its relationship to other system components.



**Figure 1. MOLDW and other OpenMolGRID system components.**

What is interesting from a Grid perspective is the requirement (1) to harvest data from public repositories into a protected Grid computing environment (the OpenMolGRID), and (2) to incorporate physically distributed data transformations, so-called *descriptor calculations*, into the logically integrated data warehouse. Descriptor calculations are fundamental to *in silico* molecular modeling. Some descriptors are calculated from the three-dimensional structure information, e.g. molecular volume, quantum chemical descriptors. Specialized 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, MOLDW will only update an entry and recompute descriptors if the entry in the underlying database has been modified, avoiding needless computation. MOLDW 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 to carry out the required integration and transformations.

Currently, various aspects of MOLDW and its interoperation both with the visible web and the OpenMolGRID system have been designed and implemented. These include the logical and physical data model of the central storage (realized on a PostgreSQL relational database platform), the database access tool (DBAT), and certain aspects of the ELT (extract, load, transform) [9] processes. Some more advanced warehouse access and query tools (e.g. fingerprinting, substructure search) are subject to being developed in the near future.

### 4. Modules for QSPR/QSAR Modeling

The QSPR/QSAR models are designed by finding relationships between property/activity and molecular structures. This process involves various tasks that are carried out at different stages of a complicated workflow. A typical workflow starts with the extraction of a training set with the experimental property/activity values from the data source (e.g. data warehouse, database, file system). Normally, the prerequisite for the model development is the calculation of molecular descriptors [10] [11], which are used to represent molecular structures in the model. The descriptor calculation itself can be a multi-step process and depend on the generation of 3-dimensional coordinates and carrying out quantum-chemical calculations. Currently thousands of different molecular descriptors are available and various data mining techniques (MLR, PCA, PLS, ANN etc) can be used to select the significant descriptors that have causal relationship with the modeled property or activity.

Each of the tasks described above can be performed with different software packages that are often incompatible with each other. However, the most optimal design of the predictive models requires the combined application of multiple software packages. This problem is addressed within the OpenMolGRID infrastructure by the development of UNICORE compliant applications that adapt existing software modules for the design of predictive QSPR/QSAR

models. These OpenMolGRID applications can be then combined to carry out complex workflows. As described in Section 2, each application consists of two parts – the plugin to the UNICORE Client and the Abstract Resource Interface. This architecture allows different software packages to be used when performing one specific task in the workflow.

A set of programs are integrated into OpenMolGRID to demonstrate its capabilities: The MOLGEO [12] program is adapted to generate 3D coordinates for molecular structures. This task is needed because all quantum chemical and most molecular descriptor calculation programs require the 3D representation of molecular structures as an input. However, the 2D representation of a molecular structure is very convenient for sketching molecular structures manually. In addition, most chemical databases contain 2D representations only. The MOPAC [13] program is adapted for performing the 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 molecular descriptor calculation and QSPR/QSAR model building modules are adapted from the CODESSA PRO program [14]. The molecular descriptors are derived directly from the molecular structure and the results of the quantum chemical calculations. The QSPR/QSAR models are derived with MLR, PCR, and PLS methods.

In addition, new molecular engineering tools are developed for the computer-aided construction of molecular structures with predefined chemical properties or biological activities. These tools will make it possible to explore large chemical space in a cost effective way to find potential candidates for new drugs, chemicals, or materials. The generation of new molecular structures is based on a library of fragment structures. Using that library, various structure generation algorithms can construct a huge number of candidate structures. The candidate structures are then validated using the previously developed predictive models and a small subset of molecules that match the target properties or activities is selected for further investigation.

## 5. Conclusions

OpenMolGRID (Open Computing Grid for Molecular Science and Engineering) will be one of the first realizations of the Grid technology in drug design. The system is designed to create QSPR/QSAR models and use them to predict biological activities or ADME (absorption, distribution, metabolism, and excretion)

related properties. OpenMolGRID is based on the adaptation and integration of existing, widely accepted, relevant computing tools and data sources, using the UNICORE infrastructure, to make a solid foundation for the next step molecular engineering tools. Using the implemented data warehouse technology, the system will be suitable to collect data from geographically distributed, heterogeneous sources.

Currently the system is being developed and a number of system components are implemented. A fully integrated test system will be operational in August 2004. The final system will be capable of solving molecular engineering problems on a large-scale. In particular the system will facilitate the discovery of novel compounds with favorable properties by analyzing millions of structures in a reasonable time. The OpenMolGRID system will be tested in real-life situations against 30,000 human fibroblast cytotoxicity data measured for novel and diverse structures. It will be validated by modeling the toxicity of chemicals for relevant ecological endpoints and will be used to identify potential anticancer agents.

## 6. Acknowledgement

We acknowledge the financial support from the European Commission for the OpenMolGRID project. The contract number is IST-2001-37238.

## 7. References

- [1] Hansch, C. & Leo, A. (1995). Exploring QSAR, Fundamentals and Applications in Chemistry and Biology, ACS, Washington, DC, Chapters 6 & 11.
- [2] McKinney, J.D.; Richard, A.; Waller, C.; Newman, M.C.; Gerberick, F. The practice of structure activity relationships (SAR) in toxicology. *Toxicol. Sci.*, 2000, 56, pp. 8-17.
- [3] Katritzky, A R; Petrukhin, R; Tatham, D; Basak, S; Benfenati, E; Karelson, M; Maran, U Interpretation of quantitative structure-property and -activity relationships, *J. Chem. Inf. Comp. Sci.*, 2001, 41, pp. 679-685.
- [4] Mazzatorta, P.; Vračko, M.; Benfenati, E. ANVAS: Artificial Neural Variables Adaptation System for descriptor selection, *J. Comput. Aid. Mol. Des.*, 2003, 17, pp. 335-346.
- [5] Schutz, T.W.; Cronin., M.T.D.; Walker, J.; Aptula, A. Quantitative structure-activity relationships (QSARs) in toxicology: a historical perspective. *J. Mol. Struct.-Theochem*, 2003, 622, pp. 1-22.
- [6] Walker, J.D.; Schultz, T.W.; Structure-activity relationships for predicting ecological effects of chemicals. In Hoffman D., Rattner B.A., Burton G. Jr., Cairns J. Jr.,

eds., Handbook of Ecotoxicology, 2nd ed. CRC. Boca Raton, FL, USA, pp. 893-910.

[7] Schutz, T.W.; Cronin, M.T.D. Essential and desirable characteristics of ecotoxicity quantitative structure-activity relationships. *Environ. Toxicol. Chem.*, 2003, 22, pp. 599-607.

[8] Romberg, M. The UNICORE Grid Infrastructure; Scientific Programming Special Issue on Grid Computing, 2002, 10, pp. 149-158.

[9] Moss, L.; Adelman, S. Data Warehousing Methodology, *Journal of Data Warehousing*, 2000, 5, pp. 23-31.

Karelson, M. *Molecular Descriptors in QSAR/QSPR*. John Wiley & Sons, New York, 2000.

[10] Todeschini, R.; Consonni, V. *Handbook of Molecular descriptors*. Wiley-VCH: Weinheim, 2000.

[11] Katritzky, A.R.; Gordeeva, E.V.; Shcherbukhin, V.V.; Zefirov, N.S. Rapid Conversion of Molecular Graphs to 3D Representation using the MOLGEO Program. *J. Chem. Inf. Comput. Sci.*, 1993, 33, pp. 102-111.

[12] Stewart, J.J. MOPAC: a semiempirical molecular orbital program, *J. Comput. Aid. Mol. Des.*, 1990, 4, pp 1-45.  
<http://www.codessa-pro.com>