OpenMolGRID – A UNICORE-based System for Molecular Science and Engineering

Uko Maran

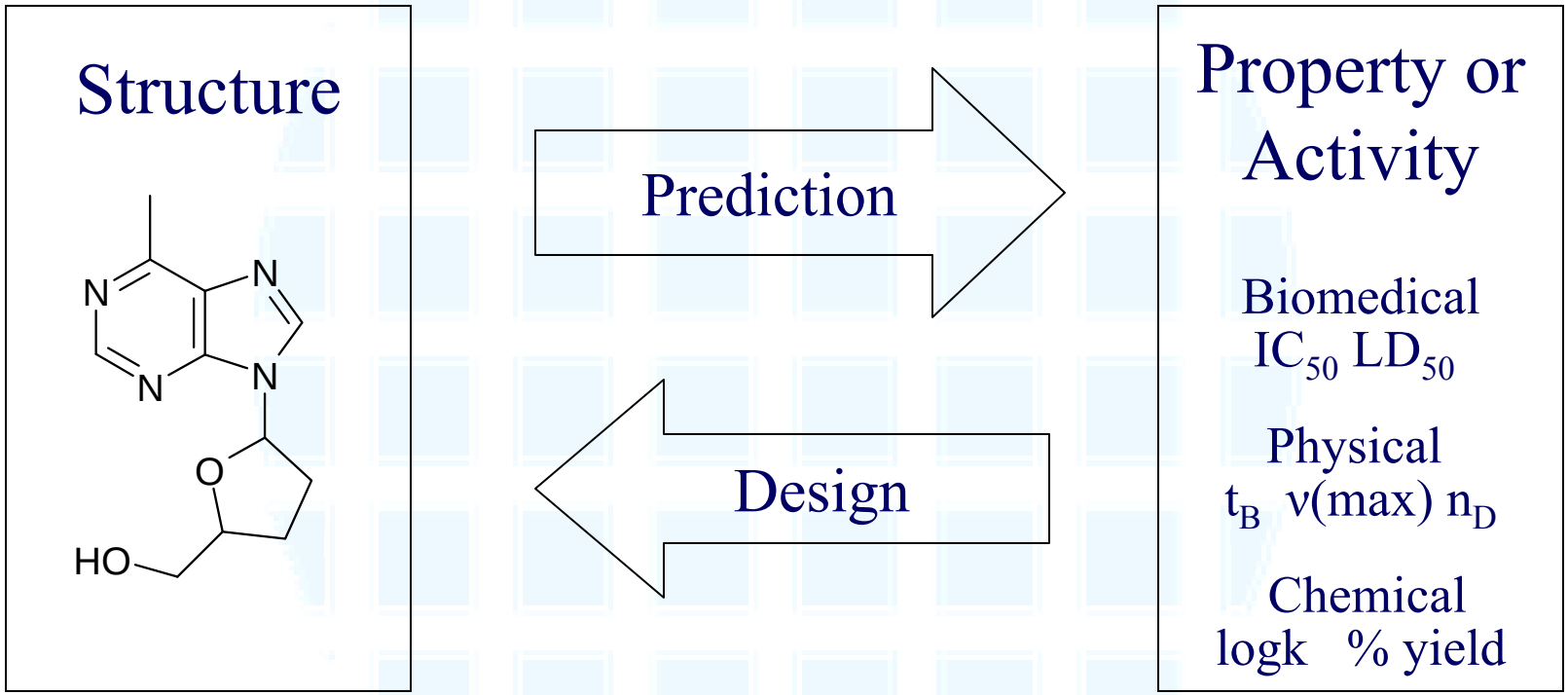
University of Tartu

uko.maran@ut.ee

Content

- Molecular engineering
- What is OpenMolGRID?
- Contributions to UNICORE
- Example
- Concluding remarks
- Chemical applications in the Grid
- …

General application framework: Molecular Engineering



What is OpenMolGRID?

Open Computing Grid for Molecular Science and Engineering

System prototype to deal with
large-scale molecular engineering problems

Specific objective of the project was to
automatise, integrate and speed-up the drug-discovery pipeline using Grid technology

OpenMolGRID Team

www.openmolgrid.org

Forschungzentrum Jülich, Germany
University of Tartu, Estonia
University of Ulster, Northern Ireland
Mario Negri Institute, Italy
ComGenex, Inc., Hungary

Subcontractors:
OpenMolConsulting, Germany
Politecnico di Milano, Italy

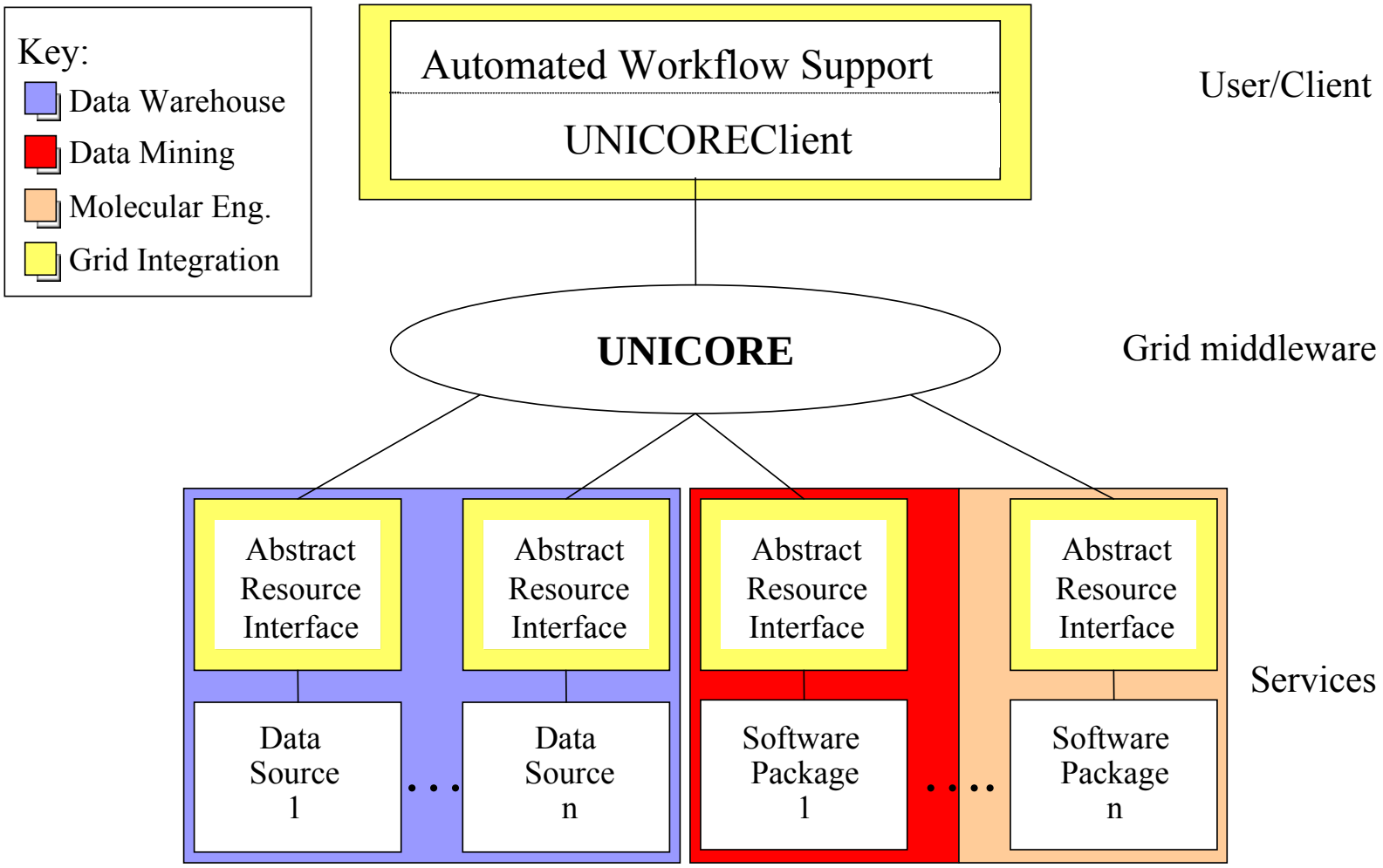
Sponsorship:
IST-2001-37238
(EC-FP5: OpenMolGRID)

Information Society Technologies

People behind OpenMolGRID

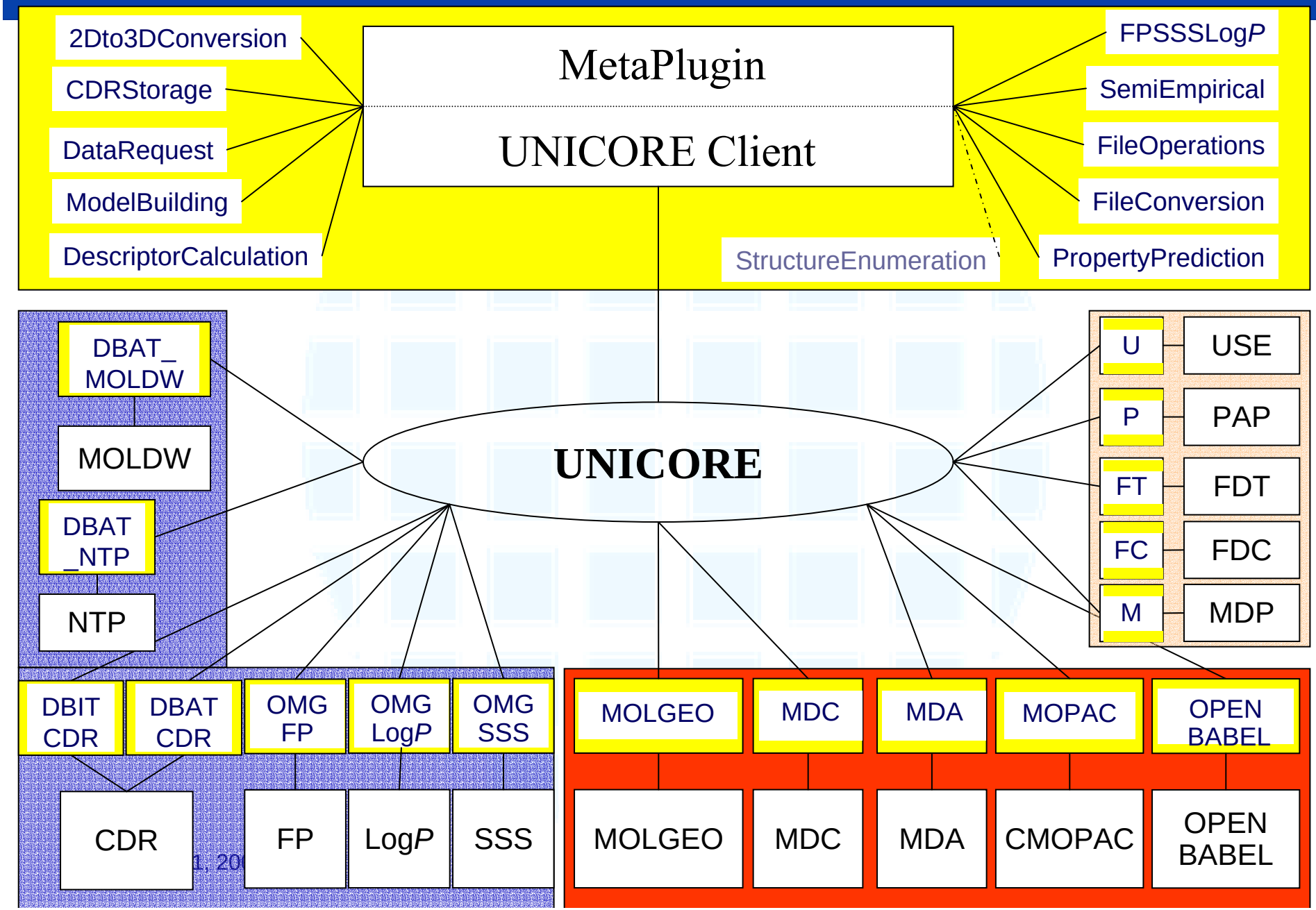
- **Forschungzentrum Jülich**, *Germany*
  - Lidia Kirtchakova, Andre Latour, Mathilde Romberg, Bernd Schuller
- **University of Tartu**, *Estonia*
  - Andre Lomaka, Iiris Kahn, Mati Karelson, Uko Maran, Sulev Sild
- **University of Ulster**, *Northern Ireland*
  - Werner Dubitzky, Mykola Galushka, Jean Jing, Jesus Lopez, Damian McCourt, Rachael Tuaim, Brian Sturgeon, Lynsay Wright
- **Mario Negri Institute**, *Italy*
  - Emilio Benfenati, Mosé Casalegno, Paolo Mazzatorta
- **ComGenex, Inc.**, *Hungary*
  - Istvan Bagyi, Tamas Csokona, Ferenc Darvas, Robert Ferenzi, Peter Hliva, Anna Kelemen, Peter Kormos, Akos Papp, Éva Wikonkál
- **OpenMolConsulting**, *Germany*
  - Geerd Diercksen
- **Politecnico di Milano**, *Italy*
  - Giuseppina Gini

OpenMolGRID Architecture



OpenMolGRID
Open Computing GRID for Molecular Science and Engineering

Integrated software



Integrated research/application fields

- Data warehousing
- Chemical structure conversion (2D to 3D)
- Quantum chemical calculations
- Molecular descriptor calculation
- QSPR/QSAR model building
- Chemical structure engineering
- **Grid technologies**

Solutions

- **Orchestration of scientific applications - Process automation**
  Integrate your scientific applications into automated workflows
- **Chemical Data Management**
  Seamless access to distributed data resources
- **Seamless QSAR/QSPR**
  Grid-enabled solution for modeling large and complex data sets
- **Molecular Engineering**
  Computer aided design of new compounds
- **Standardization of QSAR/QSPR protocols**
  Predict (bio) chemical activity/property with standardized models
- **...**

Contributions to UNICORE

- OpenMolGRID workflow support
- OpenMolGRID command-line interface (CLI)

https://sourceforge.net/projects/unicore/

Workflows Specification: XML

- XML schema - allows the high level definition of workflows
- Defined scientific processes are mapped to UNICORE job objects
- Core elements: ***task and dependency***
  - ***Dependency*** element defines relationship between two tasks
  - ***Task*** defines parameters for each independent application
- …

Sild, Maran, Romberg, Schuller, Benfenati OpenMolGRID: Using Automated Workflows in Grid Computing Environement. In *Advances in Grid Computing*, LNCS3470 (EGC 2005), pp464-473, 2005.

Workflow Processing: MetaPlugin

- Parses XML workflow
- Creates UNICORE jobs
- Assigns target systems (*vsite*) and resources
- Automatically created tasks:
  - Data transfer from one system to other
  - Data conversion between jobs
  - Data splitting, distribution and joining
- Defines the graph of task dependencies (example will follow)

Example XML workflow

```
<?xml version="1.0"?>
<!-- Model development for Solubility in Water -->
<workflow>
<task name="2Dto3Dconversion" .../>
</task>

<task name="SemiempiricalCalculation" identifier="MOPAC_OPT" id="2"
export="false" split="true" splitterTask="SplitStructureList"
joinerTask="JoinStructureLists">
<option name="keywords" value="AM1 NOINTER MMOK GNORM=0.1 EF"/>
</task>

<task name="SemiempiricalCalculation" identifier="MOPAC_PCalc" .../>
</task>

<task name="DescriptorCalculation" identifier="DescCalc" ...>
</task>

<task name="ModelBuilding" identifier="ModelBuild" ...>
<localInput source="H:\Unicore\test\Solub-data-water.plf" .../>
</task>

<dependency pred="1" succ="2"/><!-- 2D-3D to MOP1 -->
<dependency pred="2" succ="3"/><!-- MOP1 to MOP2 -->
<dependency pred="3" succ="4"/><!-- MOP2 to DC -->
<dependency pred="4" succ="5"/><!-- DC to MB -->
</workflow>
```

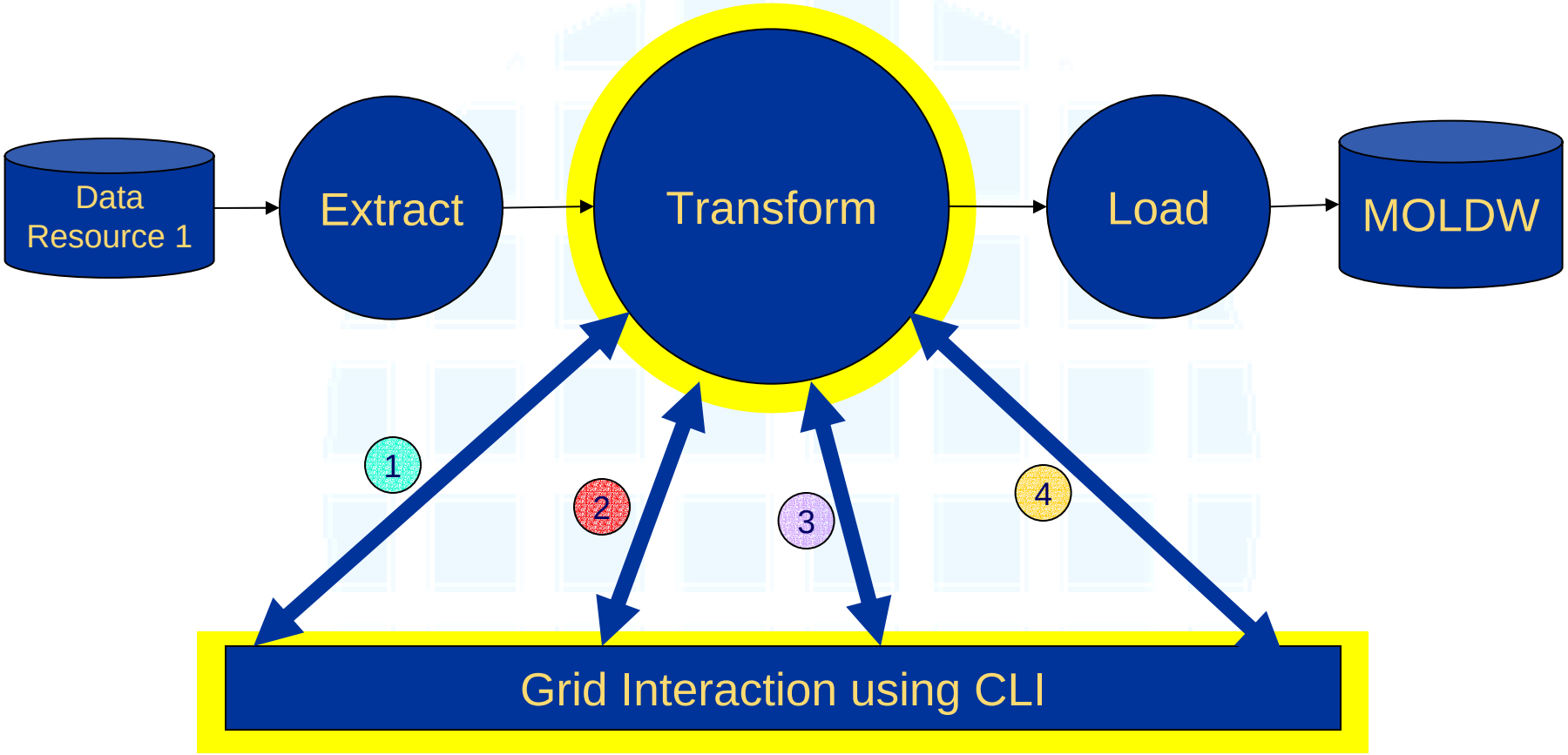
Sild, Maran, Romberg, Schuller, Benfenati OpenMolGRID: Using Automated Workflows in Grid Computing Environement. In *Advances in Grid Computing*, LNCS3470 (EGC 2005), pp464-473, 2005.

Command Line Interface (CLI)

- Unicore lacked a tool that allowed to make use of Grid resources from within applications (batch processing)
- CLI offers AJO generation function that builds job dynamically from an XML workflow description (suns jobs, monitors them, fetches the results)
- Is based on MetaPlugin and uses full OpenMolGRID metadata layer (workflows in GUI client and CLI re inter-changeable)
- …

Schuller, Romberg, Kirtchakova Application driven Grid developments in the OpenMolGRID Project. In *Advances in Grid Computing*, LNCS3470 (EGC 2005), pp23-29, 2005.

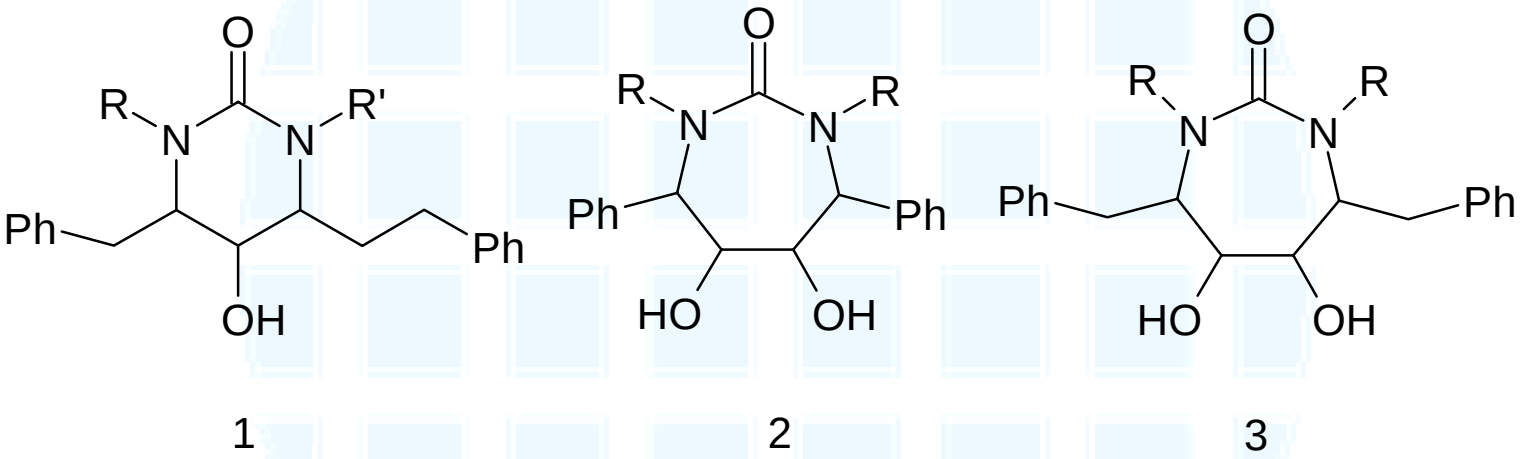
Molecular Data Warewouse (MOLDW) Transformation Process



1. Descriptor calculation for 2D-structures
2. 2D to 3D Conversion
3. Semi-empirical structure optimisation
4. Descriptor calculation for 3D-structures

Modelling HIV-1 Protease Inhibitors 1/4

Efficient inhibition of aspartyl proteased enzyme can decrease HIV-1 *via* the production of non-infectious viral particles and this prevents the further propagation of the virus

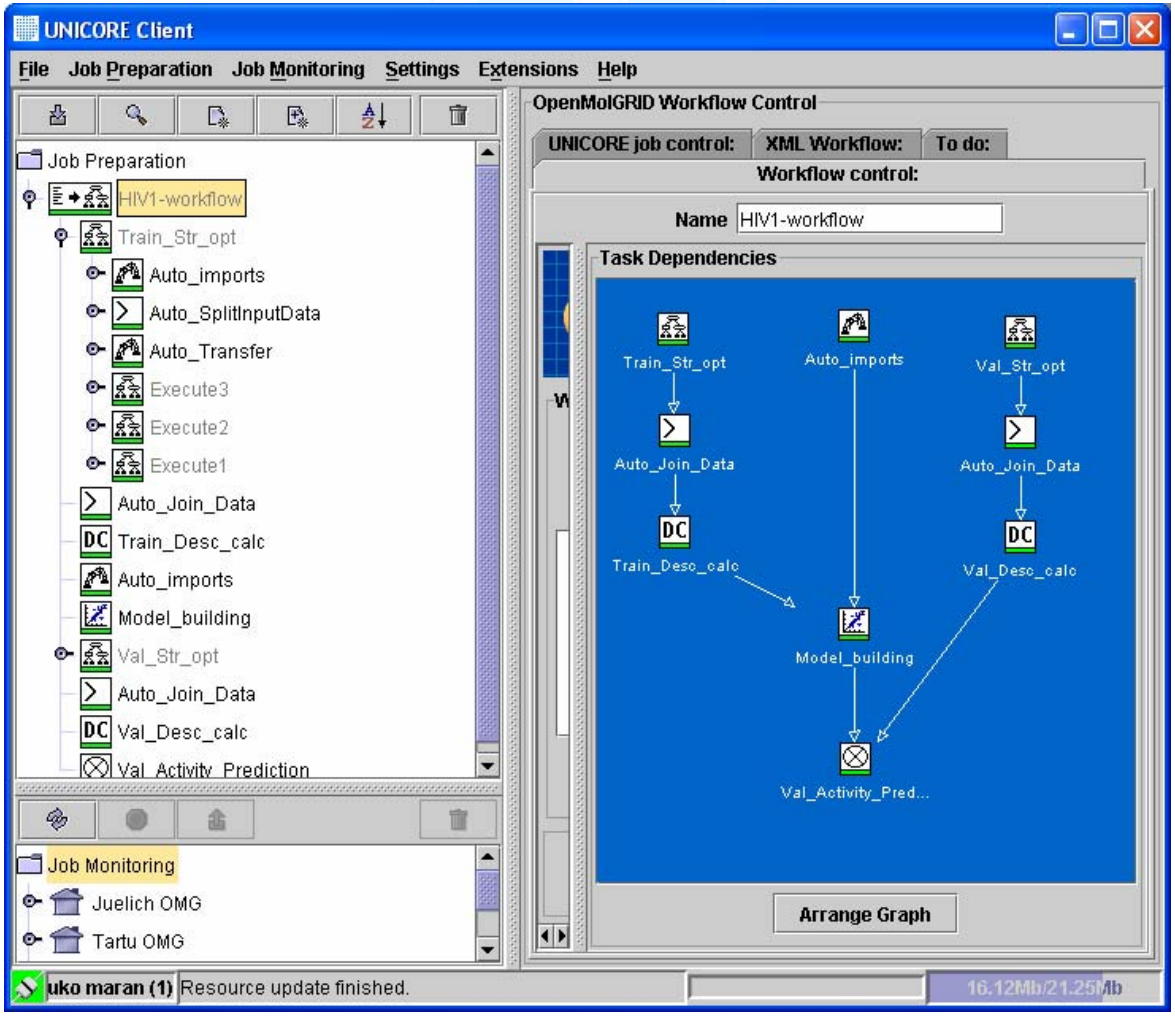


1 2 3

- cluster-based factor analysis for splitting training and validation data

Maran, Sild, Kahn, Takkis Mining of the Chemical Information in GRID Environment. *Future Generation Computer Systems* (submitted)

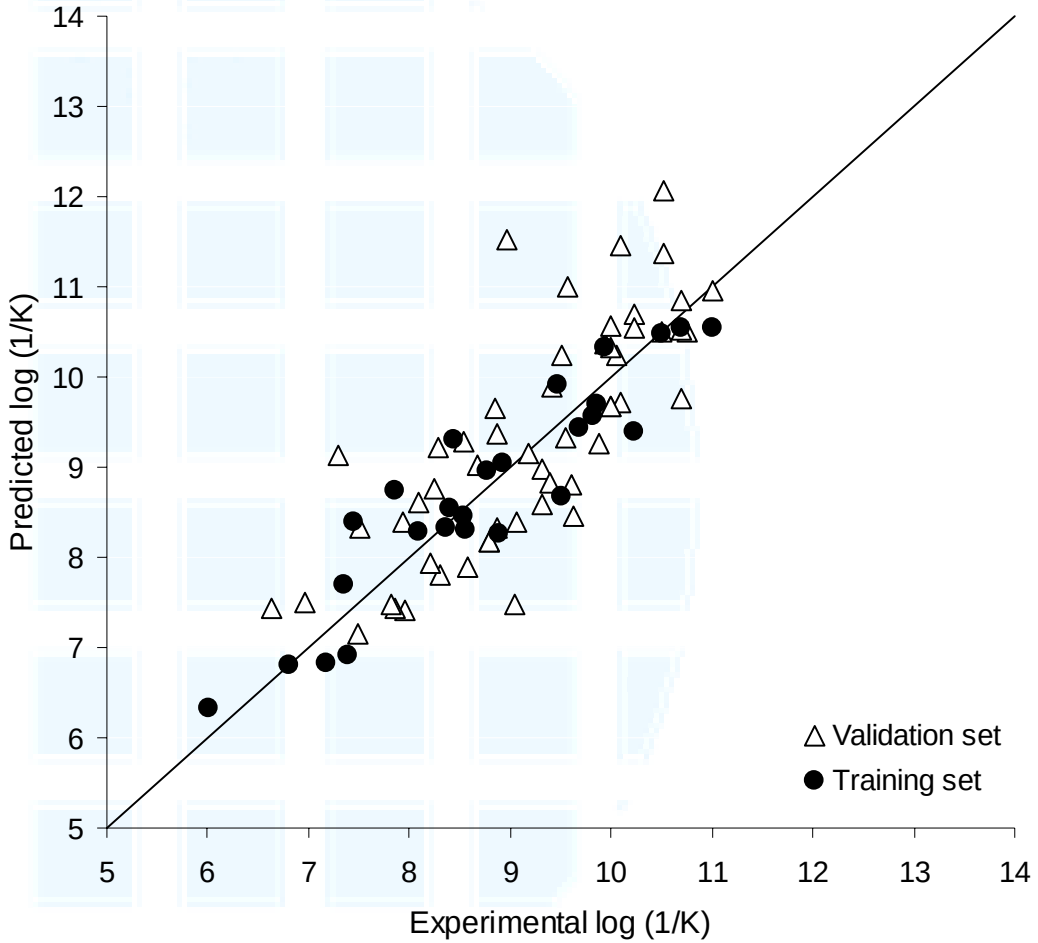
Modelling HIV-1 Protease Inhibitors 2/4



Maran, Sild, Kahn, Takkis Mining of the Chemical Information in GRID Environment.
Future Generation Computer Systems (submitted)

Modelling HIV-1 Protease Inhibitors 3/4

- Training set
  - R^2=0.86
  - s^2=0.51
  - R^2_{cv}=0.81
- Validation set
  - s^2=0.67



Maran, Sild, Kahn, Takkis Mining of the Chemical Information in GRID Environment. *Future Generation Computer Systems* (submitted)

Modelling HIV-1 Protease Inhibitors 4/4

- The improvement of the time factor of the present modelling task due to the grid integration:
  - 1 DAY: experienced user, no grid integration, standalone applications, single CPU, manual conversions and transfer of the data between different applications;
  - 1 Hour: experienced user, grid integration, automated workflow, single CPU;
  - About 10 minutes: experienced user, grid integration, automated workflow, distributed computational resources.

Maran, Sild, Kahn, Takkis Mining of the Chemical Information in GRID Environment. *Future Generation Computer Systems* (submitted)

Molecular engineering workflow

Need compounds with property or activity for predefined values

Fragment Libraray

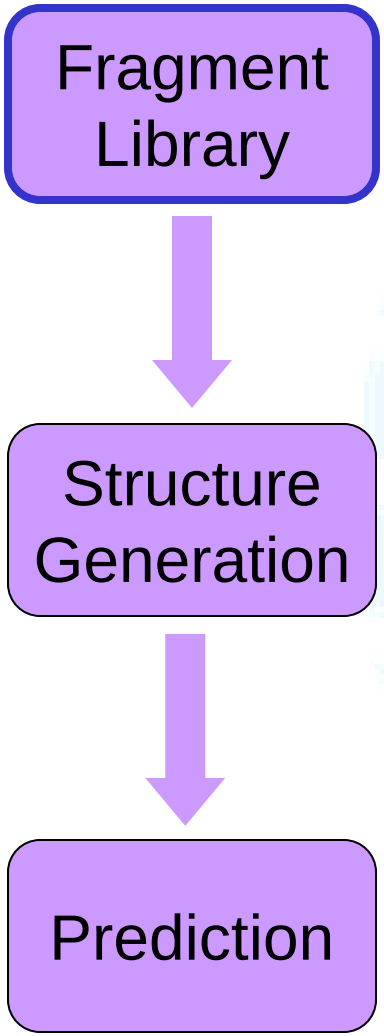
↓

Structure Generation

↓

Property of activity prediction

Fragment Library



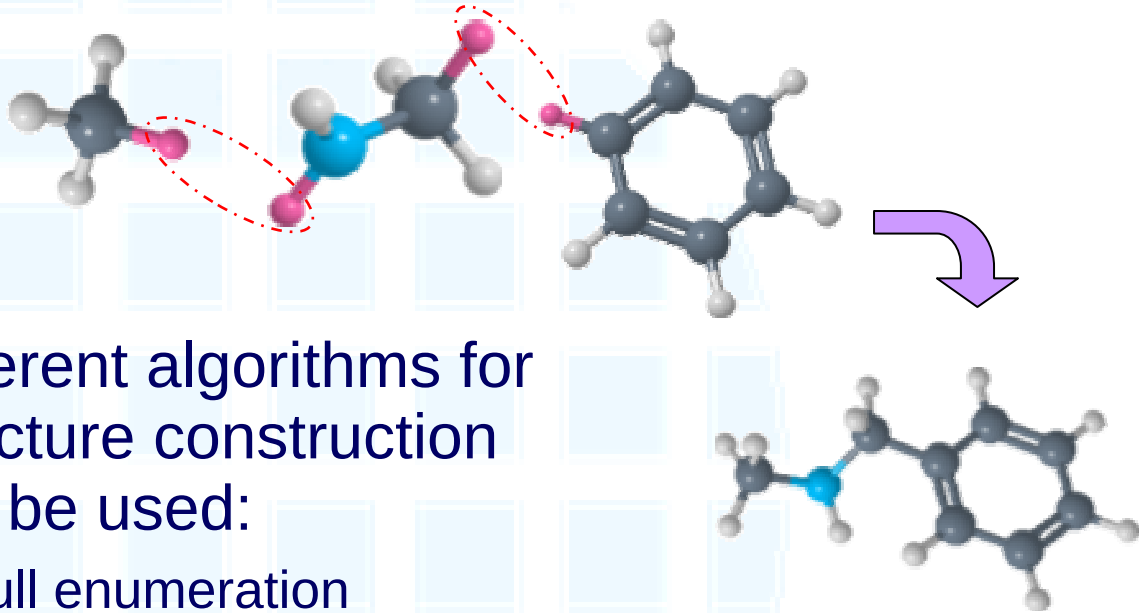
- Fragments are stored in the custom data repository and are accessed like normal molecules
- Both 2D and 3D representations are supported
- Stores fragment descriptors that can be used for rapid prediction of molecular descriptor values

Structure Generation

Fragment Library

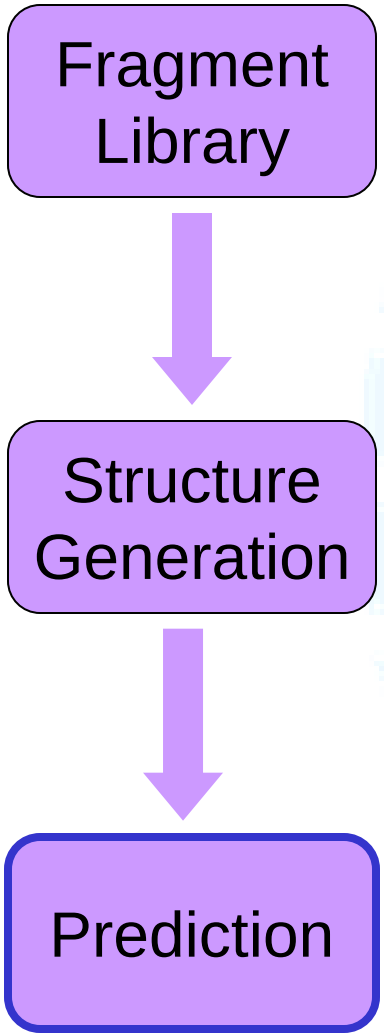
Structure Generation

Prediction



- Different algorithms for structure construction can be used:
  - full enumeration
  - stochastic methods
- At first level the candidate structures are filtered by using pre-calculated fragment descriptors.

Prediction



- For the candidate structures exact molecular descriptors are calculated using workflows (including 2Dto3D conversion, semi-empirical calculations, etc.).
- Using existing QSAR/QSPR models the properties and activities are predicted.
- The best candidates are selected for the final analysis in the lab.

Experiences and further expectations

- UNICORE is well suited to integrate applications, but we are very much look forward for new developments (GS, etc.)
- Limitations can be reached (network quality, large number of tasks, files large than 2GB, etc.)
- Management of users (or VO) is not easy
- Abstract Interface definitions not fully exploited (custom formats)
- A lot of room for more flexible application integration (restarting workflows with changed parameters from the middle)
- Prototype is working and can be used for process automation (more testing, …)
- Different expectations lead to misunderstandings
- Interdisciplinarity – there‘s much to be learnt from each other
- …

General target areas and interests today

- Drug discovery,
- Chemical design,
- Material design (nanomaterials),
- Molecular modelling applications in Life Sciences,
- Problems and tasks where the time factor in decision making support is critical

OpenMolGRID
Open Computing GRID for Molecular Science and Engineering

Chemical applications in the grid

Middleware	Software application	Grid application framework	Reference
Globus	DOCK	VLAB, Nimrod/G	[1, 2]
	Gamess	Nimrod/G	[3]
	Autodock	WISDOM	[4]
	FLEXX	WISDOM	[4]
	Gaussian98	QC Grid	[5]
	WIEN2k	ASKALON, CoG	[6]
	NAMD	BioCoRe	[7]
GridMP	THINK	Screensaver Lifesaver project	[8]
	LigandFit	Screensaver Lifesaver project	[8]
Entropia	Autodock	AIDS@Home	[9]
Condor	MOPAC 2003	WWMM	[10]
UNICORE	CPMD		[11]
	Gaussian98	BioGRID	[12]
	Gamess	BioGRID	[12]
	Amber	BioGRID	[12]
	PDB database	BioGRID	[12]
	Entrez database	BioGRID	[12]
	MOLGEO	OpenMolGRID	[13]
	MOPAC 7	OpenMolGRID	[13]
	CODESSA Pro/MDC	OpenMolGRID	[13]
	CODESSA Pro/MDA	OpenMolGRID	[13]
	NTP database	OpenMolGRID	[13, 14]
	ECOTOX database	OpenMolGRID	[13, 14]

Sulev Sild, Uko Maran, Andre Lomaka, Mati Karelson
Open Computing Grid for Molecular Science and Engineering. *J. Chem. Inf. Model.* (submitted)

The END

www.openmolgrid.org

Thank you!