Grid-enabled drug discovery to address neglected diseases

Dr. Marc Zimmermann
Department of Bioinformatics
Fraunhofer Institute for Algorithms and Scientific Computing (SCAI)

Nicolas Jacq
Computing Platform for Life Science
Corpuscular Physics Laboratory of Clermont-Ferrand
CNRS/IN2P3

WISDOM : Wide In Silico Docking On Malaria

Scientific objectives

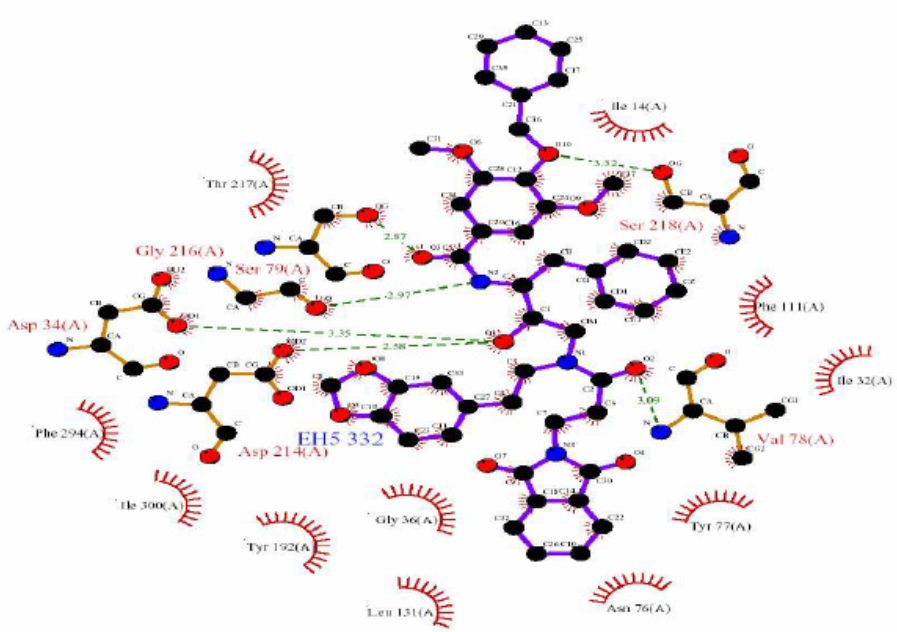
Demonstrate to the research communities active in the area of drug discovery the relevance of grid infrastructures

Deployment of a CPU consuming application generating large data flows to test the grid infrastructure and services of the BioMed VO.

Method

- Large scale molecular docking on malaria target
  to test million of compounds with
  several docking softwares.

- Docking is about scoring the potential binding
  of a protein target to a library of
  compounds



Goals of the Data Challenge (DC) on *in silico* drug discovery

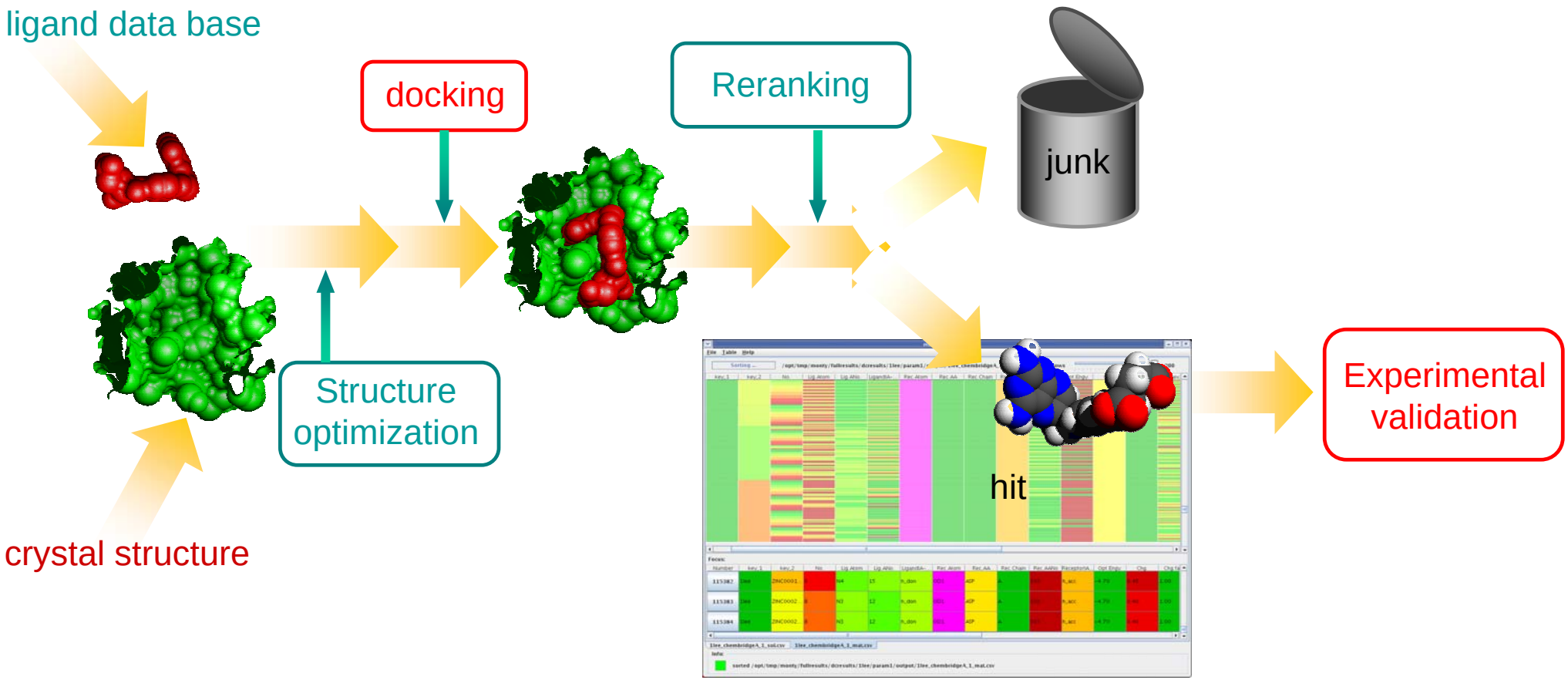
Biological Goal : proposition of new drug candidates addressed to neglected diseases

- The target : *Plasmepsin* is a promising aspartic protease target involved in the hemoglobin degradation of *P. falciparum*. 5 different structures are prepared
- The compounds database : *ZINC* is an open source library of 3,3 millions selected compounds. They are made available by chemistry companies

Biomedical informatics goal : deployment of *in silico* virtual screening on the grid

- The software : *Autodock* is an open source algorithm, *FlexX* is a commercial algorithm available for this data challenge during 3 weeks

Dataflow and workflow in a virtual screening



Success with virtual screening

Dihydrofolate reductase inhibitor (1992)

HIV-protease (1992)

Phospholypase A2 (1994)

FKBP-12 (1995)

Thrombine (1996)

Abl-SH3 (1996)

Trypsine, streptavidine, phosphrylase nucleotid

ex : virtual screening of the trypsine, 2h of computing : 153 compounds, 2 inhibitors

Overview on neglected diseases

Infectious diseases kill 14 million people each year, more than 90% of whom are in the developing world.

Access to treatment is problematic

- the medicines are unaffordable,
- some have become ineffective due to drug resistance,
- and others are not appropriately adapted to specific local conditions and constraints.

Drug discovery for a neglected disease

Lack of ongoing or well coordinated R&D

- Research often takes place in university or government labs
- Development is almost exclusively done by the pharmaceutical and biotech industry
- Critical point is the launching of clinical trials for promising candidate drugs.

Producing more drugs for neglected diseases requires

- building a focussed, disease-specific R&D agenda including short-, mid- and long-term projects.
- a public-private partnership through efficient, secure and trusted collaborations that aim to improve access to drugs and stimulate discovery of easy-to-use, affordable, effective drugs.

Virtual organisation

Motivate and gather together :

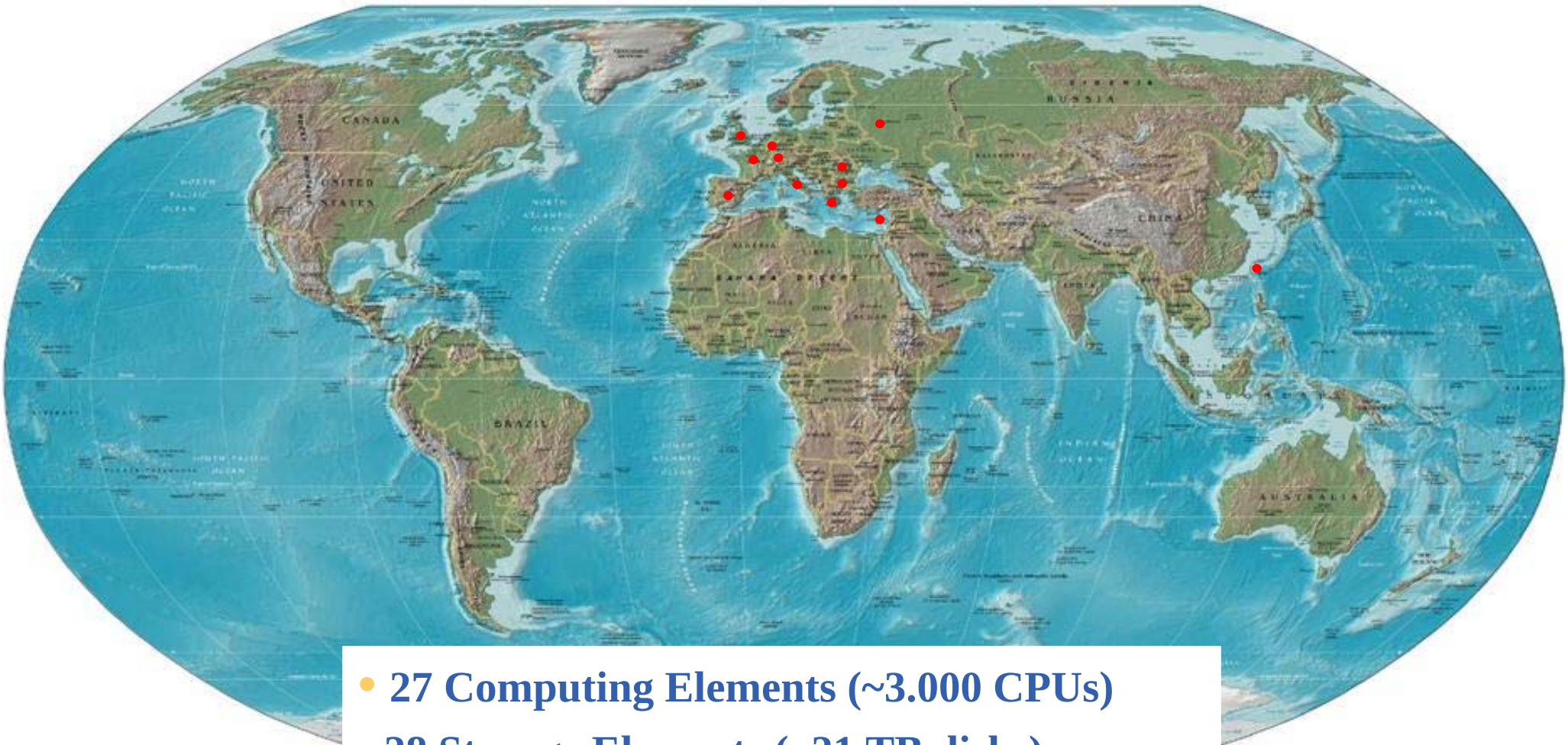
- drug designers to identify new targets and drugs
- healthcare centers involved in clinical tests and collecting patent information
- healthcare centers collecting patent information
- organizations involved in distributing existing treatments
- informatics technology developers
- computing and computer science centers
- biomedical laboratories working on vaccines, genomes of the virus and/or the parasite and/or the parasite vector

Collaborative environment

A PharmaGrid will support such processes as:

- search of new drug targets through post-genomics requiring data management and computing
- massive docking to search for new drugs requiring high performance computing and data storage
- handling of clinical tests and patient data requiring data storage and management
- overseeing the distribution of the existing drugs requiring data storage and management
- trusted exchange of IP, possibly auction-mediated

The BioMed VO at a glance



- 27 Computing Elements (~3.000 CPUs)
- 28 Storage Elements (~21 TB disks)
- in 12 countries

Selection criteria of potential drugs

Potential drugs

- 28 million compounds currently known
- Drug company biologists screen up to 1 million compounds against target using ultra-high throughput technology

- Chemists select 50-100 compounds for follow-up
- Chemists work on these compounds, developing new, more potent compounds

- Pharmacologists test compounds for pharmacokinetic and toxicological profiles
- 1-2 compounds are selected as potential drugs

Target

- Well known organism
- Multiple crystal structures
- Multiple bound inhibitors
- Structural similarity between
- multiple species

Inhibitors

- The one more selective
- Acts on multiple targets
- The one with active in low quantities
- Shows good pharmacokinetics properties
- Good pharmacodynamic properties

Data challenge scenario

	Scenario
Duration	28 days
CPU time	11 years CPU
Grid performance	50%
Max number of CPU used	1,008
Number of grid jobs (20h)	12,215
Storage	2*6 TB
Docking workflow description	
Number of software / targets / compounds / parameters settings	2 / 5+3 / 500,000 / 4 = **32 mio dockings**
Objective	Selection of the best hits with short analysis

FlexX running time : 1 mn
F. output size : 1MB
F. job output size : 1.2GB
F. job compressed output size : 250MB

Autodock running time : 2.5 mn
A. output size : 1MB
A. job output size : 0,5GB
A. job compressed output size : 100MB

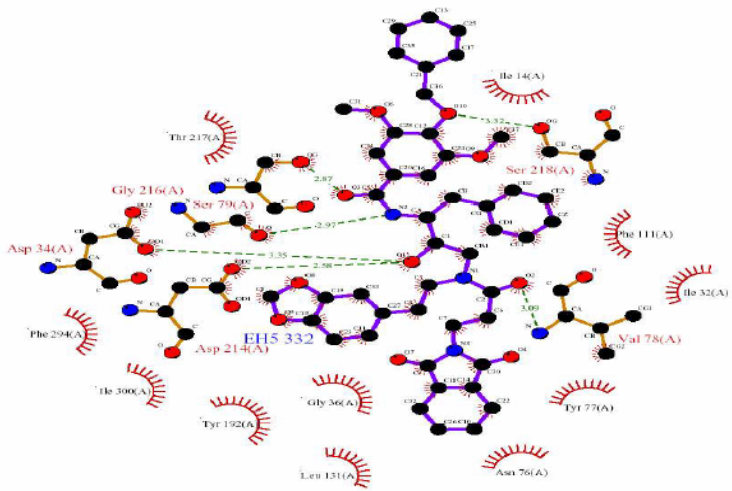
Output analysis

Ranked scoring lists:

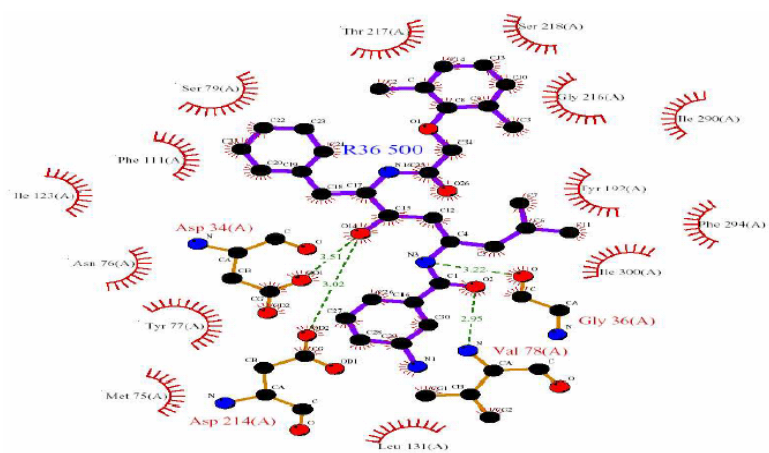
- Sorting and assembly of data from the Grid
- Offline post filtering
- Clustering of similar conformations
- Doing statistics on the score distribution
- Checking pharmacophoric points of each conformation
- Re-ranking for interesting compounds

Comparing:

- Tools
- Parameters
- Targets (and water molecules)
- Ligand scaffolds

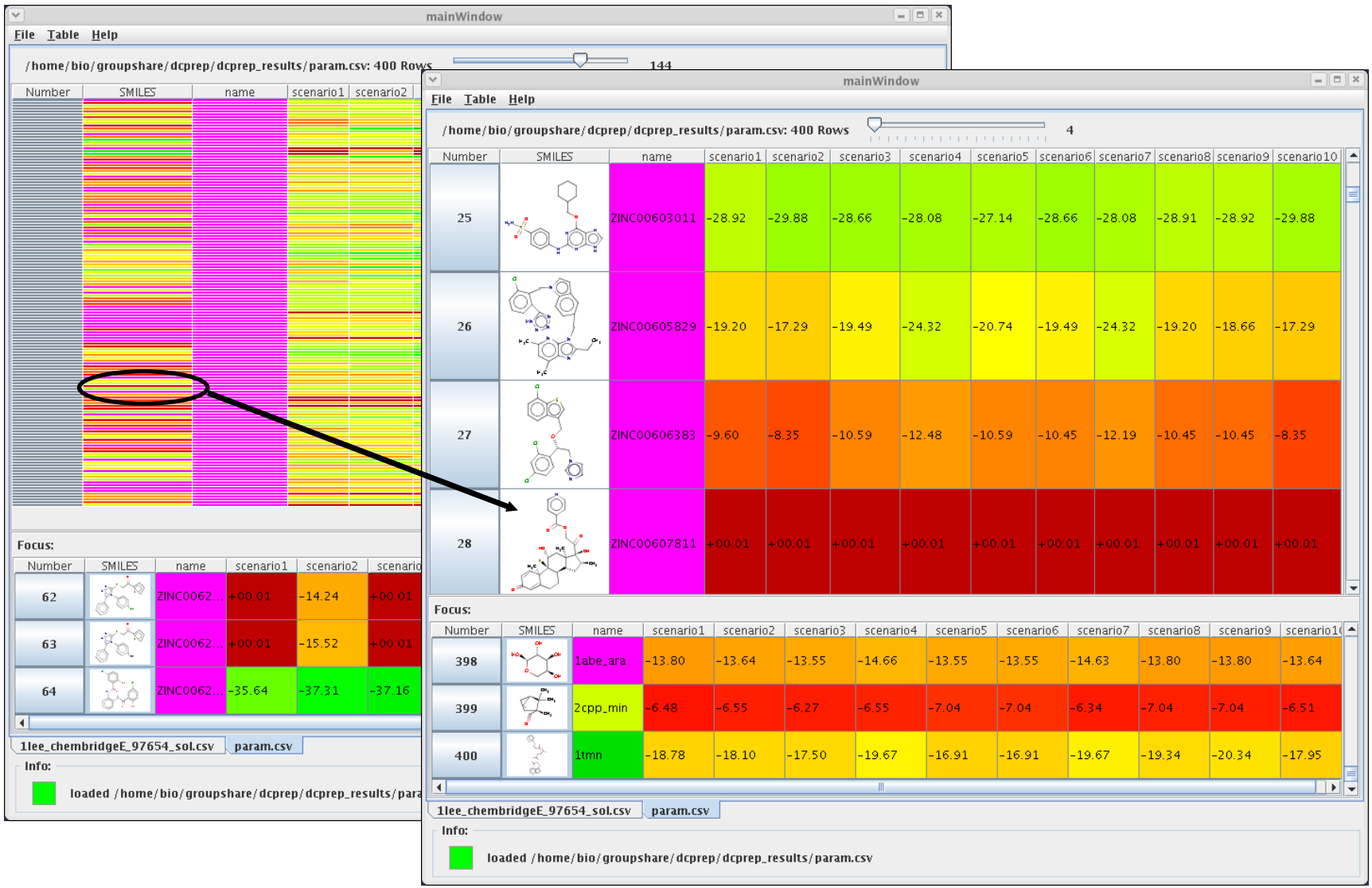


Ligand plot of 1LF3 with inhibitor EH5 332



Ligand plot of 1LEE with inhibitor R36 500

VS Explorer as tool for gridscale ranking lists



Follow-up of the DC

The best hits found by post-treatment will be published and available on a permanent grid storage via a portal

A knowledge space will be progressively build around these results
- to extract and process the most interesting information
- to enrich the data with the results found later by other *in silico* drug discovery processes

Next scenarios will be able to use the gLite middleware

The *in silico* drug discovery will be further extend
- to include more precise molecular dynamics computations using quantum chemistry software like NAMD

Acknowledgement

IN2P3/CNRS

Nicolas Jacq
Jean Salzeman
Vincent Breton

Fraunhofer SCAI

Marc Zimmermann
Astrid Maaß
Horst Schwichtenberg
Martin Hofmann

BioSoveIT

Marcus Gastreich
Holger Claussen