Relevance for Pharma Research

SCAI

Fraunhofer Institute
Algorithms and
Scientific Computing

Prof. Dr. Martin Hofmann-Apitius
Head of the Department of Bioinformatics
Fraunhofer Institute for Algorithms and Scientific Computing (SCAI)
and
Bonn-Aachen International Center for Information Technology (B-IT) /
Friedrich-Wilhelms-University of Bonn

GRID Computing for Pharma R&D

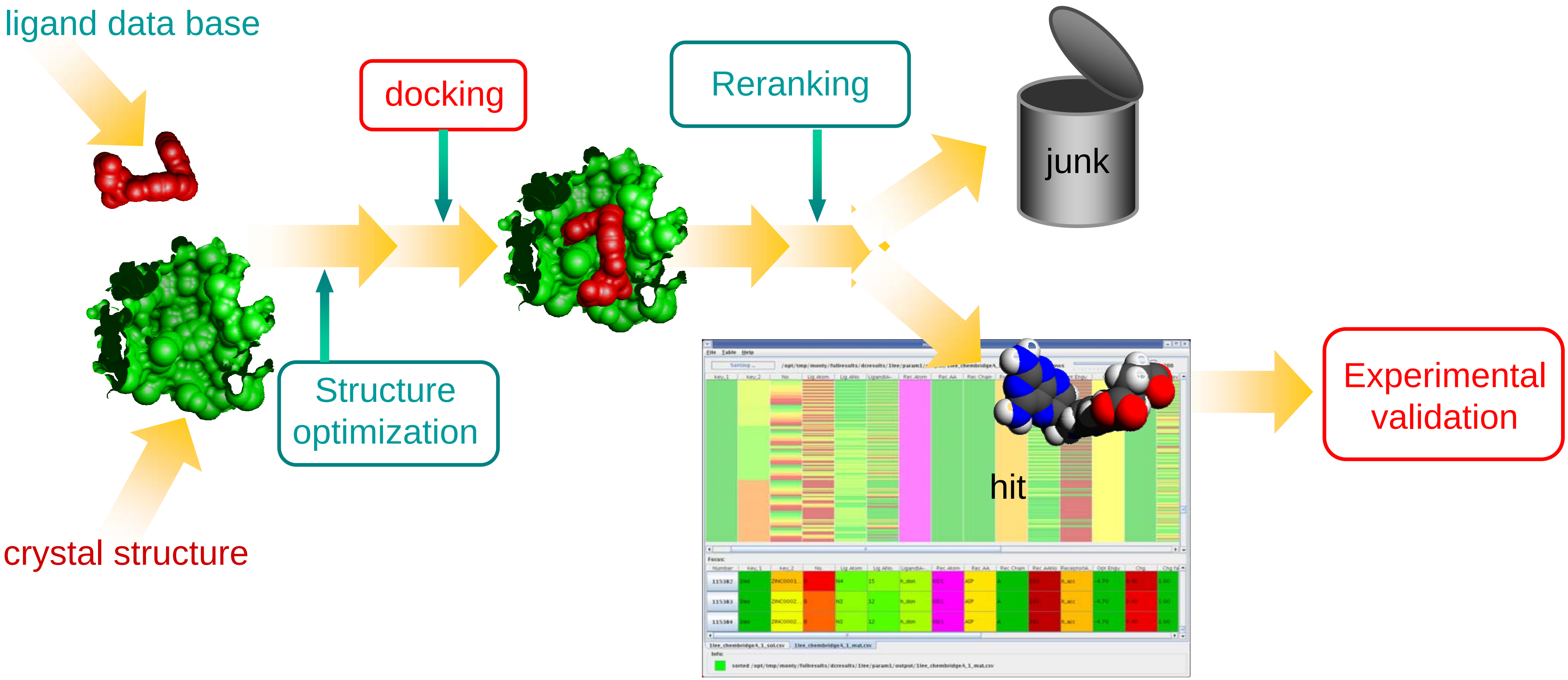
- Distributed Virtual Screening – the WISDOM Example
  - Basics of Structure-based Virtual Screening
  - Docking on the GRID: WISDOM and follow-up
- Text Mining on the GRID – Information Extraction for Scientific & Competitive Intelligence
  - Scientific & Competitive Intelligence
  - Distributed information extraction from scientific literature

Archivierungsangaben

Basics of Structure-based Virtual Screening

Archivierungsangaben

Dataflow and Workflow in a Virtual Screening Experiment



Focus:

Number	Key_1	Key_2	No	Lig Atom	Lig ANo	LigandA-	Rec Atom	Rec AA	Rec Chain	Rec AANo	ReceptorA	Opt Engy	Chg	Chg fa
115382	1lee	ZINC0001...	[illegible]	N4	15	h_don	OD1	ASP	A	[illegible]	h_acc	-4.70	[illegible]	1.00
115383	1lee	ZINC0002...	[illegible]	N3	12	h_don	OD1	ASP	A	[illegible]	h_acc	-4.70	[illegible]	1.00
115384	1lee	ZINC0002...	[illegible]	N3	12	h_don	OD1	ASP	A	[illegible]	h_acc	-4.70	[illegible]	1.00

1lee_chembridgeA_1_sol.csv | 1lee_chembridgeA_1_mat.csv

Info:

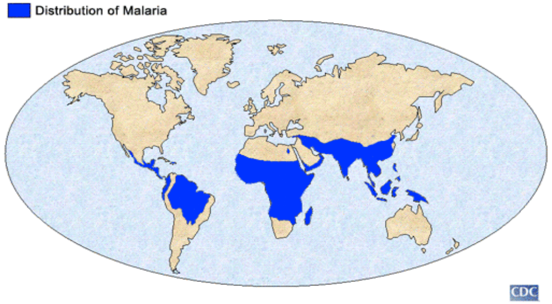
🟩 sorted /opt/tmp/monty/fullresults/dcresults/1lee/param1/output/1lee_chembridgeA_1_mat.csv

Structure – based Virtual Screening on the GRID

- **Not** a new idea
- Simple task farming approach possible
- Routine procedure at Novartis and other pharma front runners
- In EnterpriseGRID – solutions typically based on proprietary middleware platforms (e.g. United Devices (UD); Plattform Computing)
- Success stories available e.g. identification of cyclin dependent kinase inhibitors published by Novartis
- Close interaction between *in silico* and “wet“ laboratory world required

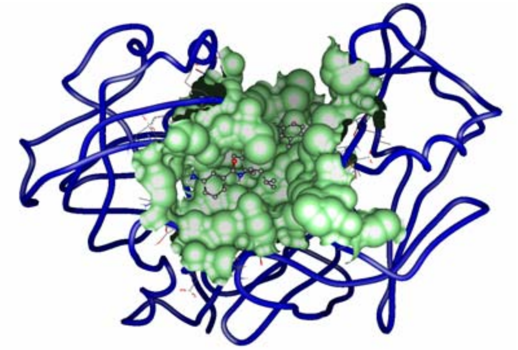
Docking on the GRID: WISDOM and follow-up

WISDOM : Wide In Silico Docking On Malaria



Biological goal

Proposition of new inhibitors for a family of proteins produced by Plasmodium *falciparum*



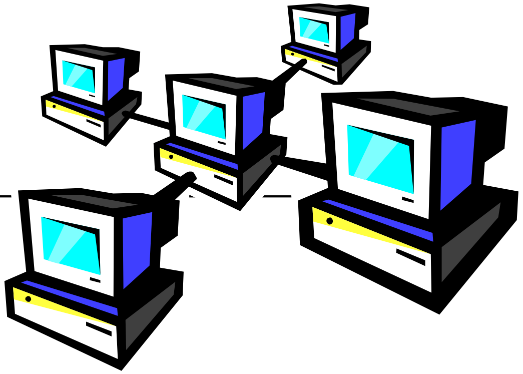
Biomedical informatics goal

Deployment of *in silico* virtual screening on the grid

Grid goal

Deployment of a CPU consuming application generating large data flows to test the grid operation and services

→ "*data challenge*"



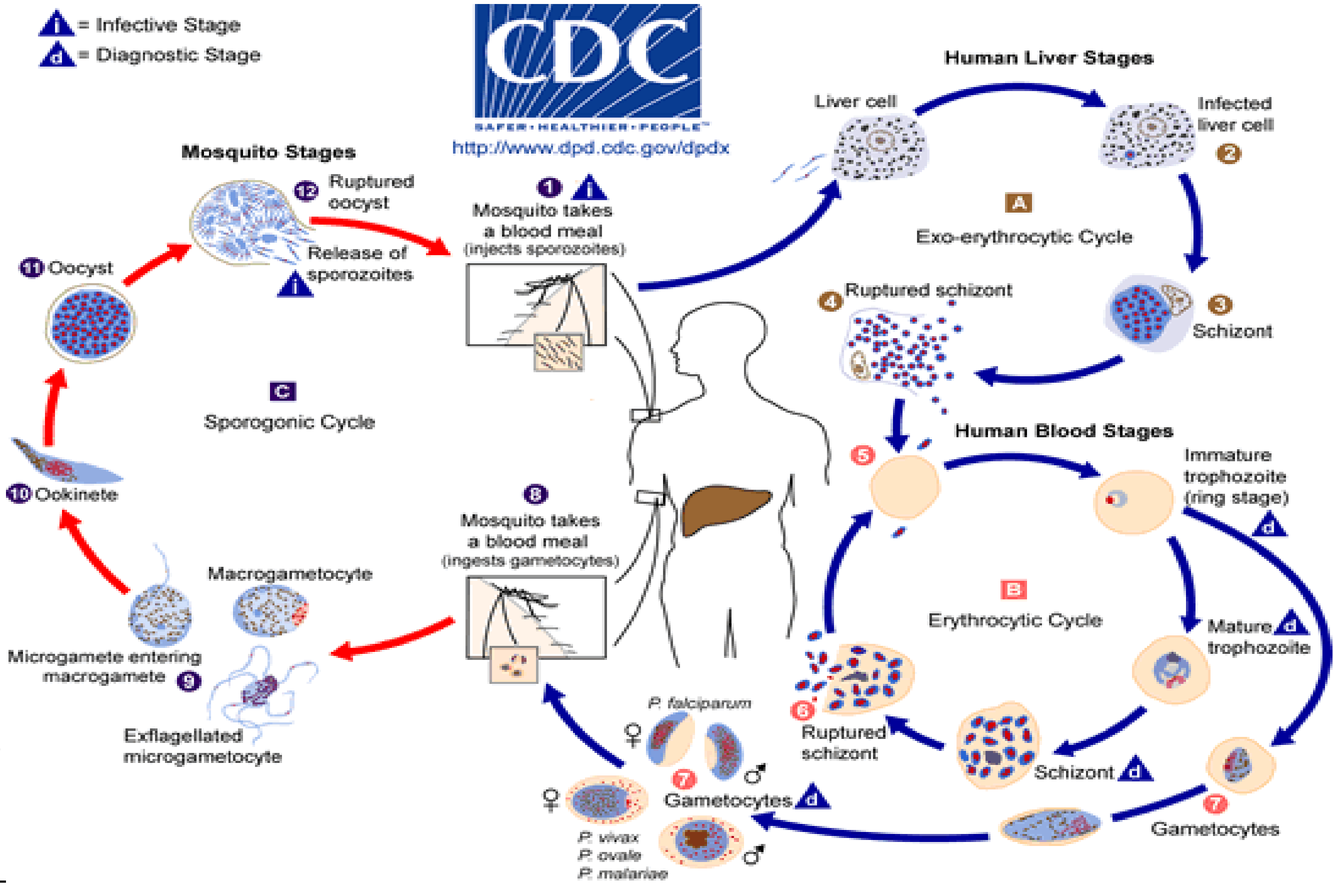
Introduction to the Disease : Malaria

~300 million people worldwide are affected

1-1.5 million people die every year

Widely spread

Caused by protozoan parasites of the genus *Plasmodium*



Archivierungsangaben

Strong Need for New Drugs against Malaria (WHO)

Drug resistance has emerged for all classes of antimalarials except artemisinins.

- ❑ Resistance to chloroquine, the cheapest and the most widely used drug, is spreading in almost all the endemic countries.
- ❑ Resistance to the combination of sulfadoxine-pyrimethamine which was already present in South America and in South-East Asia is now emerging in East Africa

All countries that experience resistance to conventional monotherapies should use ACTs (artemisinin-based combination therapies)

But there is even the threat of resistance to artemisinin too, as it is already observed in murine Plasmodium *yoelii*

Archivierungsangaben

Identification of New *Plasmodium* Targets

There is consensus that substantial scientific effort is needed to identify new targets for anti-malaria drugs

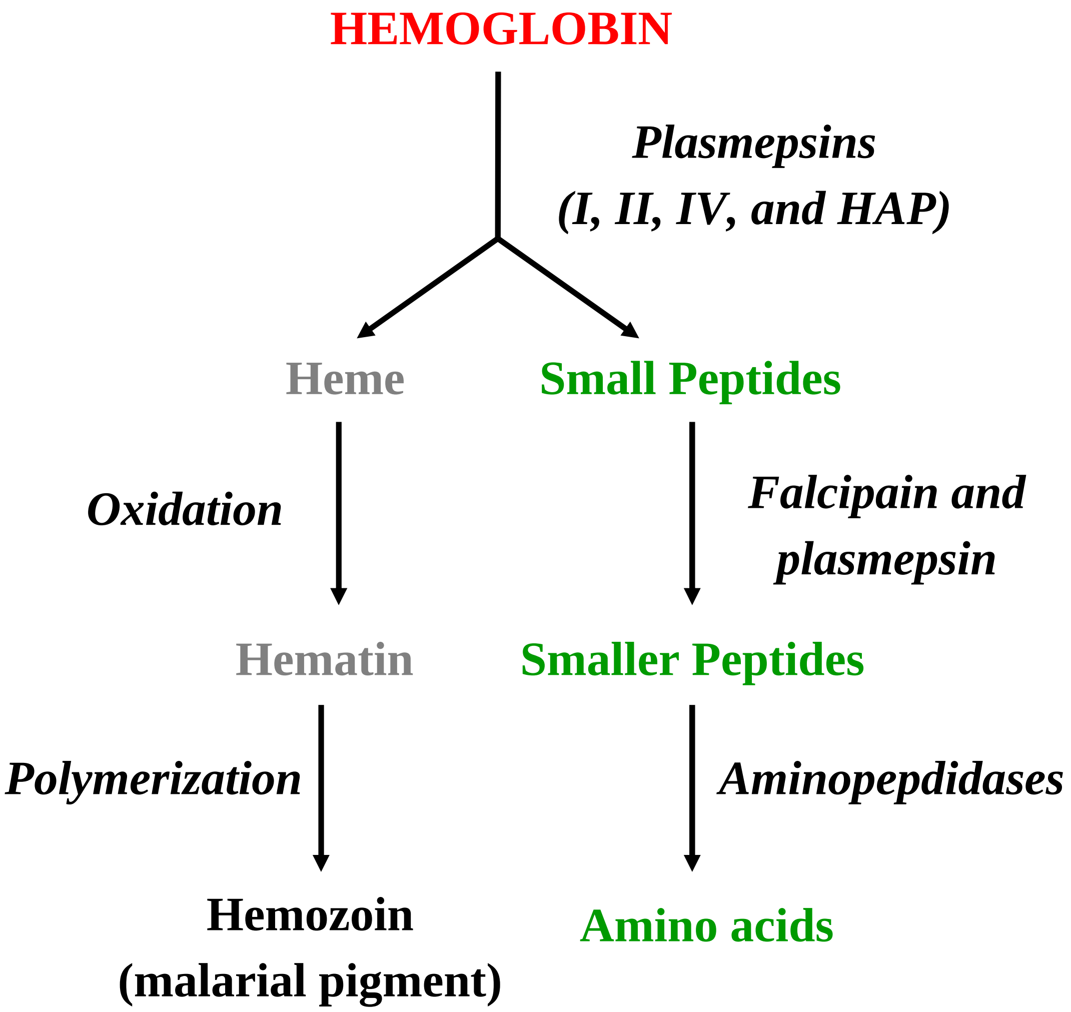
With the advent of the *Plasmodium* genome, many targets came into light

The potential anti-malarial drug targets are broadly classified into three categories, and each category has many individual targets.

- ❑ Targets involved in human hemoglobin degradation (proteases)
- ❑ Targets involved in parasite metabolism (Folate, phospholipid…)
- ❑ Targets engaged in parasite membrane transport and signalling (choline carrier etc).

WISDOM focuses on hemoglobin metabolism and especially on Plasmepsin II and Plasmepsin IV

Plasmepsins and Their Role in Human Hemoglobin Degradation



Plasmepsins are involved in hemoglobin degradation inside the food vacuole during the erythrocytic phase of the parasite life cycle.

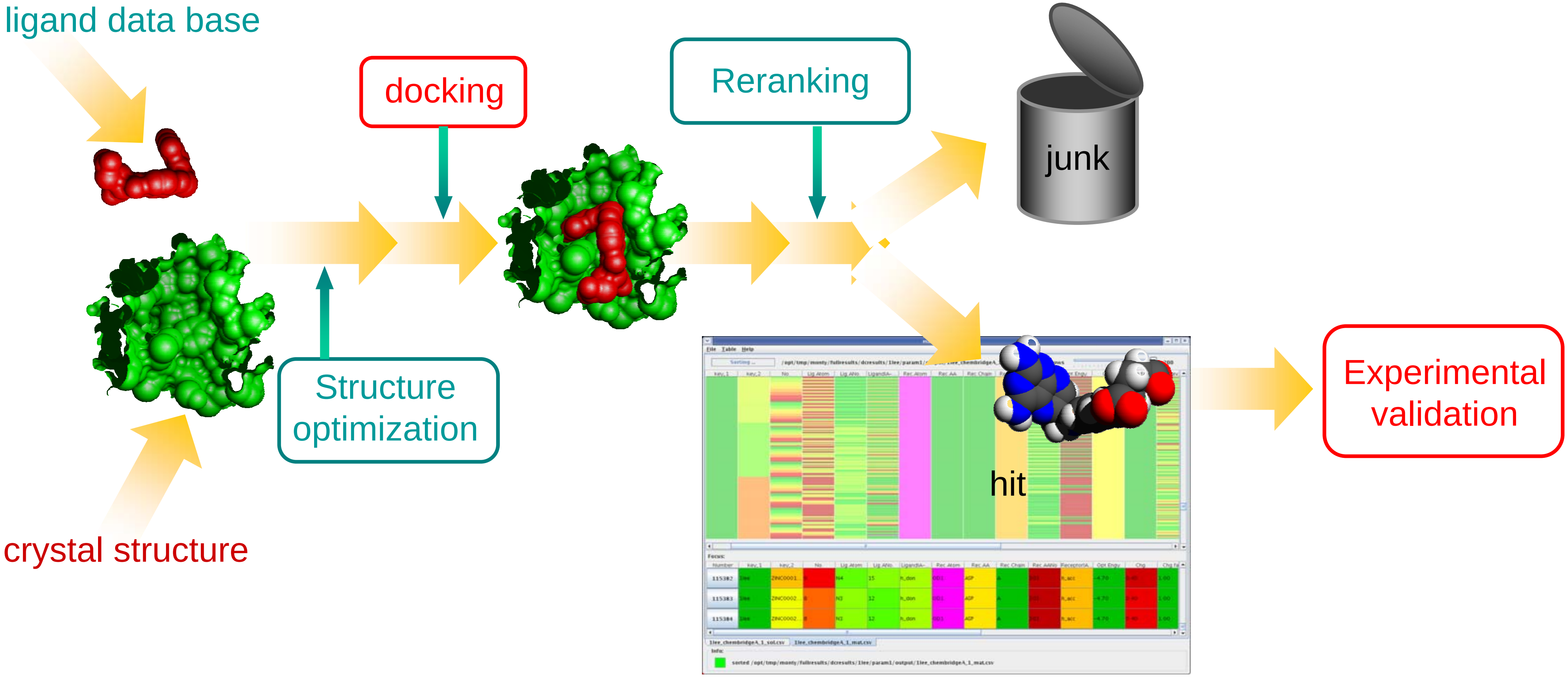
Plasmepsins are present in all of the four species of *Plasmodium* causing the disease in human

Sequence homology between the different plasmepsins is high (65-70%)

Sequence homology with its nearest human aspartic protease neighbour is fortunately low (35%)

Crystallographic data of plasmepsins are available in PDB

Dataflow and Workflow in Virtual Screening (by Docking)



ligand data base

crystal structure

Structure optimization

docking

Reranking

junk

hit

Experimental validation

EGEE, the World´s Largest Grid Infrastructure

Started in 2004, +70 partners in the world

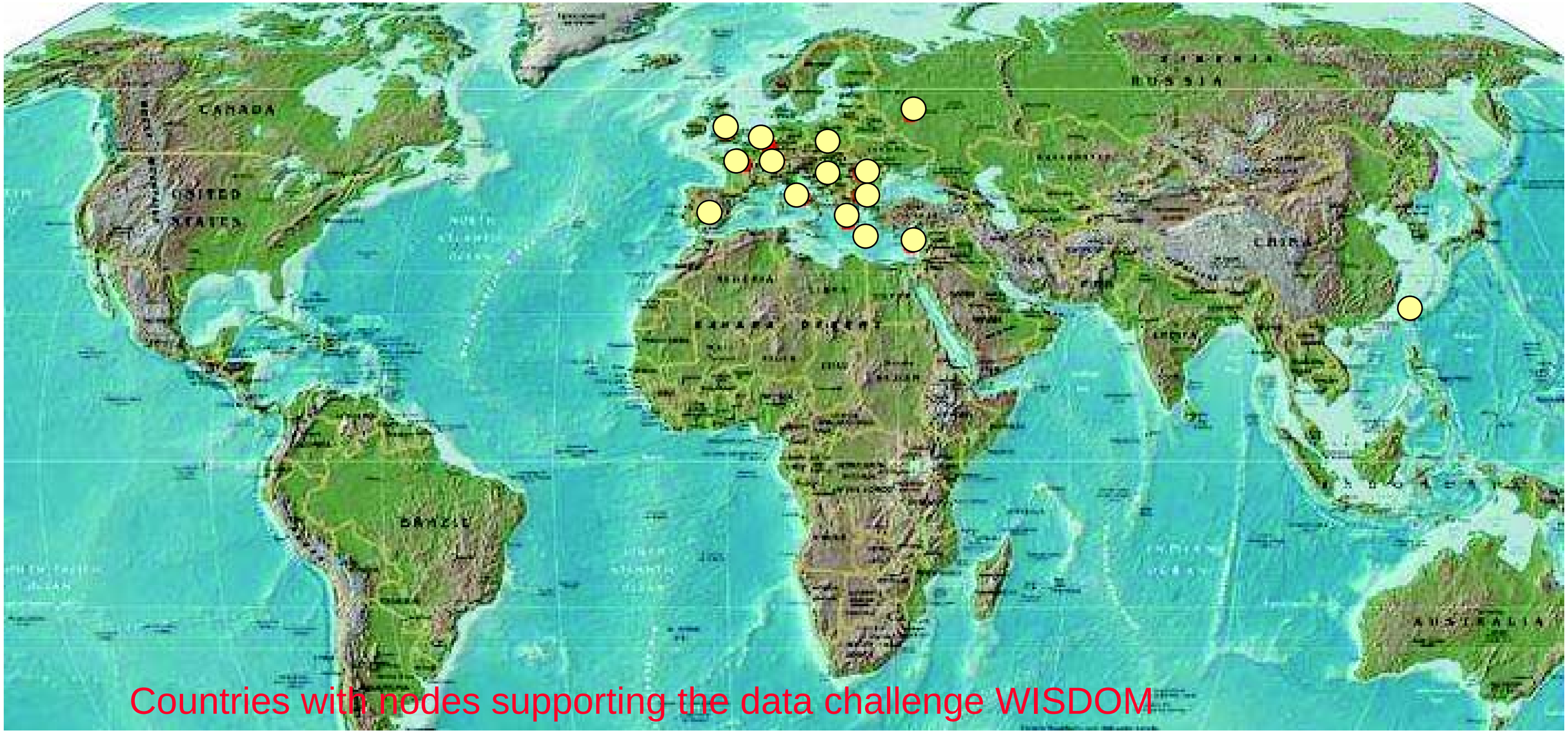
Project leader : CERN

6 scientific domains with >20 applications deployed

170 grid nodes, 17000 CPUs, several PetaBytes of data, 10000 jobs by day

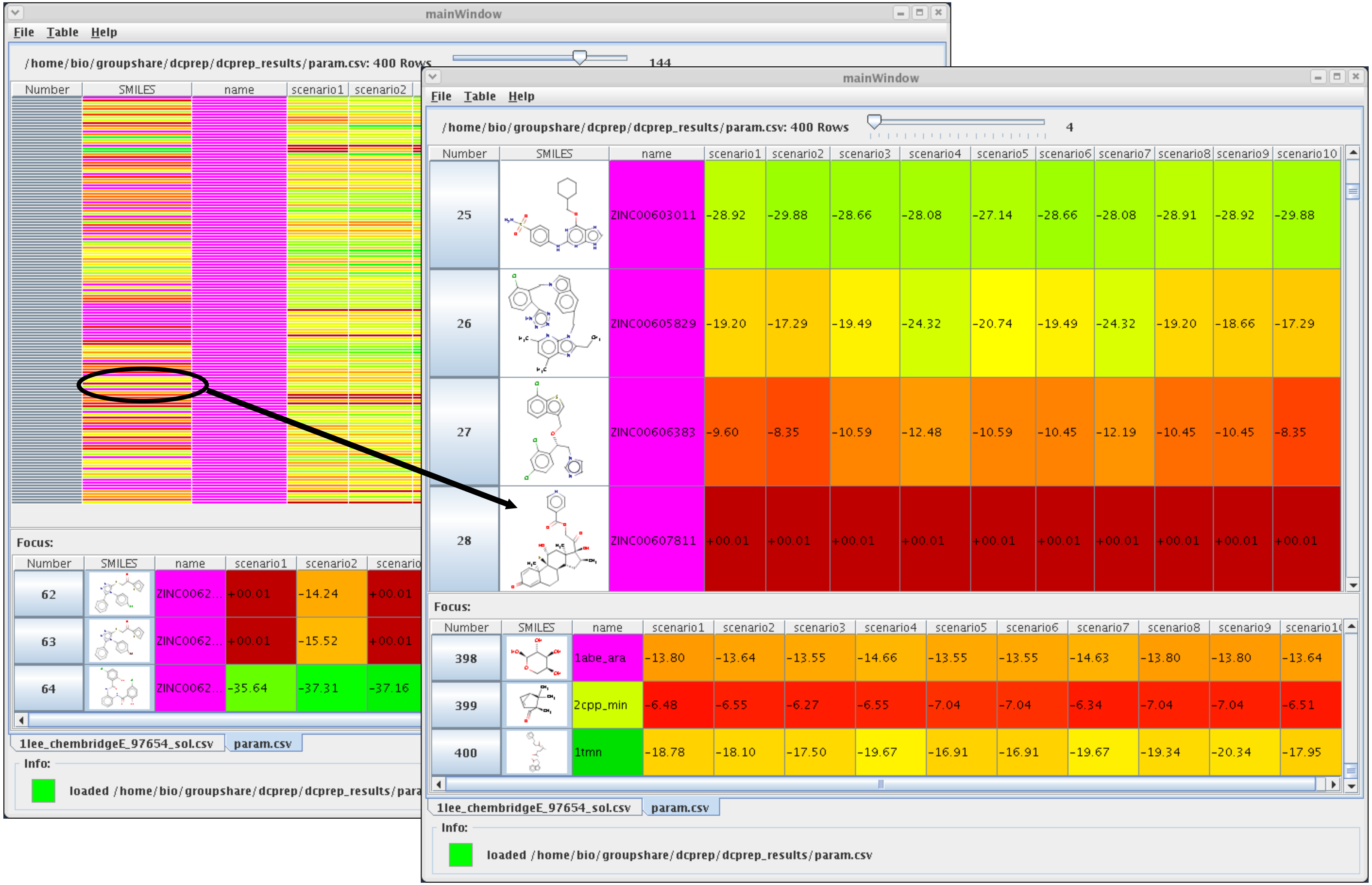
BioMed VO

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



Countries with nodes supporting the data challenge WISDOM

VS Explorer: a Tool for Analyzing “Grid Scale” Ranking Lists



mainWindow
File Table Help
/home/bio/groupshare/dcprep/dcprep_results/param.csv: 400 Rows
144
Number
SMILES
name
scenario1
scenario2

Focus:

Number	SMILES	name	scenario1	scenario2	scenario
62		ZINC0062...	+00.01	-14.24	+00.01
63		ZINC0062...	+00.01	-15.52	+00.01
64		ZINC0062...	-35.64	-37.31	-37.16

1lee_chembridgeE_97654_sol.csv
param.csv
Info:

loaded /home/bio/groupshare/dcprep/dcprep_results/para

mainWindow
File Table Help
/home/bio/groupshare/dcprep/dcprep_results/param.csv: 400 Rows
4

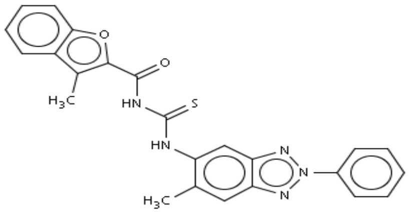
Number	SMILES	name	scenario1	scenario2	scenario3	scenario4	scenario5	scenario6	scenario7	scenario8	scenario9	scenario10
25		ZINC00603011	-28.92	-29.88	-28.66	-28.08	-27.14	-28.66	-28.08	-28.91	-28.92	-29.88
26		ZINC00605829	-19.20	-17.29	-19.49	-24.32	-20.74	-19.49	-24.32	-19.20	-18.66	-17.29
27		ZINC00606383	-9.60	-8.35	-10.59	-12.48	-10.59	-10.45	-12.19	-10.45	-10.45	-8.35
28		ZINC00607811	+00.01	+00.01	+00.01	+00.01	+00.01	+00.01	+00.01	+00.01	+00.01	+00.01

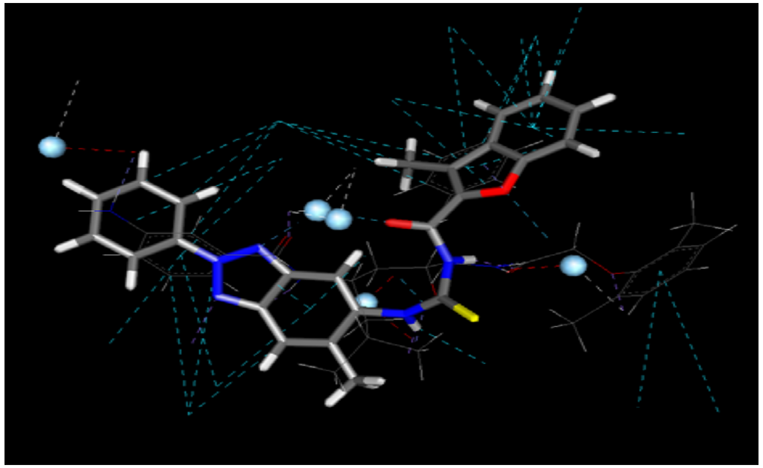
Focus:

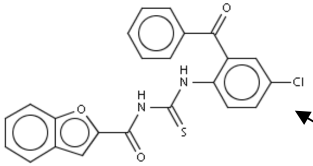
Number	SMILES	name	scenario1	scenario2	scenario3	scenario4	scenario5	scenario6	scenario7	scenario8	scenario9	scenario1(
398		1abe_ara	-13.80	-13.64	-13.55	-14.66	-13.55	-13.55	-14.63	-13.80	-13.80	-13.64
399		2cpp_min	-6.48	-6.55	-6.27	-6.55	-7.04	-7.04	-6.34	-7.04	-7.04	-6.51
400		1tmn	-18.78	-18.10	-17.50	-19.67	-16.91	-16.91	-19.67	-19.34	-20.34	-17.95

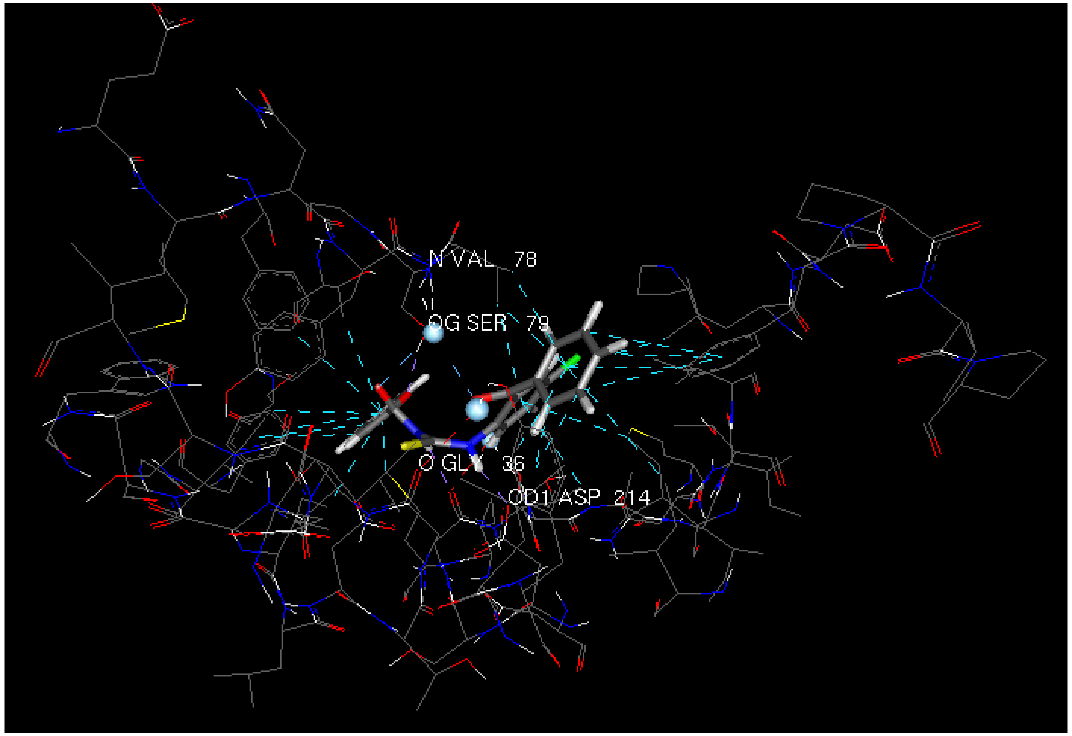
loaded /home/bio/groupshare/dcprep/dcprep_results/param.csv

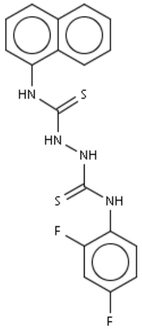
Compounds for MD - Thiourea compounds

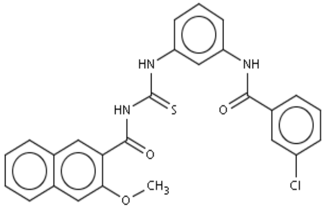




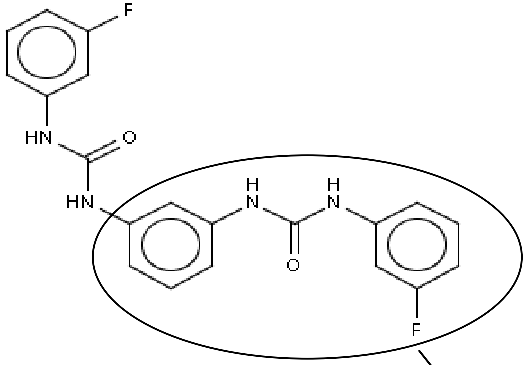


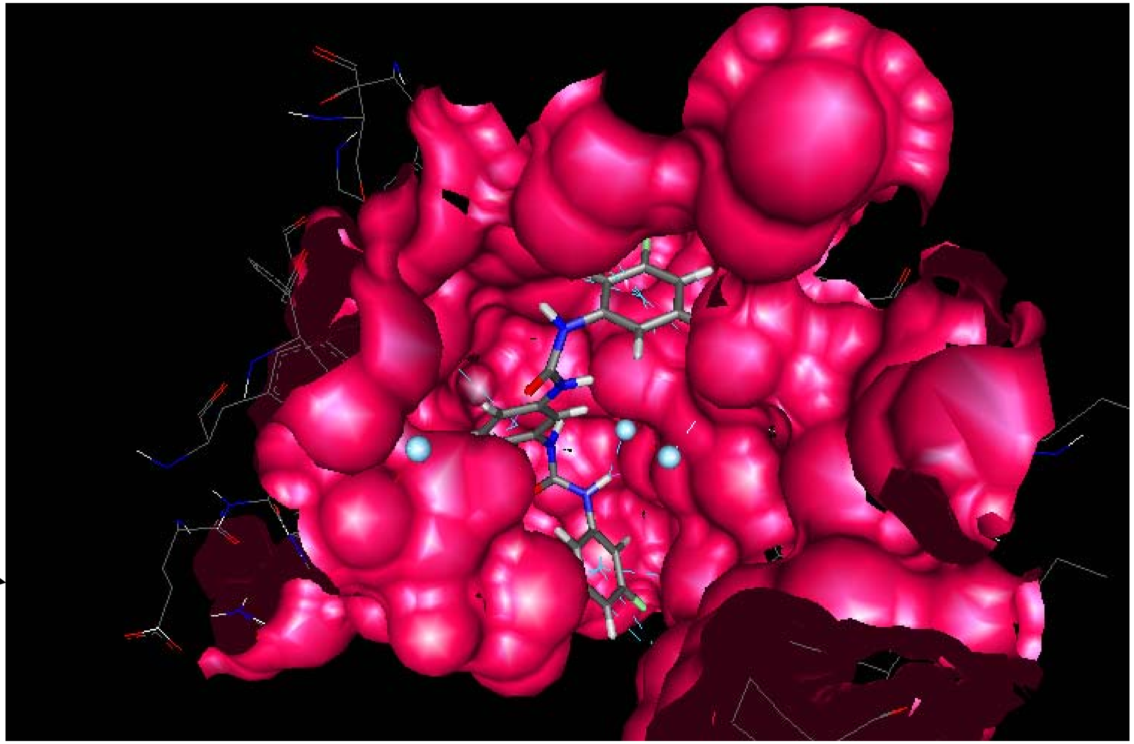


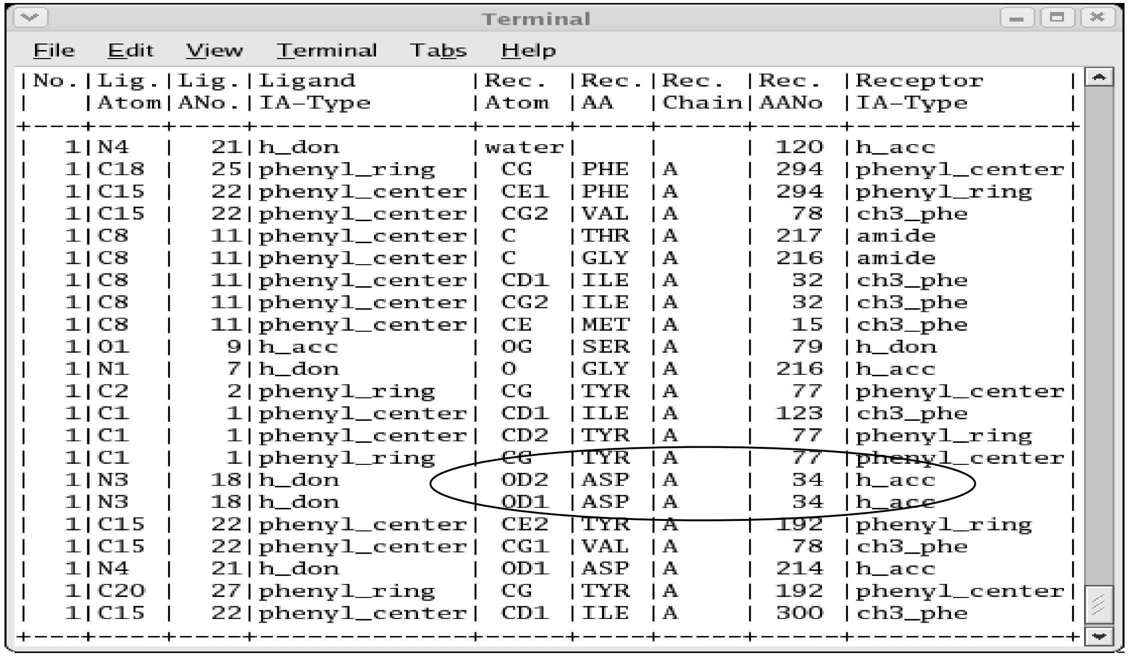




Compounds for MD-Urea compounds



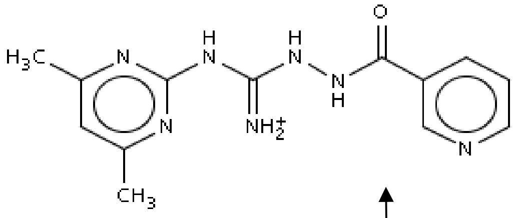


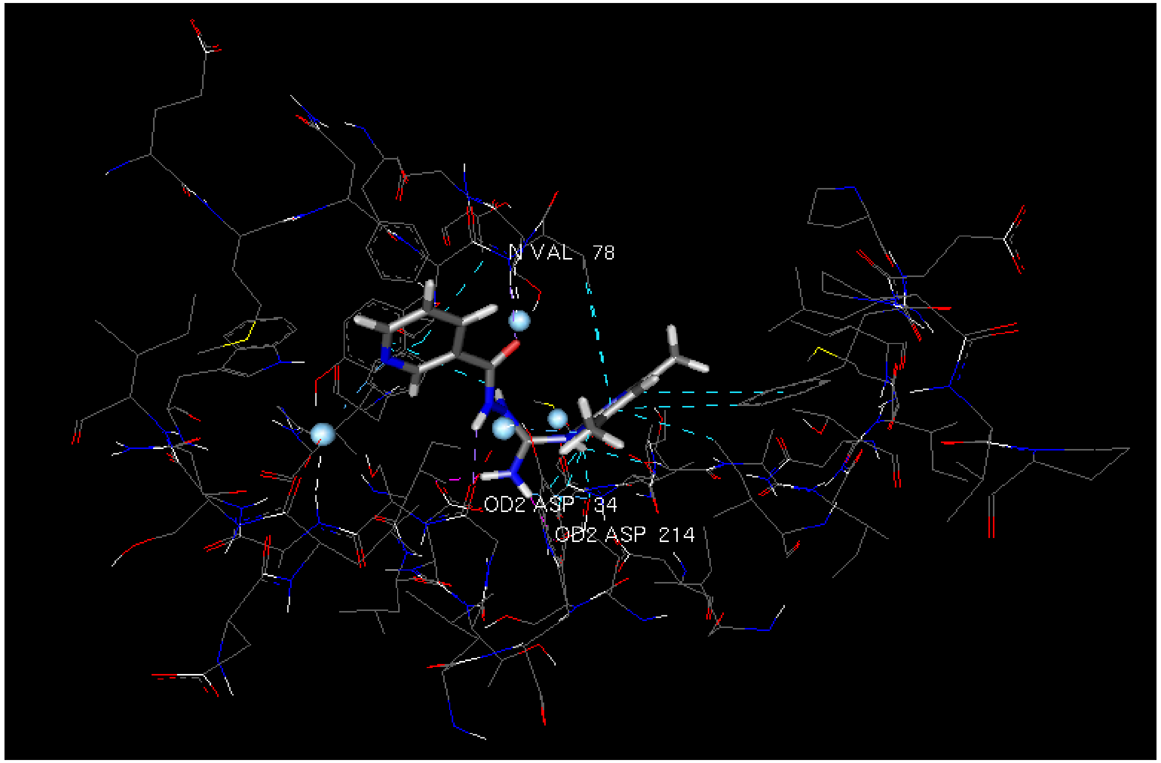


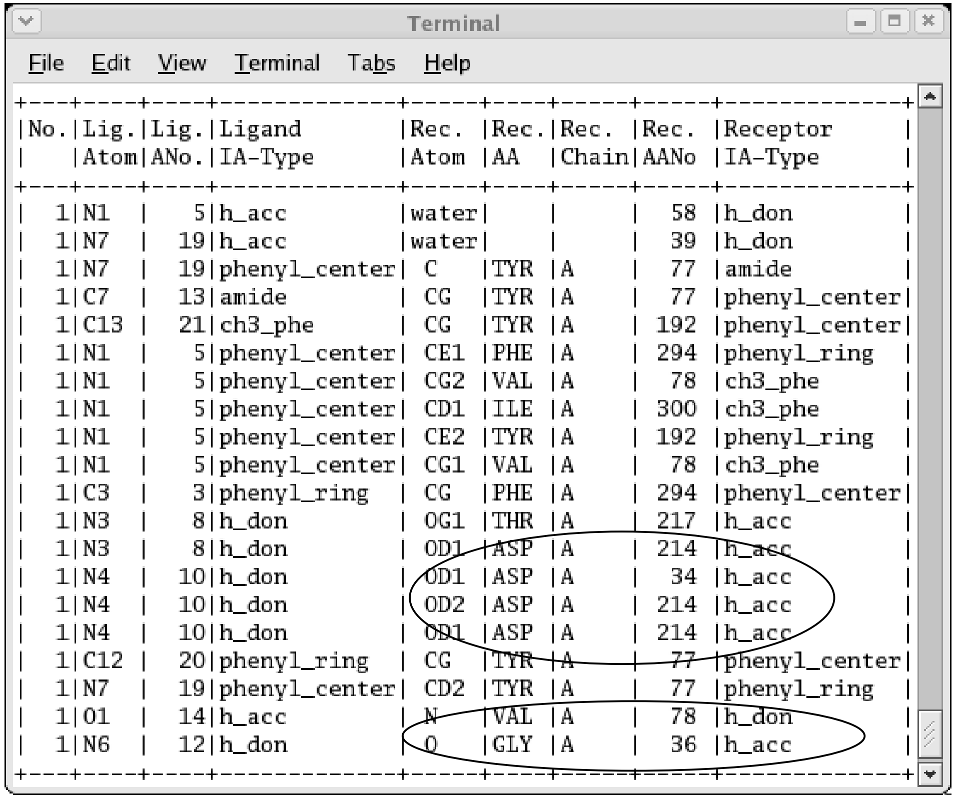
```
|No.|Lig.|Lig.|Ligand       |Rec. |Rec.|Rec. |Rec. |Receptor     |
|   |Atom|ANo.|IA-Type      |Atom |AA  |Chain|AANo |IA-Type      |
+----+-----+-----+-------------+------+-----+------+------+-------------+
|   1|N4   |   21|h_don        |water |     |      |   120|h_acc        |
|   1|C18  |   25|phenyl_ring  | CG   |PHE  |A     |   294|phenyl_center|
|   1|C15  |   22|phenyl_center| CE1  |PHE  |A     |   294|phenyl_ring  |
|   1|C15  |   22|phenyl_center| CG2  |VAL  |A     |    78|ch3_phe      |
|   1|C8   |   11|phenyl_center| C    |THR  |A     |   217|amide        |
|   1|C8   |   11|phenyl_center| C    |GLY  |A     |   216|amide        |
|   1|C8   |   11|phenyl_center| CD1  |ILE  |A     |    32|ch3_phe      |
|   1|C8   |   11|phenyl_center| CG2  |ILE  |A     |    32|ch3_phe      |
|   1|C8   |   11|phenyl_center| CE   |MET  |A     |    15|ch3_phe      |
|   1|O1   |    9|h_acc        | OG   |SER  |A     |    79|h_don        |
|   1|N1   |    7|h_don        | O    |GLY  |A     |   216|h_acc        |
|   1|C2   |    2|phenyl_ring  | CG   |TYR  |A     |    77|phenyl_center|
|   1|C1   |    1|phenyl_center| CD1  |ILE  |A     |   123|ch3_phe      |
|   1|C1   |    1|phenyl_center| CD2  |TYR  |A     |    77|phenyl_ring  |
|   1|C1   |    1|phenyl_ring  | CG   |TYR  |A     |    77|phenyl_center|
|   1|N3   |   18|h_don        | OD2  |ASP  |A     |    34|h_acc        |
|   1|N3   |   18|h_don        | OD1  |ASP  |A     |    34|h_acc        |
|   1|C15  |   22|phenyl_center| CE2  |TYR  |A     |   192|phenyl_ring  |
|   1|C15  |   22|phenyl_center| CG1  |VAL  |A     |    78|ch3_phe      |
|   1|N4   |   21|h_don        | OD1  |ASP  |A     |   214|h_acc        |
|   1|C20  |   27|phenyl_ring  | CG   |TYR  |A     |   192|phenyl_center|
|   1|C15  |   22|phenyl_center| CD1  |ILE  |A     |   300|ch3_phe      |
+----+-----+-----+-------------+------+-----+------+------+-------------+
```

Note: Diphenyl urea compounds are well in agreement with literature (Walter Reed compounds)

Compounds for MD- Guanidino compounds







```
+---+----+----+-------------+-----+----+-----+-----+-------------+
|No.|Lig.|Lig.|Ligand       |Rec. |Rec.|Rec. |Rec. |Receptor     |
|   |Atom|ANo.|IA-Type      |Atom |AA  |Chain|AANo |IA-Type      |
+---+----+----+-------------+-----+----+-----+-----+-------------+
|  1|N1  |   5|h_acc        |water|    |     |  58 |h_don        |
|  1|N7  |  19|h_acc        |water|    |     |  39 |h_don        |
|  1|N7  |  19|phenyl_center| C   |TYR |A    |  77 |amide        |
|  1|C7  |  13|amide        | CG  |TYR |A    |  77 |phenyl_center|
|  1|C13 |  21|ch3_phe      | CG  |TYR |A    | 192 |phenyl_center|
|  1|N1  |   5|phenyl_center| CE1 |PHE |A    | 294 |phenyl_ring  |
|  1|N1  |   5|phenyl_center| CG2 |VAL |A    |  78 |ch3_phe      |
|  1|N1  |   5|phenyl_center| CD1 |ILE |A    | 300 |ch3_phe      |
|  1|N1  |   5|phenyl_center| CE2 |TYR |A    | 192 |phenyl_ring  |
|  1|N1  |   5|phenyl_center| CG1 |VAL |A    |  78 |ch3_phe      |
|  1|C3  |   3|phenyl_ring  | CG  |PHE |A    | 294 |phenyl_center|
|  1|N3  |   8|h_don        | OG1 |THR |A    | 217 |h_acc        |
|  1|N3  |   8|h_don        | OD1 |ASP |A    | 214 |h_acc        |
|  1|N4  |  10|h_don        | OD1 |ASP |A    |  34 |h_acc        |
|  1|N4  |  10|h_don        | OD2 |ASP |A    | 214 |h_acc        |
|  1|N4  |  10|h_don        | OD1 |ASP |A    | 214 |h_acc        |
|  1|C12 |  20|phenyl_ring  | CG  |TYR |A    |  77 |phenyl_center|
|  1|N7  |  19|phenyl_center| CD2 |TYR |A    |  77 |phenyl_ring  |
|  1|O1  |  14|h_acc        | N   |VAL |A    |  78 |h_don        |
|  1|N6  |  12|h_don        | O   |GLY |A    |  36 |h_acc        |
+---+----+----+-------------+-----+----+-----+-----+-------------+
```

Note: Satisfied all criteria, good binding mode, interactions to key residues, good score, appropriate descriptors.

Conclusions

- Virtual Screening is a straightforward approach to use the GRID in the pharma context
- Virtual Screening has been successfully used in EnterpriseGRIDs in the pharma industry and recently also on a large eScience infrastructure, the EGEE GRID
- Novel, promising candidate structures for the development of new anti-malarial drugs have been identified using GRID-based virtual screening
- WISDOM has initialized a series of follow-up projects that address other diseases such as avian bird flue
- The relevance of virtual screening approaches on the GRID has been proven; uptake by small and medium size pharma companies is still too slow

Text – Mining on the GRID –

Information Extraction for Scientific & Competitive Intelligence

Archivierungsangaben

Scientific & Competitive Intelligence

What is Scientific & Competitive Intelligence ?

- Scientific and competitive intelligence are terms coined for the application of automated information mining methods
- Information mining ranges from improved document retrieval to full blown information extraction
- Goal of the pharmaceutical industry is to make sure that **all** relevant information is at hand when a decision about a drug development project is to be made
- Consequently, scientific and competitive intelligence encompasses not only text mining in PubMed abstracts, but extends to patent literature and business news streams

Archivierungsangaben

Protein Name Recognition

Multiple names for one gene

Ambiguous names in databases

Ambiguous acronyms

Common word names

Multi-word terms

Spelling variants

Permutations

Nested protein names

F12A

Neuronectin, GMEM, tenascin,
HXB, cytotactin, hexabrachion

p21, EPO, large T antigen

WAS, STEP, iCE, StAR

Interleukin 1 alpha
Tumor necrosis factor beta

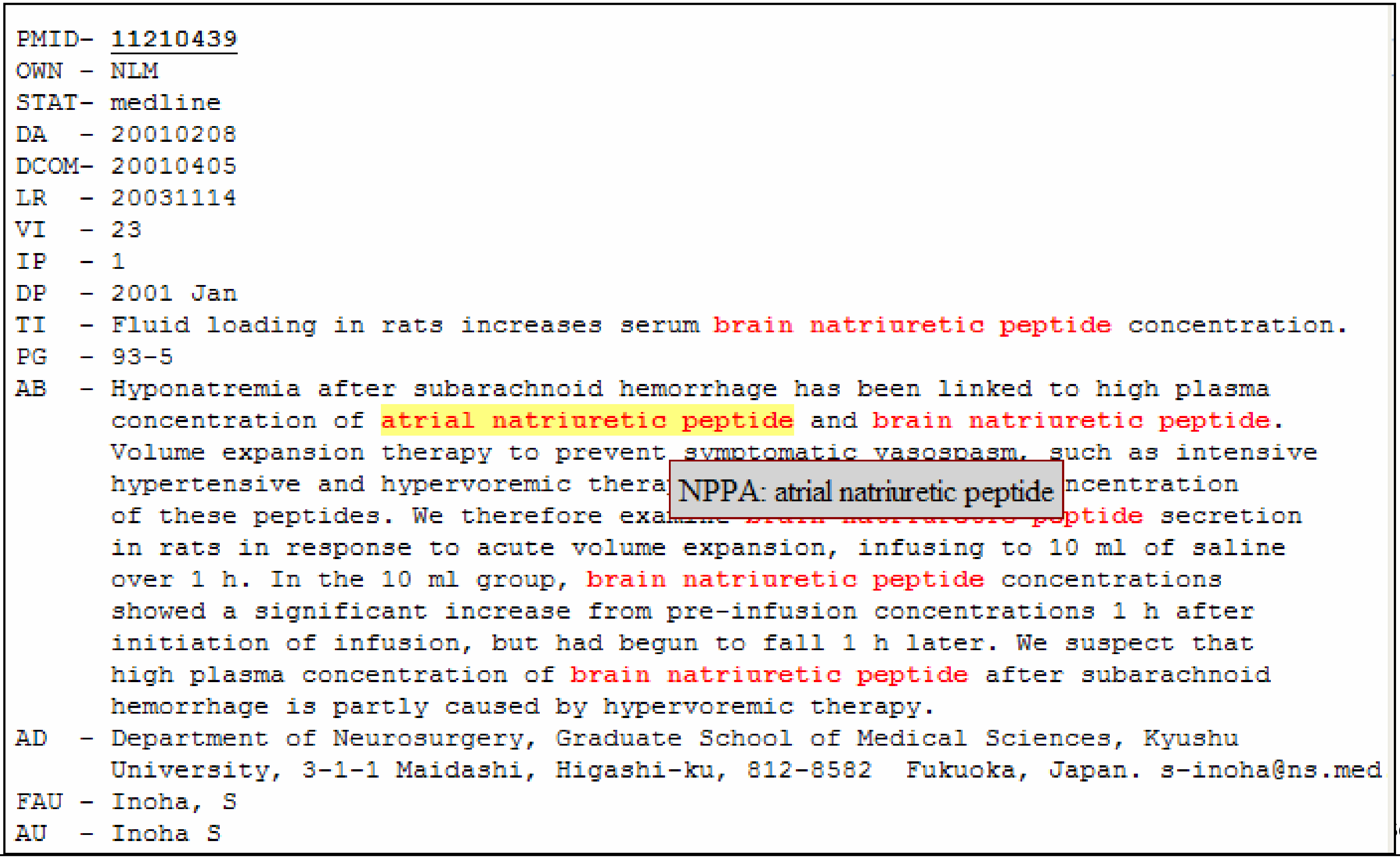
COL1A1

Collagen, type I, alpha 1
Collagen alpha 1(I) chain
Alpha 1 collagen
Alpha-1 type I collagen

TNF receptor 1
collagen, type I, alpha receptor

Archivierungsangaben

Dynamically Labelled Text:



```
PMID- 11210439
OWN - NLM
STAT- medline
DA  - 20010208
DCOM- 20010405
LR  - 20031114
VI  - 23
IP  - 1
DP  - 2001 Jan
TI  - Fluid loading in rats increases serum brain natriuretic peptide concentration.
PG  - 93-5
AB  - Hyponatremia after subarachnoid hemorrhage has been linked to high plasma
      concentration of atrial natriuretic peptide and brain natriuretic peptide.
      Volume expansion therapy to prevent symptomatic vasospasm, such as intensive
      hypertensive and hypervoremic thera                         ncentration
      of these peptides. We therefore exa                   ptide secretion
      in rats in response to acute volume expansion, infusing to 10 ml of saline
      over 1 h. In the 10 ml group, brain natriuretic peptide concentrations
      showed a significant increase from pre-infusion concentrations 1 h after
      initiation of infusion, but had begun to fall 1 h later. We suspect that
      high plasma concentration of brain natriuretic peptide after subarachnoid
      hemorrhage is partly caused by hypervoremic therapy.
AD  - Department of Neurosurgery, Graduate School of Medical Sciences, Kyushu
      University, 3-1-1 Maidashi, Higashi-ku, 812-8582  Fukuoka, Japan. s-inoha@ns.med
FAU - Inoha, S
AU  - Inoha S
```

NPPA: atrial natriuretic peptide

Archivierungsangaben

Chemical Name Recognition

Dictionary names:

Brand names	Ariven, Extren, Clivarin
Organic chemical compounds	2-Acetoxybenzoic acid
Generic names, INN, USAN	Aspirin, Celecoxib, Heparin
Substance classes	secondary amine, cholin sulfates
Side groups, atoms and ions	butyl group, potassium, fluoride anion
Pharmacological and biological effects	Cyclooxygenase inhibitor, Anticoagulants

Regular expressions:

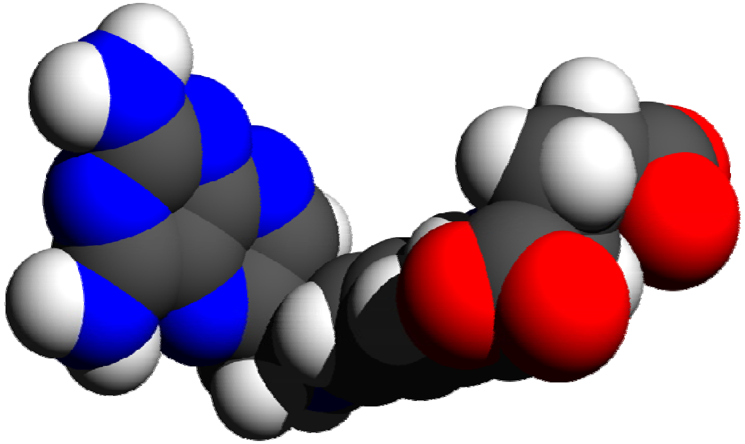
IUPAC names	N-[2-[4-[(2-oxy cyclohexyl)methyl]- …

Archivierungsangaben

Sources: ChEBi; DrugBank; MeSH; text mining

Representations of Chemical Compounds

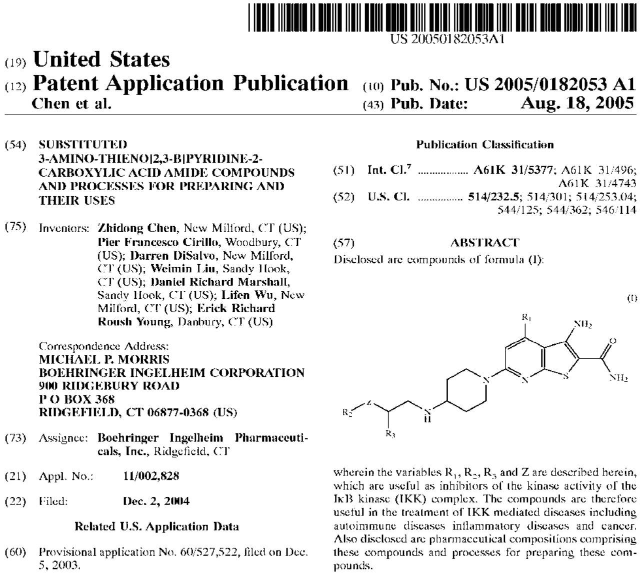
- ❑ Name (trivial, trade, brand, INN, USAN)
- ❑ Registration numbers (CAS, NCI, Beilstein)
- ❑ Formal description (sum formula, SMILES)
- ❑ Chemical nomenclature (IUPAC, CAS, InChI)
- ❑ Depictions



Archivierungsangaben

Chemical Structure Recognition – an Overview

1 Document



US 20050182053A1

(19) **United States**

(12) **Patent Application Publication**

Chen et al.

(10) **Pub. No.: US 2005/0182053 A1**

(43) **Pub. Date: Aug. 18, 2005**

(54) **SUBSTITUTED 3-AMINO-THIENO|2,3-B|PYRIDINE-2-CARBOXYLIC ACID AMIDE COMPOUNDS AND PROCESSES FOR PREPARING AND THEIR USES**

(75) Inventors: **Zhidong Chen**, New Milford, CT (US); **Pier Francesco Cirillo**, Woodbury, CT (US); **Darren DiSalvo**, New Milford, CT (US); **Weimin Liu**, Sandy Hook, CT (US); **Daniel Richard Marshall**, Sandy Hook, CT (US); **Lifen Wu**, New Milford, CT (US); **Erick Richard Roush Young**, Danbury, CT (US)

Correspondence Address:
MICHAEL P. MORRIS
BOEHRINGER INGELHEIM CORPORATION
900 RIDGEBURY ROAD
P O BOX 368
RIDGEFIELD, CT 06877-0368 (US)

(73) Assignee: **Boehringer Ingelheim Pharmaceuticals, Inc.**, Ridgefield, CT

(21) Appl. No.: **11/002,828**

(22) Filed: **Dec. 2, 2004**

Related U.S. Application Data

(60) Provisional application No. 60/527,522, filed on Dec. 5, 2003.

Publication Classification

(51) **Int. Cl.**^{7} **A61K 31/5377**; A61K 31/496; A61K 31/4743

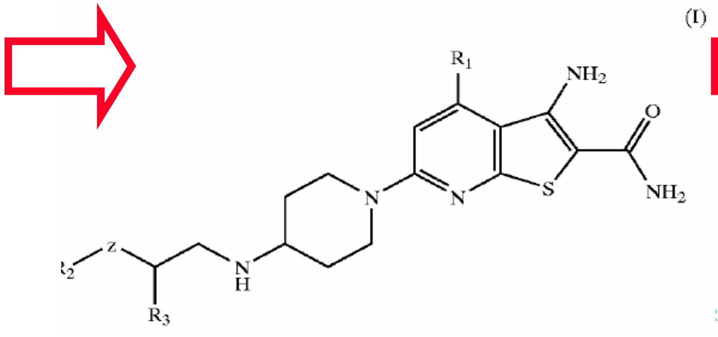
(52) **U.S. Cl.** **514/232.5**; 514/301; 514/253.04; 544/125; 544/362; 546/114

(57) **ABSTRACT**

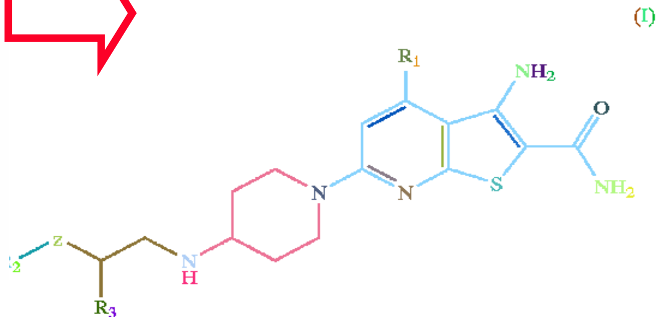
Disclosed are compounds of formula (I):

wherein the variables R_1, R_2, R_3 and Z are described herein, which are useful as inhibitors of the kinase activity of the IκB kinase (IKK) complex. The compounds are therefore useful in the treatment of IKK mediated diseases including autoimmune diseases inflammatory diseases and cancer. Also disclosed are pharmaceutical compositions comprising these compounds and processes for preparing these compounds.

2 Depiction



3 Reconstruction

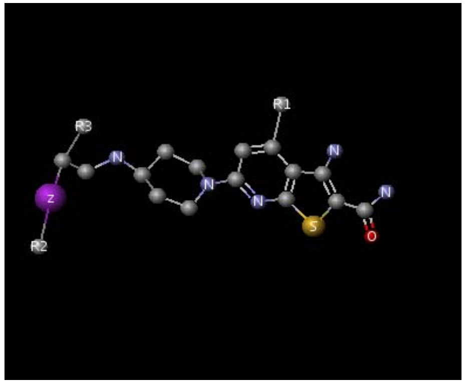


4 SDF file

```
created from
/home/marc/workspace/CSR/results/CSR/examples/US2005182053/
US2005182053_result.pnm
MZCSRv0.5010050621162D   0.00000     0.00000     0

 26 28  0     1  0  0  0  0  0999 V2000
  204.0000  102.0000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  275.0000   61.0000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  201.0000   59.0000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  422.0000  178.0000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  311.0000  164.0000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  384.0000  165.0000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  447.0000  144.0000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  383.0000  123.0000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  131.0000   60.0000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  239.0000  123.0000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  349.0000  218.0000    0.0000 R#  0  0  0  0  0  0  0  0  0  0  0  0
  447.0000  207.0000    0.0000 R#  0  0  0  0  0  0  0  0  0  0  0  0
```

5 *in silico* Chemistry



Distributed information extraction from scientific literature

Archivierungsangaben

Distributed Documents and Scaling

- ❑ MEDLINE comprises currently more than 16 million abstracts
- ❑ More and more publications available as full text (open access) and in institutional repositories
- ❑ Patent literature comprises more than 50 million full text patents; approximately 13% containing information on chemistry, biology and pharmacology
- ❑ In pharma companies, most relevant information is still available in free text (e.g. information on clinical studies; FDA / BfArM registration)
- ❑ Thousands of news streams and millions of websites comprise valuable information on innovations

Unstructured Information Management Architecture (UIMA)

❑ Service Oriented Architecture / framework proposed by IBM

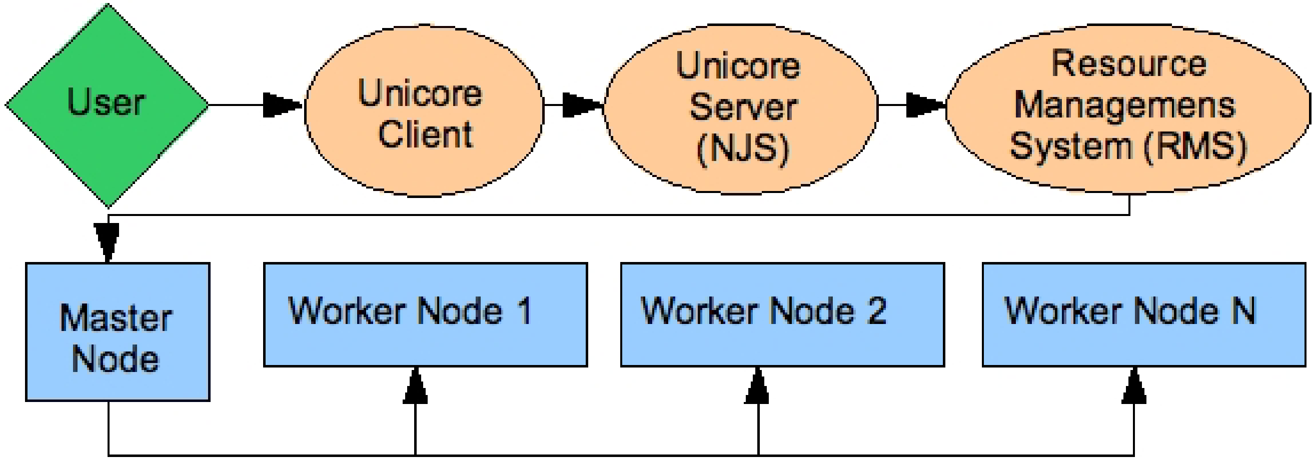
❑ Rapidly adopted by commercial and academic tool provider in the area of text mining

❑ Supports the assembly of complex annotation workflows

❑ Annotators might be entity recognition systems, part-of-speech-analysis modules and other type of unstructured information processing tools

❑ UIMA standardizes text and image mining (UIMA not restricted to pure text)

Distributed information extraction from scientific literature



Lessons Learned

- ❑ GRID is an emerging topic for the pharmaceutical industry
- ❑ The WISDOM project has demonstrated, that the GRID is well suited to support large scale virtual screening experiments; making virtual screening a “killer app” for biomedical GRIDs
- ❑ In a completely different field, namely text mining and information extraction, the GRID will enable us to deal with both, distributed documents and compute-intensive tasks
- ❑ Text Mining on the GRID might be the next biomedical GRID “killer app”

Thank you for your attention

Archivierungsangaben