“From (Text) Mining to Models: Applying Large-Scale Text Mining on Patents and Electronic Patient Records“

Fraunhofer

SCAI

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

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

High Throughput Extraction of Scientific Information from Full Text Sources:

The UIMA-HPC Project

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

UIMA-HPC

Efficient Information Extraction Workflows in many-core environments

Fraunhofer SCAI

JÜLICH FORSCHUNGSZENTRUM

scapos

TAROS Chemicals

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.

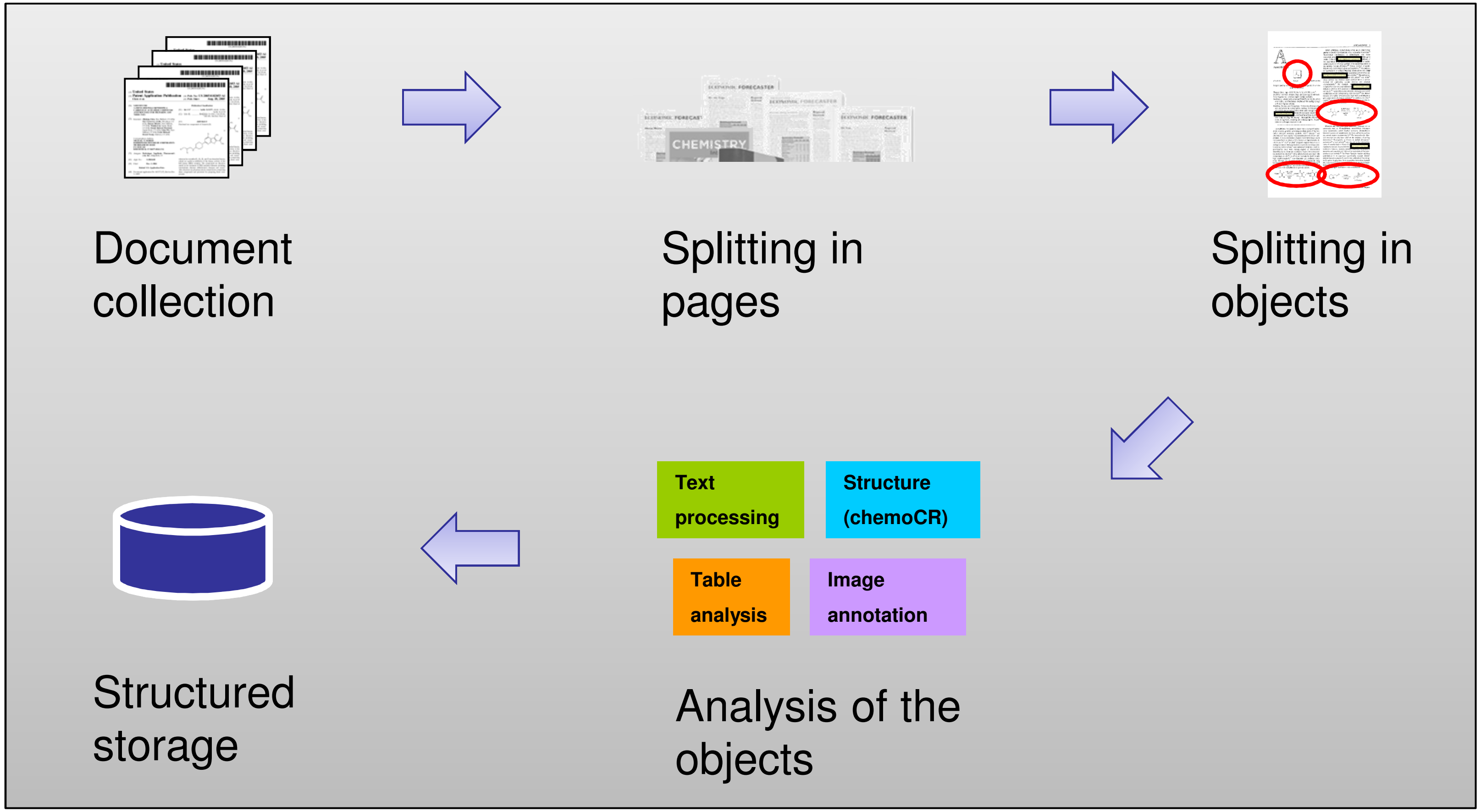
Relevant knowledge sources are not well standardized, let alone is the knowledge structured. This limits the ability to query knowledge sources.

Problem-solving approach:

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

Structured Knowledge

Use case scenario:
automatic patent structuring



Document collection

Splitting in pages

Splitting in objects

Text processing

Structure (chemoCR)

Table analysis

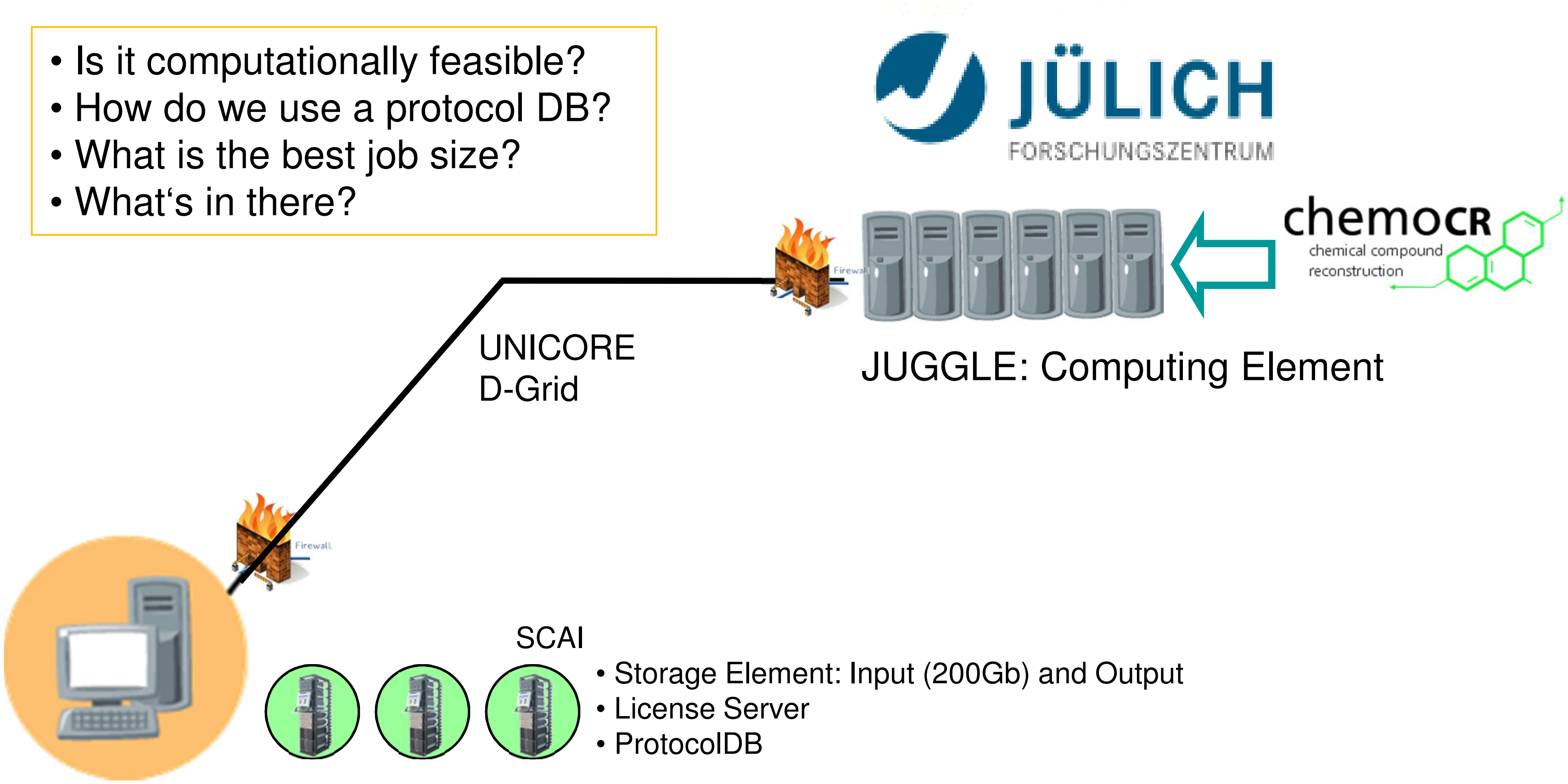
Image annotation

Analysis of the objects

Structured storage

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?



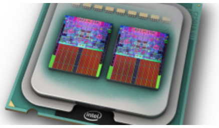
- Storage Element: Input (200Gb) and Output
- License Server
- ProtocolDB

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

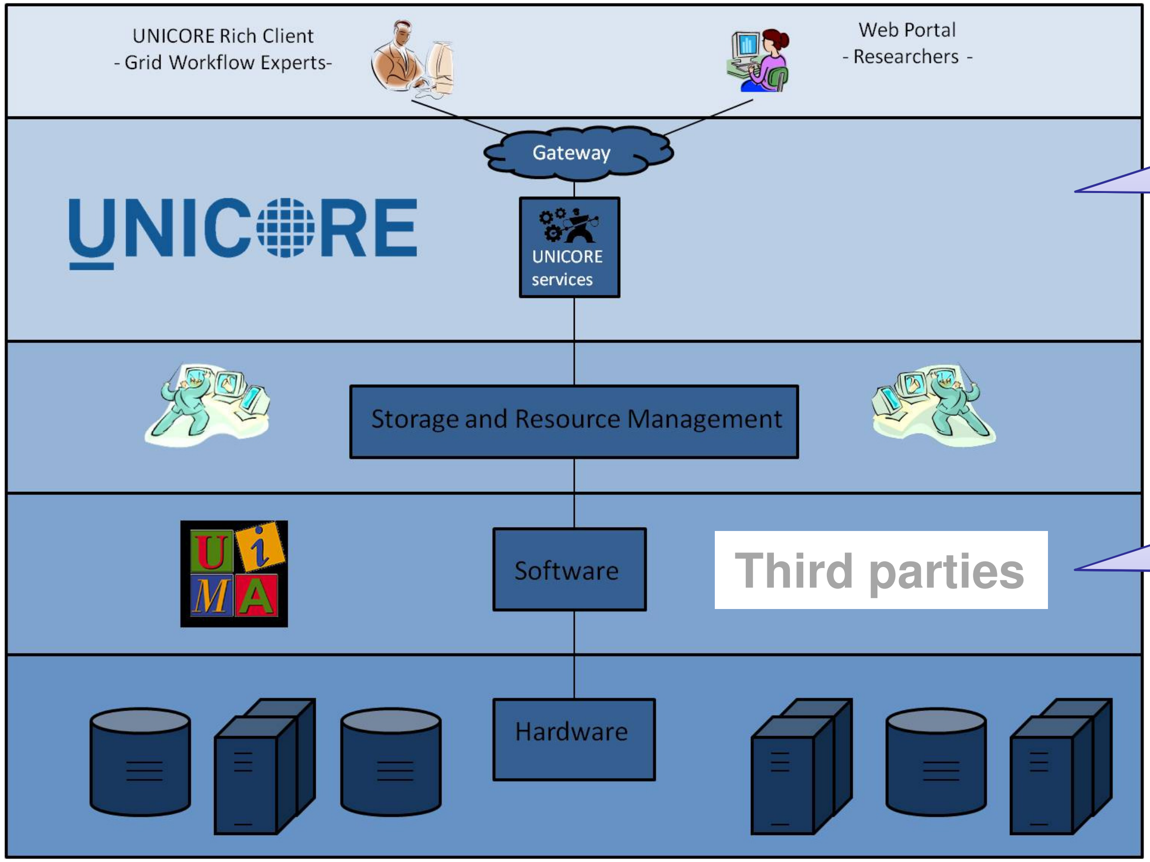


Support of clusters
- several remote services (eg SOAP)
- communication via JMX and http

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/

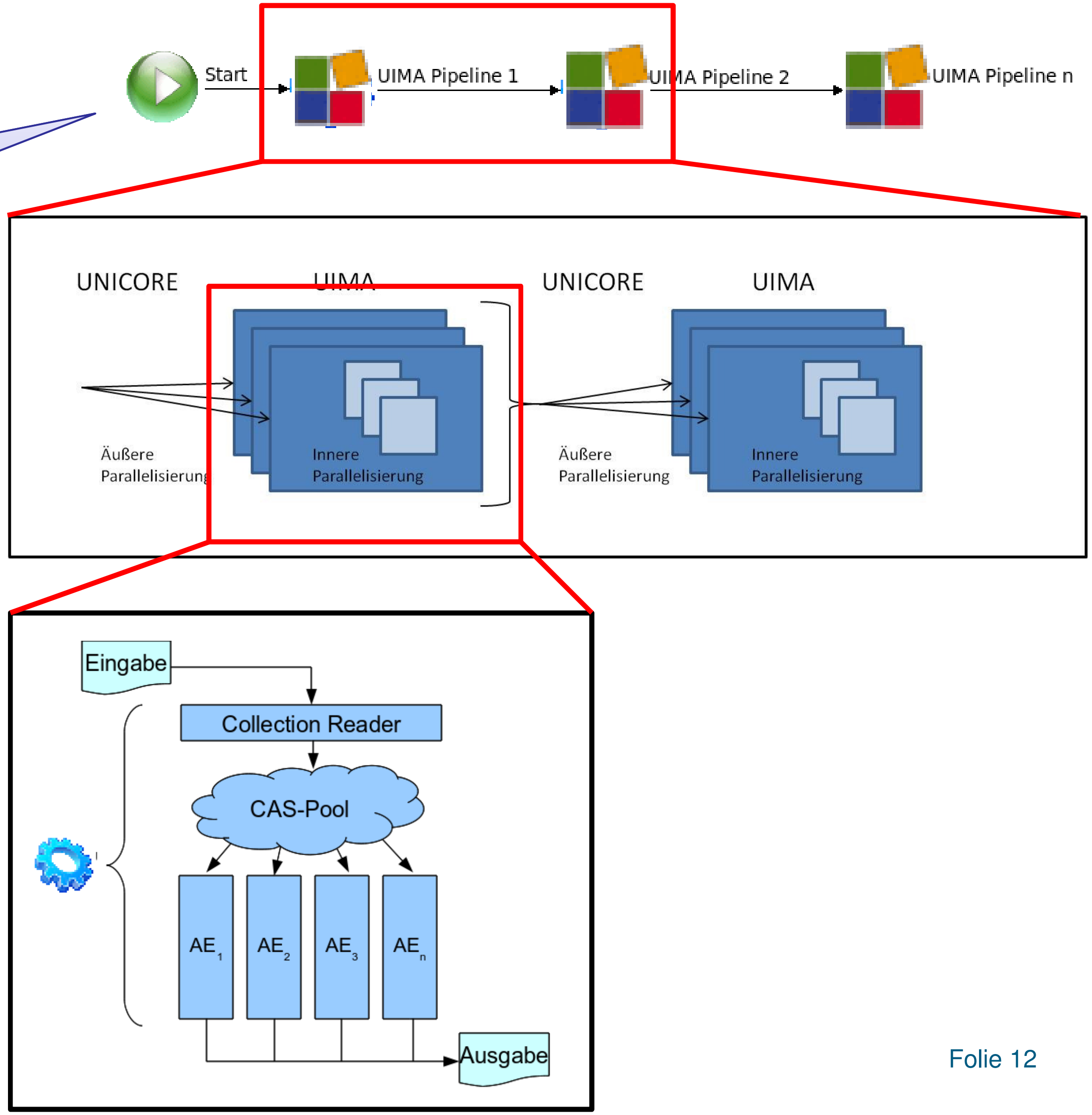
System Architecture

UIMA-HPC

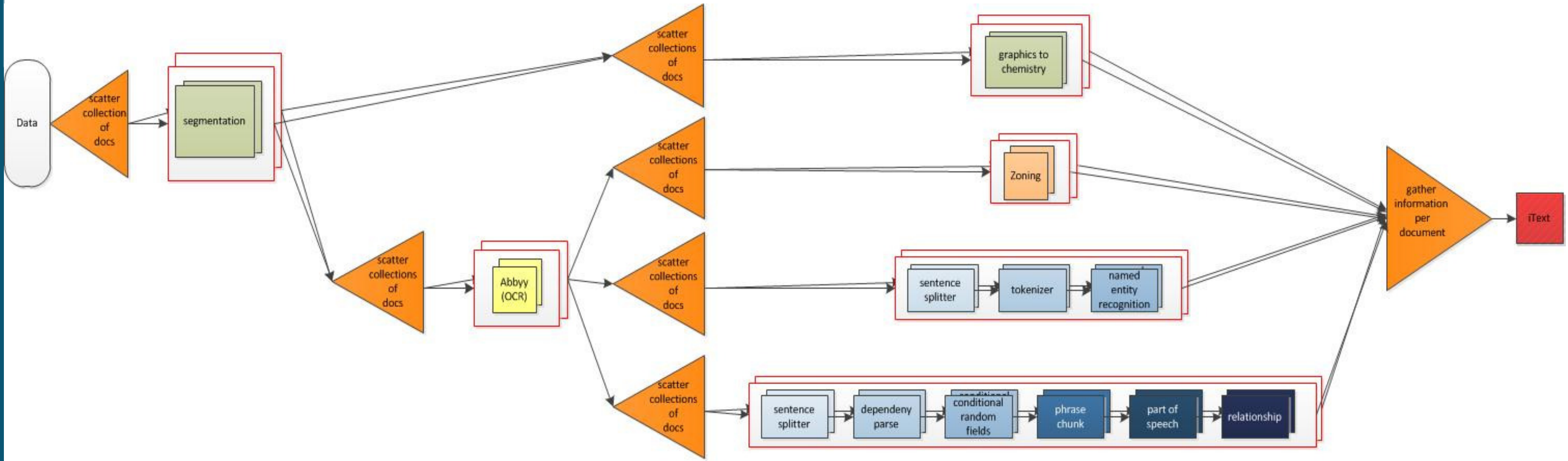
UNICORE layer

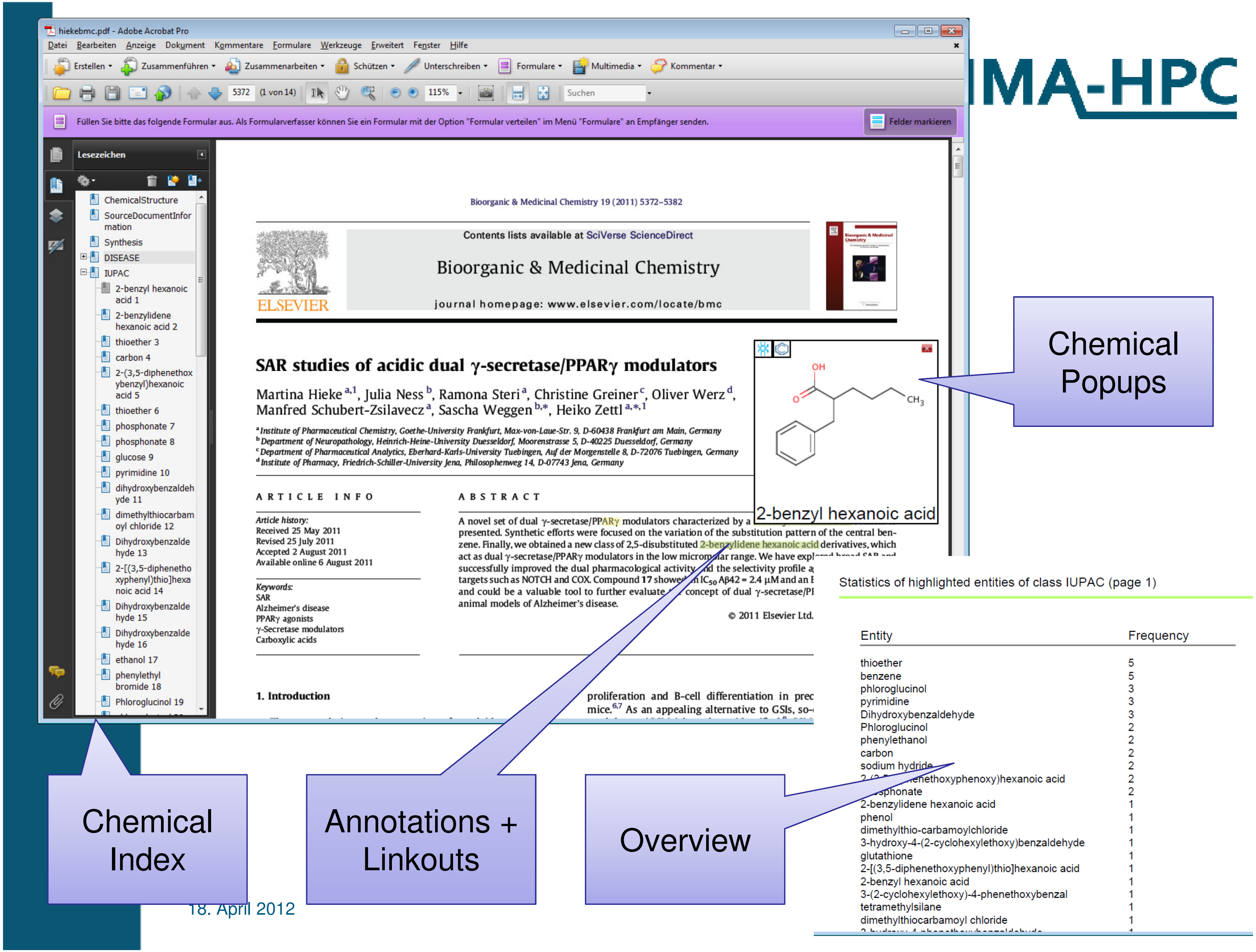
Interface layer

UIMA layer



First Prototype: PDF Annotation of Patents


Data
scatter collection of docs
segmentation
scatter collections of docs
graphics to chemistry
scatter collections of docs
Zoning
scatter collections of docs
Abbyy (OCR)
scatter collections of docs
sentence splitter
tokenizer
named entity recognition
scatter collections of docs
sentence splitter
dependeny parse
conditional random fields
phrase chunk
part of speech
relationship
gather information per document
iText


hiekebmc.pdf - Adobe Acrobat Pro
Datei Bearbeiten Anzeige Dokument Kommentare Formulare Werkzeuge Erweitert Fenster Hilfe
Erstellen · Zusammenführen · Zusammenarbeiten · Schützen · Unterschreiben · Formulare · Multimedia · Kommentar
5372 (1 von 14) 115% Suchen
Felder markieren
Lesezeichen

Füllen Sie bitte das folgende Formular aus. Als Formularverfasser können Sie ein Formular mit der Option "Formular verteilen" im Menü "Formulare" an Empfänger senden.

- ChemicalStructure
- SourceDocumentInformation
- Synthesis
- DISEASE
- IUPAC
  - 2-benzyl hexanoic acid 1
  - 2-benzylidene hexanoic acid 2
  - thioether 3
  - carbon 4
  - 2-(3,5-diphenethoxybenzyl)hexanoic acid 5
  - thioether 6
  - phosphonate 7
  - phosphonate 8
  - glucose 9
  - pyrimidine 10
  - dihydroxybenzaldehyde 11
  - dimethylthiocarbamoyl chloride 12
  - Dihydroxybenzaldehyde 13
  - 2-[(3,5-diphenethoxyphenyl)thio]hexanoic acid 14
  - Dihydroxybenzaldehyde 15
  - Dihydroxybenzaldehyde 16
  - ethanol 17
  - phenylethyl bromide 18
  - Phloroglucinol 19

Bioorganic & Medicinal Chemistry 19 (2011) 5372–5382

Contents lists available at SciVerse ScienceDirect

Bioorganic & Medicinal Chemistry

journal homepage: www.elsevier.com/locate/bmc

ELSEVIER

SAR studies of acidic dual γ-secretase/PPARγ modulators

Martina Hieke [a,1], Julia Ness [b], Ramona Steri [a], Christine Greiner [c], Oliver Werz [d], Manfred Schubert-Zsilavecz [a], Sascha Weggen [b,*], Heiko Zettl [a,*,1]

[a] Institute of Pharmaceutical Chemistry, Goethe-University Frankfurt, Max-von-Laue-Str. 9, D-60438 Frankfurt am Main, Germany
[b] Department of Neuropathology, Heinrich-Heine-University Duesseldorf, Moorenstrasse 5, D-40225 Duesseldorf, Germany
[c] Department of Pharmaceutical Analytics, Eberhard-Karls-University Tuebingen, Auf der Morgenstelle 8, D-72076 Tuebingen, Germany
[d] Institute of Pharmacy, Friedrich-Schiller-University Jena, Philosophenweg 14, D-07743 Jena, Germany

ARTICLE INFO

Article history:
Received 25 May 2011
Revised 25 July 2011
Accepted 2 August 2011
Available online 6 August 2011

Keywords:
SAR
Alzheimer's disease
PPARγ agonists
γ-Secretase modulators
Carboxylic acids

ABSTRACT

A novel set of dual γ-secretase/PPARγ modulators characterized by a presented. Synthetic efforts were focused on the variation of the substitution pattern of the central benzene. Finally, we obtained a new class of 2,5-disubstituted 2-benzylidene hexanoic acid derivatives, which act as dual γ-secretase/PPARγ modulators in the low micromolar range. We have expl successfully improved the dual pharmacological activity and the selectivity profile a targets such as NOTCH and COX. Compound **17** showed an IC_{50} Aβ42 = 2.4 μM and an and could be a valuable tool to further evaluate the concept of dual γ-secretase/P animal models of Alzheimer's disease.

© 2011 Elsevier Ltd.

1. Introduction

proliferation and B-cell differentiation in prec mice.[6,7] As an appealing alternative to GSIs, so-

2-benzyl hexanoic acid

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
3-hydroxy-4-phenethoxybenzaldehyde	1

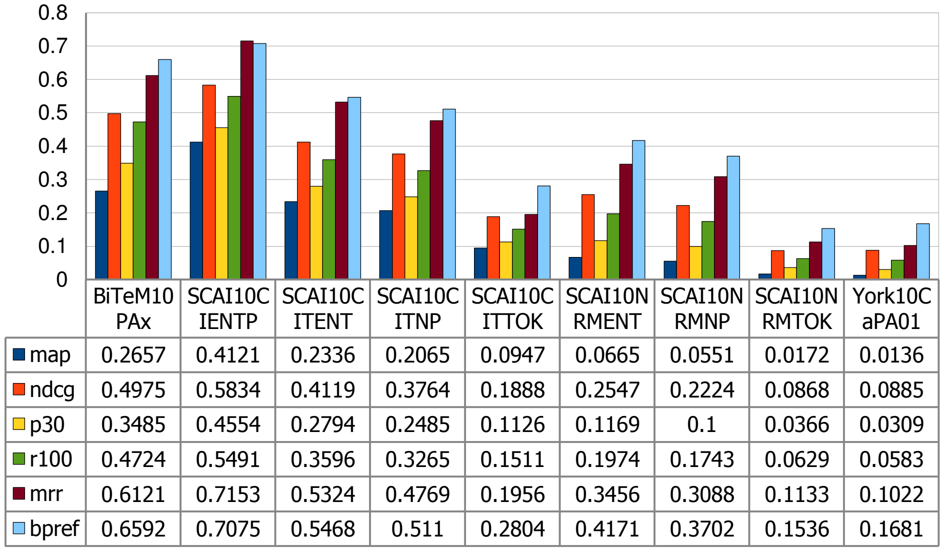
Chemical Popups

Chemical Index

Annotations + Linkouts

Overview

Performance of Fraunhofer SCAI in international Benchmarking Competitions



	BiTeM10 PAx	SCAI10C IENTP	SCAI10C ITENT	SCAI10C ITNP	SCAI10C ITTOK	SCAI10N RMENT	SCAI10N RMNP	SCAI10N RMTOK	York10C aPA01
🟦 map	0.2657	0.4121	0.2336	0.2065	0.0947	0.0665	0.0551	0.0172	0.0136
🟧 ndcg	0.4975	0.5834	0.4119	0.3764	0.1888	0.2547	0.2224	0.0868	0.0885
🟨 p30	0.3485	0.4554	0.2794	0.2485	0.1126	0.1169	0.1	0.0366	0.0309
🟩 r100	0.4724	0.5491	0.3596	0.3265	0.1511	0.1974	0.1743	0.0629	0.0583
🟥 mrr	0.6121	0.7153	0.5324	0.4769	0.1956	0.3456	0.3088	0.1133	0.1022
🟦 bpref	0.6592	0.7075	0.5468	0.511	0.2804	0.4171	0.3702	0.1536	0.1681

TRECCHEM 2009 → better not talk about it ...

TRECCHEM 2010 → winner of the Prior Art Search Task

TRECCHEM 2011 → winner of the Technology Survey Task

I2B2 challenge 2010 → rank 3 (out of 22) in the concept id task

TRECMED 2011 → rank 4 (amongst 24 participants)

Direct Usage of Unstructured Information Sources for Disease Modelling

From Medline Mining
to
Modelling Neurodegenerative Diseases

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)

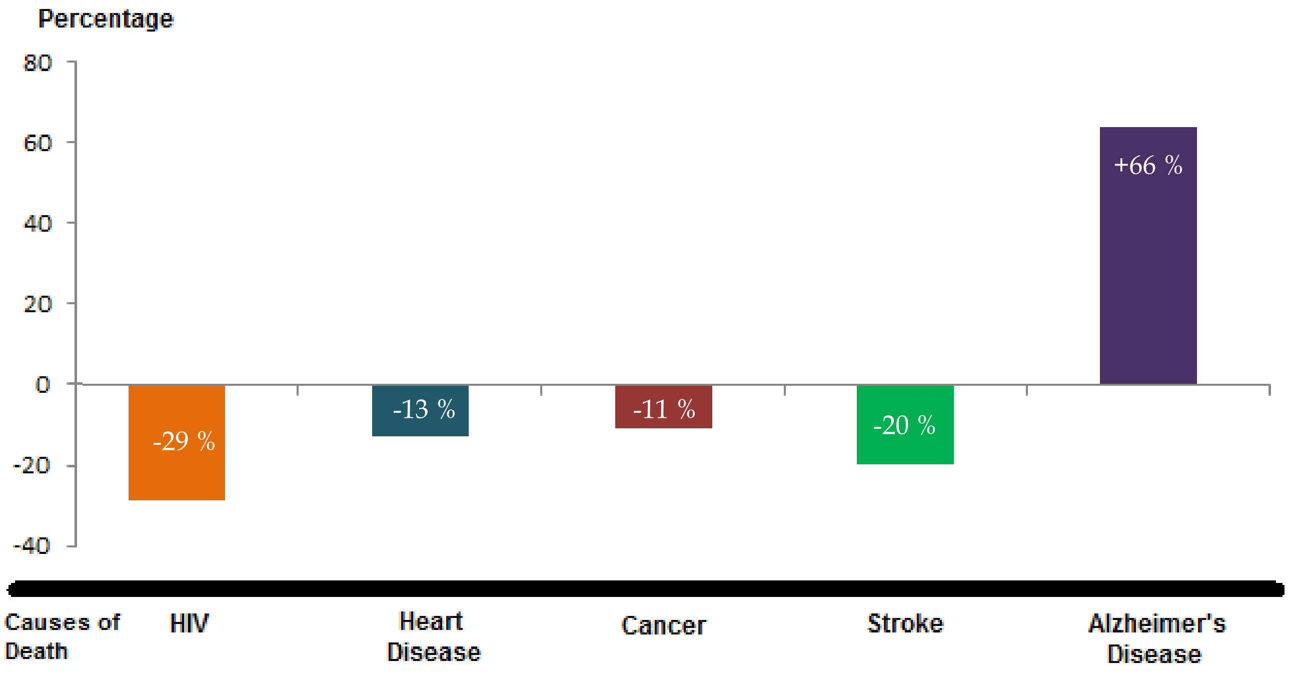
The total costs of Alzheimer is estimated to exceed 20 trillion US$ in the US in the years between 2020 - 2050. (source: Alzheimer.org). Current costs / year in the US (according to Alzheimer.org): **183 billion US$**

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

Diseases specific mortality rate

Changes in selected causes of death in USA , 2000-2010[1]



[1] www.alz.org

The Starting Conditions

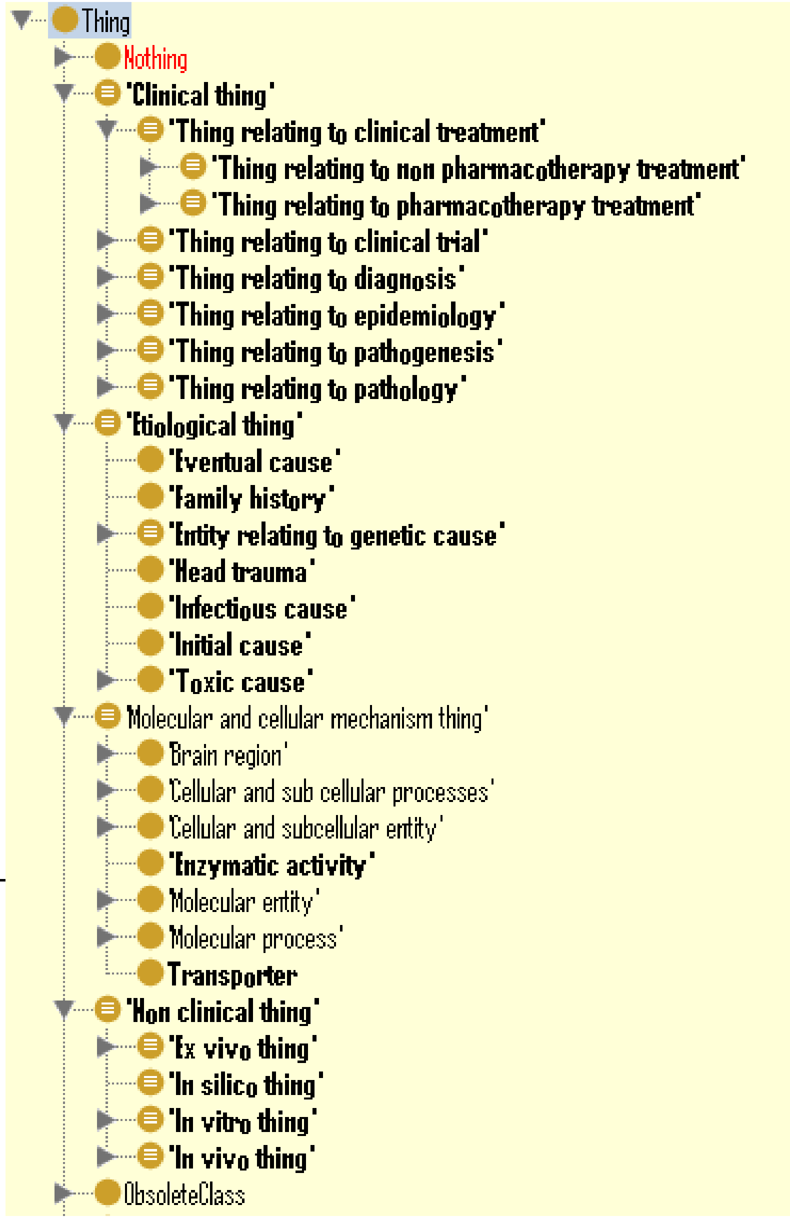
What we have:

- 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)

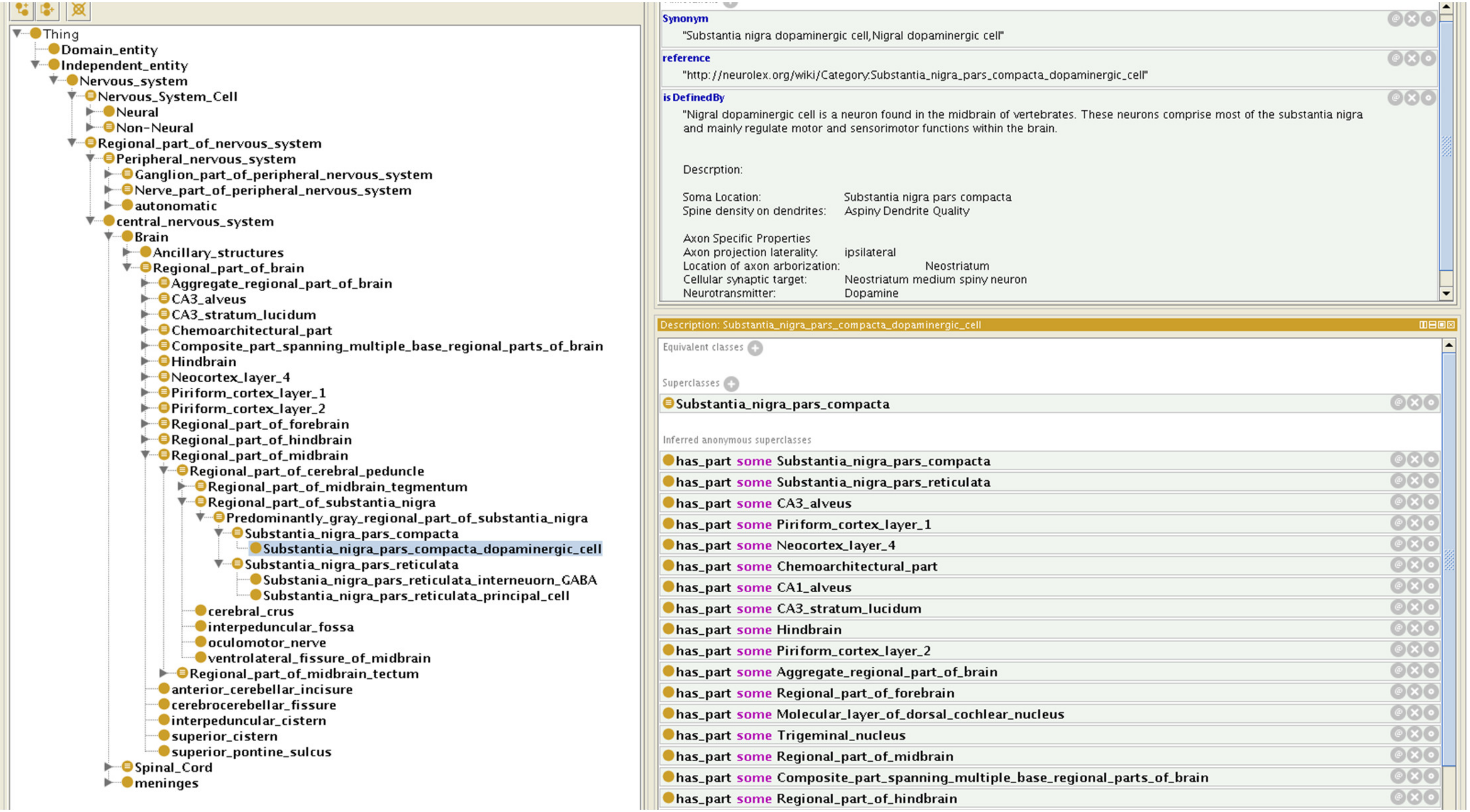
Alzheimer’s ontology:

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



- Thing
  - Nothing
  - 'Clinical thing'
    - 'Thing relating to clinical treatment'
      - 'Thing relating to non pharmacotherapy treatment'
      - 'Thing relating to pharmacotherapy treatment'
    - 'Thing relating to clinical trial'
    - 'Thing relating to diagnosis'
    - 'Thing relating to epidemiology'
    - 'Thing relating to pathogenesis'
    - 'Thing relating to pathology'
  - 'Etiological thing'
    - 'Eventual cause'
    - 'Family history'
    - 'Entity relating to genetic cause'
    - 'Head trauma'
    - 'Infectious cause'
    - 'Initial cause'
    - 'Toxic cause'
  - 'Molecular and cellular mechanism thing'
    - 'Brain region'
    - 'Cellular and sub cellular processes'
    - 'Cellular and subcellular entity'
    - 'Enzymatic activity'
    - 'Molecular entity'
    - 'Molecular process'
    - Transporter
  - 'Non clinical thing'
    - 'Ex vivo thing'
    - 'In silico thing'
    - 'In vitro thing'
    - 'In vivo thing'
  - ObsoleteClass

Brain Region and Cell-type Ontology (BRCO)



```
Thing
  Domain_entity
  Independent_entity
    Nervous_system
      Nervous_System_Cell
        Neural
        Non-Neural
      Regional_part_of_nervous_system
        Peripheral_nervous_system
          Ganglion_part_of_peripheral_nervous_system
          Nerve_part_of_peripheral_nervous_system
          autonomatic
        central_nervous_system
          Brain
            Ancillary_structures
            Regional_part_of_brain
              Aggregate_regional_part_of_brain
              CA3_alveus
              CA3_stratum_lucidum
              Chemoarchitectural_part
              Composite_part_spanning_multiple_base_regional_parts_of_brain
              Hindbrain
              Neocortex_layer_4
              Piriform_cortex_layer_1
              Piriform_cortex_layer_2
              Regional_part_of_forebrain
              Regional_part_of_hindbrain
              Regional_part_of_midbrain
                Regional_part_of_cerebral_peduncle
                  Regional_part_of_midbrain_tegmentum
                  Regional_part_of_substantia_nigra
                    Predominantly_gray_regional_part_of_substantia_nigra
                      Substantia_nigra_pars_compacta
                        Substantia_nigra_pars_compacta_dopaminergic_cell
                      Substantia_nigra_pars_reticulata
                        Substania_nigra_pars_reticulata_interneuorn_GABA
                        Substantia_nigra_pars_reticulata_principal_cell
                  cerebral_crus
                  interpeduncular_fossa
                  oculomotor_nerve
                  ventrolateral_fissure_of_midbrain
                Regional_part_of_midbrain_tectum
              anterior_cerebellar_incisure
              cerebrocerebellar_fissure
              interpeduncular_cistern
              superior_cistern
              superior_pontine_sulcus
          Spinal_Cord
          meninges
```

Synonym
"Substantia nigra dopaminergic cell,Nigral dopaminergic cell"

reference
"http://neurolex.org/wiki/Category:Substantia_nigra_pars_compacta_dopaminergic_cell"

isDefinedBy
"Nigral dopaminergic cell is a neuron found in the midbrain of vertebrates. These neurons comprise most of the substantia nigra and mainly regulate motor and sensorimotor functions within the brain.

Descrption:

Soma Location: Substantia nigra pars compacta
Spine density on dendrites: Aspiny Dendrite Quality

Axon Specific Properties
Axon projection laterality: ipsilateral
Location of axon arborization: Neostriatum
Cellular synaptic target: Neostriatum medium spiny neuron
Neurotransmitter: Dopamine

Description: Substantia_nigra_pars_compacta_dopaminergic_cell

Equivalent classes

Superclasses

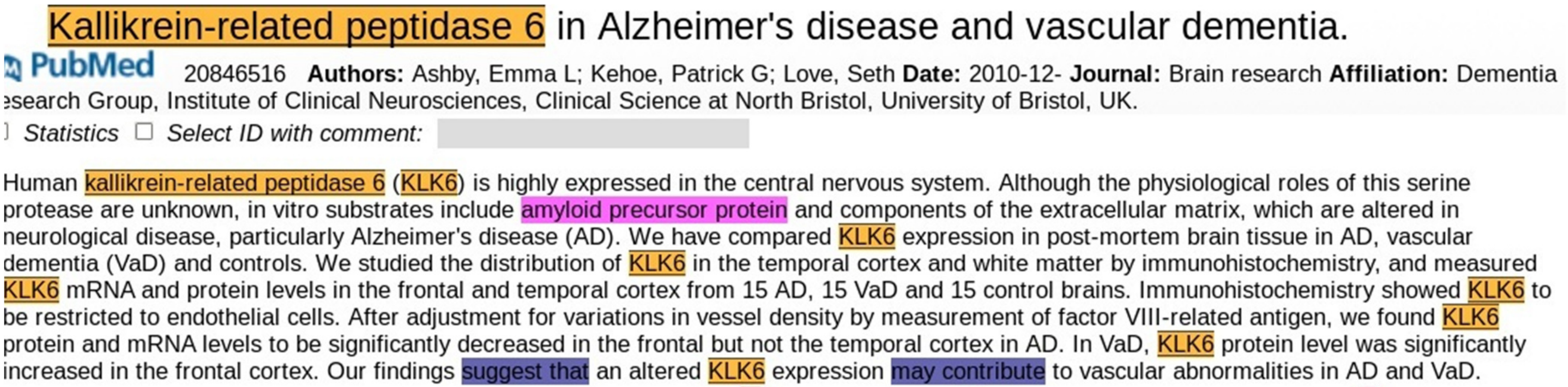
Substantia_nigra_pars_compacta

Inferred anonymous superclasses

has_part some Substantia_nigra_pars_compacta
has_part some Substantia_nigra_pars_reticulata
has_part some CA3_alveus
has_part some Piriform_cortex_layer_1
has_part some Neocortex_layer_4
has_part some Chemoarchitectural_part
has_part some CA1_alveus
has_part some CA3_stratum_lucidum
has_part some Hindbrain
has_part some Piriform_cortex_layer_2
has_part some Aggregate_regional_part_of_brain
has_part some Regional_part_of_forebrain
has_part some Molecular_layer_of_dorsal_cochlear_nucleus
has_part some Trigeminal_nucleus
has_part some Regional_part_of_midbrain
has_part some Composite_part_spanning_multiple_base_regional_parts_of_brain
has_part some Regional_part_of_hindbrain

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.

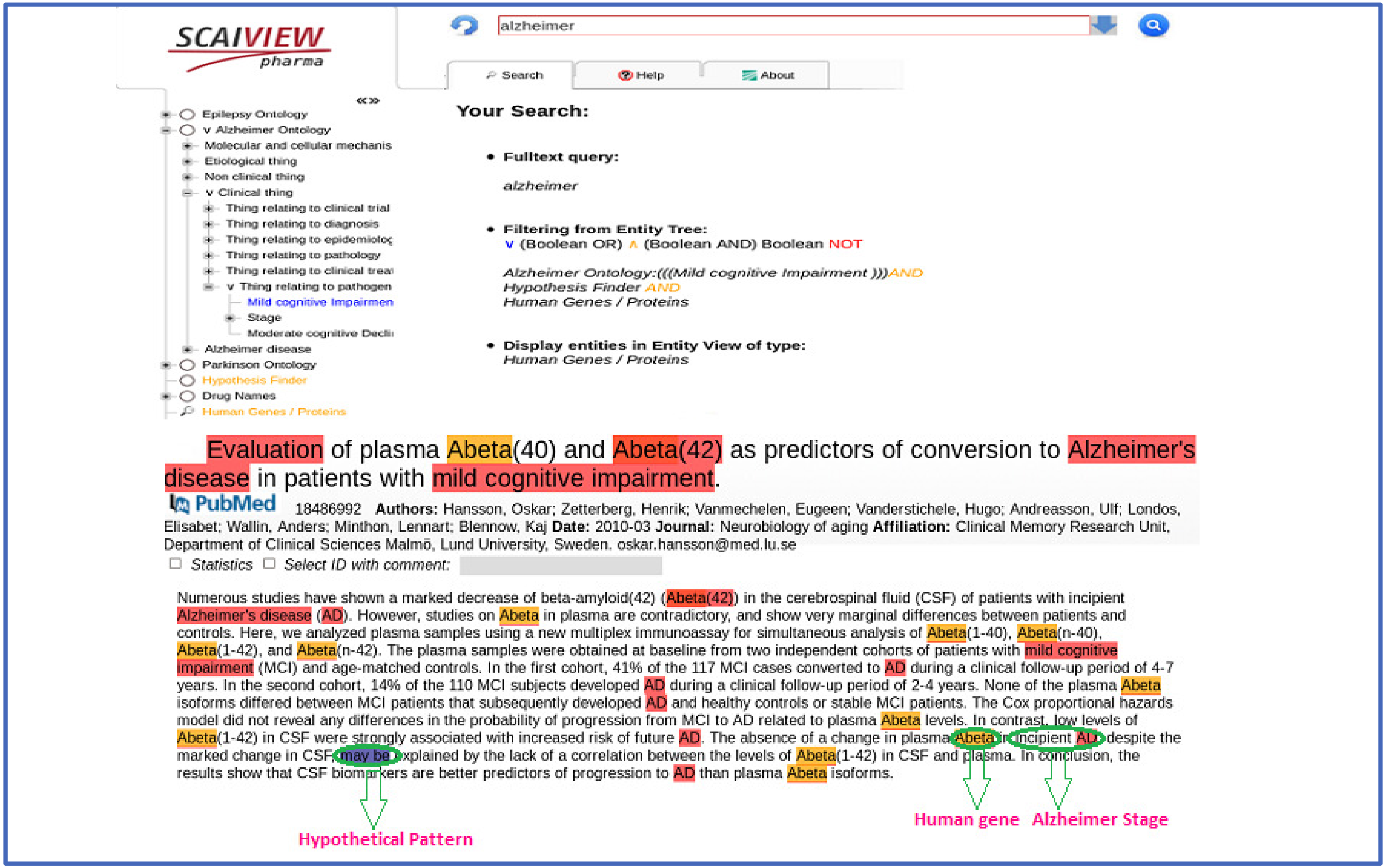
PubMed 20846516 **Authors:** Ashby, Emma L; Kehoe, Patrick G; Love, Seth **Date:** 2010-12- **Journal:** Brain research **Affiliation:** Dementia esearch Group, Institute of Clinical Neurosciences, Clinical Science at North Bristol, University of Bristol, UK.

☐ *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

Hypotheses finder∩ AD ontology ∩Human genes and proteins



Your Search:

- **Fulltext query:**

  alzheimer

- **Filtering from Entity Tree:**
  v (Boolean OR) ∧ (Boolean AND) Boolean NOT

  Alzheimer Ontology:(((Mild cognitive Impairment)))AND
  Hypothesis Finder AND
  Human Genes / Proteins

- **Display entities in Entity View of type:**
  Human Genes / Proteins

Evaluation of plasma Abeta(40) and Abeta(42) as predictors of conversion to Alzheimer's disease in patients with mild cognitive impairment.

PubMed 18486992 **Authors:** Hansson, Oskar; Zetterberg, Henrik; Vanmechelen, Eugeen; Vanderstichele, Hugo; Andreasson, Ulf; Londos, Elisabet; Wallin, Anders; Minthon, Lennart; Blennow, Kaj **Date:** 2010-03 **Journal:** Neurobiology of aging **Affiliation:** Clinical Memory Research Unit, Department of Clinical Sciences Malmö, Lund University, Sweden. oskar.hansson@med.lu.se

□ *Statistics* □ *Select ID with comment:*

Numerous studies have shown a marked decrease of beta-amyloid(42) (Abeta(42)) in the cerebrospinal fluid (CSF) of patients with incipient Alzheimer's disease (AD). However, studies on Abeta in plasma are contradictory, and show very marginal differences between patients and controls. Here, we analyzed plasma samples using a new multiplex immunoassay for simultaneous analysis of Abeta(1-40), Abeta(n-40), Abeta(1-42), and Abeta(n-42). The plasma samples were obtained at baseline from two independent cohorts of patients with mild cognitive impairment (MCI) and age-matched controls. In the first cohort, 41% of the 117 MCI cases converted to AD during a clinical follow-up period of 4-7 years. In the second cohort, 14% of the 110 MCI subjects developed AD during a clinical follow-up period of 2-4 years. None of the plasma Abeta isoforms differed between MCI patients that subsequently developed AD and healthy controls or stable MCI patients. The Cox proportional hazards model did not reveal any differences in the probability of progression from MCI to AD related to plasma Abeta levels. In contrast, low levels of Abeta(1-42) in CSF were strongly associated with increased risk of future AD. The absence of a change in plasma Abeta in incipient AD despite the marked change in CSF may be explained by the lack of a correlation between the levels of Abeta(1-42) in CSF and plasma. In conclusion, the results show that CSF biomarkers are better predictors of progression to AD than plasma Abeta isoforms.

Hypothetical Pattern

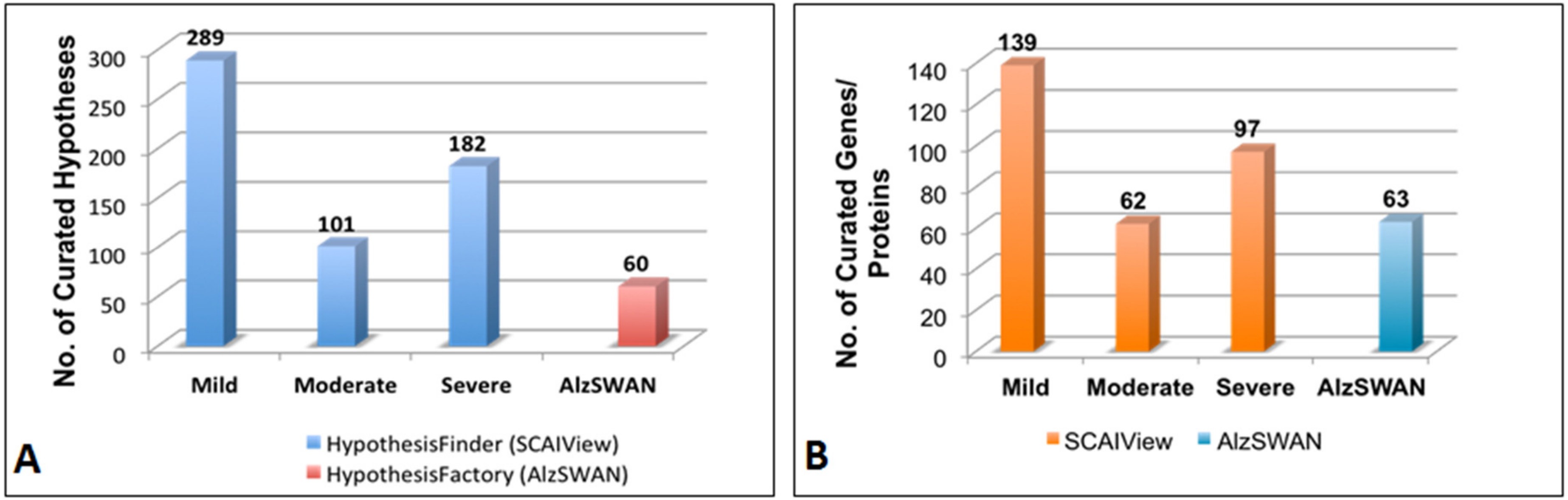
Human gene

Alzheimer Stage

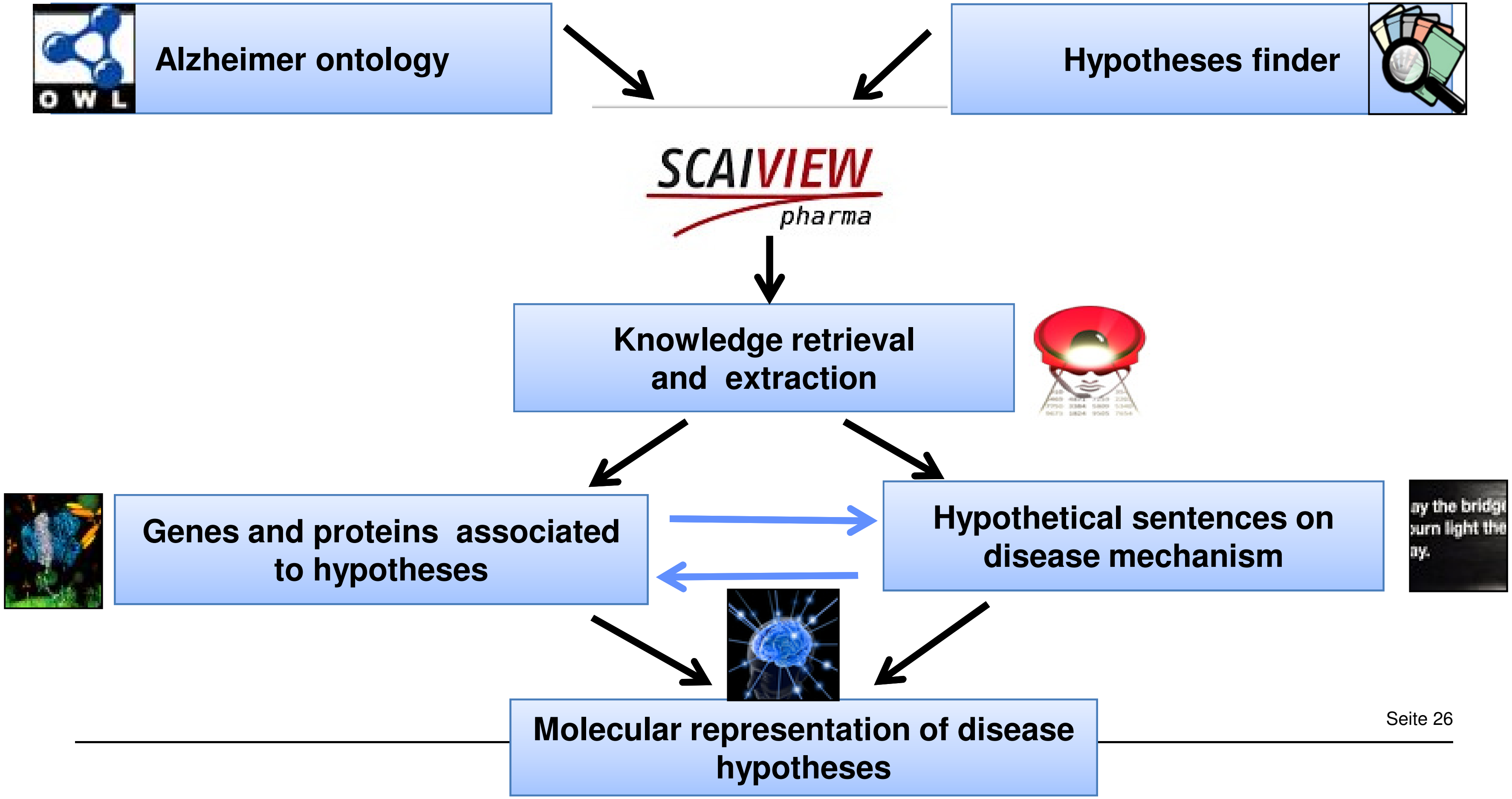
Performance of Hypotheses finder

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

Performance of Hypotheses finder



The Knowledge – Discovery Strategy



Alzheimer ontology

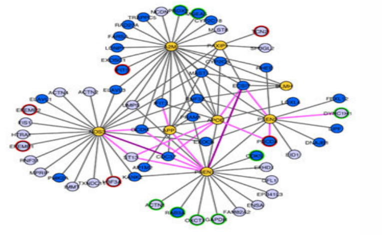
Hypotheses finder

Knowledge retrieval and extraction

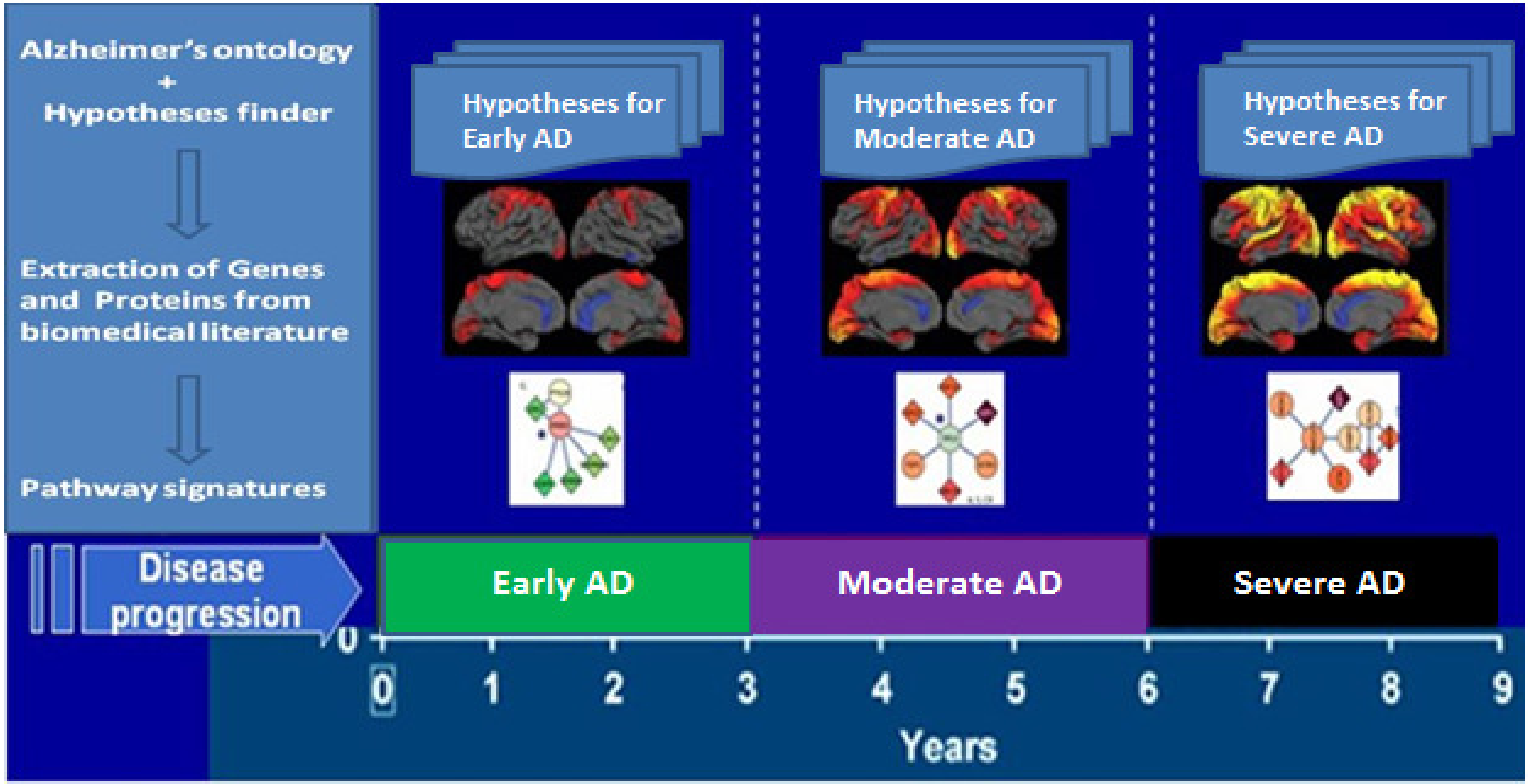
Genes and proteins associated to hypotheses

Hypothetical sentences on disease mechanism

Molecular representation of disease hypotheses

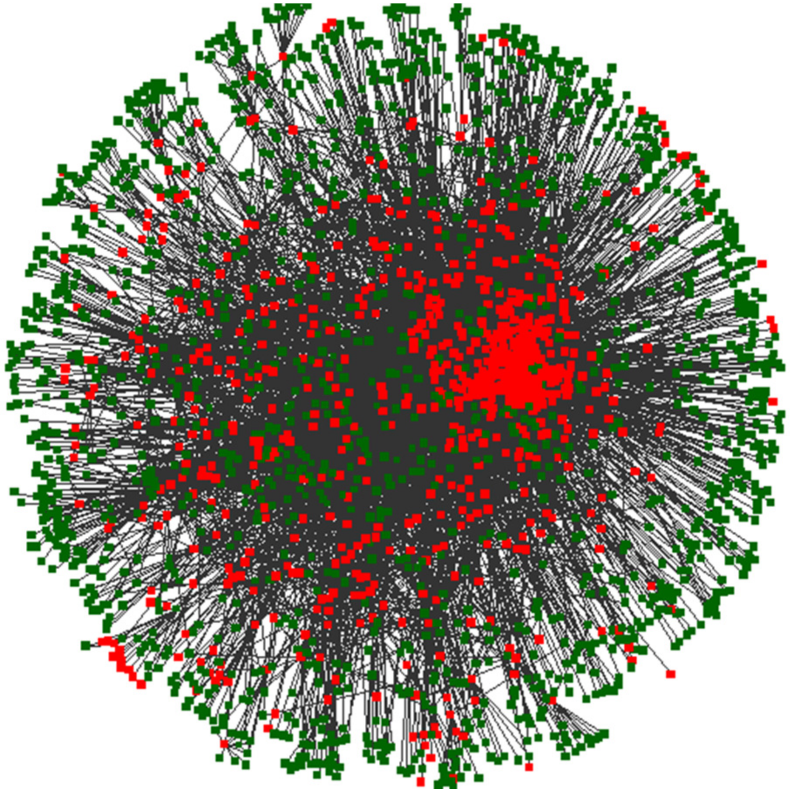


Analysis of hypotheses patterns across disease stages



Fraunhofer SCAI

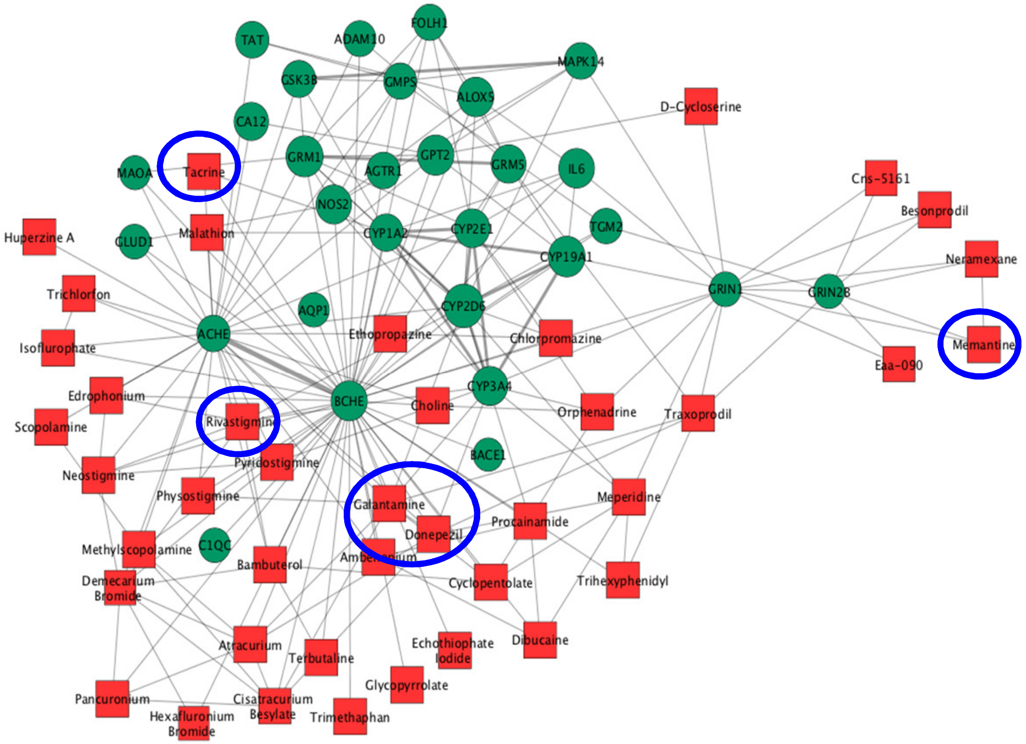
Current Work at SCAI: NDD Pharmacome



NDD Pharmacome Network:
Total nodes: 2,153
Total edges: 11,630

- 🟥 Drugs
- ● Targets

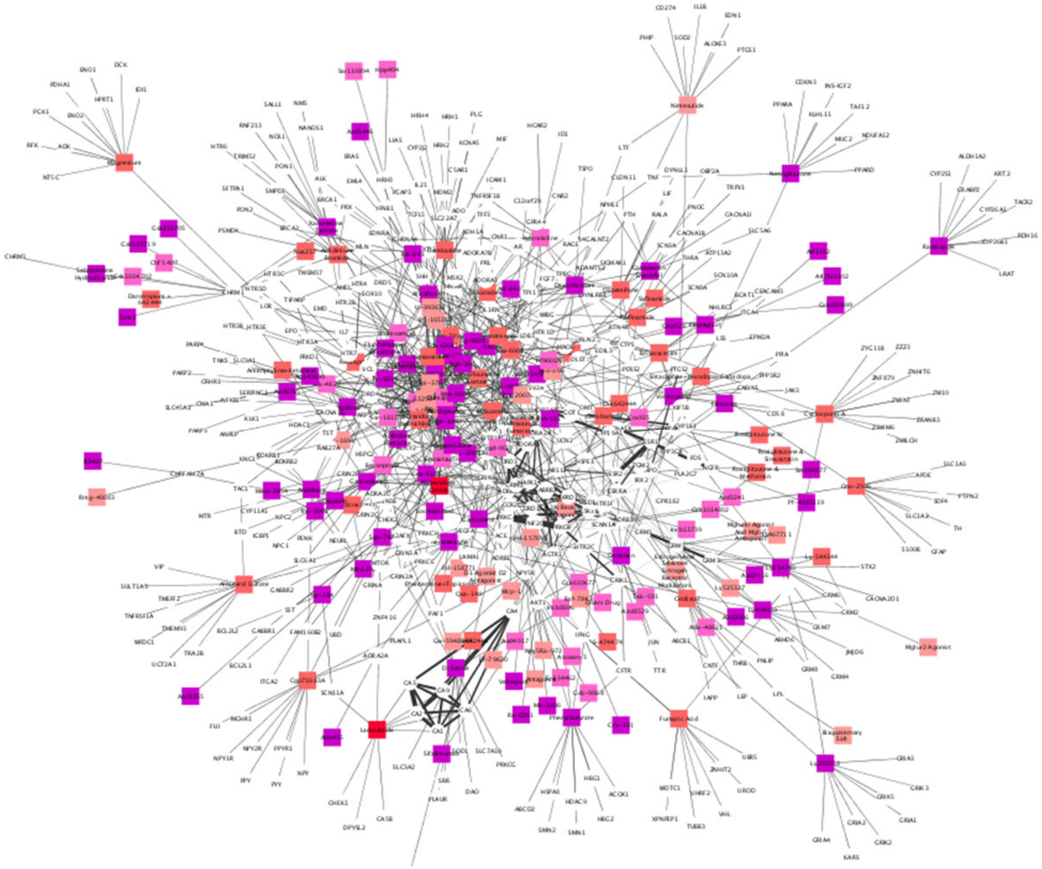
Using the Pharmacome for VS



1. Galantamine
2. Memantine
3. Rivastigmine
4. Tacrine
5. Donepezil

Drugs having structural similarity and targets having binding site similarities are connected by edges.

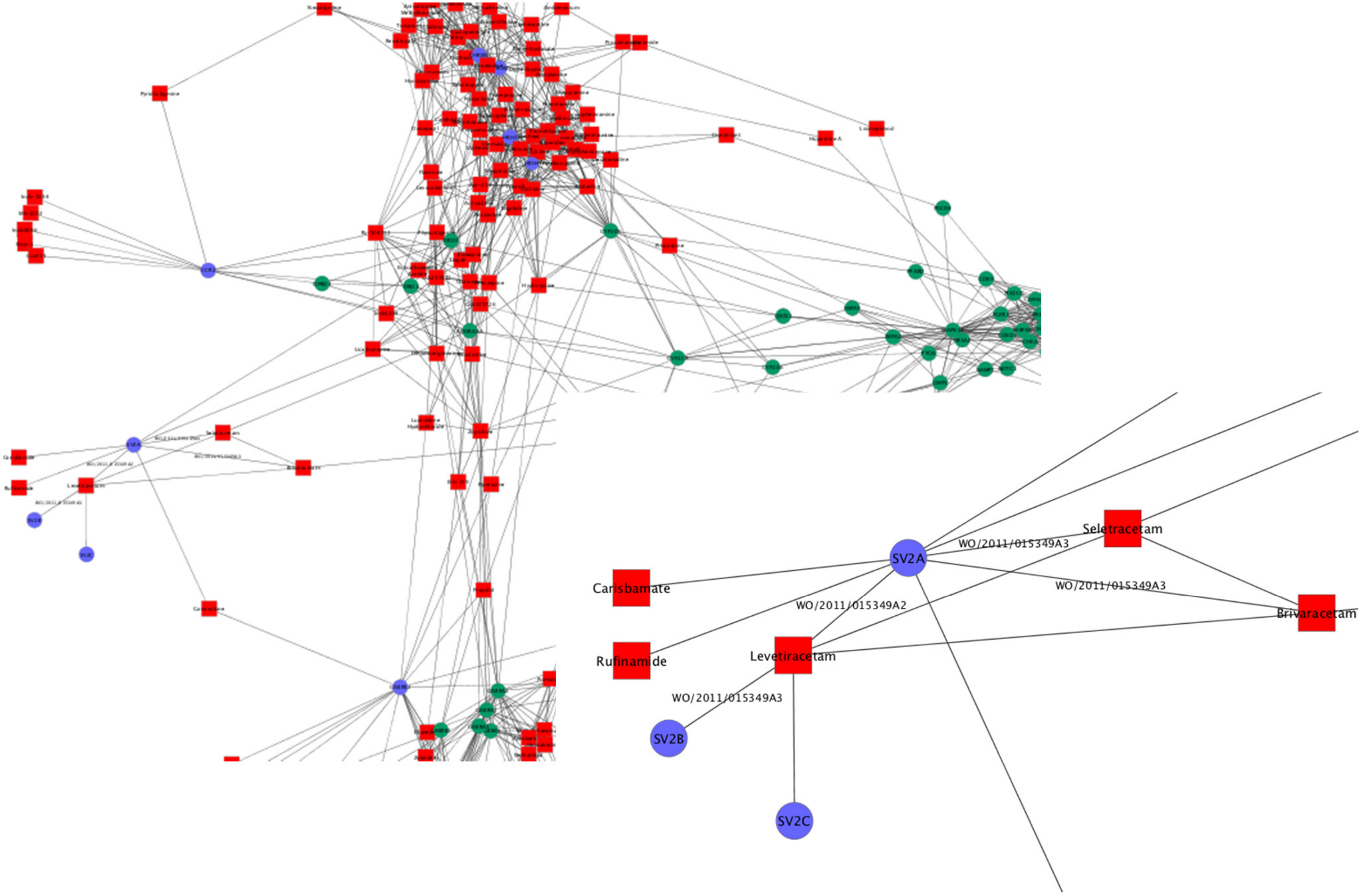
What is Competition doing? Competitive intelligence and strategic partnering



Drugs in various phases

- 🟦 Phase I
- ■ Phase II
- 🟧 Phase III
- 🟥 Phase IV

Brain Drug Portfolios in the Pharmacome



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
- We are using our competence in HPC to enable large-scale information extraction from full text documents (focus on patents)
- Internal usage of technologies (information extraction; distributed and high performance computing) for biomedical application: modelling of neurodegenerative diseases
- Information integration and aggregation in models: the brain pharmacome