OSDC OPEN SCIENCE DATA CLOUD

MoSGrid

SPONSORED BY THE Federal Ministry of Education and Research

grant: 01IG09006

MoSGrid – Molecular Simulations in a Distributed Environment

Sandra Gesing

sandra.gesing@uni-tuebingen.de

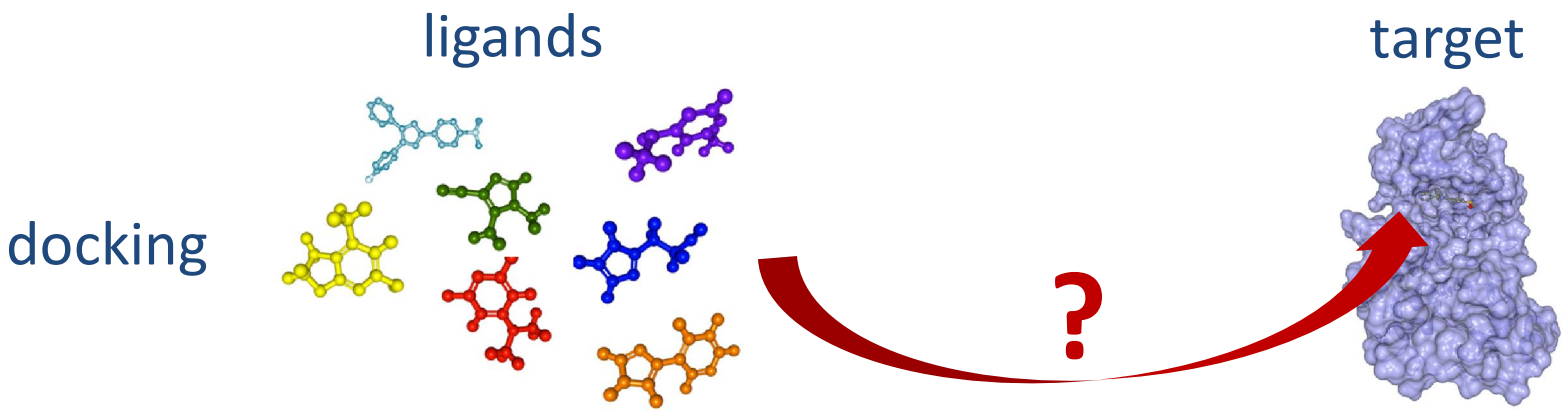
18 June 2013

edinburgh data-intensive research

Molecular Simulations and Docking

edinburgh data-intensive research

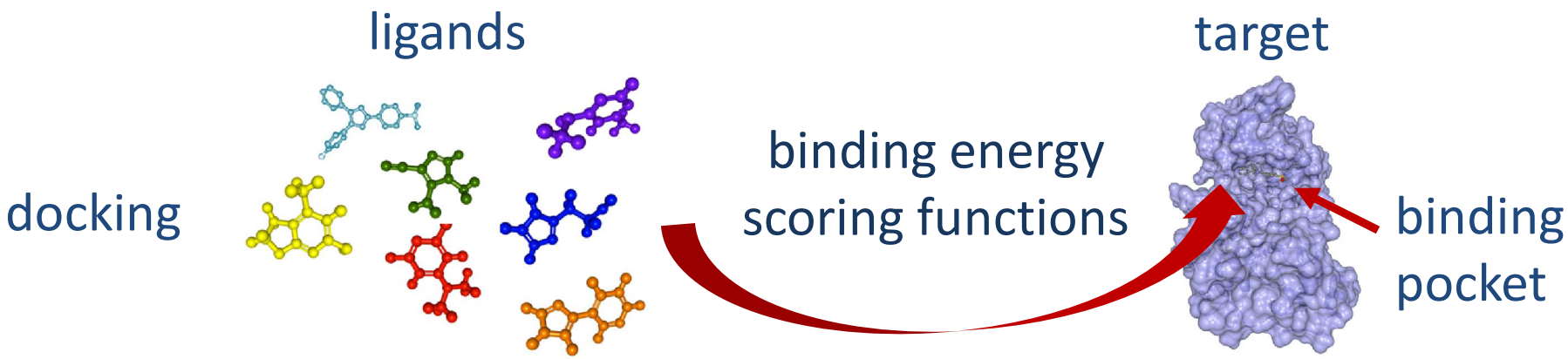
- Prediction and analysis of molecular structures
- Support by sophisticated tools and methods
- Numerous applications, e.g.
  - Materials science
  - Drug design



Molecular Simulations and Docking

edinburgh data-intensive research

- Prediction and analysis of molecular structure
- Support by sophisticated tools and methods
- Numerous applications, e.g.
  - Materials science
  - Drug design



Molecular Simulations and Docking

edinburgh data-intensive research

- Data intensive and compute intensive
- Structure databases,
  e.g., ZINC with ~20 mio structures
- Sensitive and „expensive“ data
- Distributed data management available
- DCIs (Distributed Computing Infrastructures) available

Why do researchers not use the distributed environments on a large scale?

Open Issues

edinburgh data-intensive research

- Usability of tools often limited
- Complexity of methods
- Lack of graphical user interfaces

Open Issues

edinburgh data-intensive research

- Usability of tools often limited
- Complexity of methods

```
|=====================================|
|                                     |
| Version: 1.1                        |
| build date: Jan 10 2012             |
| execution host: vomitoxin           |
| execution time: 2012-09-09, 16:39:43 (MST) |
 -------------------------------------

Available parameters are ('*' indicates mandatory parameters):
   *  -i <in.file>            input molecule file
   *  -o <out.file>           output file
      -ef <double>            error fraction; print error if fraction of invalid mols is larger
      -write_par <out.file>   write xml parameter file for this tool
      -par <in.file>          read parameters from parameter-xml-file

Available flags are:
      -ri      remove invalid molecules.
      -ut      check for unique topologies
      -nc      no not check for unique conformations
      -rm      remove input file when finished
      -help    show help about parameters and flags of this program


This tool checks all molecules of the given input file for errors. Supported formats are mol2, sdf or drf (DockResultFile, xml-based).

The following checks are done for each molecule:

    * bond-lengths may not be completely senseless (i.e. <0.7 or >2.5 Angstroem)
    * each 'molecule' in the input file may only contain one actual molecule, i.e. there may be no unconnected atoms or fragments.
    * each atom must have a valid assigned element
    * the molecule must be protonated (since this is necessary for docking/(re-)scoring).
    * 3D coordinates must be present (instead of 2D coordinates; also necessary for docking/(re-)scoring)
    * partial charges may not contain completely senseless values (>5 or <-5).
    * each conformation should appear only once within the given file, otherwise it is rejected and not written to the output file. However, if option '-ut' is used,
 molecules will instead be checked for unique topologies.

If option '-ri' is used, only those molecules that pass all those tests are written to the output file. If this option is not used, all molecules are written to outp
ut containing a property 'score_ligcheck' with a value of 1 if the molecule passed all tests or with a value of 0 if it did not pass them.


sshgw-bs[13]
```

Open Issues

edinburgh data-intensive research

- Usability of tools often limited
- Complexity of methods
- Lack of graphical user interfaces
- Workflows

a sequence of connected steps in a defined order based on their control and data dependencies

Slide copied from: Stuart Owen „Workflows with Taverna“

Open Issues

edinburgh data-intensive research

- Usability of tools often limited
- Con
- Lack
- Wo

Slide copied from: Stuart Owen „Workflows with Taverna“

Open Issues

edinburgh data-intensive research

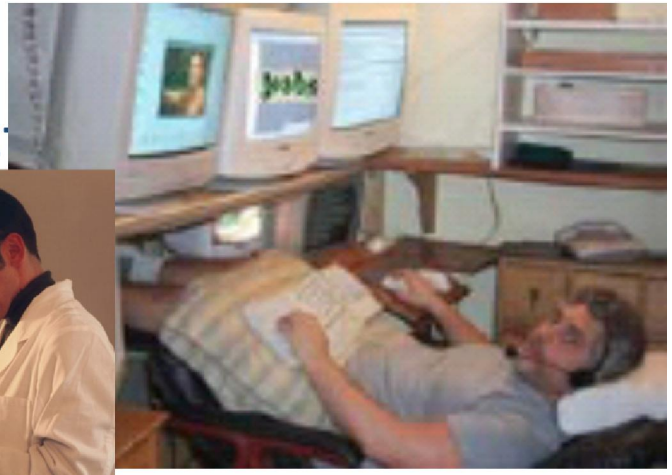
- Usability of tools often limited
- Complexity of methods
- Lack of graphical user interfaces
- Workflows
- Complexity of infrastructures
- Users are generally not IT specialists

Open Issues

edinburgh data-intensive research

- Usability of tools often limited
- Complexity of methods
- Lack of graphical user interfaces
- Workflows
- Complexity of infrastructures

not IT specialist



Open Issues

edinburgh data-intensive research

- Usability of tools often limited
- Complexity of methods
- Lack of graphical user interfaces
- Workflows
- Complexity of infrastructures
- Users are generally not IT specialists

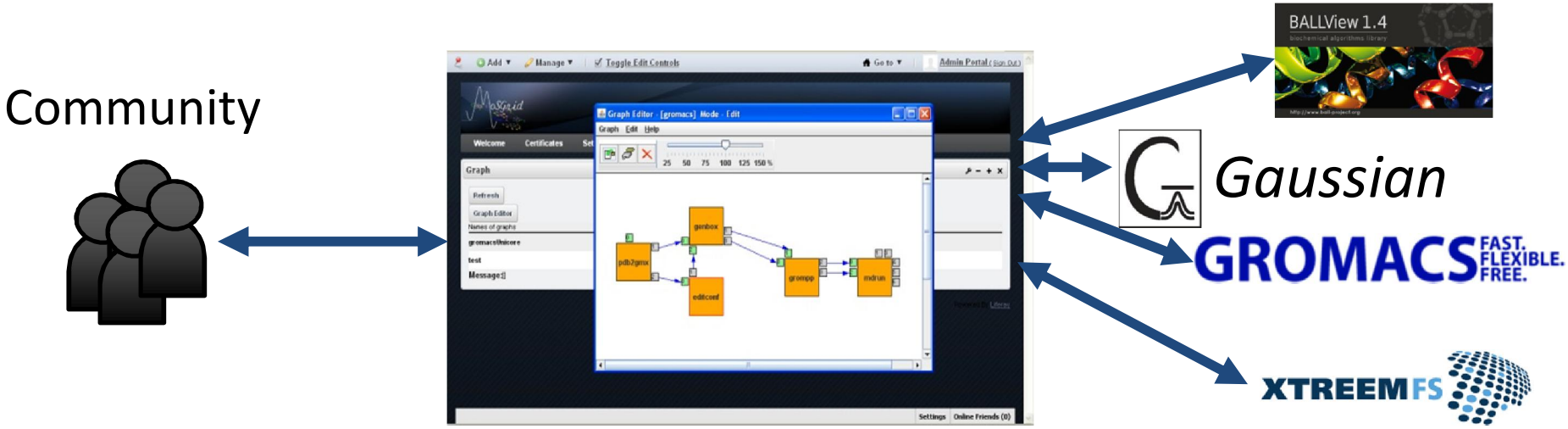
⇒ **User interfaces need to be intuitive and self-explanatory**

⇒ **Science gateways**

Science Gateways

"A Science Gateway is a community-developed set of tools, applications, and data that is integrated via a portal or a suite of applications, usually in a graphical user interface, that is further customized to meet the needs of a specific community."

TeraGrid/XSEDE



Goal of Science Gateways

edinburgh data-intensive research

Usability of software and data

"After all, usability really just means that making sure that something works well: that a person … can use the thing - whether it's a Web site, a fighter jet, or a revolving door - for its intended purpose without getting hopelessly frustrated."

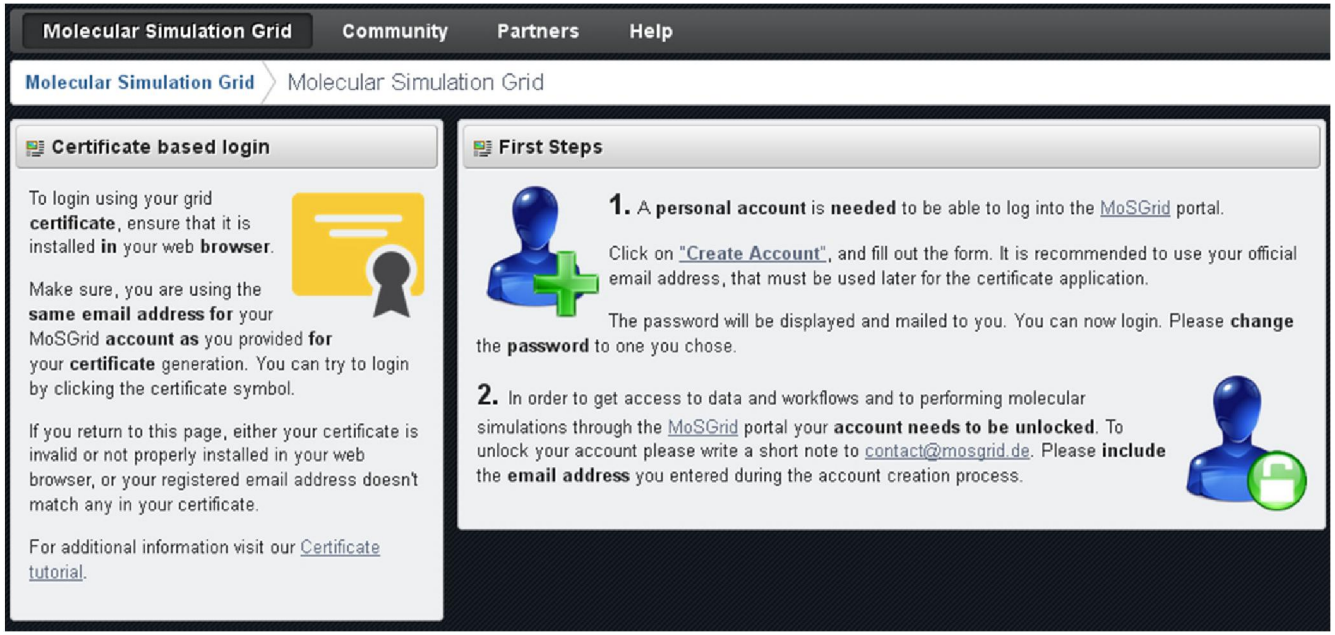
(Steve Krug in *"Don't make me think!: A Common Sense Approach to Web Usability"*, 2005)

MoSGrid

edinburgh data-intensive research

Molecular Simulation Grid

- Science gateway integrated with underlying compute and data management infrastructure
- Distributed workflow management
- Data repository
- Open source



MoSGrid

edinburgh data-intensive research

Molecular Simulation Grid

- Science gateway integrated with underlying compute and data management infrastructure
- Distributed workflow management
- Data repository
- Open source

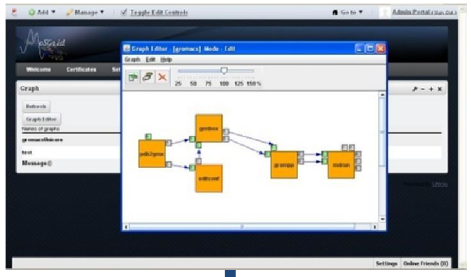
Survey of willingness to share knowledge in the community

⇒ 90% share workflows

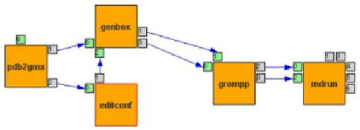
⇒ 70% share results after publication

MoSGrid in a Nutshell

edinburgh data-intensive research



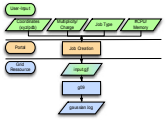
Portal
WS-PGRADE



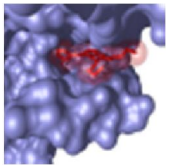
Workflow

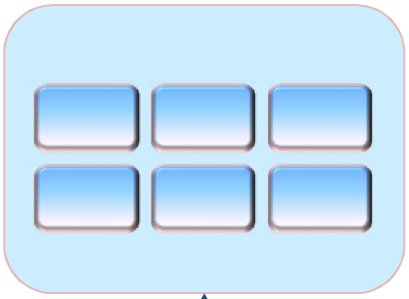
Structure

Recipe



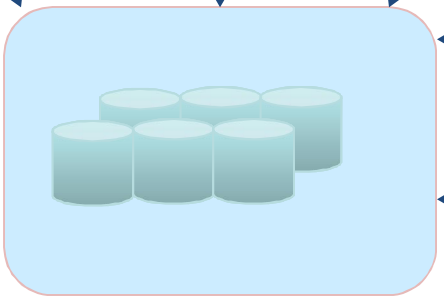
Result



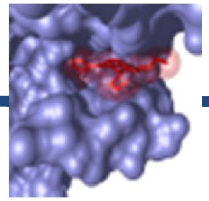


High-level middleware service level
gUSE

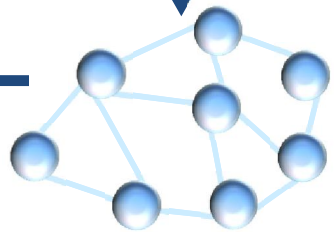
Distributed file system
XtreemFS



Data repositories



Result

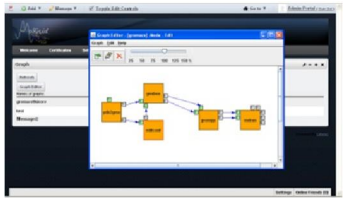


DCIs
UNICORE 6

Job and Workflow Management

edinburgh data-intensive research

grid User Support Environment

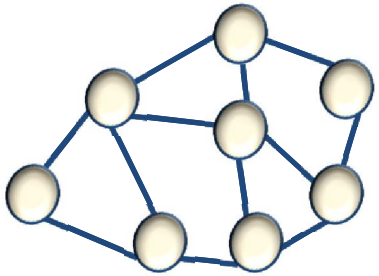


User Interface
WS-PGRADE

Workflow storage | Application repository | Information system

Workflow engine | Submitters | Logging

High-Level Middleware Service Layer
gUSE

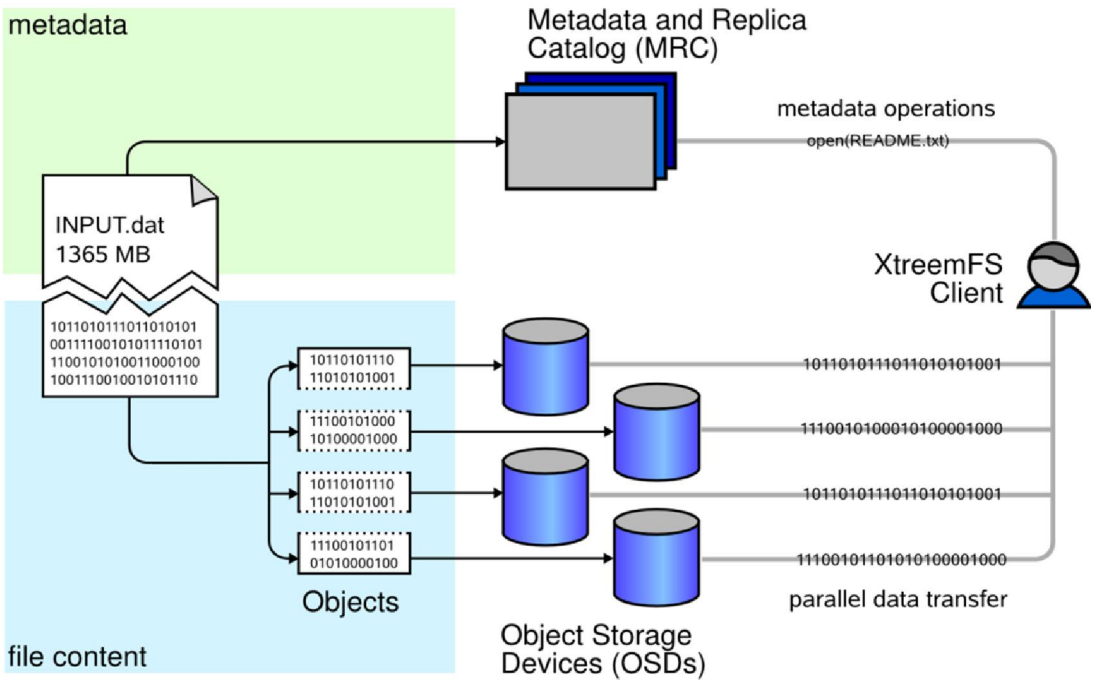


DCI Resources Middleware Layer
UNICORE 6

Distributed Data Management

edinburgh data-intensive research

- XtreemFS is an object-based grid and cloud filesystem
- Replication for availability, locality, bandwidth, and latency
- Easy integration in heterogenous environments
- UNICORE extension with URL scheme xtreemfs:// available



Data Repository

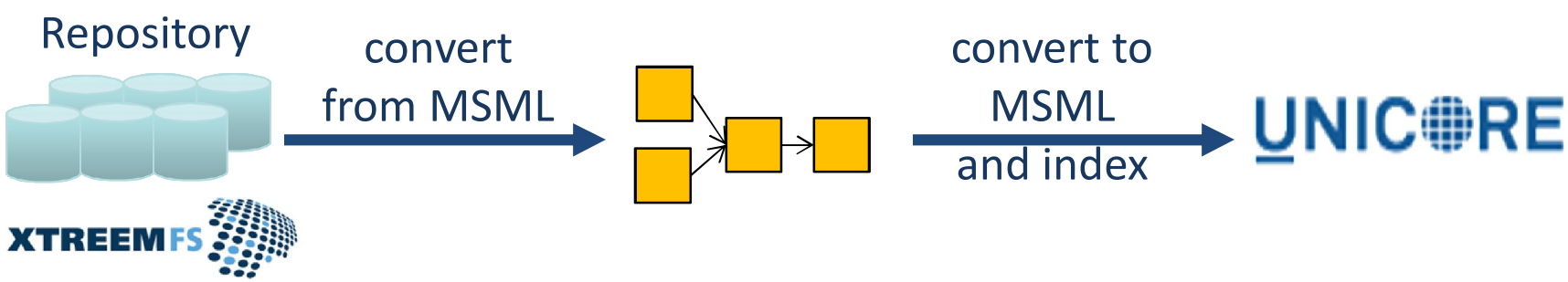
edinburgh data-intensive research

- Repository consists of data and metadata storage
- MSML (Molecular Simulation Mark-up Language)
- Subset and extension of CML (Chemical Mark-up Language)
- Unified data representation
- Used for storing structures of molecules and macromolecules, simulation descriptions, and results
- Parsers and adapters used for conversions to and from MSML

Repository – Data Management

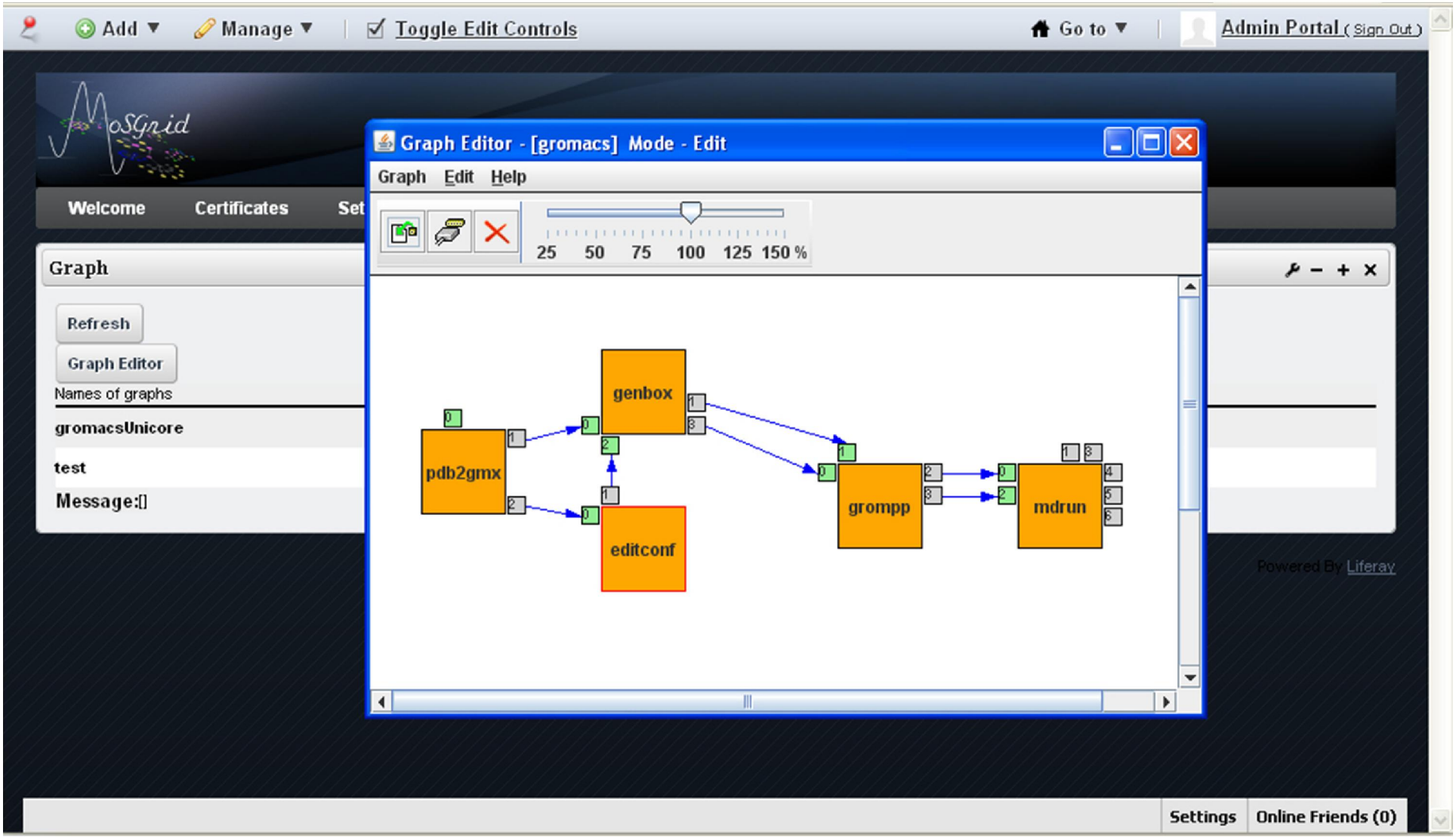
edinburgh data-intensive research

- First step in workflows: MSML to application specific input
- Results are computed
- Output converted to MSML
- Last step in workflows: MSML to JSON metadata (metadata extractor in UNICORE)
- JSON indexed for searchable results by UNICORE and LUCENE (text search engine API)



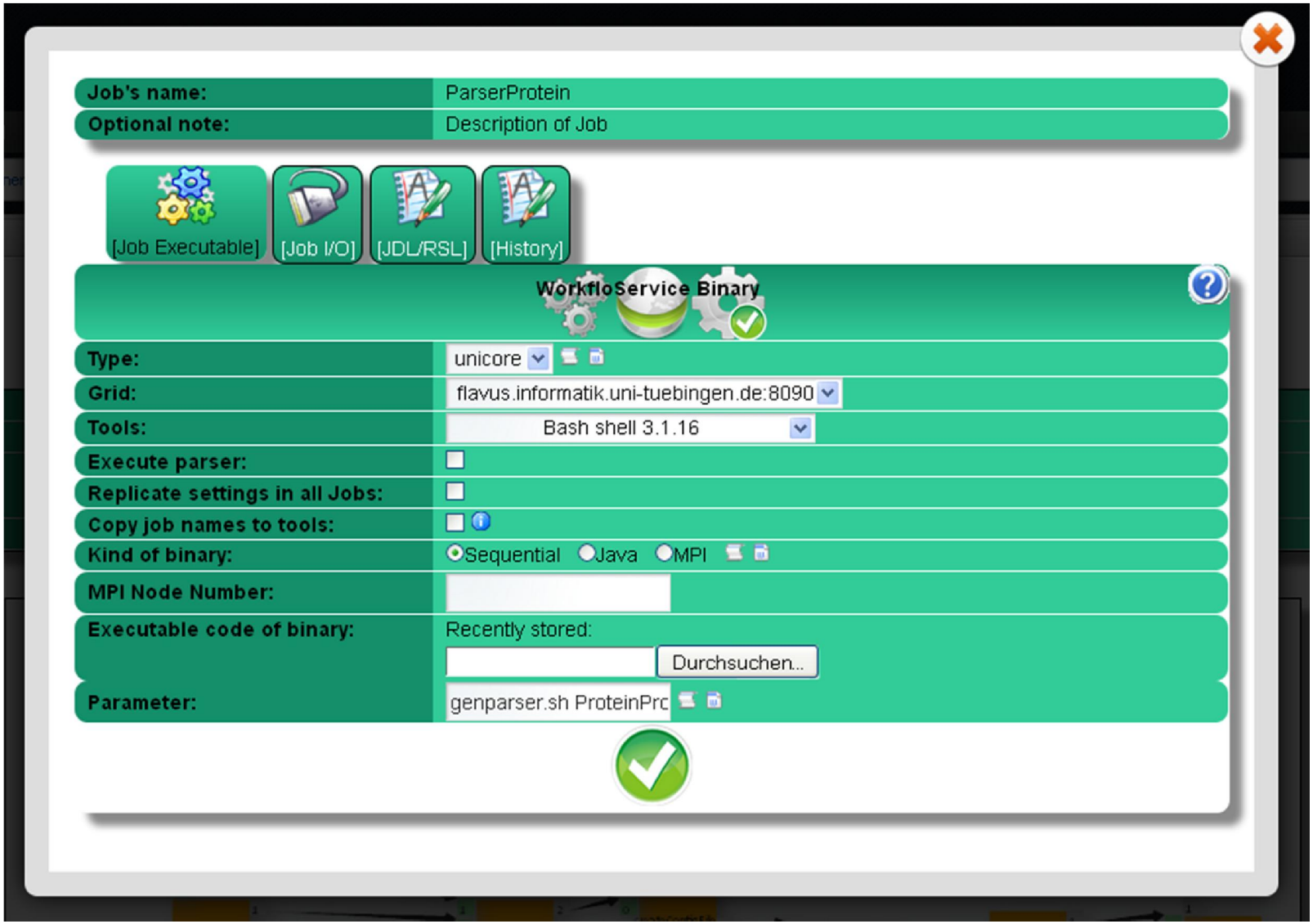
WS-PGRADE

edinburgh data-intensive research



Job Configuration

edinburgh data-intensive research



Job Configuration

edinburgh data-intensive research

Job's name: PDBCutter
Optional note: Description of Job

[Job Executable] [Job I/O] [JDL/RSL] [History]

Workflow Service Binary

Type: unicore
Grid: flavus.informatik.uni-tuebingen.de:8090
Tools: PDBCutter 1.0.0
Execute parser:
Replicate settings in all Jobs:
Copy job names to tools:
Kind of binary:
MPI Node Number:
Executable code of binary:
Parameter:

ModelCreator 1.0.0
MolCombine 1.0.0
MolDepict 1.0.0
MolFilter 1.0.0
MolPredictor 1.0.0
nwchem 6.1
obabel (OpenBabel) 2.3.1
PartialChargesCopy 1.0.0
pdb2gmx 4.5.5
PDBCutter 1.0.0
PDBDownload 1.0.0
Perl 5.8.8
PocketDetector 1.0.0
POVRay 3.5
Predictor 1.0.0
PropertyModifier 1.0.0
PropertyPlotter 1.0.0
ProteinCheck 1.0.0
ProteinProtonator 1.0.0
Python Script 2.4.2

Remote File Configuration

edinburgh data-intensive research

Domain-specific Workflows

edinburgh data-intensive research

Molecular Dynamics

- Study and simulation of molecular motion

Quantum Chemistry

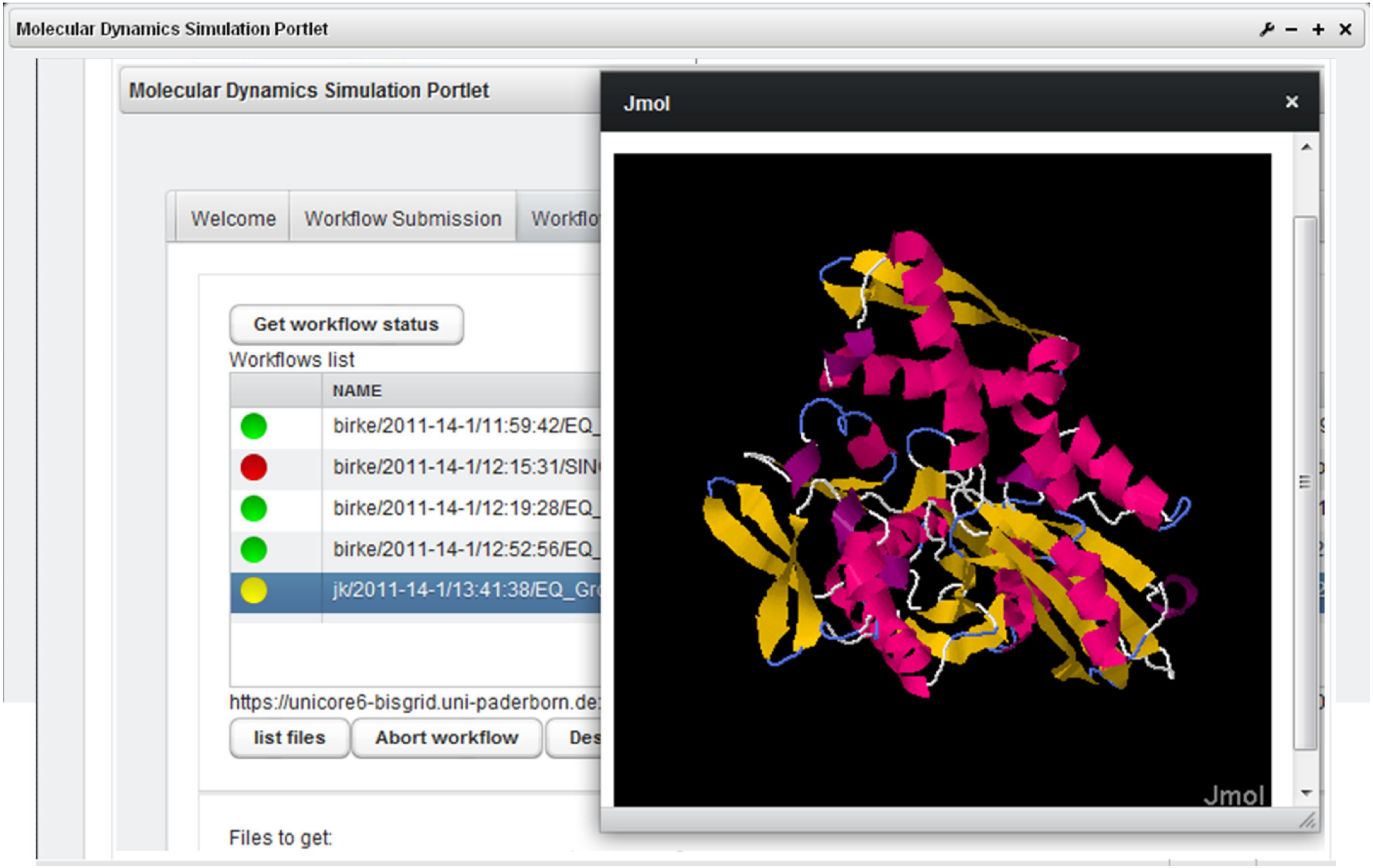
- Study and simulation of molecular electronic behavior relative to their chemical reactivity

Docking

- Main focus on evaluation of ligand-receptor interactions (e.g., for drug design)

MD Portlet

edinburgh data-intensive research



QC Portlet

edinburgh data-intensive research

Quantum C

Job View | Results View

fetch/refresh file list

Workflow Results

FILENAME	DOWNLOAD FILE
t6.log	Download
t7.log	Download
t6-thermo.txt	Download
t7-freq.csv	
t7-thermo.t	
t6-freq.csv	

Extra Files

FILENAME

Geo

Recipe

Rese

Title

```
  4        1        0        -2
  5        6        0        -1
  6        1        0        -2
  7        6        0         0
  8        1        0         0
  9        6        0         1
 10        1        0         2
 11        6        0         1
 12        1        0         2
-----------------------------------
                Distance matrix (ar
                1          2
  1  C    0.000000
  2  H    1.087570   0.000000
  3  C    1.402719   2.162403   0.00
  4  H    2.162409   2.490292   1.08
  5  C    2.429582   3.415017   1.40
  6  H    3.415025   4.313316   2.16
  7  C    2.805440   3.893010   2.42
  8  H    3.893010   4.980580   3.41
  9  C    2.429582   3.415017   2.805438   3.893016   2.429580
 10  H    3.415025   4.313316   3.893016   4.980595   3.415023
 11  C    1.402719   2.162403   2.429580   3.415023   2.805438
```

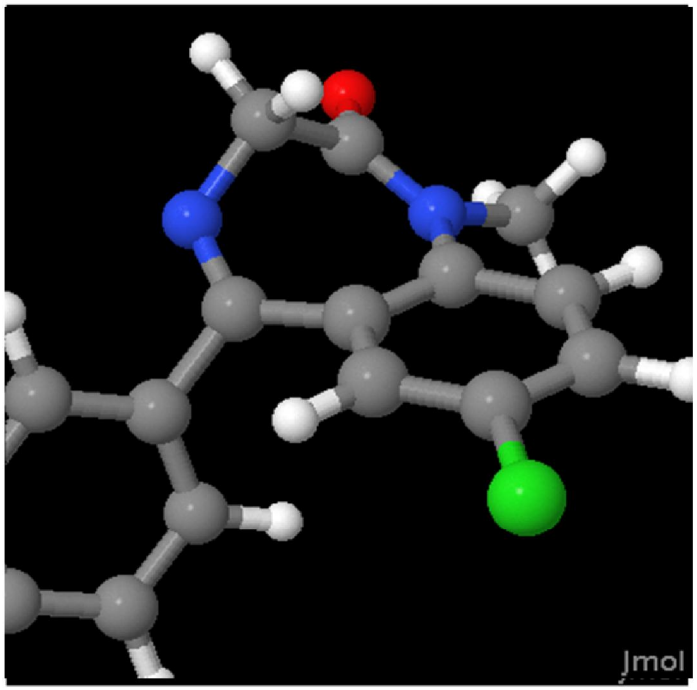
Mozilla Firefox

Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Hilfe

http://localhost:8

Meistbesuchte Se... Erste Schritte Aktuelle Nachrich...

http://localhost:...467/result/t6.log



Jmol script terminated 501x562

Docking Portlet

edinburgh data-intensive research

Docking-Portlet

Welcome | Monitoring | Debug

Workflows:

- TEST20111025
- TEST20111025
- 1EVE
  - Docking
    - recH.pdb
    - results.sorted.sdf

1EVE/ STATUS: FINISHED / selected

```
ATOM      1  N    SER A
ATOM      2  CA   SER A
ATOM      3  C    SER A
ATOM      4  O    SER A
ATOM      5  CB   SER A
ATOM      6  OG   SER A
ATOM      7  HN1  SER A
ATOM      8  HN2  SER A
ATOM      9  HN3  SER A
ATOM     10  HA   SER A
ATOM     11  HB1  SER A
ATOM     12  HB2  SER A
ATOM     13  HG   SER A
ATOM     14  N    GLU A
ATOM     15  CA   GLU A
ATOM     16  C    GLU A
ATOM     17  O    GLU A
ATOM     18  CB   GLU A
ATOM     19  CG   GLU A
ATOM     20  CD   GLU A
ATOM     21  OE1  GLU A
ATOM     22  OE2  GLU A
ATOM     23  HN   GLU A
ATOM     24  HA   GLU A
ATOM     25  HB1  GLU A
ATOM     26  HB2  GLU A
ATOM     27  HG1  GLU A
ATOM     28  HG2  GLU A   5      -7.010  86.637  39.078  1.00  0.00          H
```

Update | Delete | Download | View in Jmol

Jmol

1320400625949recH.pdb

The MoSGrid Science Gateway

edinburgh data-intensive research

Built by 17 institutes and companies

TECHNISCHE UNIVERSITÄT DRESDEN

SCHRÖDINGER

MTA SZTAKI

origenis

ZIB

BioSolve IT

technische universität dortmund

UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

ORACLE

COSMOlogic

Bayer Technology Services

GETLIG & TAR

JÜLICH

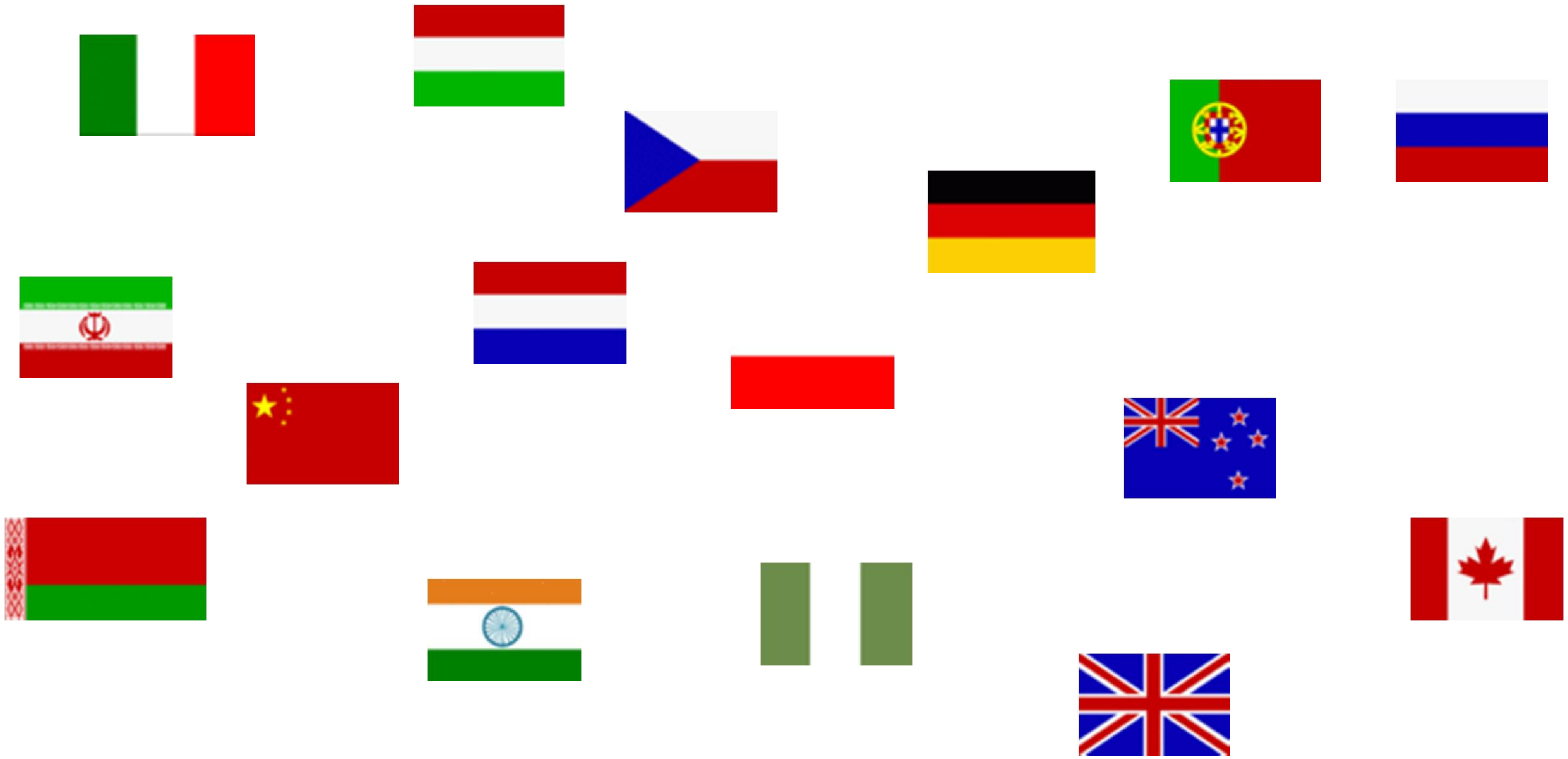
PC² PADERBORN CENTER FOR PARALLEL COMPUTING

rrzk

The MoSGrid Science Gateway

edinburgh data-intensive research

Used by 125 user groups in 16 countries

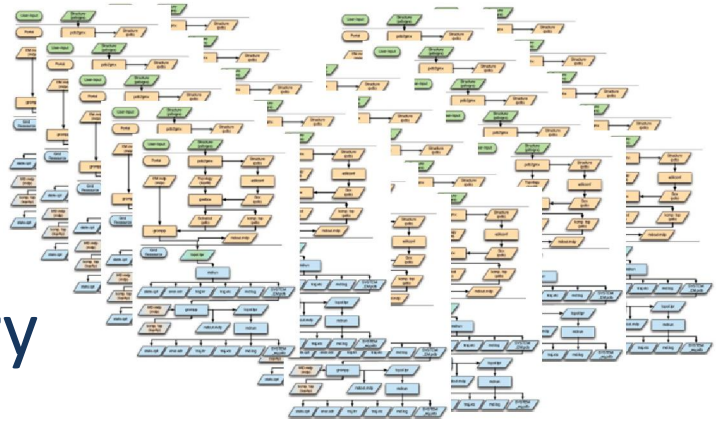


The MoSGrid Science Gateway

edinburgh data-intensive research

Currently supports

- 65 workflows in repositories
- 90 applications
- 85 GB data in central repository
- ~50,000 structures (1 – 5 GB) in user repositories



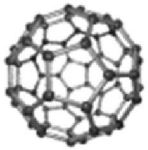
NWCHEM HIGH-PERFORMANCE COMPUTATIONAL CHEMISTRY SOFTWARE

AutoDock Vina

GROMACS FAST. FLEXIBLE. FREE.

Gaussian

FlexX



CADDSuite

Sustainability

edinburgh data-intensive research

MoSGrid ended 31.12.2012 but partners participate in

SCI-BUS (SCIentific gateway Based User Support)

- EU project 01.10.2011 – 30.09.2014
- Extension of the MoSGrid portal with an interactive molecule editor based on WebGL and a semantic search

ER-flow (Building an European Research Community through Interoperable Workflows and Data)

- EU project 01.10.2012 – 30.09.2014
- Integration of applications in SHIWA simulation platform
- Study of data exchange between workflow systems
- Community management

sandra.gesing@uni-tuebingen.de