"Large-Scale Information Extraction for Biomedical Modelling and Simulation"

Fraunhofer
SCAI

Martin Hofmann-Apitius
Head of the Department of Bioinformatics
Fraunhofer Institute for Algorithms and Scientific Computing (SCAI)

Fraunhofer Society

- Founded 1949
- Europe´s largest applied research organisation
- 60 Research Institutes
- 17.000 Employees
- Annual Budget about 1,5 Billion Euro
- Financial model: 40 % industry collaborations
  40 % public funding
  20 % institutional funding

*Joseph von Fraunhofer (1787 – 1826)
Scientist, Inventor and Entrepreneur

Fraunhofer
SCAI

Fraunhofer-Campus Schloss Birlinghoven

Institutes

- Algorithms and Scientific Computing SCAI
- Intelligent Analysis and Information Systems IAIS
- Applied Information Technology FIT

600 Scientists, 200 Students
Linked to Universities Bonn, Aachen
and Cologne

b-it

Fraunhofer
SCAI

Fraunhofer: Applied Research for Industrial Applications

Fraunhofer stands for:

- sustainable (applied) research
- focus on contract research and innovation
- bridging between excellent academic research and industrial application
- clear mission towards improving and fostering innovation
- research done with the idea in mind to generate added value in a commercial sense

Fraunhofer SCAI

SCAI Department of Bioinformatics: R&D in a nutshell

Fraunhofer SCAI Department of Bioinformatics R&D activities:

1. Information extraction in the **life sciences**:

   I. Text Mining - Recognition of named entities & relationships in text

   II. Image Mining - Reconstruction of chemical information from chemical structure depictions

2. Disease modelling (focus on neurodegenerative diseases)

3. eScience, Grid-/Cloud- Computing and HPC (Clus

Making Scientific Content available for Computing

Fraunhofer SCAI

People

Fraunhofer SCAI Department of Bioinformatics currently comprises:

- 10 scientists
- 2 scientific software developers
- 6 PhD students
- > 6 Master students

- predominantly computer scientists & biologists
- additional scientists and PhD students via University of Bonn

Fraunhofer SCAI

The History of High Performance Computing in BioMedicine at SCAI: Once upon a time ....

Fraunhofer SCAI

- WISDOM stands for Wide In Silico Docking on Malaria
  - Goals
    - Computational goals
      - To show the relevance of computational grids in biomedical applications
    - Biological goals
      - Risk of Malaria, Post genomic era, number of targets, Finding new compounds in a cost effective way
    - Method establishment
      - Establishing virtual screening technology on computational grids

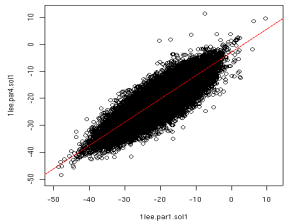
WISDOM-I and WISDOM-II

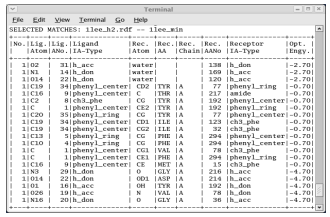
- WISDOM-I: First large scale docking against plasmepsin on computational grids
  - Achievements:
    - 80 CPU years in 45 days (Computational side)
    - Identified three novel scaffolds (biological side)
    - Docking on grids (Method establishment)
- WISDOM-II: Second assault on 4 different proteins (DHFR; Tubulin; two different Plasmepsins) implicated in malaria
  - Achievements:
    - 413 CPU years in 90 days (Computational side)
    - Result analysis is in progress (biological side)
    - Docking on grids (Method improved)

BioinfoGRID
Bioinformatics Grid Application for life science

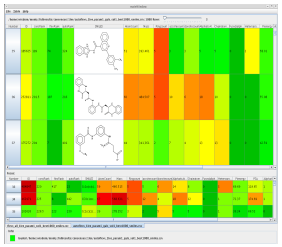
Strategies in result analysis

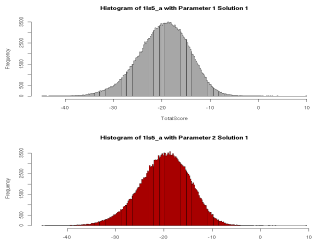
Results based on Scoring





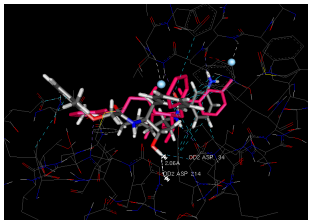
Results based on match information





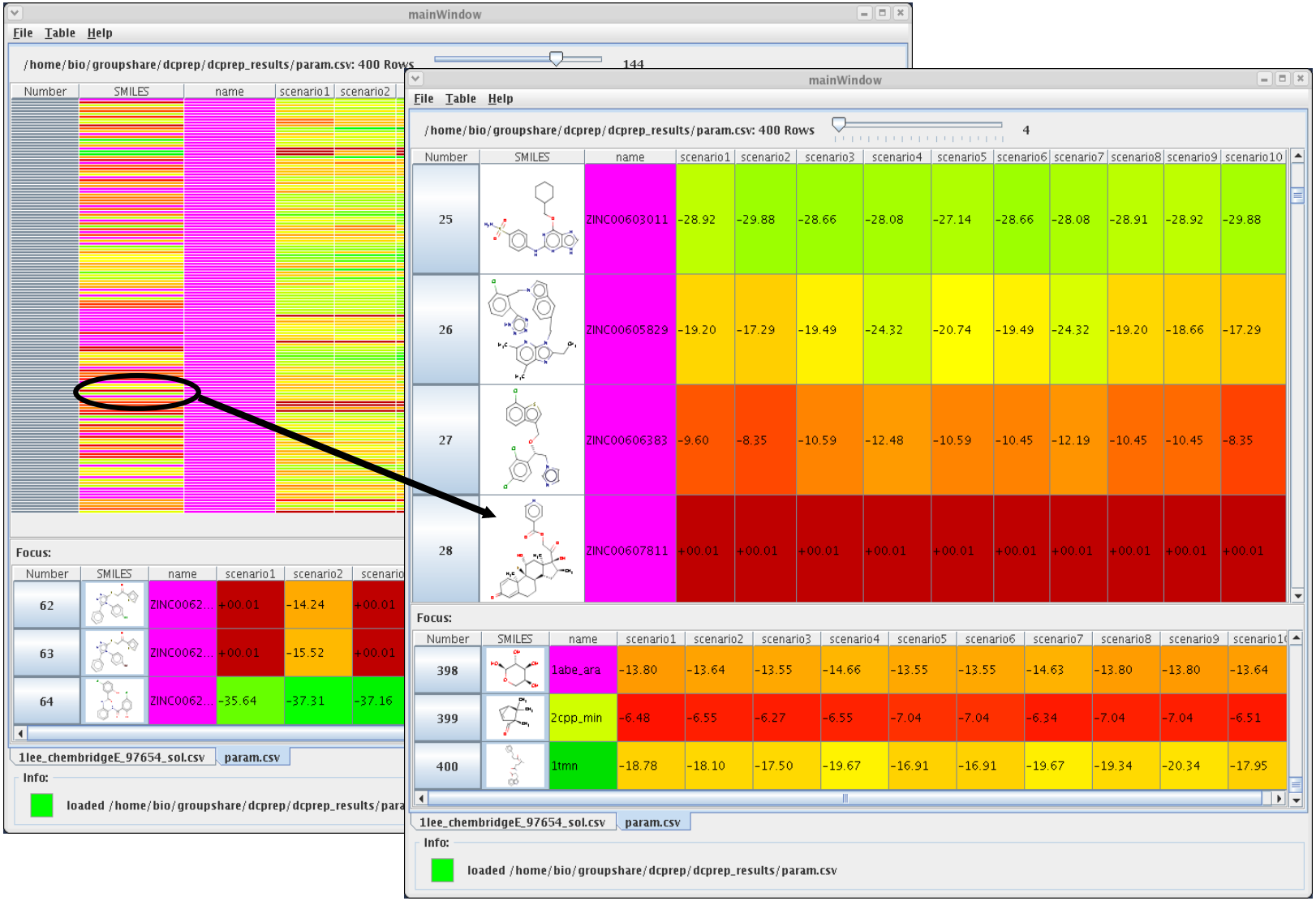
Results based on consensus scoring

Results based on different parameter settings



Results based on knowledge on binding site

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



WISDOM II achievements

Achievements

Completed the screening of 4.3 million compounds against four different targets

Identified more than 200 new “interesting chemical scaffolds” that have been tested in “wet lab assays”

Validated biological activity of 13 novel “virtual screening hits”

Submitted patents for 4 new compounds / scaffolds

Lessons Learned from WISDOM and other Workflows

1. Infrastructures such as EGEE (EGI) can be productively used for *in silico experimentation* in biomedicine

2. Post-Processing (searching where the data have been stored; which jobs have been successfully finished) of generated data (docking results) took much more time than simulating the ligand-target interaction

3. Support of complex workflow design and workflow optimisation in compute infrastructures is still a challenge

Fraunhofer SCAI

Scientific Workflow Optimization on the Grid

Sonja Holl

Jülich Supercomputing Centre (JSC)

Institute for Advanced Simulation (IAS)

Forschungszentrum Jülich, Germany

JÜLICH
FORSCHUNGSZENTRUM

Motivation

- Scientific workflows have many parameters
- Optimal settings not known a priori and data dependent
- Often several methods for the same task are available
- Parameters have constraints and non-linear dependencies
- Testing many parameter settings requires HPC
- Currently try and error based optimization, lack of tools

- Our primary goal: Optimizing the *quality* of a scientific workflow

Workflow Optimization

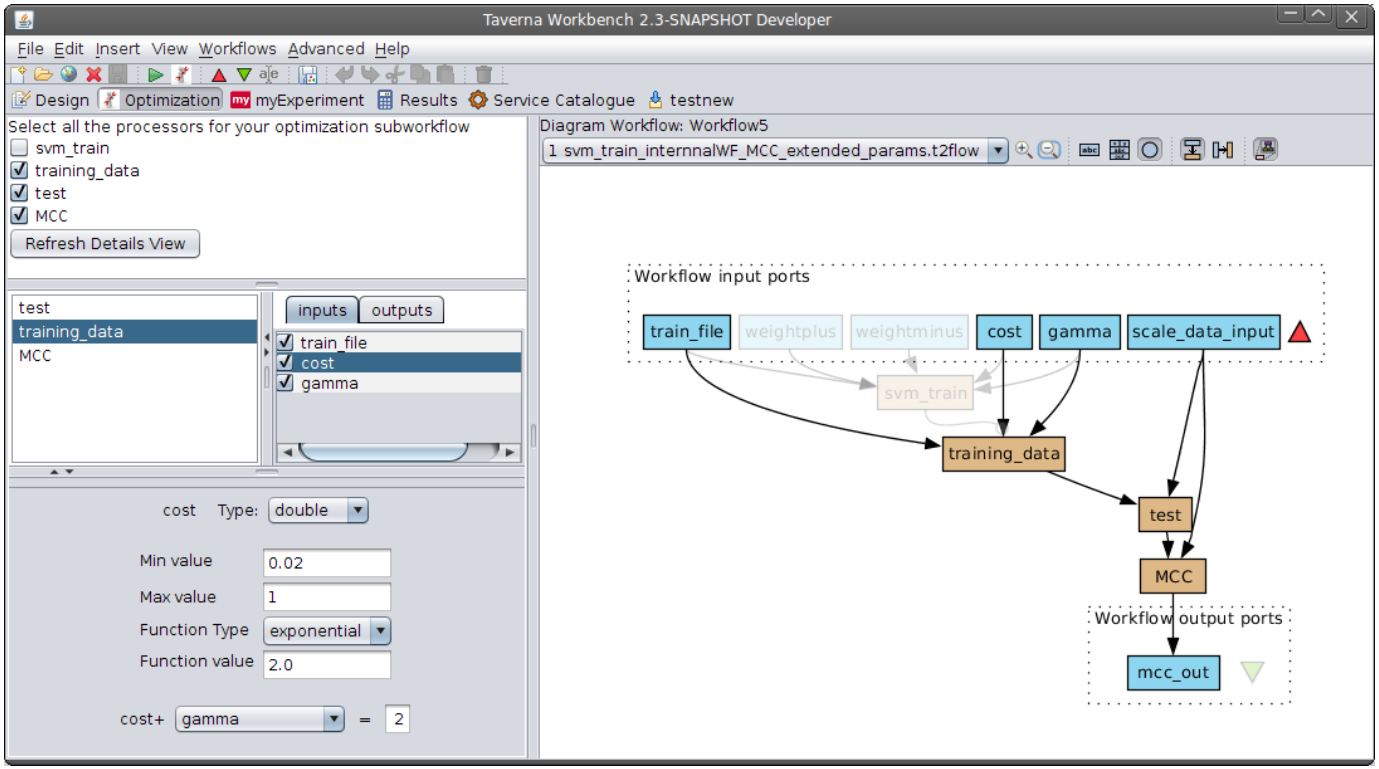
Find workflow settings that optimize the output quality

- Identify user requirements (Parameter, Constraints)
- Optimize relevant parts of a workflow
- Parameter optimization, Component set optimization
- Measure output quality (Fitness Function, User Feedback)

- Use the Grid to speed up optimization
  (UNICORE Grid middleware)

The UNICORE Taverna Optimization Plugin

- Easy, GUI-based setup of WF and optimization tasks
- Plugin enables optimization within Taverna WFMS
- Parallel execution of WF optimization runs via UNICORE
- Link to major biomedical in silico labs (caBIG; myExperiment)



Workflow Optimization – Techniques 1/2

*Find optimal **input parameters** for WF applications*

- Varying free application parameters
- Realized via evolutionary algorithm (JGAP-Framework)

*Advanced **Parameter Optimization***

- Dependent parameters
  - Discrete set of parameter values (files, flags, integer)
  - Dependencies: Fixed combinations, Mathematical dependency, Logical dependency

Workflow Optimization – Techniques 2/2

***Component Set Optimization**:*

Replace tool or algorithm with an equivalent one

- What is "equivalent“?
- Ontologies for tasks, inputs, and outputs required

Future work

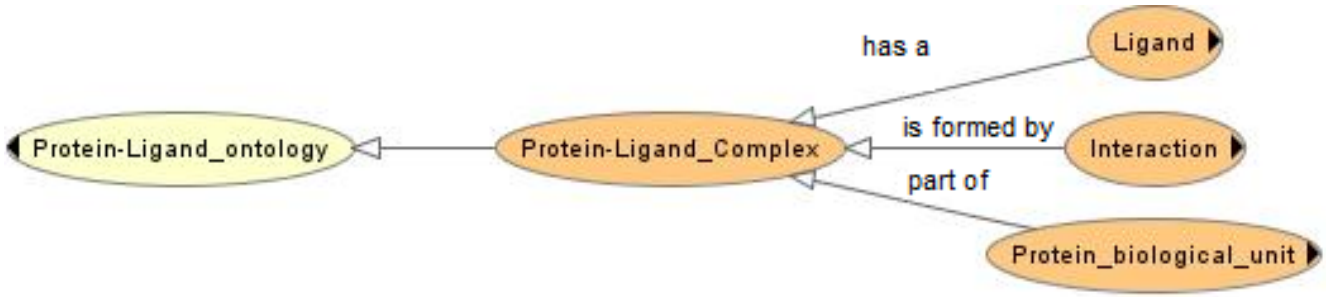
- Workflow structure optimization
- Adaptive optimization – learning from community runs
- Supervised optimization - User feedback as fitness function

Linking Methods and Algorithms to Scientific Objectives and Knowledge

The PLIO Ontology

Fraunhofer SCAI

PLIO root concepts (total concepts: 371)



- PLIO concept coverage
  - **Biophysical forces,** e.g. Van der Waals, electrostatics, etc.
  - **Interaction descriptors** e.g. pharmacophore, interaction fingerprint, etc.
  - **Experimental techniques,** e.g. NMR, X-ray, etc.
  - **Prediction and simulation methods,** e.g. MD, docking, etc.
  - **Classification of ligand activities,** e.g. biological activity, binding activity, etc.
  - **Classification of ligand modes of action,** e.g. agonist, inhibitor, etc.
  - **Classification of binding sites,** e.g. allosteric site, orthosteric site, etc.
  - **Structure-Activity-Relationships,** e.g. QSAR, COMFA, etc.

Each Concept has:

Reference

"http://www.shodor.org/chemviz/glossary.html"

Synonym

"activation energy
Energy of activation
Arrhenius activation energy
activation barrier
activation barriers
E(a)"

formula

"k=Aexp(-Ea/RT)

Ea - activation energy (Arrhenius activation energy)
k - rate constant
A - the pre-exponential factor or simply the prefactor
R - the gas constant
T - tempreture"

software

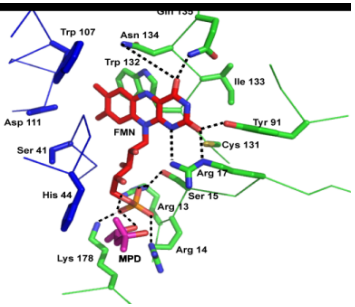
"Activation energy
Arrhenius calculation from two temperatures.
http://www.calctool.org/CALC/chem/kinetics/act_en
"

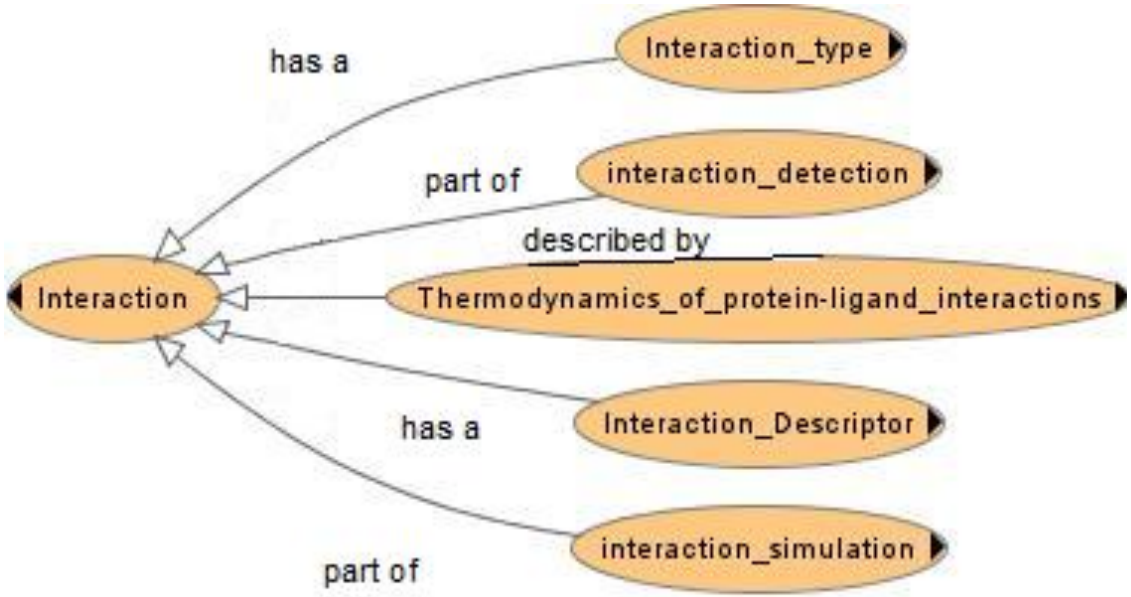
isDefinedBy

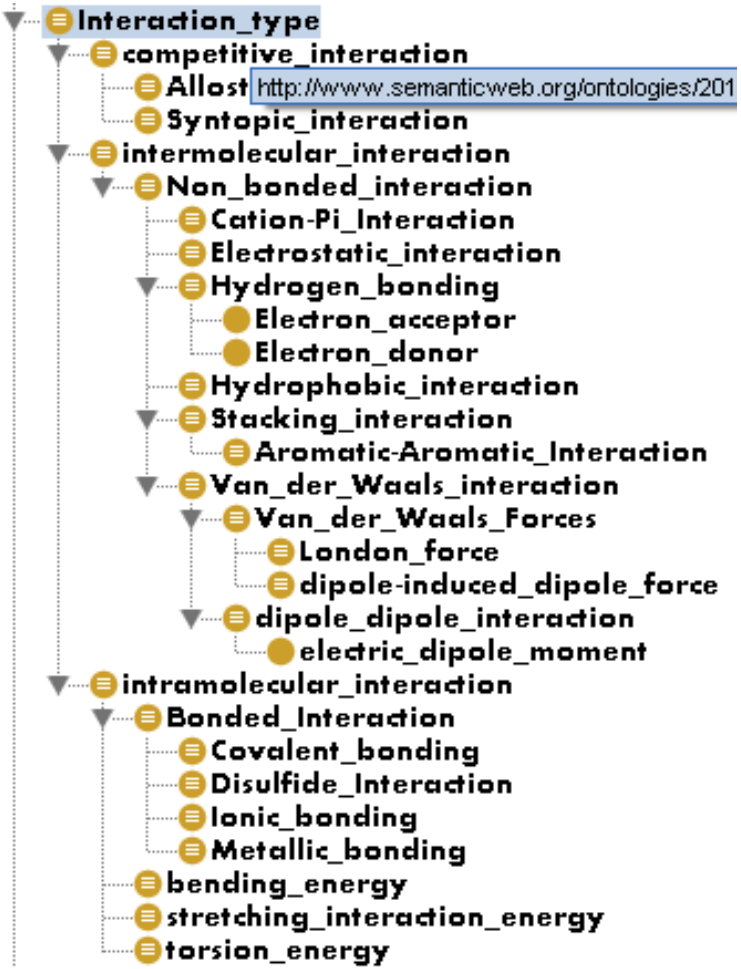
"Activation energy can be thought of as the height of the potential barrier separating two minima of potential energy (of the reactants and products of a reaction). For a chemical reaction to proceed at a reasonable rate, there should exist an appreciable number of molecules with energy equal to or greater than the activation energy.

Interaction concept

picture taken from: http://www.topsan.org/Proteins/JCSG/3ek3



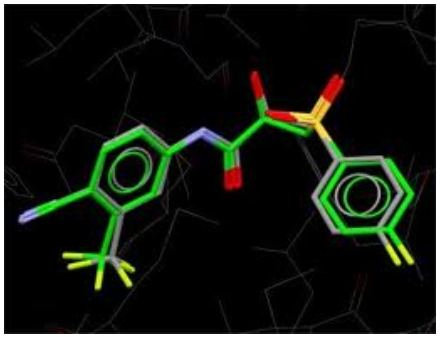


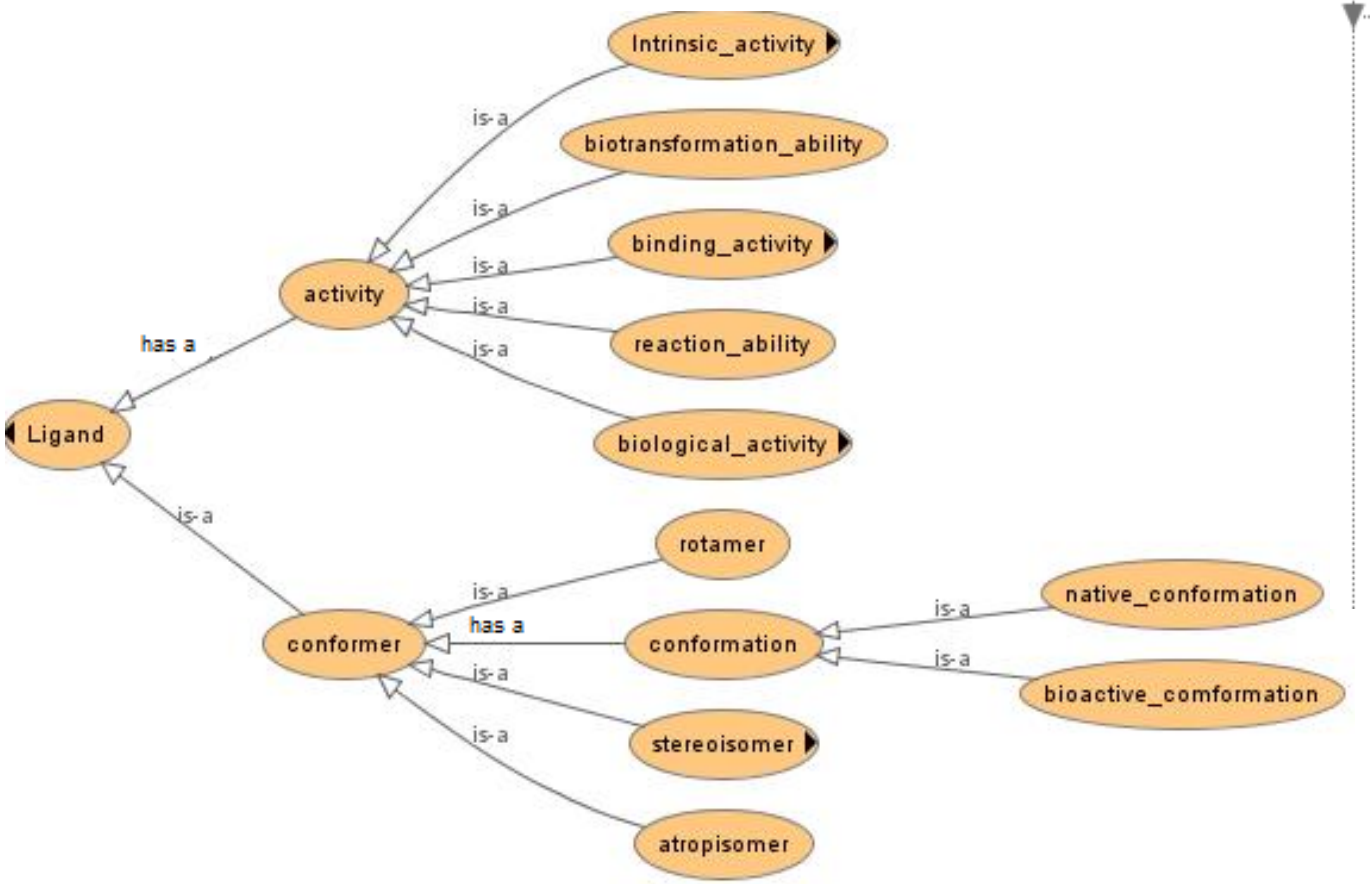


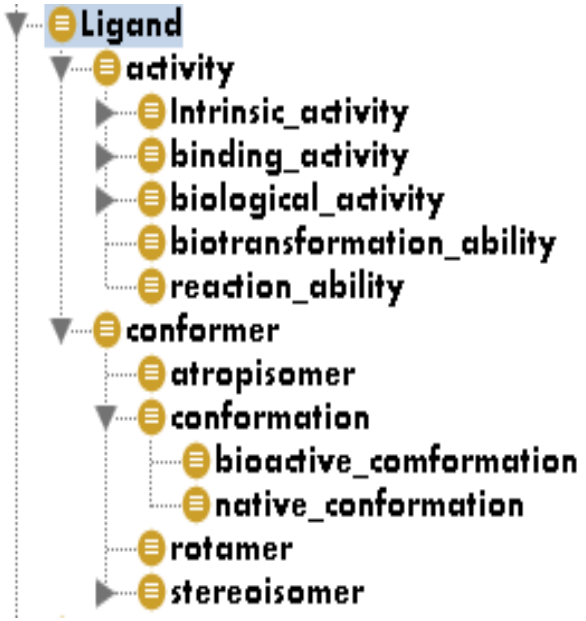
Fraunhofer SCAI

Ligand concept

picture taken from http://coolvod.net/software/927-ccdc-gold-suite-412.html



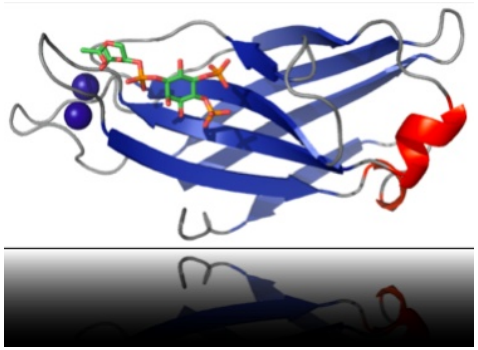


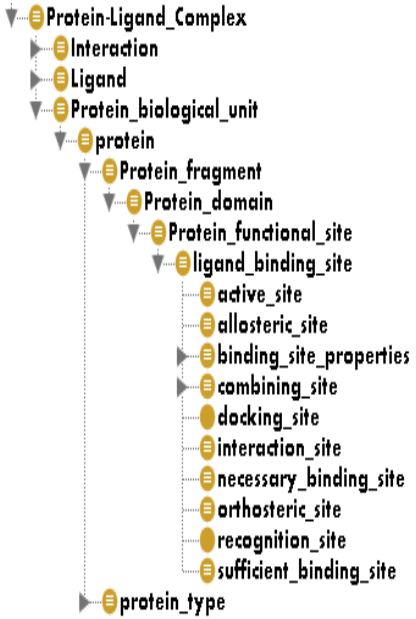


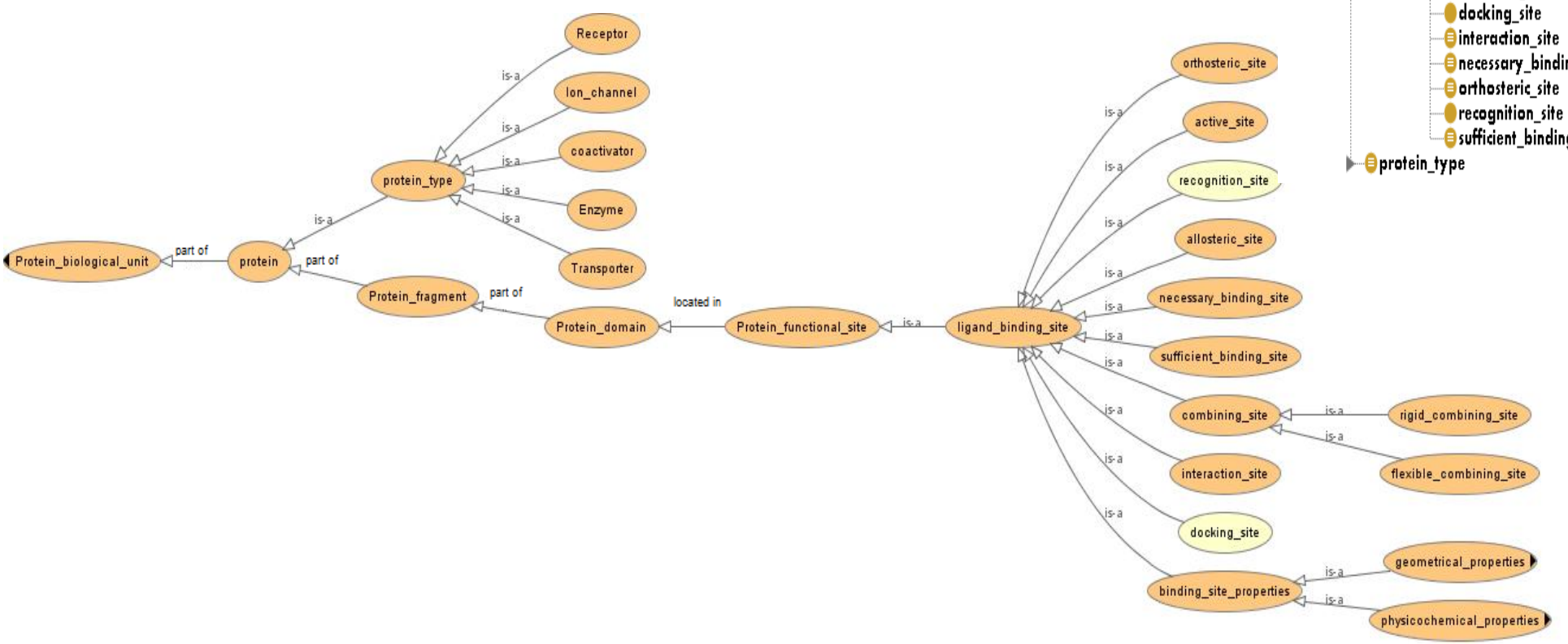
Fraunhofer SCAI

Protein concept

picture taken from: http://homepage.univie.ac.at/nicolas.coudevylle/Home.html







Fraunhofer SCAI

High Throughput Extraction of Scientific Information from Full Text Sources:

The UIMA-HPC Project

Fraunhofer SCAI

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

UIMA-HPC

Efficient Information Extraction Workflows in many-core environments

Fraunhofer SCAI

JÜLICH FORSCHUNGSZENTRUM

scapos

TAROS Chemicals

UIMA-HPC

Vision

Scientific Challenge:

The knowledge in Chemistry, Biology and Pharmaceutical Sciences growths with impressive speed. As a result, the number of publications in these areas is reaching unparalleled dimensions. However, knowledge is being communicated in non-standardised ways.

Relevant knowledge sources are not well standardized, let alone is the knowledge structured. This limits the ability to query knowledge sources.

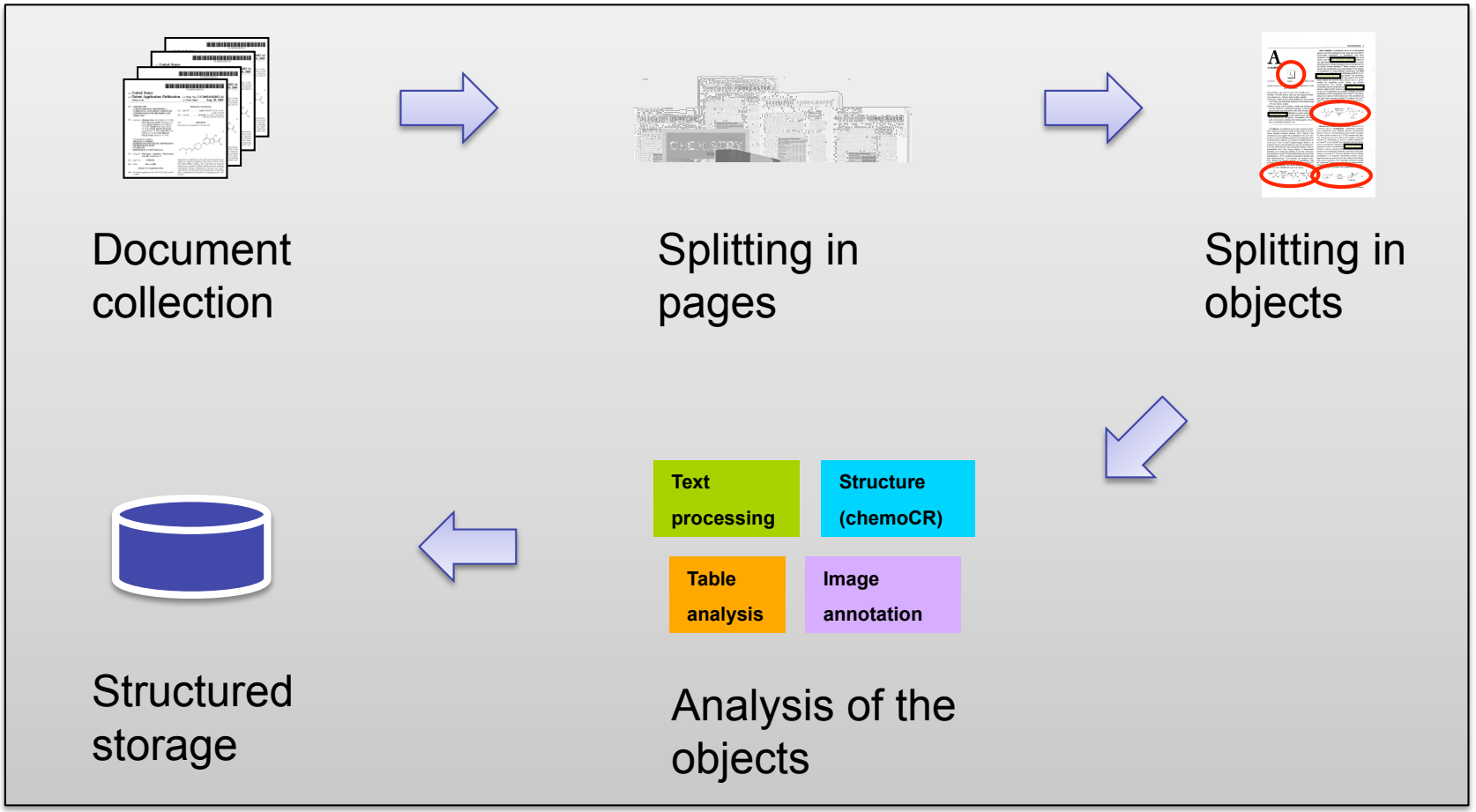
Problem-solving approach:

Development of technology that – based on HPC – allows for high throughput extraction of structured information from unstructured knowledge sources

Structured Knowledge

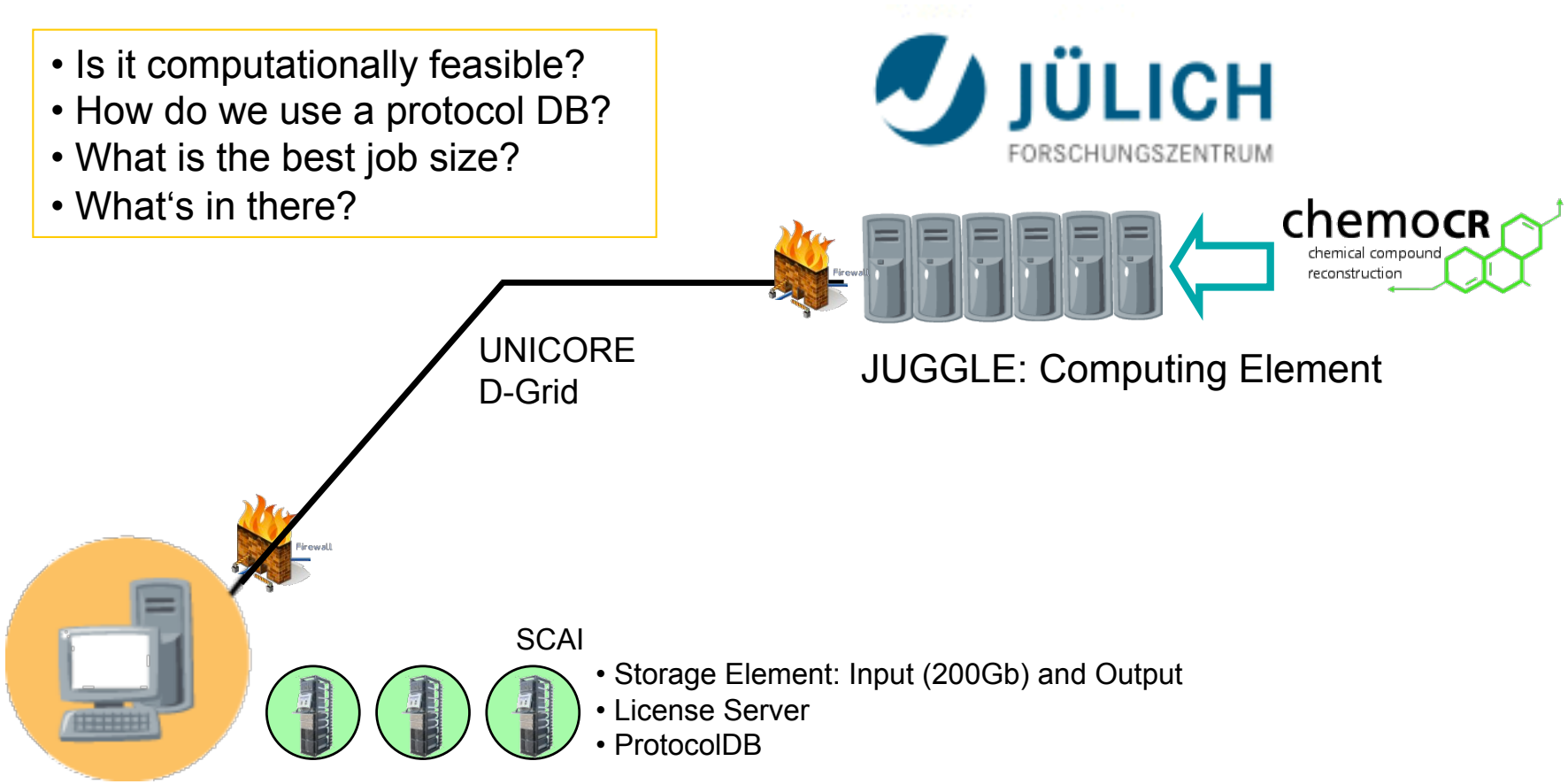
UIMA-HPC

Use case scenario:
automatic patent structuring



Previous study: The grand patent challenge 2009

- Is it computationally feasible?
- How do we use a protocol DB?
- What is the best job size?
- What's in there?

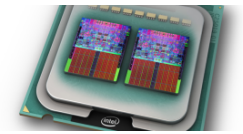


Technical Issues and Pitfalls

- User accession rights (files, scheduler, installed tools and libs, ...)
- Firewall (ports: MySQL, denial of service attack, time outs, ...)
- Missing files (NFS down, package lost, not installed, ...)
- Too many requests on license server
- Too many connections in database
- Ressources (reservation, priorities, ...)

UIMA AS in the context of HPC

Support of many-core architecture
- several instances of a service
- eff. usage of shared memory (JVM)
- asynchronous execution

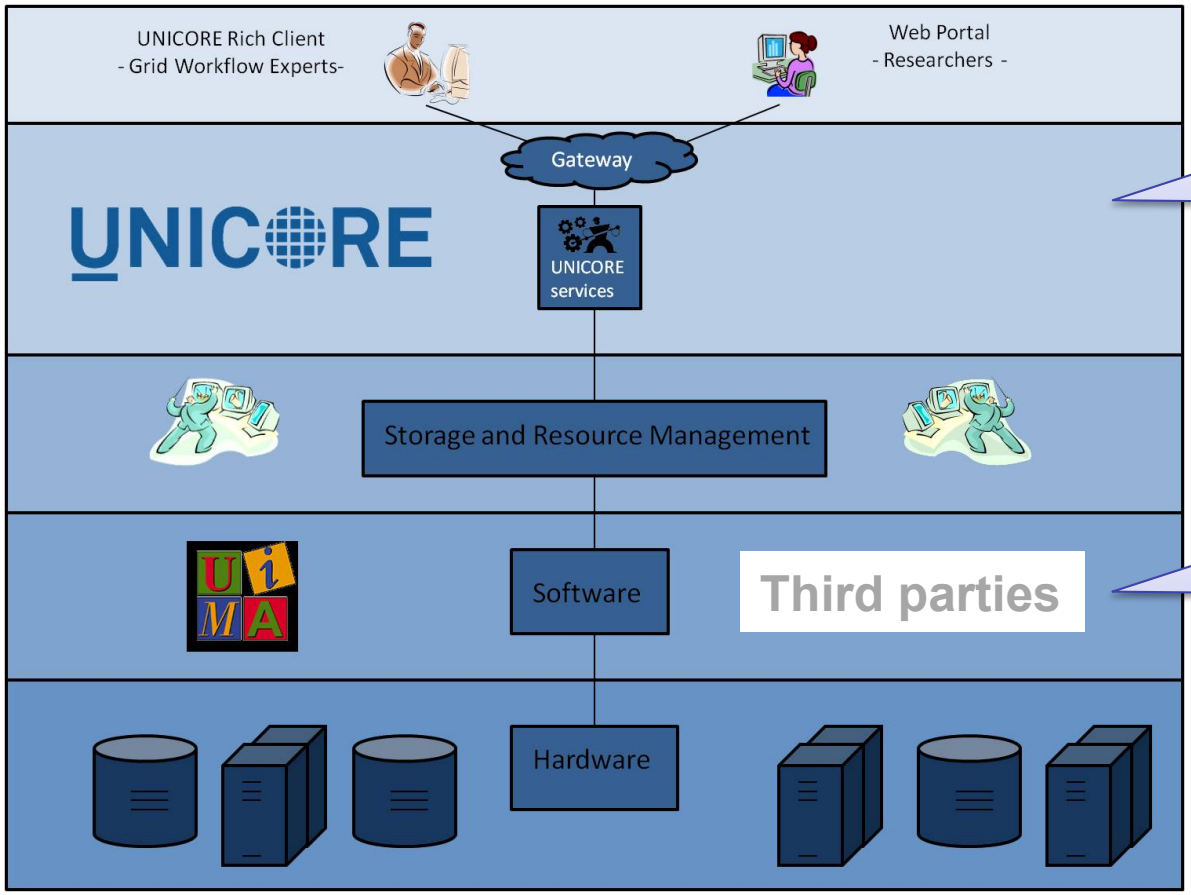


Support of clusters
- several remote services (eg SOAP)
- communication via JMX and http

Control via pre-configured parameters
- CAS pool size
- casMultiplier poolSize
- …

Manual tuning

Problem-Solving Approach



HPC

http://www.unicore.eu/

Multi Core

http://uima.apache.org/

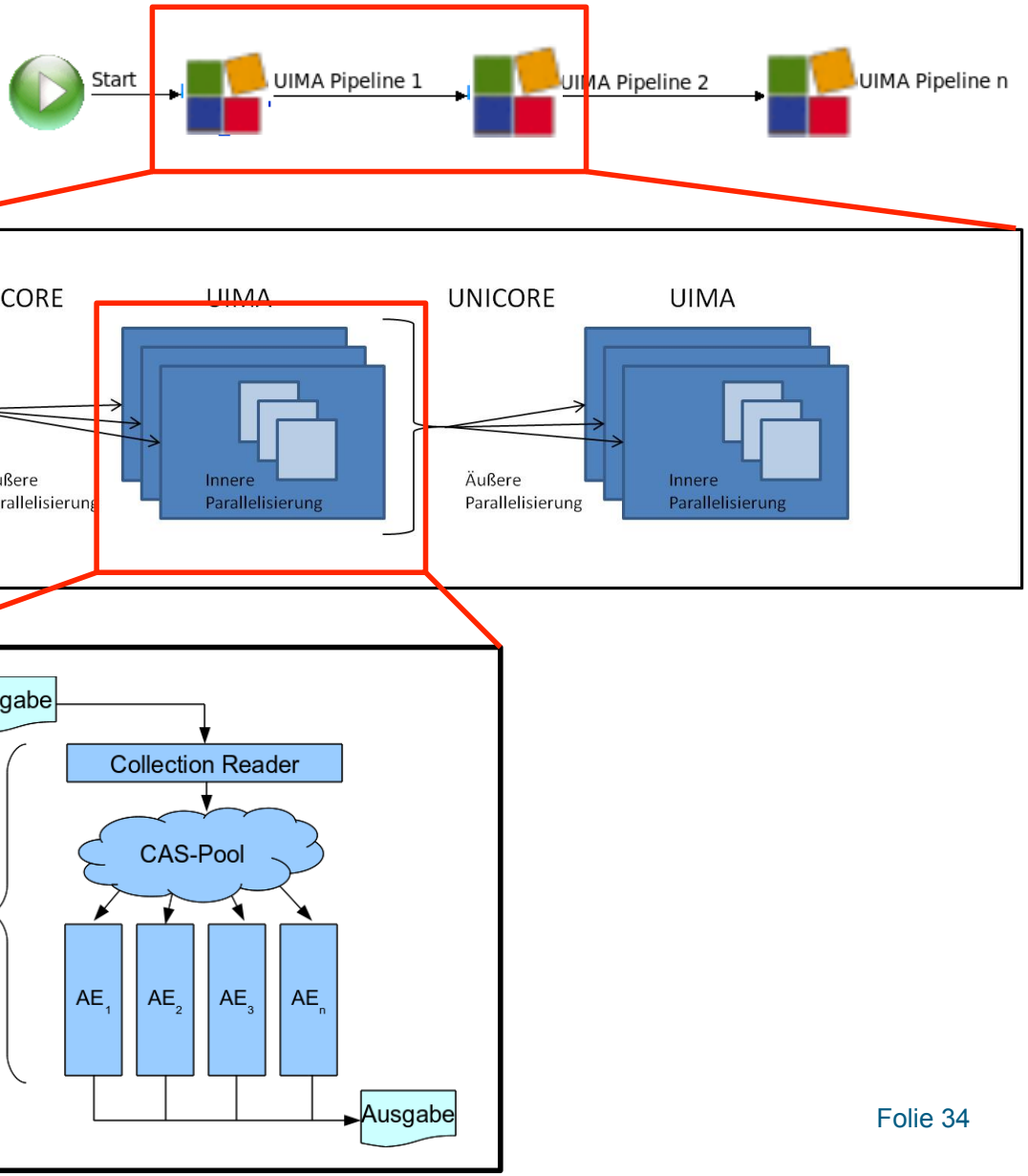
System Architecture

UIMA-HPC

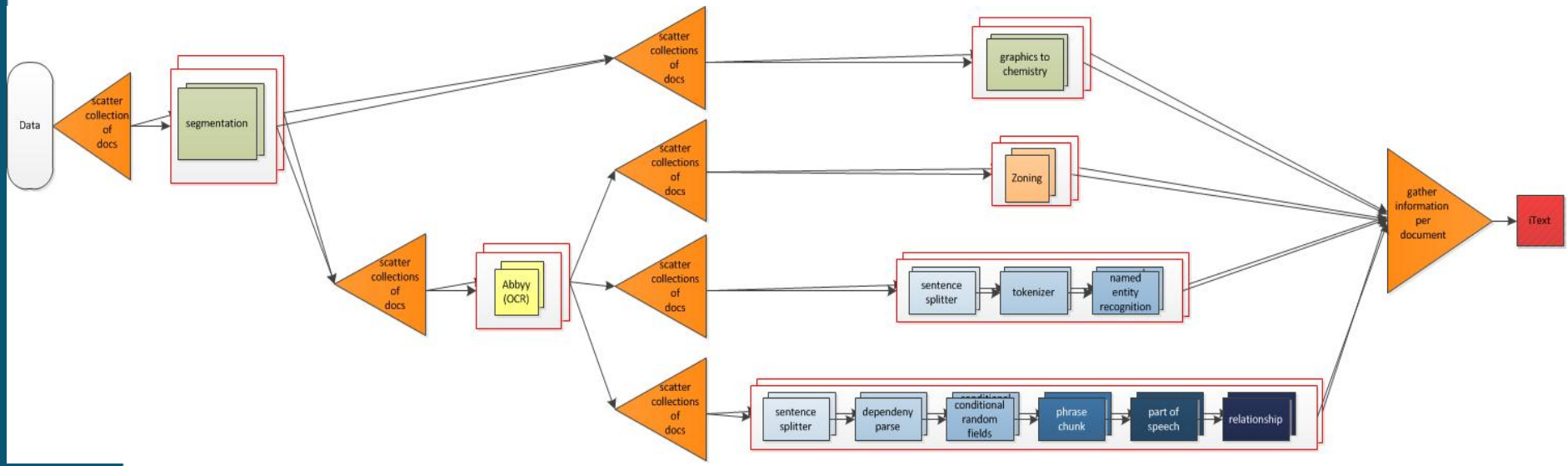
UNICORE layer

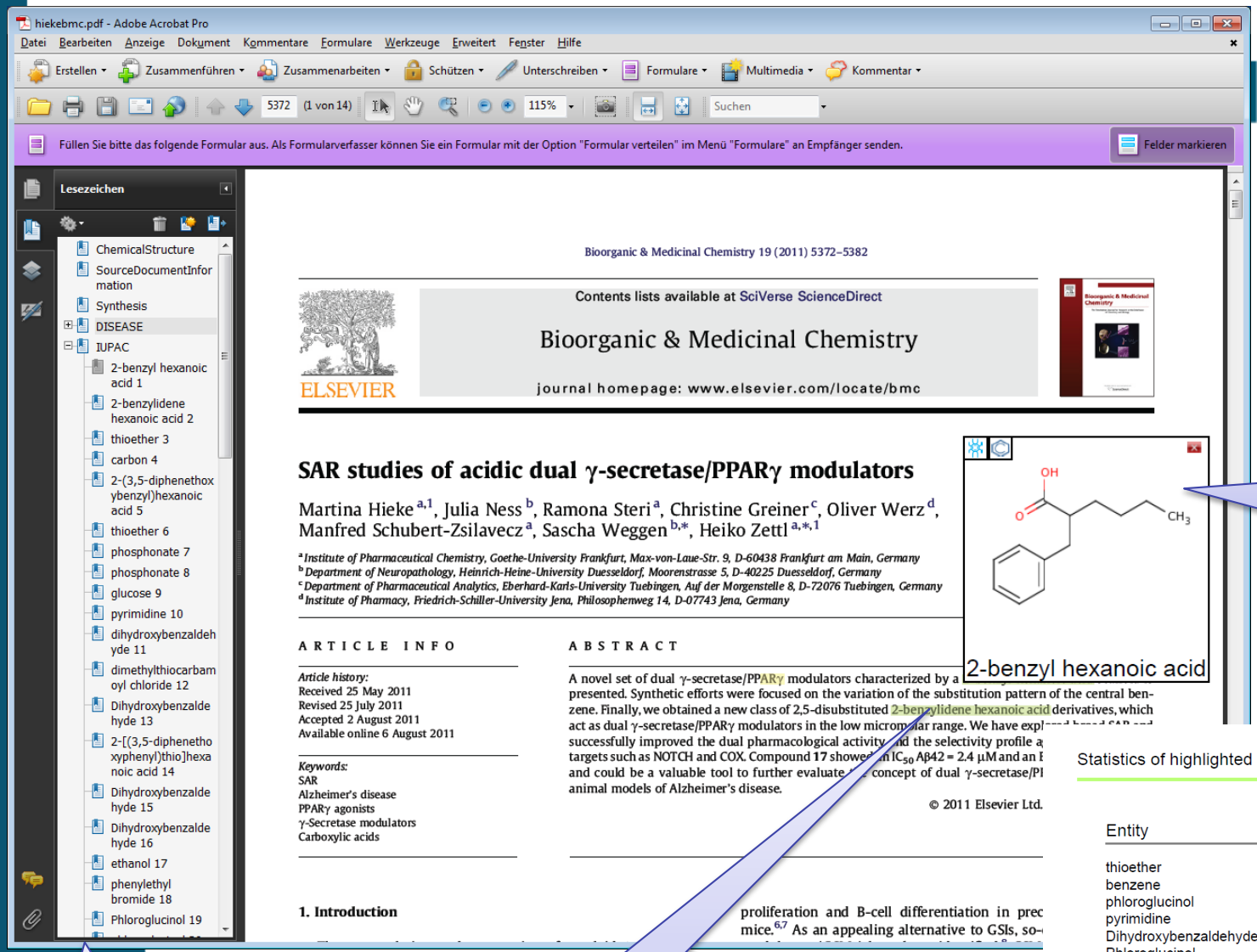
Interface layer

UIMA layer



First Prototype: PDF Annotation of Patents





Chemical Popups

Statistics of highlighted entities of class IUPAC (page 1)

Entity	Frequency
thioether	5
benzene	5
phloroglucinol	3
pyrimidine	3
Dihydroxybenzaldehyde	3
Phloroglucinol	2
phenylethanol	2
carbon	2
sodium hydride	2
2-(3,5-diphenethoxyphenoxy)hexanoic acid	2
phosphonate	2
2-benzylidene hexanoic acid	1
phenol	1
dimethylthio-carbamoylchloride	1
3-hydroxy-4-(2-cyclohexylethoxy)benzaldehyde	1
glutathione	1
2-[(3,5-diphenethoxyphenyl)thio]hexanoic acid	1
2-benzyl hexanoic acid	1
3-(2-cyclohexylethoxy)-4-phenethoxybenzal	1
tetramethylsilane	1
dimethylthiocarbamoyl chloride	1

Chemical Index

Annotations + Linkouts

Overview

Direct Usage of Unstructured Information Sources for Disease Modelling

From Medline Mining
to
Modelling Neurodegenerative Diseases

Fraunhofer
SCAI

Why Modelling of Neurodegeneration?

In 2009 the Federal Government of Germany decided to start a new research centre that focuses on translational research on neurodegenerative diseases. In fact, neurodegenerative diseases (Alzheimer, Parkinson, Multiple Sclerosis; Epilepsy; „rare“ NDDs)

The total costs of Alzheimer is estimated to exceed 20 trillion US$ in the US in the years between 2020 - 2050. (source: Alzheimer.org). Current costs / year in the US (according to Alzheimer.org): **183 billion US$**

The incidence rate of Alzheimer and other dementias is almost 50% in the population older than 85 years. Next generation will regularly have a life span of >100 years.

Fraunhofer SCAI

The Starting Conditions

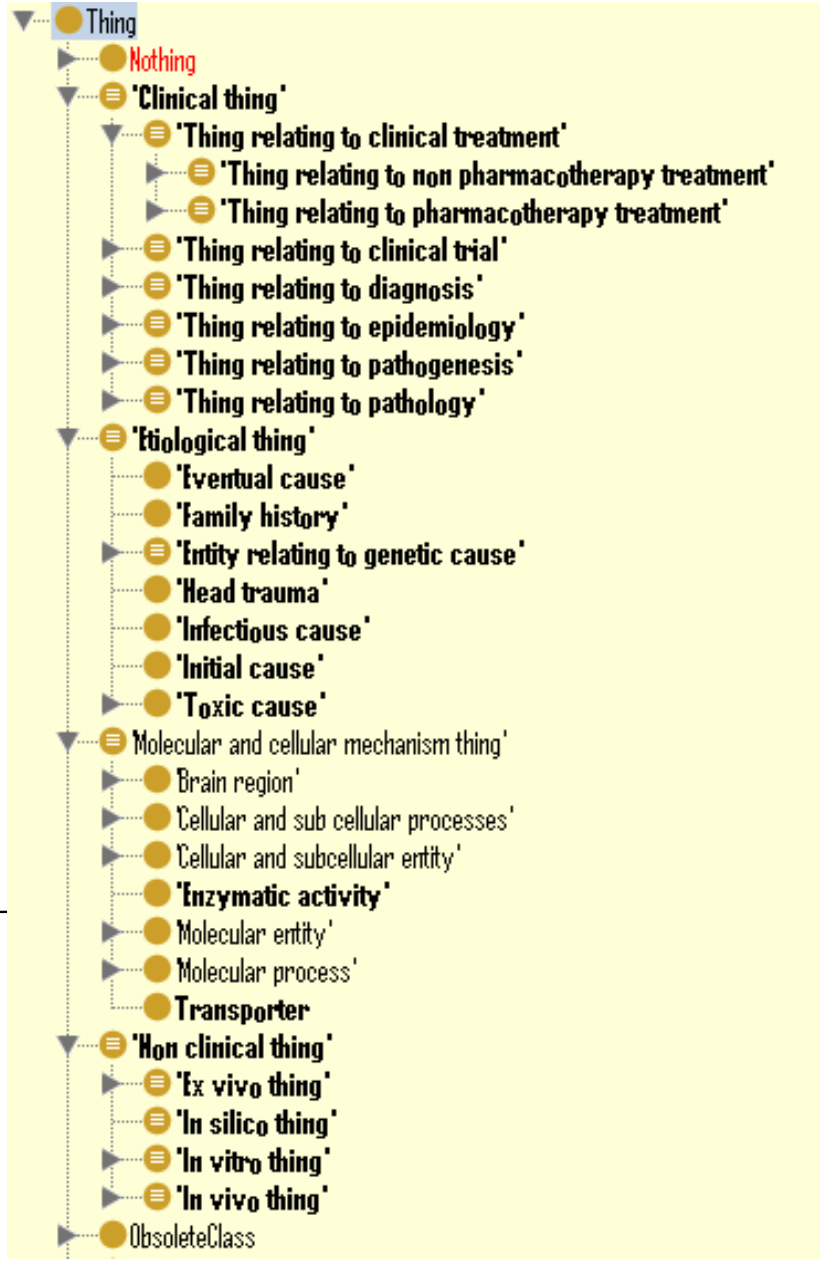
What we have:

- An ontology capturing relevant knowledge on Alzheimer´s Disease (ADO)
- An ontology representing and integrating brain regions and cell types (BRCO)
- A method for the automated identification of hypotheses in text based on regular expressions
- An excellent machinery for biomedical text mining (ProMiner) with top performing gene and protein name recognition

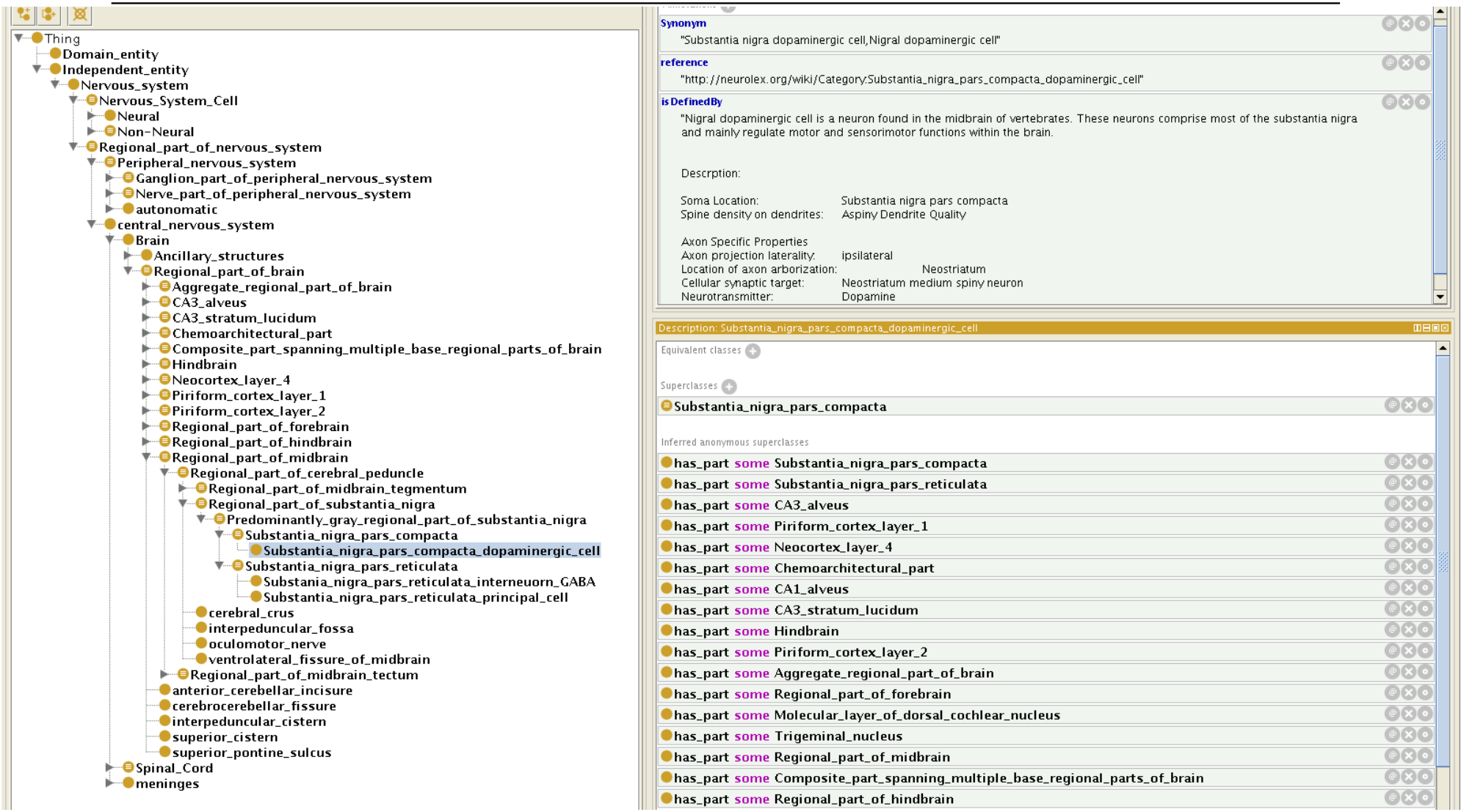
Alzheimer´s Disease Ontology (ADO)

Alzheimer's ontology:

- ❑ Captures more than 700 classes/ concepts
- ❑ BFO already implemented



Brain Region and Cell-type Ontology (BRCO)



Current state: more than 3000 concepts; more than 5000 synonyms

Expression of Speculative Statements in Scientific Text

Kallikrein-related peptidase 6 in Alzheimer's disease and vascular dementia.

PubMed 20846516 **Authors:** Ashby, Emma L; Kehoe, Patrick G; Love, Seth **Date:** 2010-12- **Journal:** Brain research **Affiliation:** Dementia esearch Group, Institute of Clinical Neurosciences, Clinical Science at North Bristol, University of Bristol, UK.

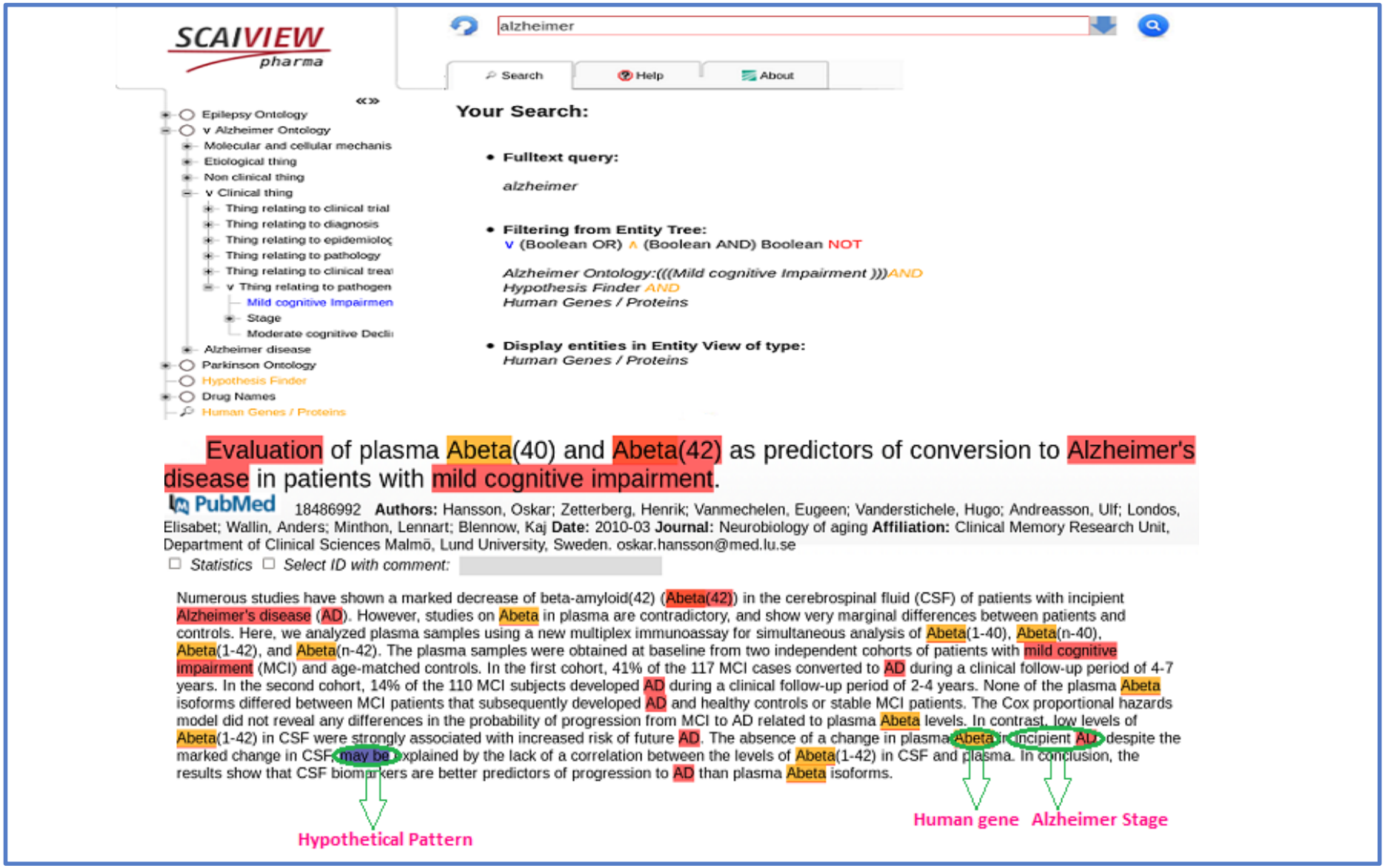
] *Statistics* ☐ *Select ID with comment:*

Human kallikrein-related peptidase 6 (KLK6) is highly expressed in the central nervous system. Although the physiological roles of this serine protease are unknown, in vitro substrates include amyloid precursor protein and components of the extracellular matrix, which are altered in neurological disease, particularly Alzheimer's disease (AD). We have compared KLK6 expression in post-mortem brain tissue in AD, vascular dementia (VaD) and controls. We studied the distribution of KLK6 in the temporal cortex and white matter by immunohistochemistry, and measured KLK6 mRNA and protein levels in the frontal and temporal cortex from 15 AD, 15 VaD and 15 control brains. Immunohistochemistry showed KLK6 to be restricted to endothelial cells. After adjustment for variations in vessel density by measurement of factor VIII-related antigen, we found KLK6 protein and mRNA levels to be significantly decreased in the frontal but not the temporal cortex in AD. In VaD, KLK6 protein level was significantly increased in the frontal cortex. Our findings suggest that an altered KLK6 expression may contribute to vascular abnormalities in AD and VaD.

Hypothesis = KLK6 + may contribute + AD

Fraunhofer SCAI

Hypotheses finder∩ AD ontology ∩Human genes and proteins

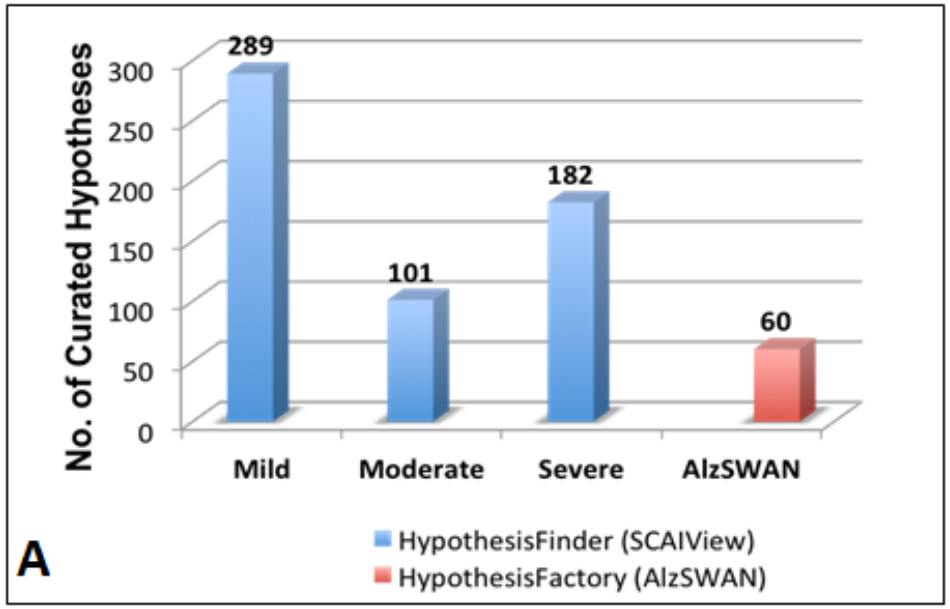


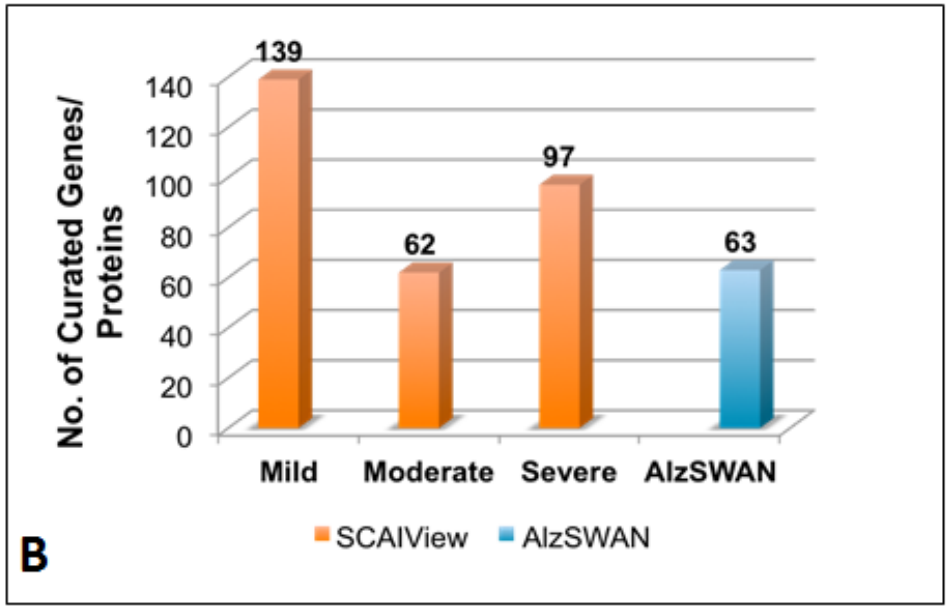
Performance of Hypotheses finder

S.No	Data type	Source	Precision	Recall	F score
1	200 abstracts related to Alzheimer's	PubMed	0.84	0.86	0.85
2	2 full text articles related to Alzheimer's	Journal of Medical Hypotheses	0.85	0.88	0.86
3	143 abstracts related to Alzheimer's	Alzswan/PubMed	0.90	0.97	0.93
4	100 abstracts related to Epilepsy	PubMed	0.96	0.91	0.94
5	100 abstracts related to Parkinson's	PubMed	0.90	0.93	0.92

Fraunhofer SCAI

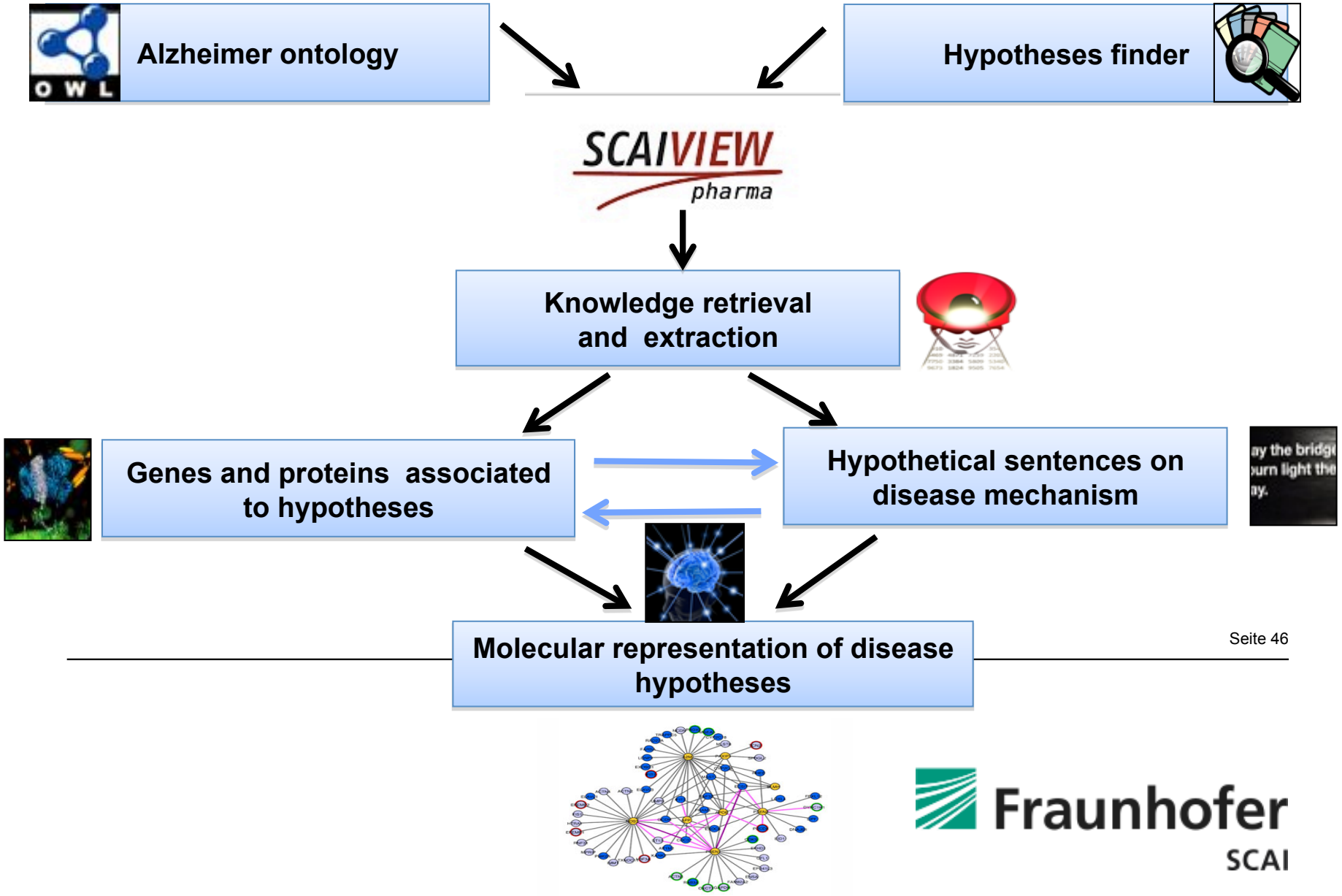
Performance of Hypotheses finder





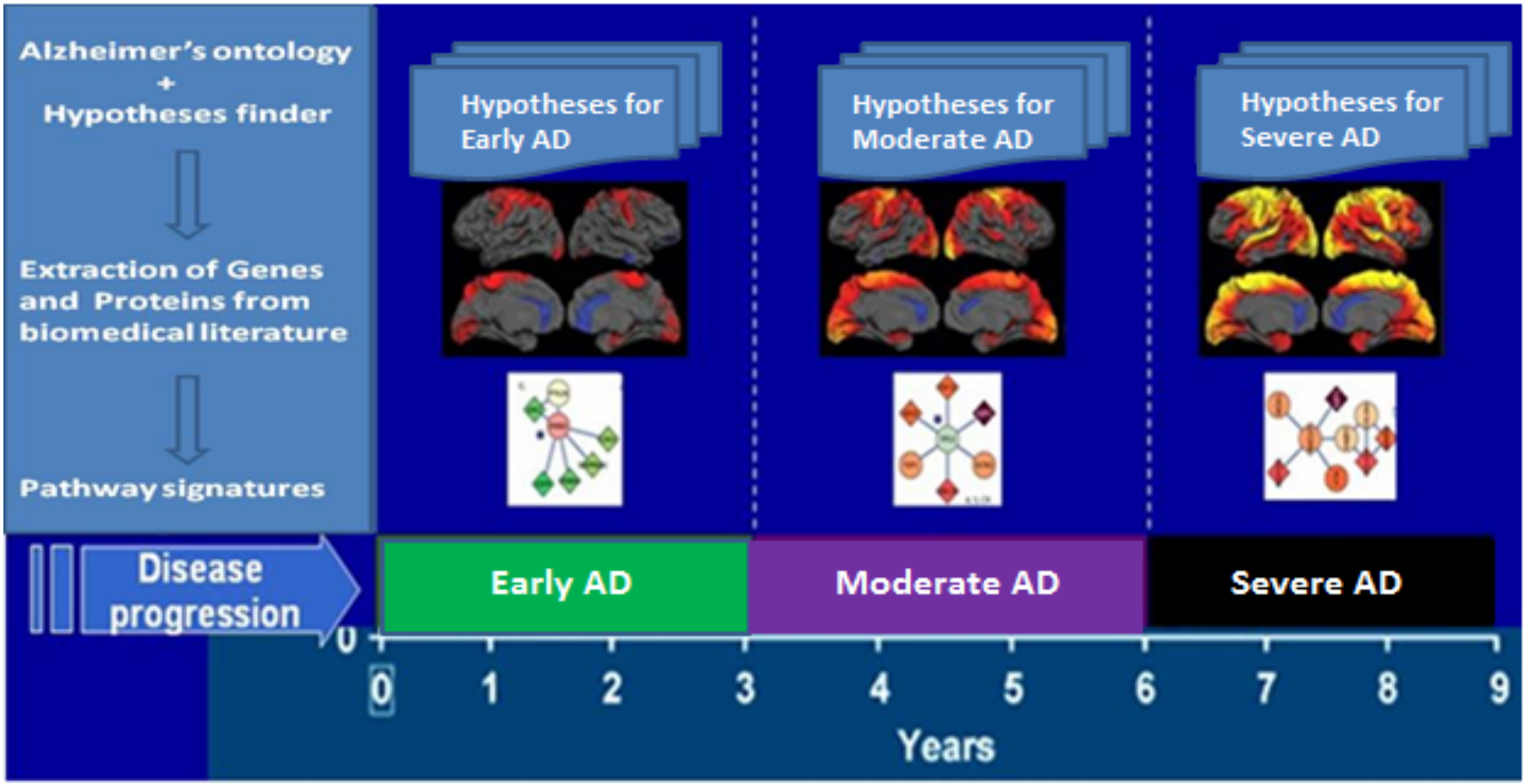
Fraunhofer SCAI

The Knowledge – Discovery Strategy



Fraunhofer
SCAI

Analysis of hypotheses patterns across disease stages



Fraunhofer SCAI

Summary

Fraunhofer SCAI Department of Bioinformatics stands for:

- Advanced technologies in text- and data mining, disease modelling in the area of neurodegeneration and high performance computing
- Trying hard to make the „eScience“ paradigm a “living experience”
- Clear dedication towards contract research and sustainable innovation
- Internal usage of technologies (information extraction; distributed and high performance computing) for biomedical application

Fraunhofer SCAI