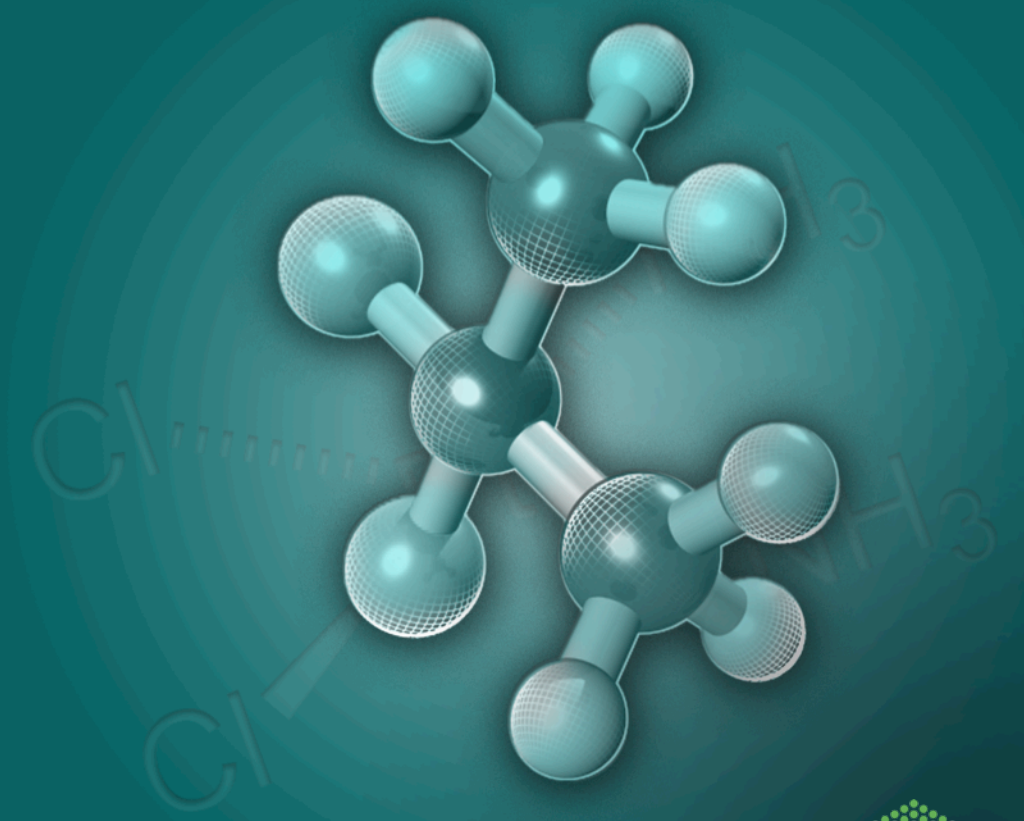
Leveraging Open Chemogenomics Data and Tools with KNIME



George Papadatos

ChEMBL Group

georgep@ebi.ac.uk

EMBL-EBI

What is EMBL-EBI?

- Europe's home for biological **data**, **services**, **research** and **training**
- A trusted data provider for the life sciences
- Part of the European Molecular Biology Laboratory, an intergovernmental research organisation
- International: 570 members of staff from 57 nations

Data resources at EMBL-EBI

Genes, genomes & variation

European Nucleotide Archive
European Variation Archive
Ensembl
Ensembl Genomes
GWAS Catalog
Metagenomics portal

Gene, protein & metabolite expression

RNA Central
ArrayExpress
Expression Atlas
Metabolights
PRIDE

Protein sequences, families & motifs

InterPro
Pfam
UniProt

Molecular structures

Protein Data Bank in Europe
Electron Microscopy Data Bank

Chemical biology

ChEMBL
SureChEMBL
ChEBI

Systems

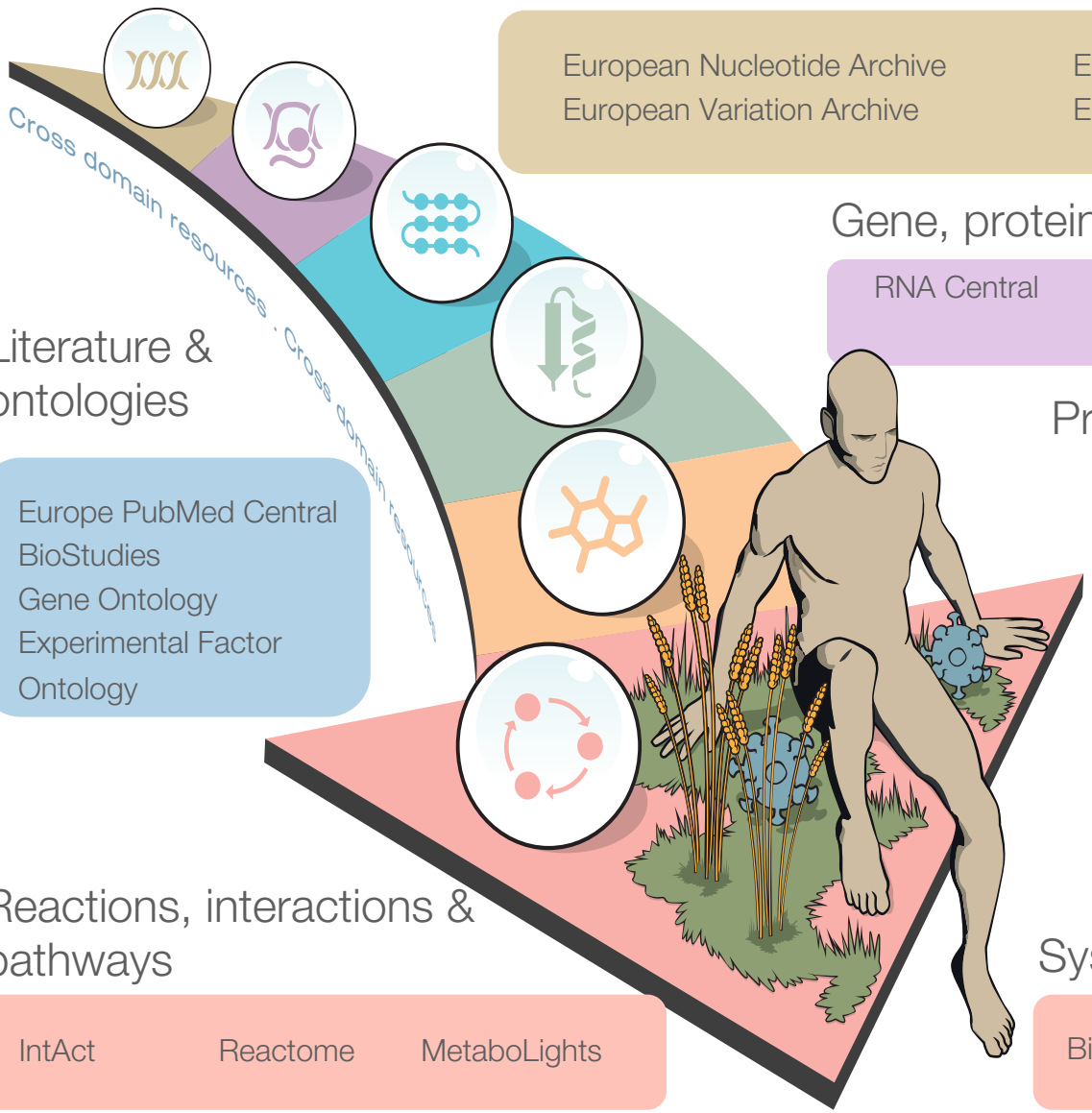
BioModels
Enzyme Portal
BioSamples

Reactions, interactions & pathways

IntAct
Reactome
MetaboLights

Literature & ontologies

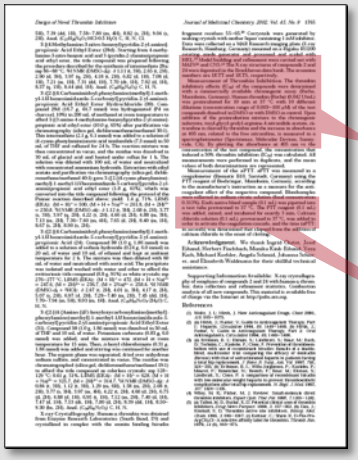
Europe PubMed Central
BioStudies
Gene Ontology
Experimental Factor Ontology

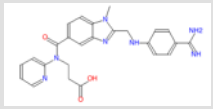


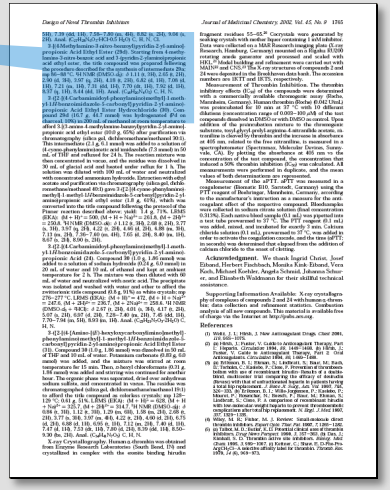
ChEMBL: Data for drug discovery

1. Scientific facts

3. Insight, tools and resources for translational drug discovery







>Thrombin
MAHVRGLQLPGCLALAALCSLVHSQHVFLAPQQARSLLQRVRRANTFLEEVRKGNLE
RECVEETCSYEEAFEALESSTATDVFWAKYTACETARTPRDKLAACLEGNCAEGLGT
NYRGHVNITRSGIECQLWRSRYPHKPEINSTTHPGADLQENFCRNPDSSTTGPWCYT
TDPTVRRQECSIPVCGQDQVTVAMTPRSEGSSVNLSPPLEQCVPDRGQQYQGRLAVT
THGLPCLAWASAQAKALSKHQDFNSAVQLVENFCRNPDGDEEGVWCYVAGKPGDFG
CDLNYCEEAVEEETGDGLDEDSDRAIEGRTATSEYQTFFNPRTFGSGEADCGLRPL
EKKSLEDKTERELLESYIDGRIVEGSDAEIGMSPWQVMLFRKSPQELLCGASLISDR
WVLTAAHCLLYPPWDKNFTENDLLVRIGKHSRTRYERNIEKISMLEKIYIHPRYNWR
ENLDRDIALMKLKKPVAFSDYIHPVCLPDRETAASLLQAGYKGRVTGWGNLKE
NVGKGQPSVLQVVNLPIVERPVCKDSTRIRITDNMFCAGYKPDEGKRGDACEGD
PFVMKSPFNNRWYQMGIVSWGEGCDRDGKYGFYTHVFRLLKWIQKVIDQFGE

Compound

K_i = 4.5nM

Assay/Target

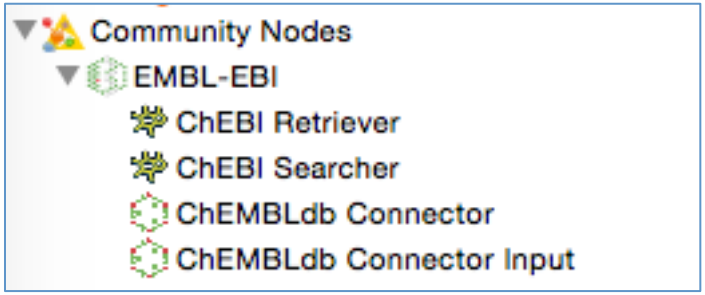
Bioactivity data

APTT = 11 min.

2. Organization, integration, curation and standardization of pharmacology data

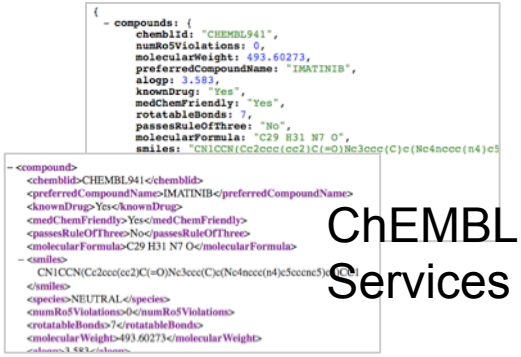
KNIME at the EBI

- KNIME nodes to access ChEBI and ChEMBL databases
  - Trusted community nodes
  - Workflows on Examples server
- Method development and use cases
- Provide KNIME training to scientists and researchers
  - Wellcome Trust drug discovery courses, EMBL courses
- CDK community nodes support



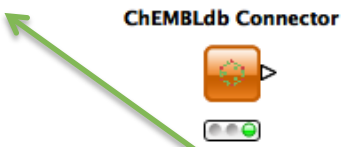
https://tech.knime.org/book/embl-ebi-nodes-for-knime-trusted-extension

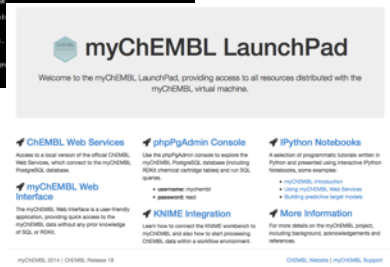
KNIME and ChEMBL



ChEMBL Web Services

14M bioactivities
1.5M structures





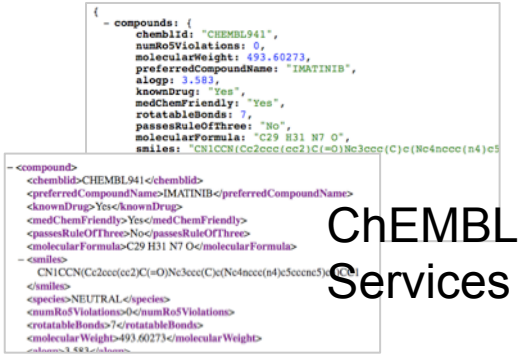
Virtual Machine

Local access to ChEMBL data and services

UniChem Web Services

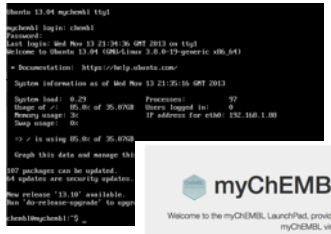
Access ~110M structures from 27 sources

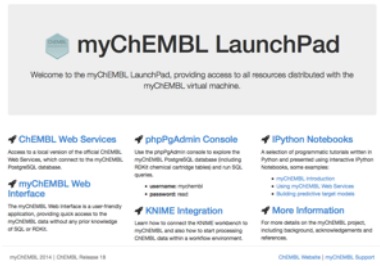
KNIME and ChEMBL



ChEMBL Web Services

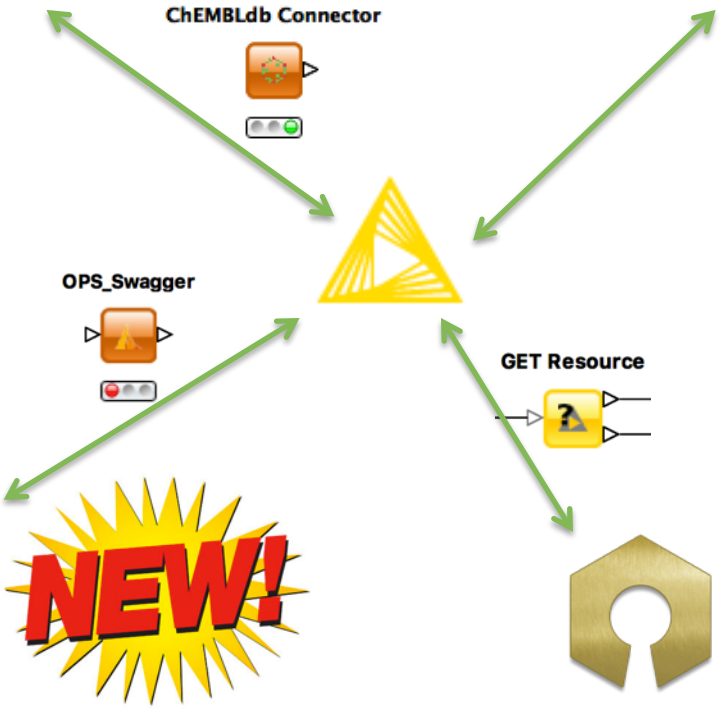
14M bioactivities
1.5M structures





Virtual Machine

Local access to ChEMBL data and services



Patent Annotations

4M patent documents
14M structures
260M annotations

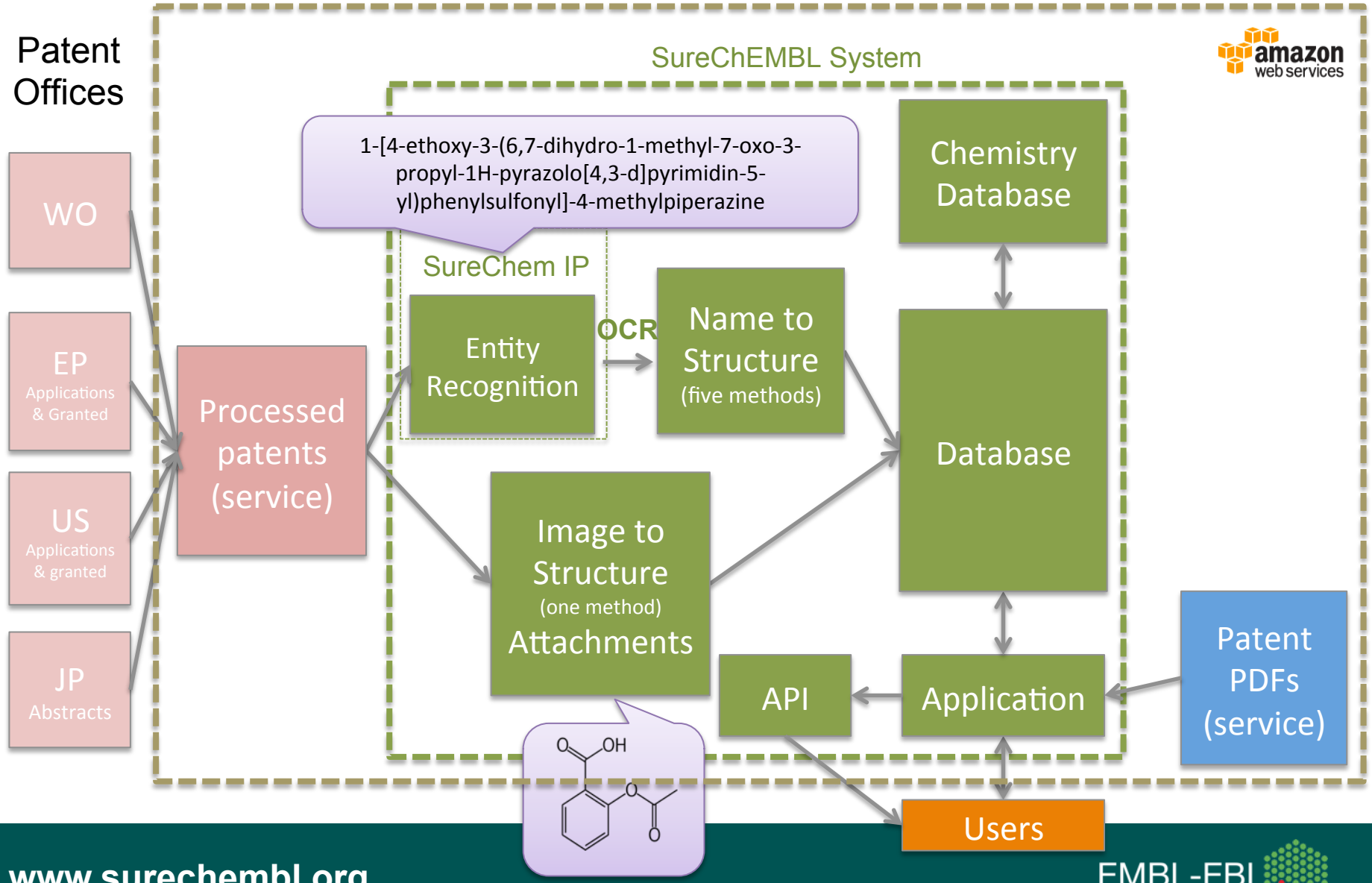
UniChem Web Services

Access ~110M structures from 27 sources

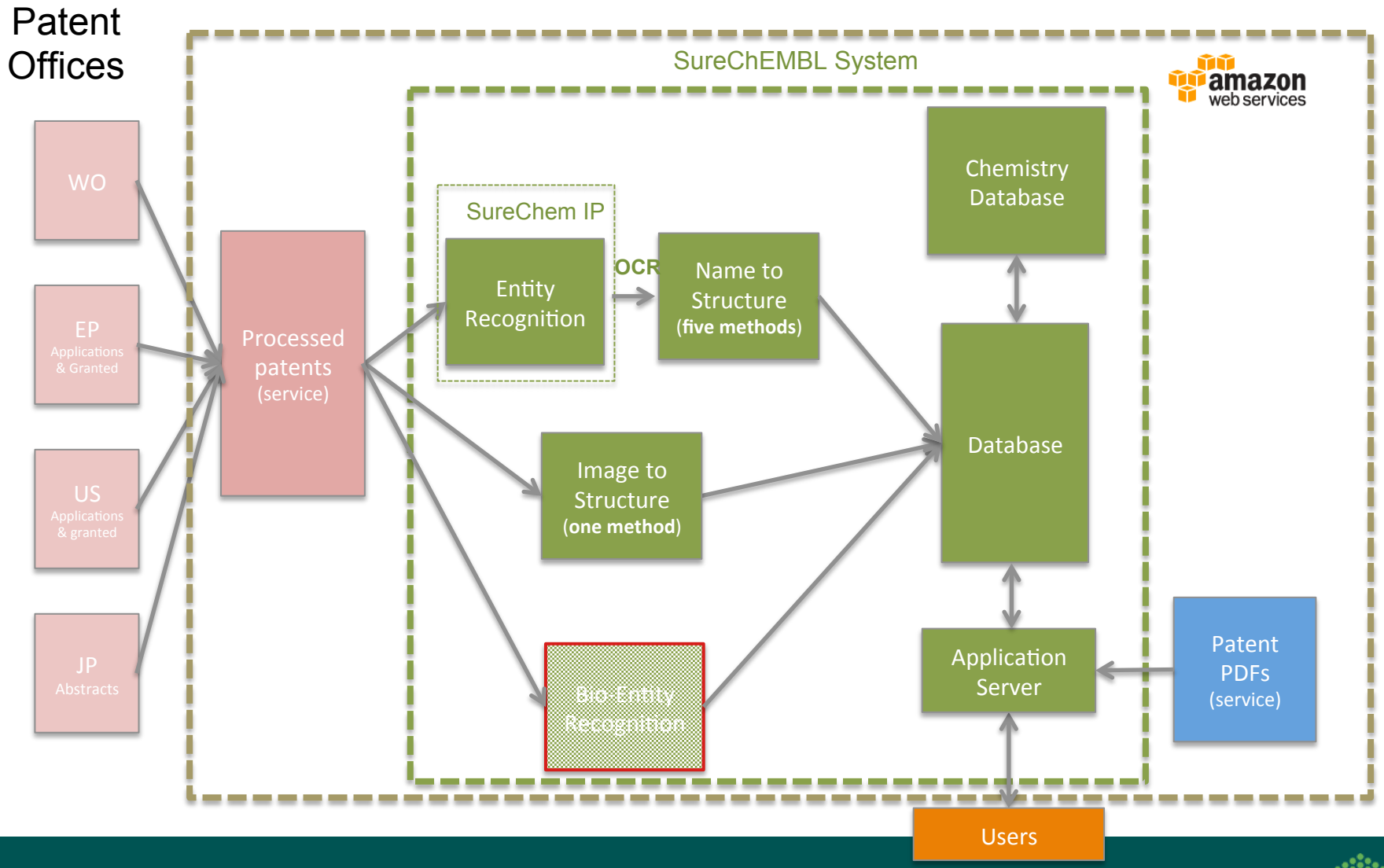
Why looking at patent documents?

- Patent filing and searching
  - Legal, financial and commercial incentives & interests
  - Prior art, novelty, freedom to operate searches
  - Competitive intelligence
- Unprecedented wealth of knowledge
  - Most of the knowledge will never be disclosed anywhere else
    - Compounds, scaffolds, reactions
    - Biological targets, diseases, indications
  - Average lag of 2-4 years between patent document and journal publication disclosure for chemistry, 4-5 for biological targets

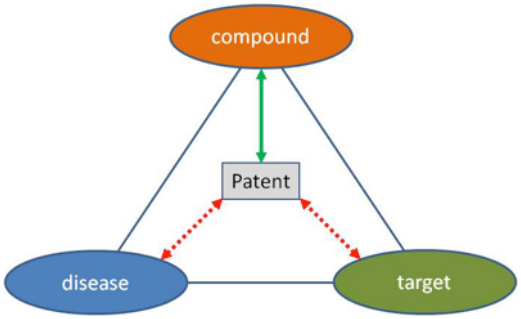
SureChEMBL data processing

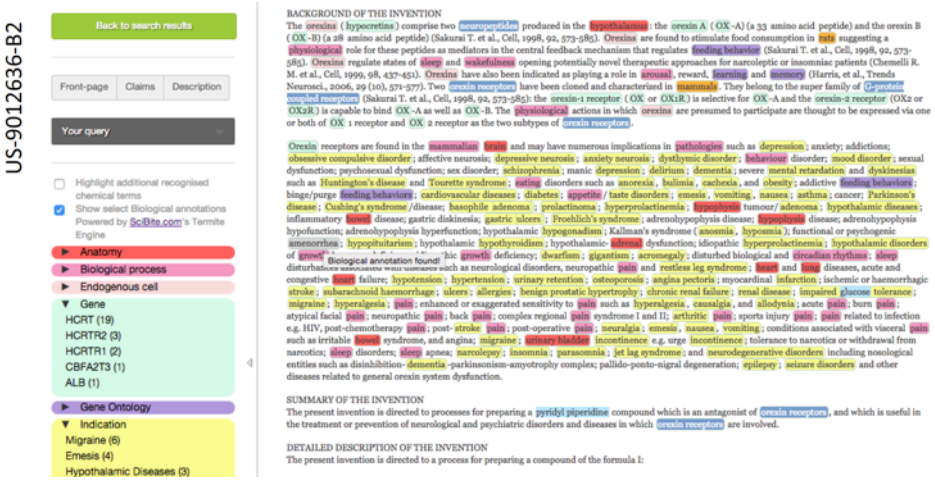


SureChEMBL data processing v2



SureChEMBL bioannotation





- SciBite's Termite text-mining engine run on 4M life-science patents from SureChEMBL corpus
- Genes (identified by HGNC symbols) and diseases (identified by MeSH IDs) annotated
- Section/frequency information annotated (e.g., in title, abstract, claims, total frequency)
- Relevance score (0-3) to flag important chemical and biological entities and remove noise

Relevance scoring – genes/diseases

- Various features used:
  - Term frequency
  - Position (title, abstract, figure, caption, table)
  - Frequency distribution
- Scores range from 0 – 3
  - 3 – most important entities in the patent
  - 2 – important entities in the patent
  - 1 – mentioned entities in the patent
  - 0 – ambiguous entity/likely annotation error

Relevance scoring - compounds

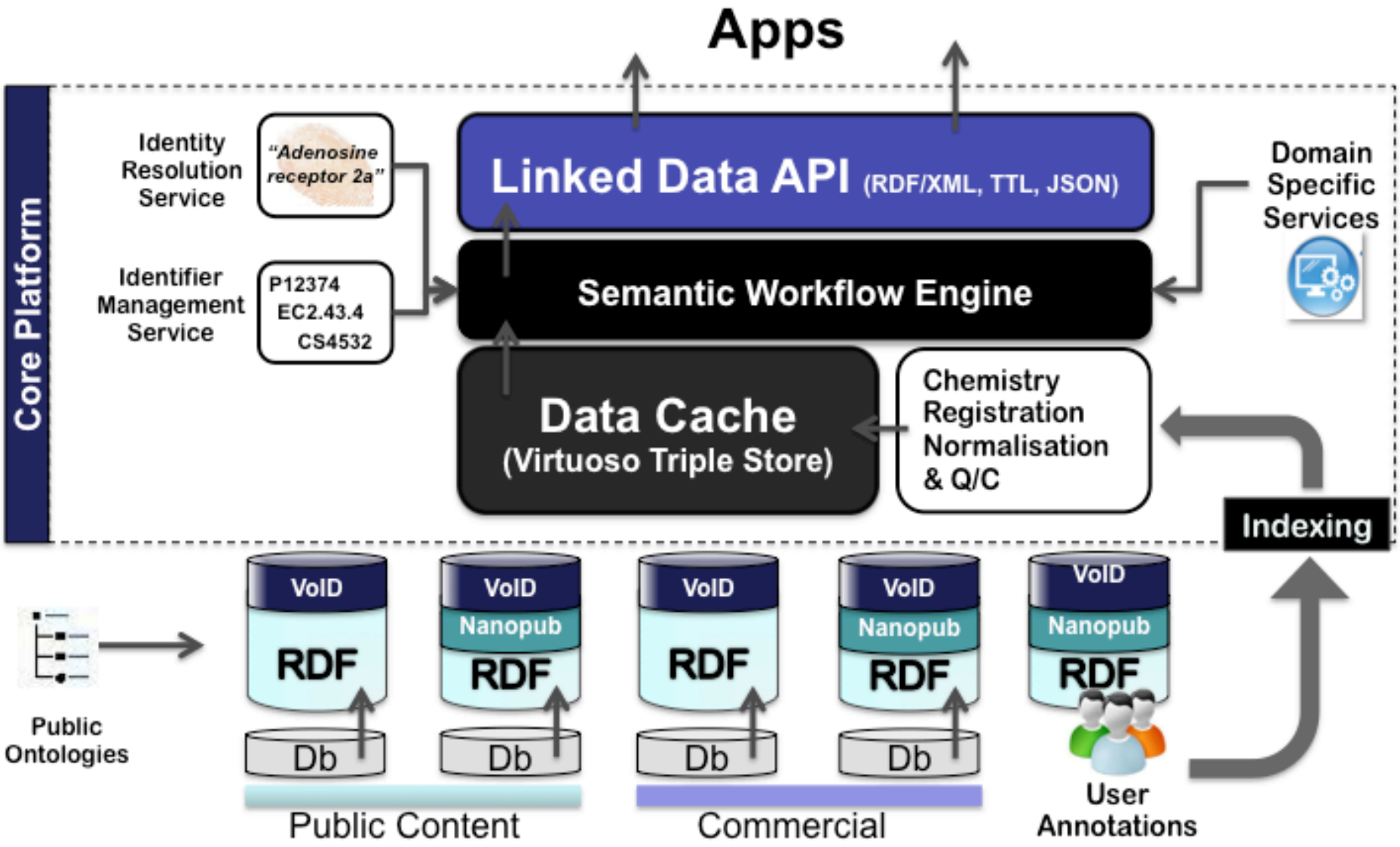
- Main assumptions for relevance:
  1. Very frequent compounds are irrelevant (but if drug-like then that's OK)
  2. Compounds with busy chemical space around them are interesting
     - Use distribution of close analogues (NNs) among compounds found in the **same patent family**
- Scores range from 0 – 3
  - **3 – highest number of NNs: most important entities in the patent**
  - **2 – important entities in the patent**
  - 1 – few NNs: mentioned entities in the patent
  - 0 – singletons or trivial entities, most likely errors or reagents, solvents, substituents

Hatori, K., Wakabayashi, H., & Tamaki, K. (2008). JCIM, 48(1), 135–142. doi:10.1021/ci7002686
Tyrchan, C., Boström, J., Giordaneto, F., Winter, J., & Muresan, S. (2012). JCIM, 52(6), 1480–1489. doi:10.1021/ci3001293

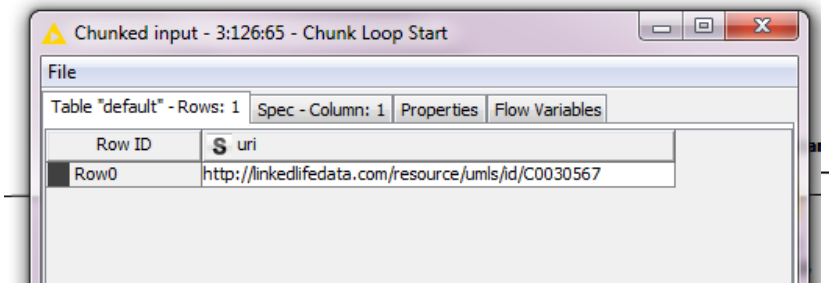
Gotchas & out of scope

- No Markush extraction
- No natural language processing (e.g., ‘compound **x** is an **inhibitor** of target **y**’)
- No extraction of bioactivities
- No chemistry search (yet)
- Patent coverage stops in April 2015
  - Incremental updates TBD
- Patent calls still in dev
  - Old scripts / workflows may break

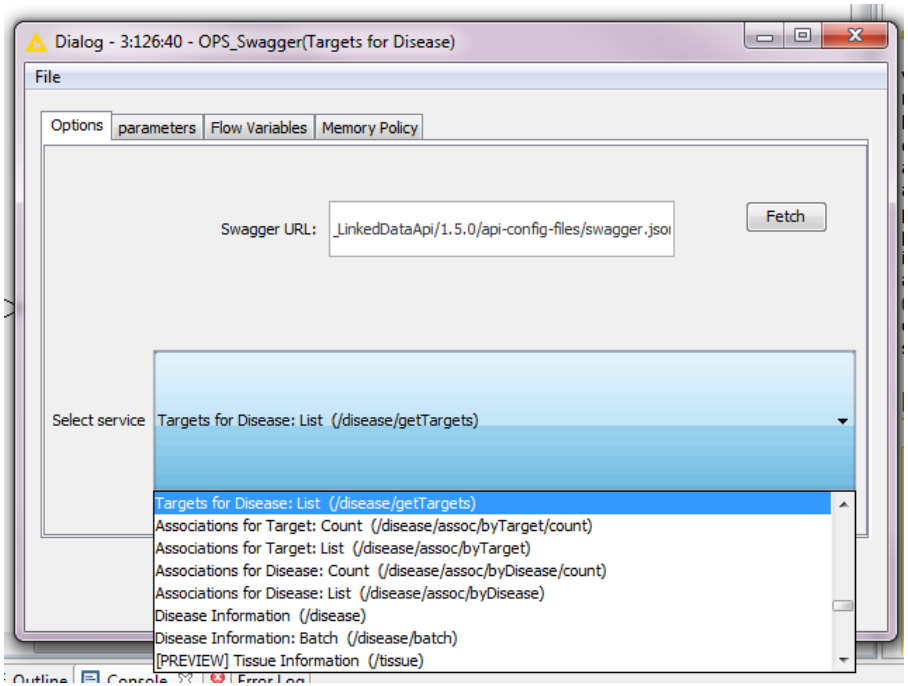
Open PHACTS Architecture

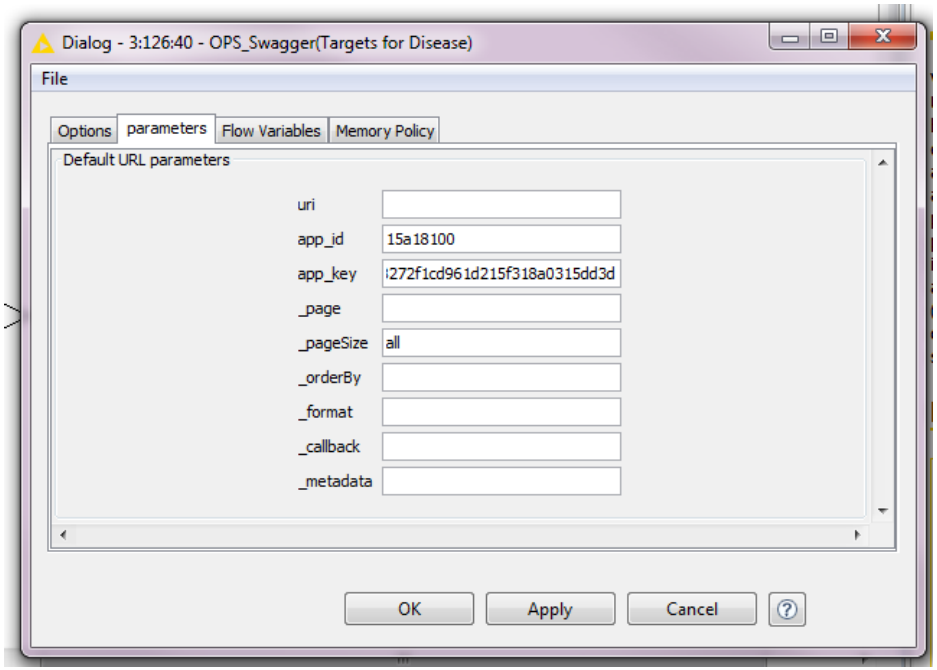


The Open PHACTS node

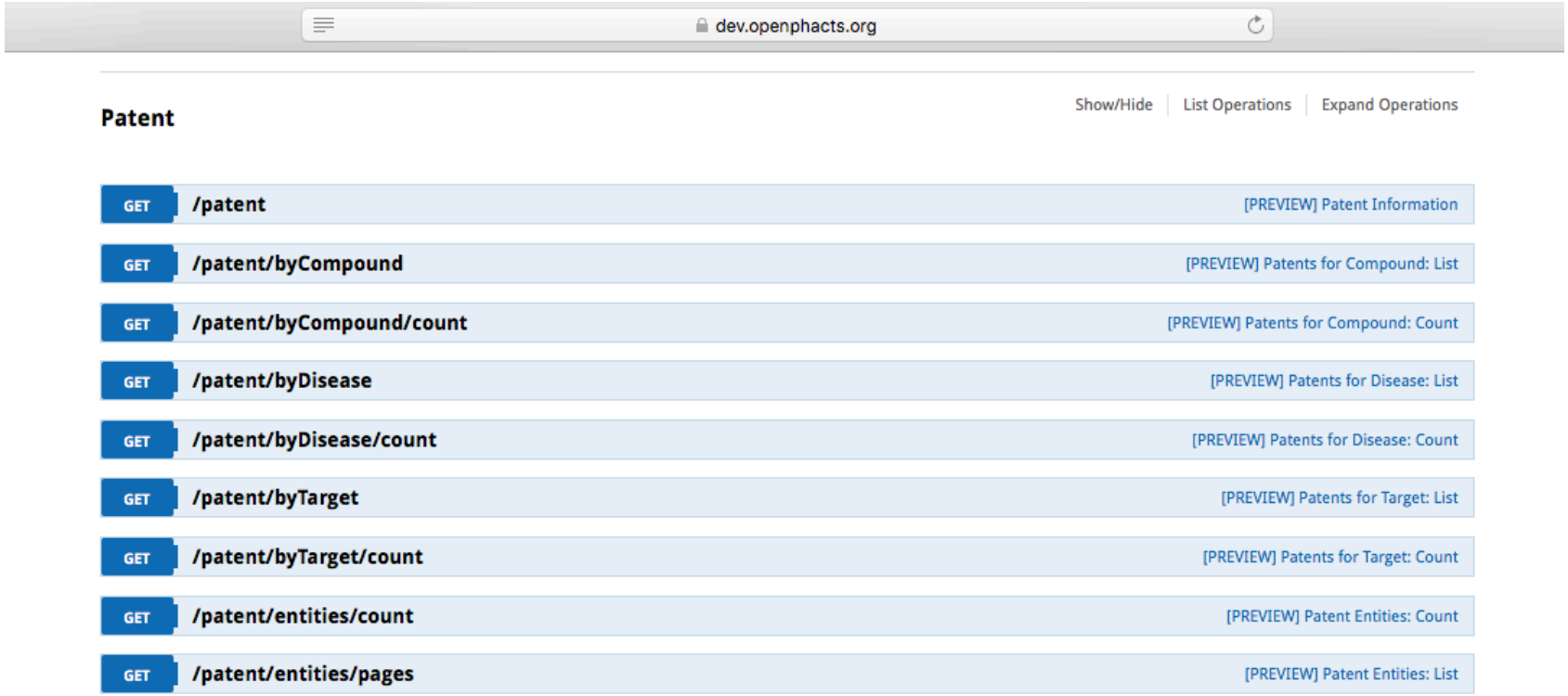


executable API call to KREST nodes



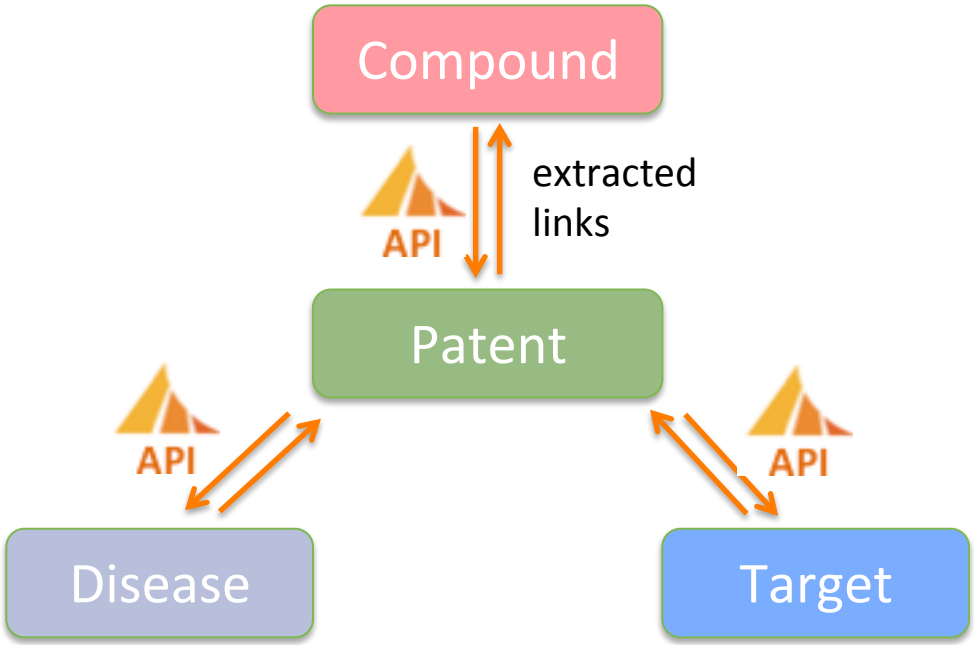


Open PHACTS Patent API

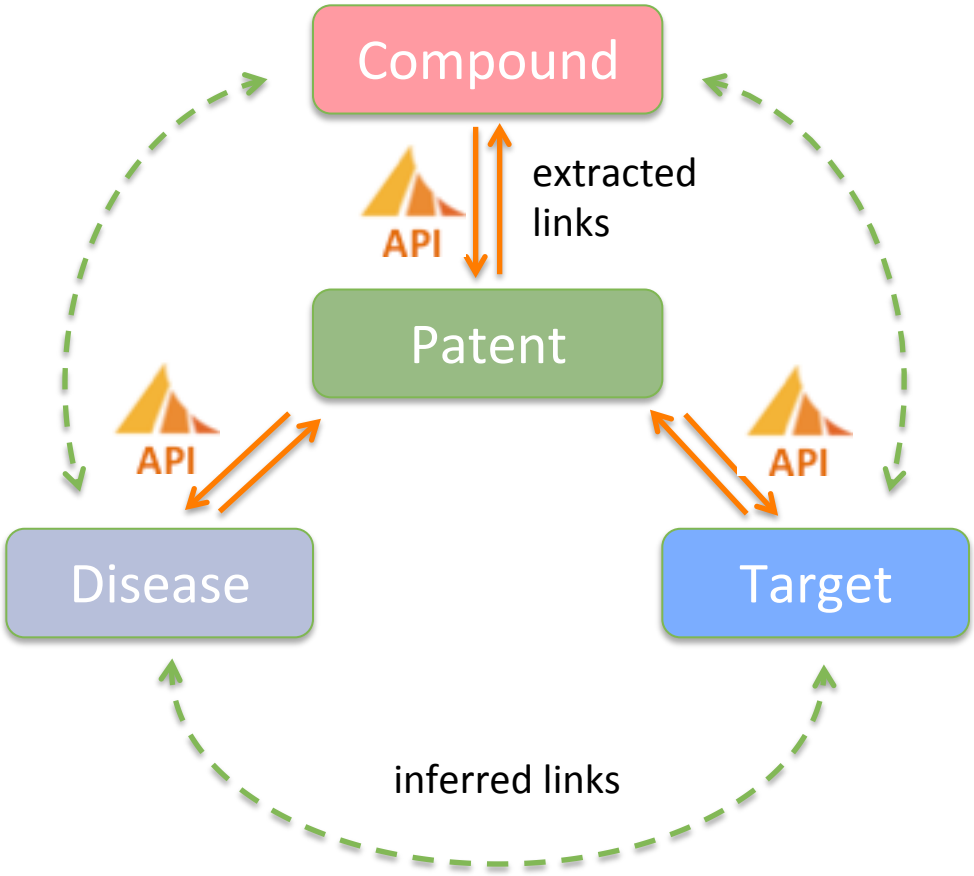


https://dev.openphacts.org/docs/develop

Open PHACTS Patent API

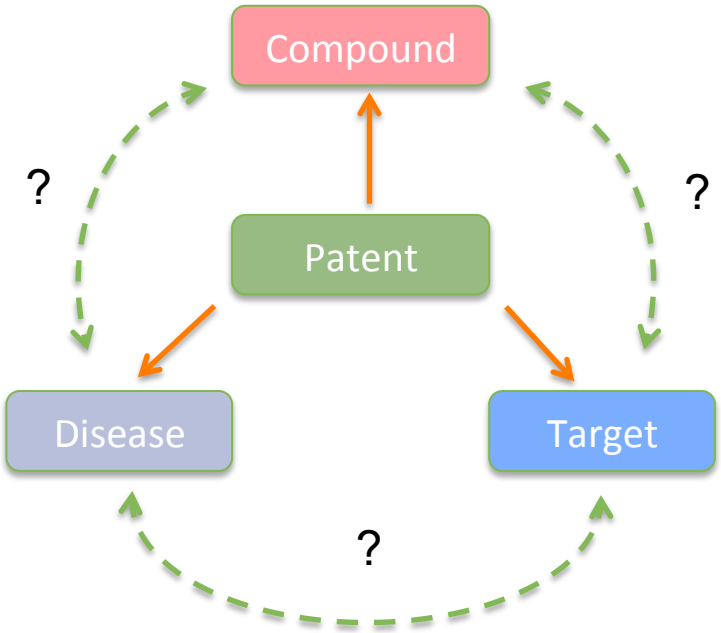


Open PHACTS Patent API



Use case #1: Patent to Entities

1. From a patent get compounds, genes and diseases
2. Filter to remove noise
   - Frequency and relevance score
3. Process and visualise



(12) **United States Patent**
Walter

(10) **Patent No.: US 7,718,693 B2**
(45) **Date of Patent: May 18, 2010**

(54) **RECEPTOR ANTAGONISTS AND THEIR METHODS OF USE**

(75) Inventor: **Daryl Simon Walter**, Harlow (GB)

(73) Assignee: **Glaxo Group Limited**, Greenford, Middlesex (GB)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

(21) Appl. No.: **11/772,977**

(22) Filed: **Jul. 3, 2007**

(65) **Prior Publication Data**

US 2008/0009541 A1 Jan. 10, 2008

(30) **Foreign Application Priority Data**

Jul. 6, 2006	(GB)	0613473.8
Nov. 15, 2006	(GB)	0622825.8
Mar. 19, 2007	(GB)	0705263.2
Jun. 13, 2007	(GB)	0711439.0

(51) **Int. Cl.**
A61K 31/4015 (2006.01)
C07D 207/04 (2006.01)

WO	WO 00/33788 A2	6/2000
WO	WO 05/27882 A1	3/2005
WO	WO 08/003697 A1	1/2008

OTHER PUBLICATIONS

Angier et al.; "Some N-Substituted-5-oxo-2-pyrrolidinecarboxamides"; Journal of Organic Chemistry; 1956; vol. 21, No. 12; pp. 1540-1543.
Zhao et al., *Bioorganic & Medicinal Chemistry*, 7(8), pp. 1647-1654 (1999).
Møss et al., *International Journal of Pharmaceutics*, 52(3), pp. 255-263 (1989).
Tye et al., *Organic & Biomolecular Chemistry*, vol. 2, pp. 813-815 (2004).
Harriman, *Tetrahedron Letters*, 38(32), pp. 5591-5594 (1997).
Bundgaard et al., *Biochemical Society Transactions*, 17(5), pp. 947-949 (1989).
Bundgaard et al., *Journal of Pharmaceutical Sciences*, 78(2), pp. 122-126 (1989).
Møss et al., *Acta Pharmaceutica Nordica*, 4(4), pp. 301-308 (1992).
International Search Report for corresponding application PCT/EP2007/056675 (published as WO 08/00369A1).
Pharmaprojects Review (Jan. 2009) http://www.pharmaprojects.com/therapy_analysis/purin_P2X7_0109.htm.
Chemcats listing of CAS # 1001389-19-0 available from ChemDiv, Inc., San Diego, CA (Jul. 1, 2008).

Primary Examiner—Golam M M Shameem
(74) *Attorney, Agent, or Firm*—Duke M. Fitch; Kathryn L. Sieburth; Lorraine B. Ling

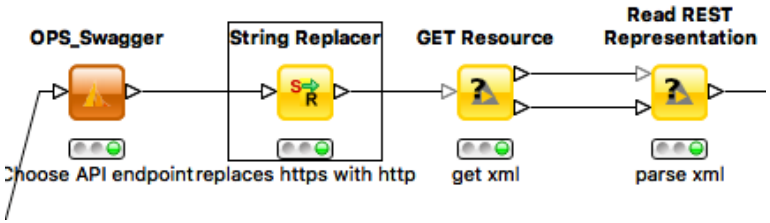
US-7718693-B2

Use case #1: Patent to Entities

- Patent URI:
  - http://rdf.ebi.ac.uk/resource/surechembl/patent/US-7718693-B2
- API call:

GET /patent/entities/pages

- 586 entities back

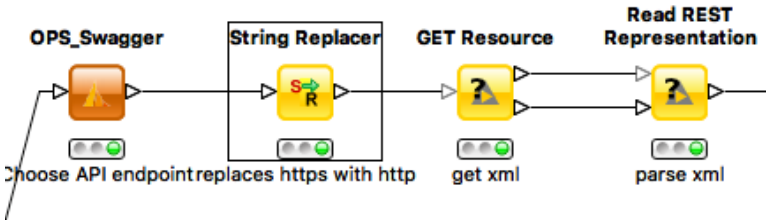


Use case #1: Patent to Entities

- Patent URI:
  - http://rdf.ebi.ac.uk/resource/surechembl/patent/US-7718693-B2
- API call:

GET /patent/entities/pages

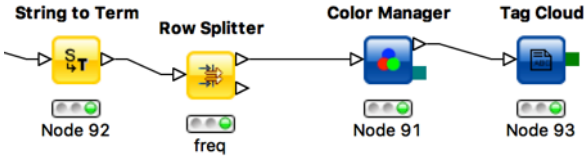
- 586 entities back



item.href	type.prefLabel	smiles	relevanceScore	frequency	occursInPatentTitle	occursInPatentClaims	occursInPatentDescription
http://rdf.ebi.ac.uk/resource/surechembl/molecule/SCHEMBL766149	Molecule	COC(=O)C1CCC(=O)N1C1=CC=CC=C1	1	4	false	false	true
http://rdf.ebi.ac.uk/resource/surechembl/indication/D018771	Disease		2	4	false	false	true
http://rdf.ebi.ac.uk/resource/surechembl/molecule/SCHEMBL288651	Molecule	OC(=O)[C@@H]1CCC(=O)N1CC1=CC=CC=C1	1	8	false	false	true
http://rdf.ebi.ac.uk/resource/surechembl/indication/D018805	Disease		1	1	false	false	true
http://rdf.ebi.ac.uk/resource/surechembl/indication/D018781	Disease		1	1	false	false	true
http://rdf.ebi.ac.uk/resource/surechembl/molecule/SCHEMBL960	Molecule	ClC(=O)OCC1=CC=CC=C1	0	1	false	false	true
http://rdf.ebi.ac.uk/resource/surechembl/molecule/SCHEMBL8956	Molecule	N1N=NC2=C1C=CC=C2	0	1	false	false	true
http://rdf.ebi.ac.uk/resource/surechembl/molecule/SCHEMBL8249	Molecule	NC1=CC=C(C=C1)C(O)=O	0	1	false	false	true
http://rdf.ebi.ac.uk/resource/surechembl/molecule/SCHEMBL2578506	Molecule	CC(C)N1[C@@H](CCC1=O)C(=O)NCC1=CC=...	2	2	false	false	true
http://rdf.ebi.ac.uk/resource/surechembl/molecule/SCHEMBL2586270	Molecule	CCN1[C@@H](CCC1=O)C(=O)NCC1=CC=C(C...	2	1	false	false	true
http://rdf.ebi.ac.uk/resource/surechembl/molecule/SCHEMBL2583583	Molecule	CC(C)(C)OC(=O)CN=C1C[C@@H]2C[C@H](C2(...	1	4	false	false	true
http://rdf.ebi.ac.uk/resource/surechembl/indication/D017116	Disease		1	1	false	false	true

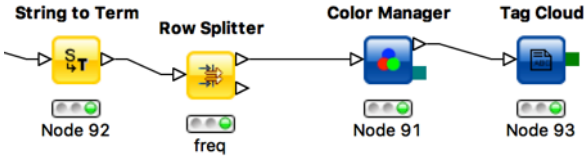
1) Look at target and disease entities

- 178 target and disease entities
- Filter: Relevance score >= 2 → 23 remain
- Visualise in tag cloud by frequency



1) Look at target and disease entities

- 178 target and disease entities
- Filter: Relevance score >= 2 → 23 remain
- Visualise in tag cloud by frequency



Does it make sense?

Does it make sense?

Immunologic Deficiency Syndromes
Motor Neuron Disease Bone Diseases
Myocardial Ischemia Visceral Pain
Carcinoma Neuropathic Pain Osteoarthritis
purinergic receptor P2X, ligand-gated ion channel, 7
neurodegenerative disease Pain Inflammation Neuralgia
Nephritis Dementia Alzheimer Disease Hyperalgesia
Paresthesia Rheumatoid Arthritis Headache
Arthralgia
Mild Cognitive Impairment

(57) **ABSTRACT**

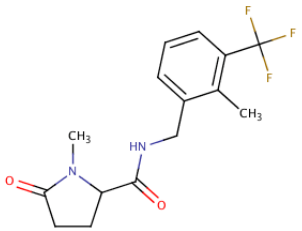
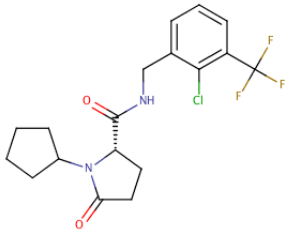
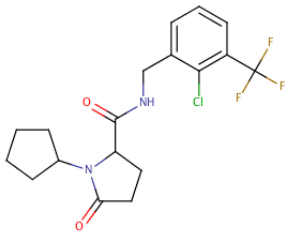
The present invention relates to novel oxo-prolinamide derivatives of formula (I) which modulate P2X7 receptor function and are capable of antagonizing the effects of ATP at the P2X7 receptor

and the use of such compounds or pharmaceutical compositions thereof in the treatment of disorders mediated by the P2X7 receptor, for example pain, inflammation and neurodegeneration.

US-7718693-B2

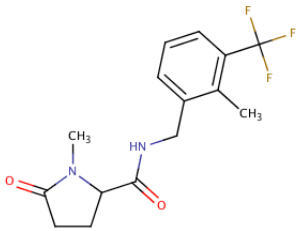
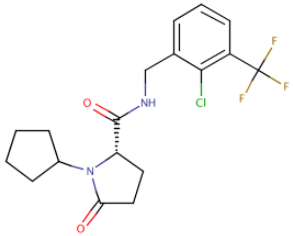
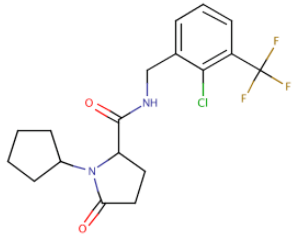
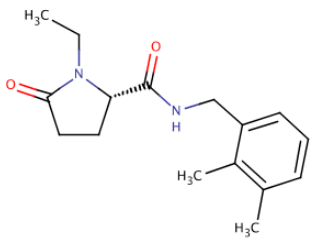
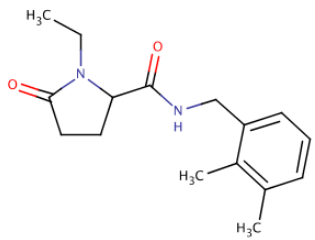
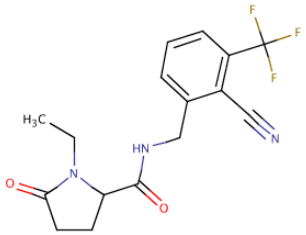
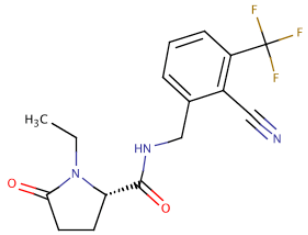
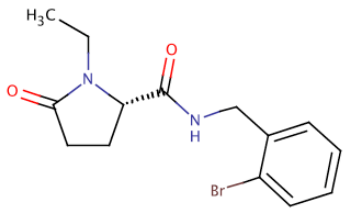
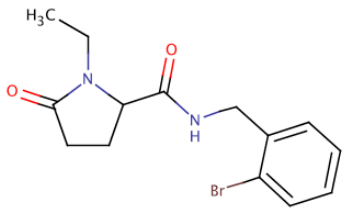
2) Look at compound entities

- 408 compound entities
- Filter: Relevance score >= 1 → 201 remain
- Calculate properties

item.href	type.prefLabel	relevanceScore	smiles
http://rdf.ebi.ac.uk/resource/surechembl/molecule/SCHEMBL2578453	Molecule	2	
http://rdf.ebi.ac.uk/resource/surechembl/molecule/SCHEMBL2584628	Molecule	2	
http://rdf.ebi.ac.uk/resource/surechembl/molecule/SCHEMBL2584624	Molecule	2	

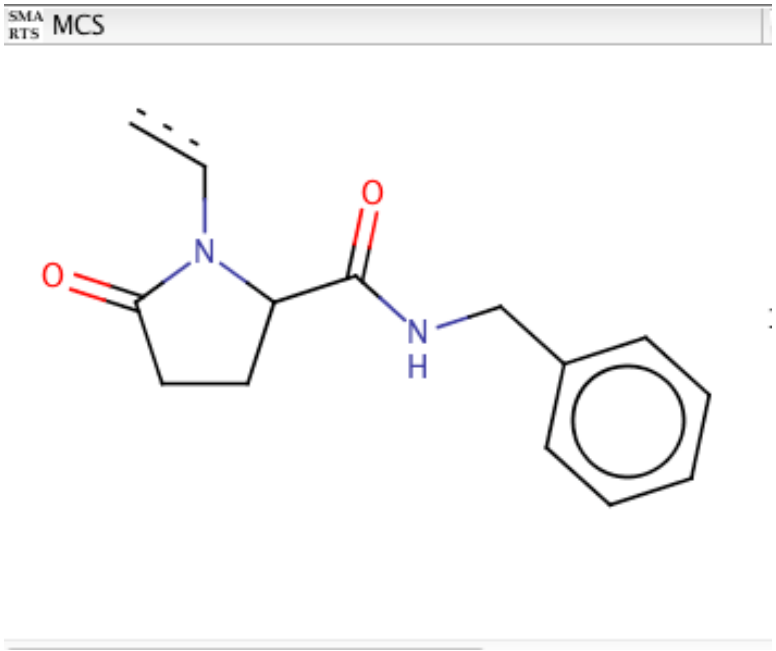
2) Look at compound entities

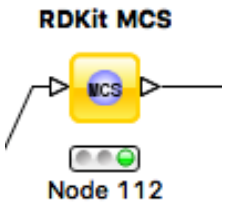
- 408 compound entities
- Filter: Relevance score >= 1 → 201 remain
- Calculate properties

item.href	type.prefLabel	relevanceScore	smiles
http://rdf.ebi.ac.uk/resource/surechembl/molecule/SCHEMBL2578453	Molecule	2	
http://rdf.ebi.ac.uk/resource/surechembl/molecule/SC			   
http://rdf.ebi.ac.uk/resource/surechembl/molecule/SC			   

Does it make sense?

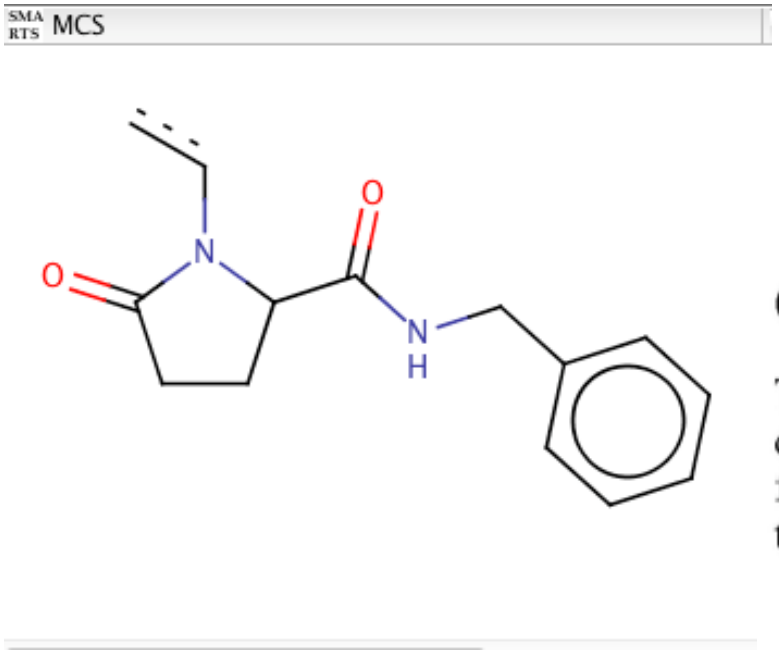
- Calculate MCS

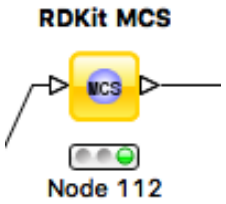




Does it make sense?

- Calculate MCS

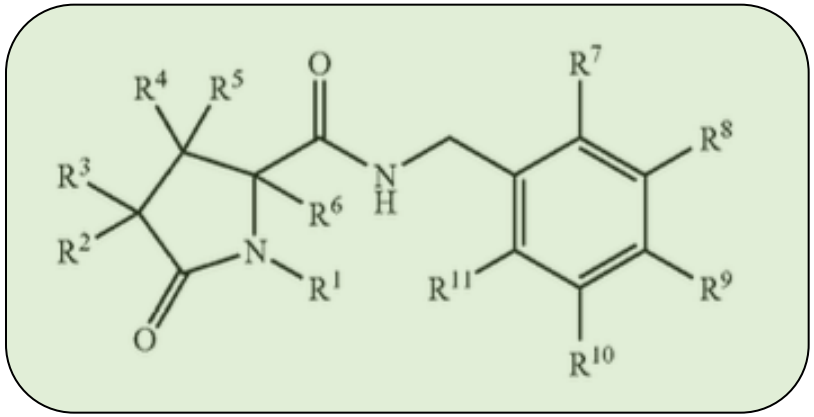




US-7718693-B2

(57) **ABSTRACT**

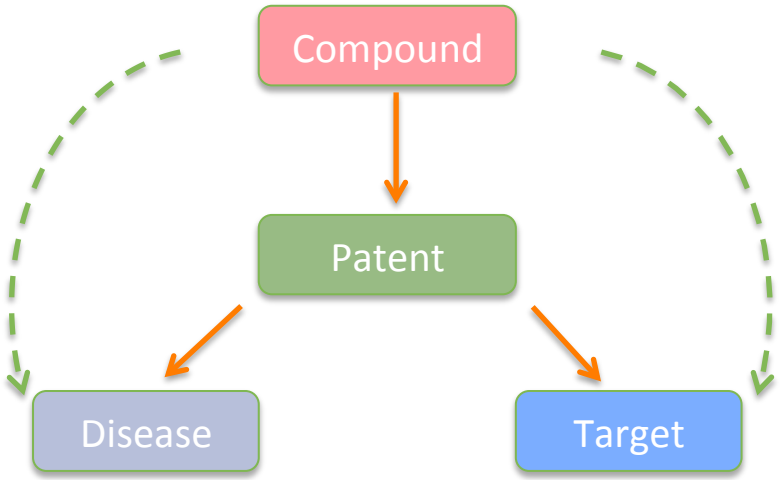
The present invention relates to novel oxo-prolinamide derivatives of formula (I) which modulate P2X7 receptor function and are capable of antagonizing the effects of ATP at the P2X7 receptor



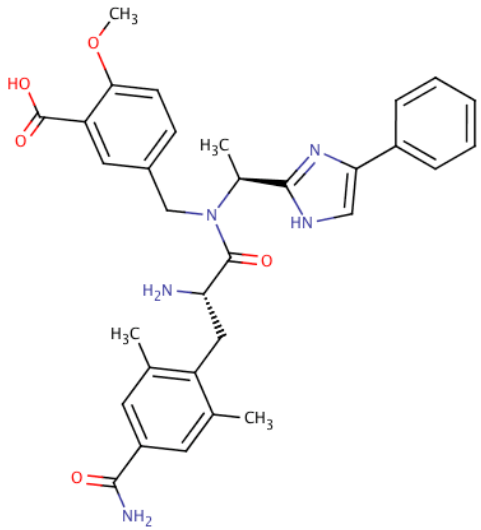
(I)

Use case #2: Drug targets & indications for compound

1. Search patents for a compound (approved drug)
2. Filter to remove noise
   - Frequency, relevance score and classification code
3. For remaining patents, get disease and target entities
4. Filter to remove noise
   - Frequency and relevance score
5. Visualise results



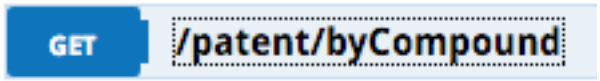
Eluxadoline (JNJ-27018966, VIBERZI)

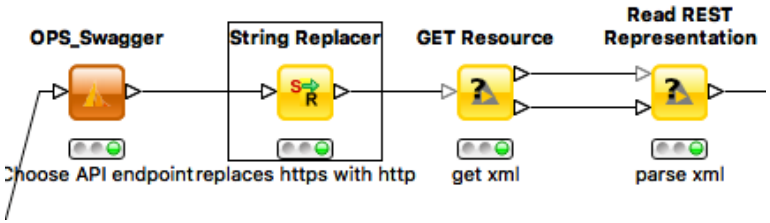


CHEMBL2159122
FDA Approval: 2015

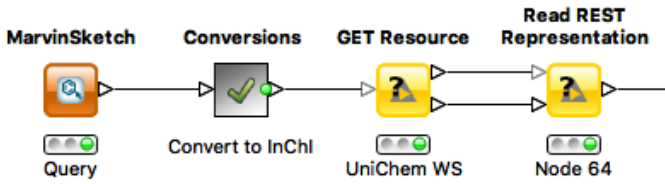
1) Get patents for Eluxadoline

- UniChem call → SCHEMBL12971682
- Compound URI:
  - http://rdf.ebi.ac.uk/resource/surechembl/molecule/SCHEMBL12971682
- API call:





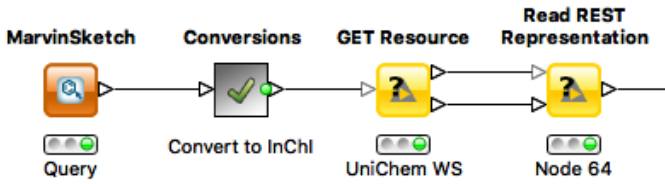
  - Relevance score >=1 → 17 patents (**patentome**):

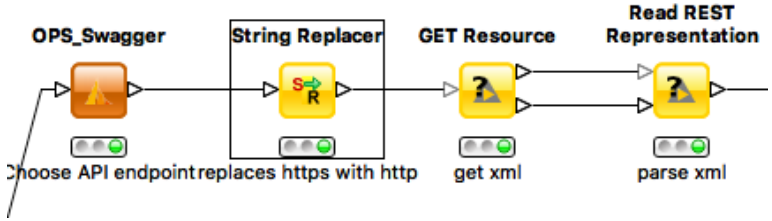


1) Get patents for Eluxadoline

- UniChem call → SCHEMBL12971682
- Compound URI:
  - http://rdf.ebi.ac.uk/resource/surechembl/molecule/SCHEMBL12971682
- API call:

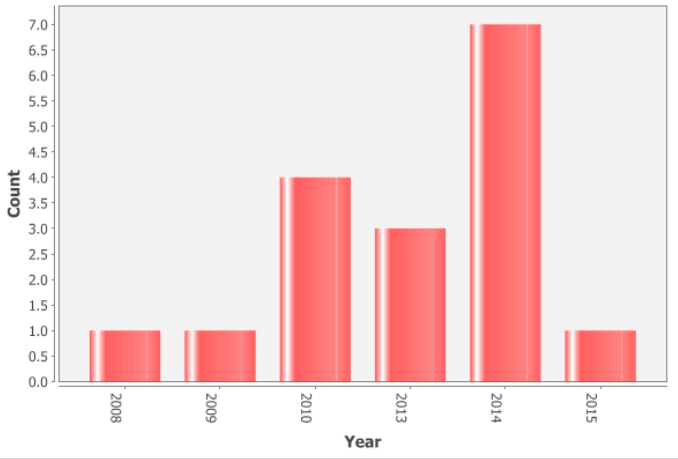
GET /patent/byCompound





- Relevance score >=1 → 17 patents (**patentome**):

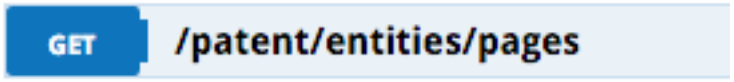
frequency	datePublis...	title
3	24.Apr.2008	NOVEL COMPOUNDS AS OPIOID RECEPTOR MODULATORS
1	08.Dec.2009	2-tert-Butoxycarbonylamino-3-(4-carbamoyl-2,6-dimethylphe...
1	11.Feb.2010	PROCESS FOR THE PREPARATION OF OPIOD MODULATORS
5	22.Jun.2010	Compounds as opioid receptor modulators
2	31.Aug.2010	Compounds as opioid receptor modulators
3	23.Dec.2010	NOVEL COMPOUNDS AS OPIOID RECEPTOR MODULATORS
6	01.Jan.2013	Compounds as opioid receptor modulators
4	11.Apr.2013	NOVEL COMPOUNDS AS OPIOID RECEPTOR MODULATORS
1	23.Oct.2013	Opioid receptor modulators
1	09.Jan.2014	Novel crystals and process of making 5-(-METHYL)-2-METHOX...
8	06.Feb.2014	NOVEL COMPOUNDS AS OPIOID RECEPTOR MODULATORS
1	08.Apr.2014	Crystals and process of making 5-({(2-amino-3-(4-carbamoyl-...
8	08.Jul.2014	Compounds as opioid receptor modulators
8	11.Sep.2014	NOVEL COMPOUNDS AS OPIOID RECEPTOR MODULATORS
3	18.Sep.2014	OPIOID RECEPTOR MODULATOR DOSAGE FORMULATIONS
1	02.Oct.2014	OPIOID RECEPTOR MODULATOR DOSAGE FORMULATIONS
1	21.Jan.2015	PROCESS FOR THE PREPARATION OF OPIOID MODULATORS



2) Get target and disease entities for patents

- API call:

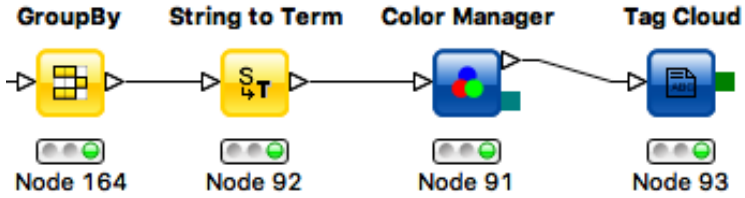
- Classification codes for patents
  - Filter: A61 and C07* → 11 patents remaining
- API call:



  - Filter: Relevance score >= 2, Frequency >= 2 → 2 targets, 21 diseases
- Visualise with tag clouds by frequency

* http://web2.wipo.int/classifications/ipc/ipcpub/

Results



Relevant targets:

opioid receptor, delta 1

opioid receptor, mu 1

Relevant diseases:

Diabetes Mellitus
Postherpetic Neuralgia
Inflammatory Bowel Diseases
Irritable Bowel Syndrome Inflammation
Gastrointestinal Diseases
Chronic Pain Diarrhea Pain Constipation
Neuralgia Hyperalgesia Crohn Disease
Stroke Neuropathic Pain Ulcerative Colitis
Facial Pain Trigeminal Neuralgia Acute Pain
Peripheral Neuropathies Ileus

Does it make sense?

HIGHLIGHTS OF PRESCRIBING INFORMATION
These highlights do not include all the information needed to use VIBERZI safely and effectively. See full prescribing information for VIBERZI.

VIBERZI (eluxadoline) tablets, for oral use, C-X
Initial U.S. Approval: 2015

----------INDICATIONS AND USAGE----------
VIBERZI is a mu-opioid receptor agonist, indicated in adults for the treatment of irritable bowel syndrome with diarrhea (IBS-D). (1)

----------DOSAGE AND ADMINISTRATION----------

- The recommended dosage in adults is 100 mg twice daily taken with food. (2)
- The recommended dosage is 75 mg twice daily taken with food in patients who:
  - do not have a gallbladder (2, 5.1)
  - are unable to tolerate the 100 mg dose (2, 6.1)
  - are receiving concomitant OATP1B1 inhibitors (2, 7)
  - have mild or moderate hepatic impairment (2, 8.6)
- Discontinue VIBERZI in patients who develop severe constipation for more than 4 days (2)
- If a dose is missed, take the next dose at the regular time; do not take 2 doses at once (2)

http://www.accessdata.fda.gov/drugsatfda_docs/label/2015/206940s000lbl.pdf

Does it make sense?

HIGHLIGHTS OF PRESCRIBING INFORMATION
These highlights do not include all the information needed to use VIBERZI safely and effectively. See full prescribing information for VIBERZI.

VIBERZI (eluxadoline) tablets, for oral use, C-X
Initial U.S. Approval: 2015

----------INDICATIONS AND USAGE----------
VIBERZI is a mu-opioid receptor agonist, indicated in adults for the treatment of irritable bowel syndrome with diarrhea (IBS-D). (1)

----------DOSAGE AND ADMINISTRATION----------

- The recommended dosage in adults is 100 mg twice daily taken with food. (2)
- The recommended dosage is 75 mg twice daily taken with food in patients who:
  - do not have a gallbladder (2, 5.1
  - are unable to tolerate the 100 mg
  - are receiving concomitant OATI
  - have mild or moderate hepatic ir
- Discontinue VIBERZI in patients who constipation for more than 4 days (2)
- If a dose is missed, take the next dose take 2 doses at once (2)

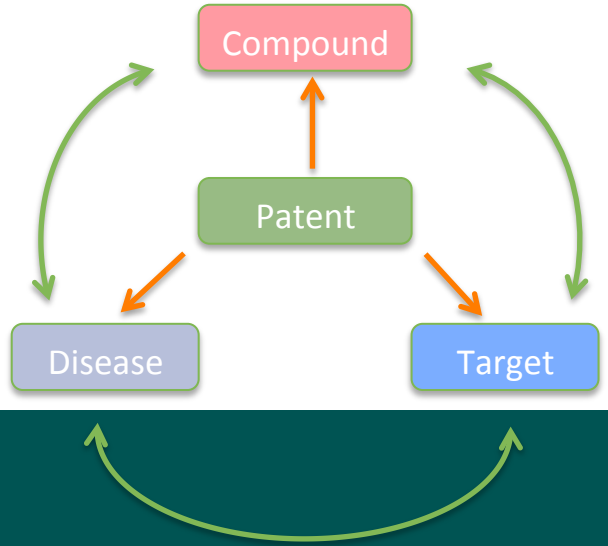
opioid receptor, delta 1
opioid receptor, mu 1

Diabetes Mellitus
Postherpetic Neuralgia
Inflammatory Bowel Diseases
Irritable Bowel Syndrome Inflammation
Gastrointestinal Diseases
Chronic Pain Diarrhea Pain Constipation
Neuralgia Hyperalgesia Crohn Disease
Stroke Neuropathic Pain Ulcerative Colitis
Facial Pain Trigeminal Neuralgia Acute Pain
Peripheral Neuropathies Ileus

http://www.accessdata.fda.gov/drugsatfda_docs/label/2015/206940s000lbl.pdf

Take home message

- It is now possible to extract and interlink the *key* structures, scaffolds, targets and diseases from med. chem. patent corpus *automatically*
  - By high-throughput text-mining only
  - Thanks to simple heuristics (relevance scores and frequency)
  - Using the Open PHACTS API and **KNIME**
  - For the first time in a free resource in such scale



What next: Other ideas and use cases

- Target validation / Druggability
  - For a target, get me all related and relevant diseases
    - Compare with DisGeNET / CTTV, etc.
  - Any known patented scaffolds for my target?
    - Start a pharmacophore hypothesis for patent busting
- Novelty checking / Due diligence
  - What do we know about this scaffold / compound?
- Add ChEMBL pharmacology, pathway information
- Large-scale data mining
  - Annotated patent chemogenomics space and predictive models in KNIME
    - Anyone?

Availability

- API calls will be released to production soon – available for testing now:
  - https://dev.openphacts.org/docs/develop
  - KNIME workflows available on request
- SureChEMBL annotations licensed under CC BY-SA
- SciBite annotations licensed under CC BY-NC-SA
- Check out the Open PHACTS workshop on Friday 9am

W612–W620 *Nucleic Acids Research, 2015, Vol. 43, Web Server issue* *Published online 16 April 2015*
doi: 10.1093/nar/gkv352

ChEMBL web services: streamlining access to drug discovery data and utilities

Mark Davies, Michał Nowotka, George Papadatos, Nathan Dedman, Anna Gaulton, Francis Atkinson, Louisa Bellis and John P. Overington*

Nucleic Acids Research Advance Access published November 7, 2013

Nucleic Acids Research, 2013, 1–8
doi:10.1093/nar/gkt1031

The ChEMBL bioactivity database: an update

A. Patrícia Bento, Anna Gaulton, Anne Hersey, Louisa J. Bellis, Jon Chambers, Mark Davies, Felix A. Krüger, Yvonne Light, Lora Mak, Shaun McGlinchey, Michal Nowotka, George Papadatos, Rita Santos and John P. Overington*

Chambers *et al. Journal of Cheminformatics* 2014, **6**:43
http://www.jcheminf.com/content/6/1/43

Journal of Cheminformatics

DATABASE **Open Access**

UniChem: extension of InChI-based compound mapping to salt, connectivity and stereochemistry layers

Jon Chambers*, Mark Davies, Anna Gaulton, George Papadatos, Anne Hersey and John P Overington

Bioinformatics Advance Access published November 20, 2013

Database and ontologies

myChEMBL: A virtual machine implementation of open data and cheminformatics tools

Rodrigo Ochoa, Mark Davies, George Papadatos, Francis Atkinson, John P. Overington*
European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, CB10 1SD, UK.

Papadatos *et al. Journal of Cheminformatics* 2014, **6**:40
http://www.jcheminf.com/content/6/1/40

Journal of Cheminformatics

SOFTWARE **Open Access**

A document classifier for medicinal chemistry publications trained on the ChEMBL corpus

George Papadatos†, Gerard JP van Westen*†, Samuel Croset, Rita Santos, Simone Trubian and John P Overington

Challenges **2014**, *5*, 444-449; doi:10.3390/challe5020444

OPEN ACCESS
challenges
ISSN 2078-154
www.mdpi.com/journal/challenge

Communication

ChEMBL Beaker: A Lightweight Web Framework Providing Robust and Extensible Cheminformatics Services

Michał Nowotka, Mark Davies, George Papadatos and John P. Overington *

EMBL-EBI

Nucleic Acids Research Advance Access published November 17, 2015

Nucleic Acids Research, 2015 1
doi: 10.1093/nar/gkv1253

SureChEMBL: a large-scale, chemically annotated patent document database

George Papadatos[1], Mark Davies[1], Nathan Dedman[1], Jon Chambers[1], Anna Gaulton[1], James Siddle[2], Richard Koks[2], Sean A. Irvine[3], Joe Pettersson[4], Nicko Goncharoff[2,*], Anne Hersey[1,*] and John P. Overington[1,*]

[1]European Molecular Biology Laboratory, European Bio
Campus, Hinxton, Cambridgeshire CB10 1SD, UK, [2]Di
3240, New Zealand and [4]McKinsey & Company, Londo

F1000Research

F1000Research 2016, 5(ELIXIR):160 Last updated: 11 FEB 2016

CrossMark click for updates

RESEARCH ARTICLE

Patterns of database citation in articles and patents indicate long-term scientific and industry value of biological data resources [version 1; referees: awaiting peer review]

David Bousfield[1,2], Johanna McEntyre[3], Sameer Velankar[3], George Papadatos[3],
-Hyub Kim[3], Florian Graef[3], Vid Vartak[3],

rmatics Institute (EMBL-EBI), Wellcome Genome Campus, Cambridge, UK

Senger *et al. J Cheminform (2015) 7:49*
DOI 10.1186/s13321-015-0097-z

Journal of Cheminformatics

RESEARCH ARTICLE Open Access

CrossMark

Managing expectations: assessment of chemistry databases generated by automated extraction of chemical structures from patents

Stefan Senger[1*], Luca Bartek[1], George Papadatos[2] and Anna Gaulton[2]

EMBL-EBI

feature

CrossMark

Drug discovery FAQs: workflows for answering multidomain drug discovery questions

Christine Chichester[1,5], christine.chichester@isb-sib.ch, Daniela Digles[2,5], Ronald Siebes[3], Antonis Loizou[3], Paul Groth[3] and Lee Harland[4]

Scientific competency questions as the basis for semantically enriched open pharmacological space development

Kamal Azzaoui[1], Edgar Jacoby[14], Stefan Senger[2], Emiliano Cuadrado Rodríguez[3], Mabel Loza[3], Barbara Zdrazil[4], Marta Pinto[4], Antony J. Williams[5], Victor de la Torre[6], Jordi Mestres[7], Manuel Pastor[7], Olivier Taboureau[8], Matthias Rarey[9], Christine Chichester[10], Steve Pettifer[11], Niklas Blomberg[12,a], Lee Harland[13], Bryn Williams-Jones[13] and Gerhard F. Ecker[4]

Acknowledgements

- ChEMBL and SureChEMBL
  - Anna Gaulton
  - Mark Davies
  - Nathan Dedman
  - James Siddle
  - Anne Hersey
- SciBite
  - Lee Harland
- Open PHACTS consortium
  - Nick Lynch
  - Daniela Digles
  - Antonis Loizou
- EBI alumni
  - Edmund Duesbury
  - Stephan Beisken

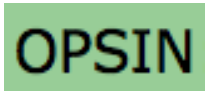
innovative medicines initiative

wellcome trust Strategic Award

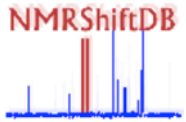
EMBL-EBI

The Open PHACTS Foundation

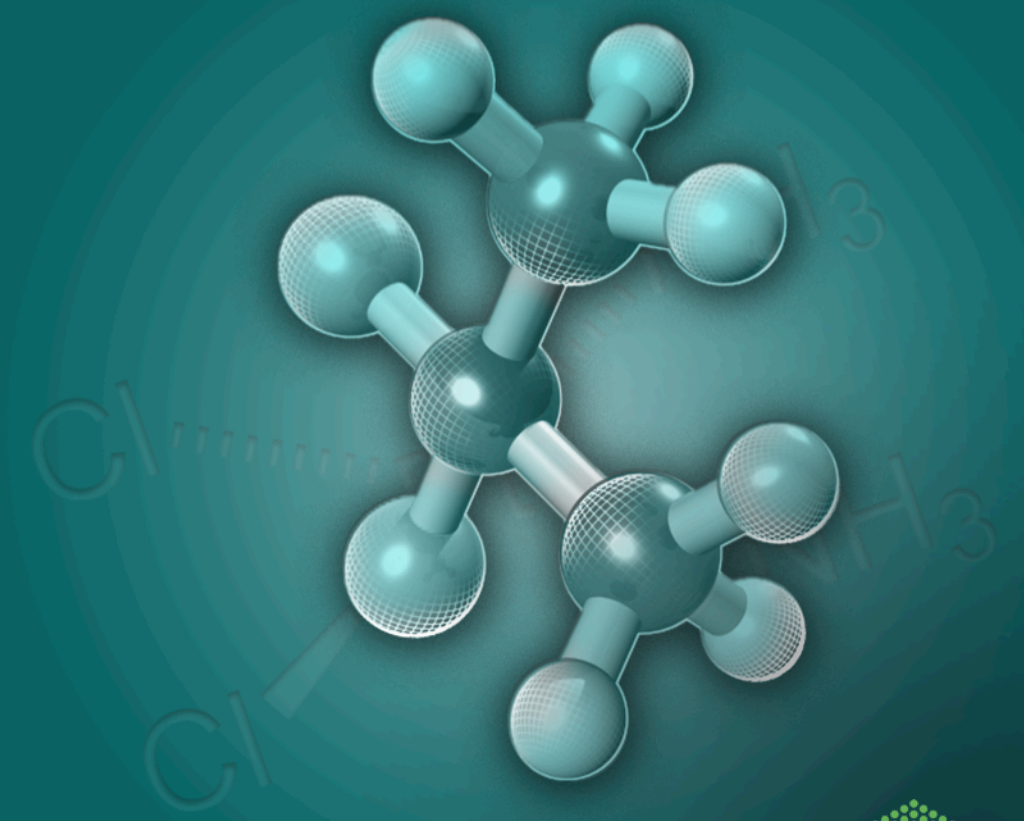
Technology partners



The Data Are Out There



Leveraging Open Chemogenomics Data and Tools with KNIME



George Papadatos

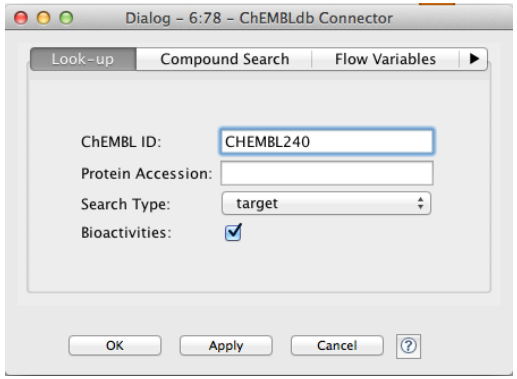
ChEMBL Group

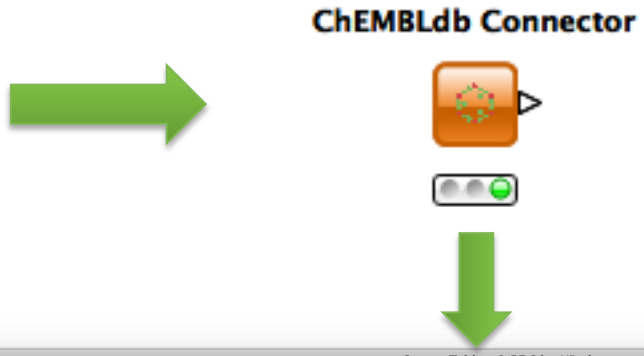
georgep@ebi.ac.uk

EMBL-EBI

Back-up slides

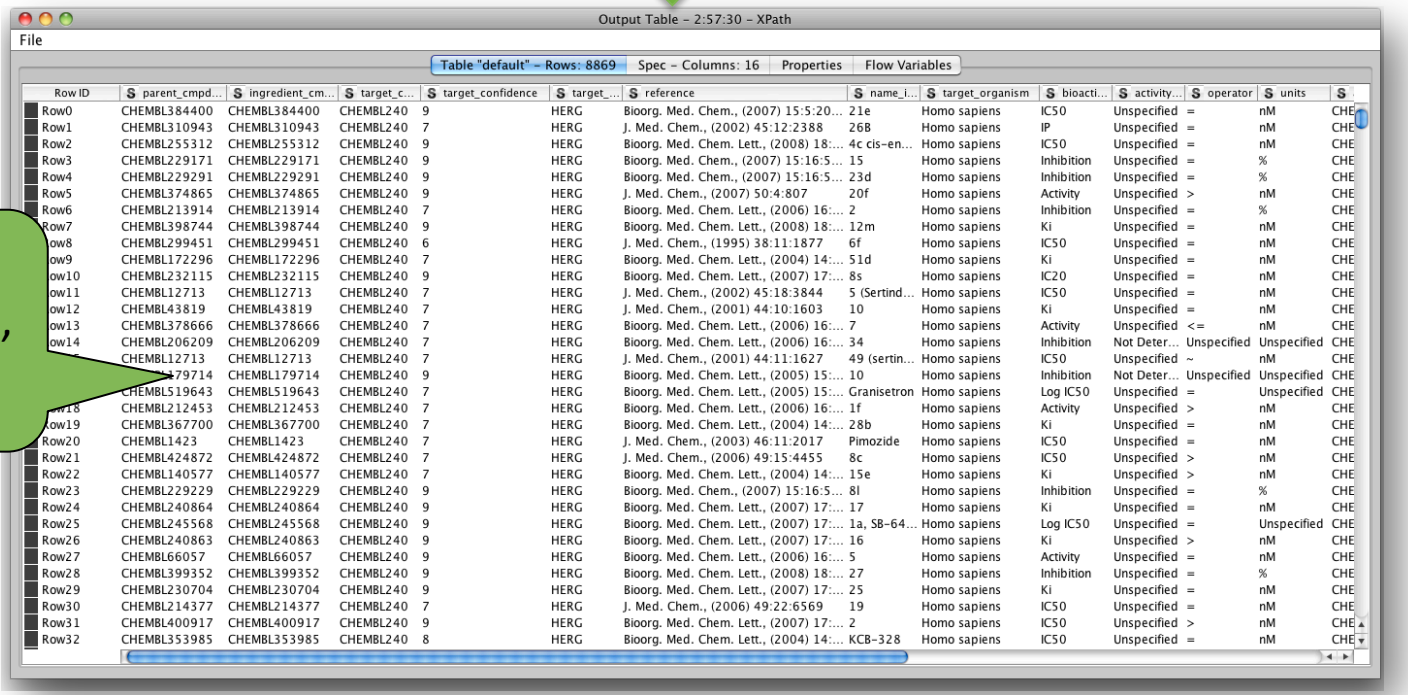
Example: All bioactivities for hERG





All bioactivities for hERG

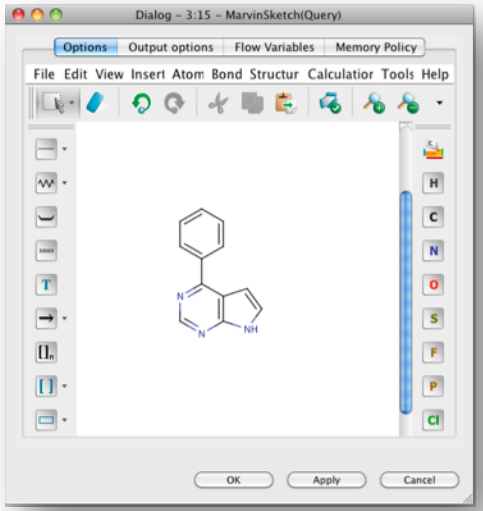
Activity value, assay description, compound, reference



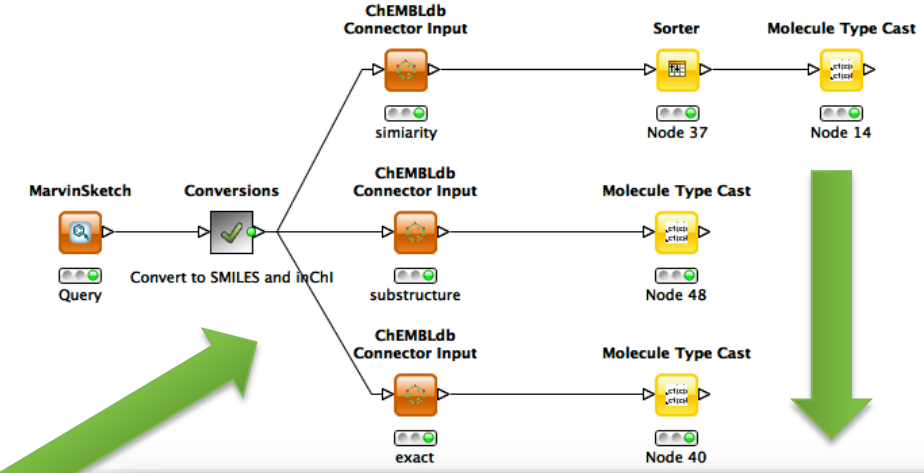
Row ID	parent_cmpd...	ingredient_cm...	target_c...	target_confidence	target_...	reference	name_i...	target_organism	bioacti...	activity...	operator	units	
Row0	CHEMBL384400	CHEMBL384400	CHEMBL240	9	HERG	Bioorg. Med. Chem., (2007) 15:5:20...	21e	Homo sapiens	IC50	Unspecified	=	nM	CHE
Row1	CHEMBL310943	CHEMBL310943	CHEMBL240	7	HERG	J. Med. Chem., (2002) 45:12:2388	26B	Homo sapiens	IP	Unspecified	=	nM	CHE
Row2	CHEMBL255312	CHEMBL255312	CHEMBL240	9	HERG	Bioorg. Med. Chem. Lett., (2008) 18:...	4c cis-en...	Homo sapiens	IC50	Unspecified	=	nM	CHE
Row3	CHEMBL229171	CHEMBL229171	CHEMBL240	9	HERG	Bioorg. Med. Chem., (2007) 15:16:5...	15	Homo sapiens	Inhibition	Unspecified	=	%	CHE
Row4	CHEMBL229291	CHEMBL229291	CHEMBL240	9	HERG	Bioorg. Med. Chem., (2007) 15:16:5...	23d	Homo sapiens	Inhibition	Unspecified	=	%	CHE
Row5	CHEMBL374865	CHEMBL374865	CHEMBL240	9	HERG	J. Med. Chem., (2007) 50:4:807	20f	Homo sapiens	Activity	Unspecified	>	nM	CHE
Row6	CHEMBL213914	CHEMBL213914	CHEMBL240	7	HERG	Bioorg. Med. Chem. Lett., (2006) 16:...	2	Homo sapiens	Inhibition	Unspecified	=	%	CHE
Row7	CHEMBL398744	CHEMBL398744	CHEMBL240	9	HERG	Bioorg. Med. Chem. Lett., (2008) 18:...	12m	Homo sapiens	Ki	Unspecified	=	nM	CHE
ow8	CHEMBL299451	CHEMBL299451	CHEMBL240	6	HERG	J. Med. Chem., (1995) 38:11:1877	6f	Homo sapiens	IC50	Unspecified	=	nM	CHE
ow9	CHEMBL172296	CHEMBL172296	CHEMBL240	7	HERG	Bioorg. Med. Chem. Lett., (2004) 14:...	51d	Homo sapiens	Ki	Unspecified	=	nM	CHE
ow10	CHEMBL232115	CHEMBL232115	CHEMBL240	9	HERG	Bioorg. Med. Chem. Lett., (2007) 17:...	8s	Homo sapiens	IC20	Unspecified	=	nM	CHE
ow11	CHEMBL12713	CHEMBL12713	CHEMBL240	7	HERG	J. Med. Chem., (2002) 45:18:3844	5 (Sertind...	Homo sapiens	IC50	Unspecified	=	nM	CHE
ow12	CHEMBL43819	CHEMBL43819	CHEMBL240	7	HERG	J. Med. Chem., (2001) 44:10:1603	10	Homo sapiens	Ki	Unspecified	=	nM	CHE
ow13	CHEMBL378666	CHEMBL378666	CHEMBL240	7	HERG	Bioorg. Med. Chem. Lett., (2006) 16:...	7	Homo sapiens	Activity	Unspecified	<=	nM	CHE
ow14	CHEMBL206209	CHEMBL206209	CHEMBL240	7	HERG	Bioorg. Med. Chem. Lett., (2006) 16:...	34	Homo sapiens	Inhibition	Not Deter...	Unspecified	Unspecified	CHE
	CHEMBL12713	CHEMBL12713	CHEMBL240	7	HERG	J. Med. Chem., (2001) 44:11:1627	49 (sertin...	Homo sapiens	IC50	Unspecified	~	nM	CHE
	179714	CHEMBL179714	CHEMBL240	9	HERG	Bioorg. Med. Chem. Lett., (2005) 15:...	10	Homo sapiens	Inhibition	Not Deter...	Unspecified	Unspecified	CHE
	CHEMBL519643	CHEMBL519643	CHEMBL240	7	HERG	Bioorg. Med. Chem. Lett., (2005) 15:...	Granisetron	Homo sapiens	Log IC50	Unspecified	=	Unspecified	CHE
w18	CHEMBL212453	CHEMBL212453	CHEMBL240	7	HERG	Bioorg. Med. Chem. Lett., (2006) 16:...	1f	Homo sapiens	Activity	Unspecified	>	nM	CHE
ow19	CHEMBL367700	CHEMBL367700	CHEMBL240	7	HERG	Bioorg. Med. Chem. Lett., (2004) 14:...	28b	Homo sapiens	Ki	Unspecified	=	nM	CHE
Row20	CHEMBL1423	CHEMBL1423	CHEMBL240	7	HERG	J. Med. Chem., (2003) 46:11:2017	Pimozide	Homo sapiens	IC50	Unspecified	=	nM	CHE
Row21	CHEMBL424872	CHEMBL424872	CHEMBL240	7	HERG	J. Med. Chem., (2006) 49:15:4455	8c	Homo sapiens	IC50	Unspecified	>	nM	CHE
Row22	CHEMBL140577	CHEMBL140577	CHEMBL240	7	HERG	Bioorg. Med. Chem. Lett., (2004) 14:...	15e	Homo sapiens	Ki	Unspecified	>	nM	CHE
Row23	CHEMBL229229	CHEMBL229229	CHEMBL240	9	HERG	Bioorg. Med. Chem., (2007) 15:16:5...	8l	Homo sapiens	Inhibition	Unspecified	=	%	CHE
Row24	CHEMBL240864	CHEMBL240864	CHEMBL240	9	HERG	Bioorg. Med. Chem. Lett., (2007) 17:...	17	Homo sapiens	Ki	Unspecified	=	nM	CHE
Row25	CHEMBL245568	CHEMBL245568	CHEMBL240	9	HERG	Bioorg. Med. Chem. Lett., (2007) 17:...	1a, SB-64...	Homo sapiens	Log IC50	Unspecified	=	Unspecified	CHE
Row26	CHEMBL240863	CHEMBL240863	CHEMBL240	9	HERG	Bioorg. Med. Chem. Lett., (2007) 17:...	16	Homo sapiens	Ki	Unspecified	>	nM	CHE
Row27	CHEMBL66057	CHEMBL66057	CHEMBL240	9	HERG	Bioorg. Med. Chem. Lett., (2006) 16:...	5	Homo sapiens	Activity	Unspecified	=	nM	CHE
Row28	CHEMBL399352	CHEMBL399352	CHEMBL240	9	HERG	Bioorg. Med. Chem. Lett., (2008) 18:...	27	Homo sapiens	Inhibition	Unspecified	=	%	CHE
Row29	CHEMBL230704	CHEMBL230704	CHEMBL240	9	HERG	Bioorg. Med. Chem. Lett., (2007) 17:...	25	Homo sapiens	Ki	Unspecified	=	nM	CHE
Row30	CHEMBL214377	CHEMBL214377	CHEMBL240	7	HERG	J. Med. Chem., (2006) 49:22:6569	19	Homo sapiens	IC50	Unspecified	=	nM	CHE
Row31	CHEMBL400917	CHEMBL400917	CHEMBL240	9	HERG	Bioorg. Med. Chem. Lett., (2007) 17:...	2	Homo sapiens	IC50	Unspecified	>	nM	CHE
Row32	CHEMBL353985	CHEMBL353985	CHEMBL240	8	HERG	Bioorg. Med. Chem. Lett., (2004) 14:...	KCB-328	Homo sapiens	IC50	Unspecified	=	nM	CHE

EMBL-EBI

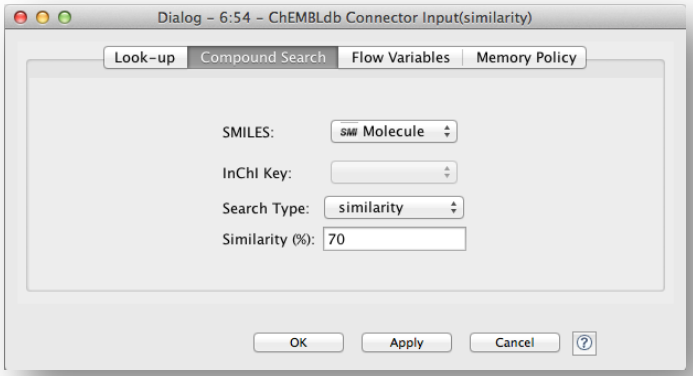
Example: Compound searching in ChEMBL

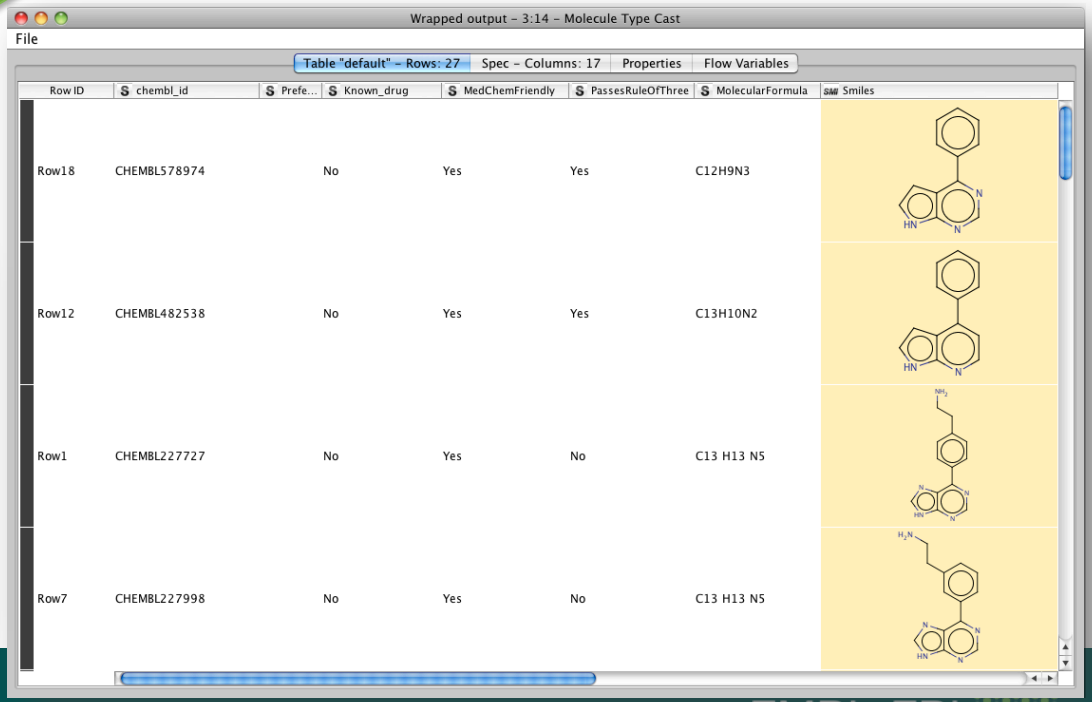


Query



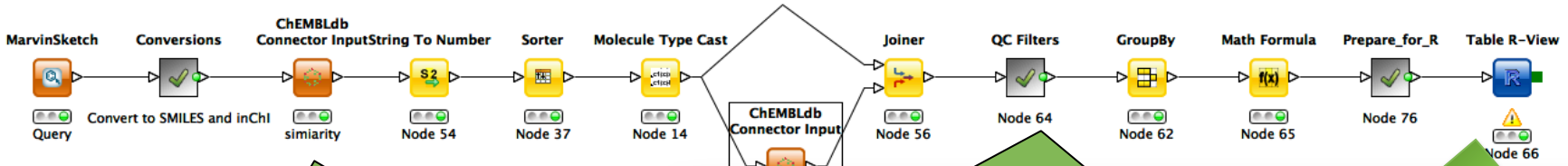
List of NNs





EMBL-EBI

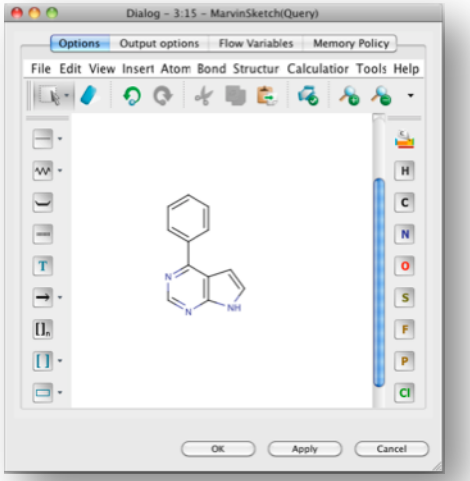
Example: Polypharmacology profile



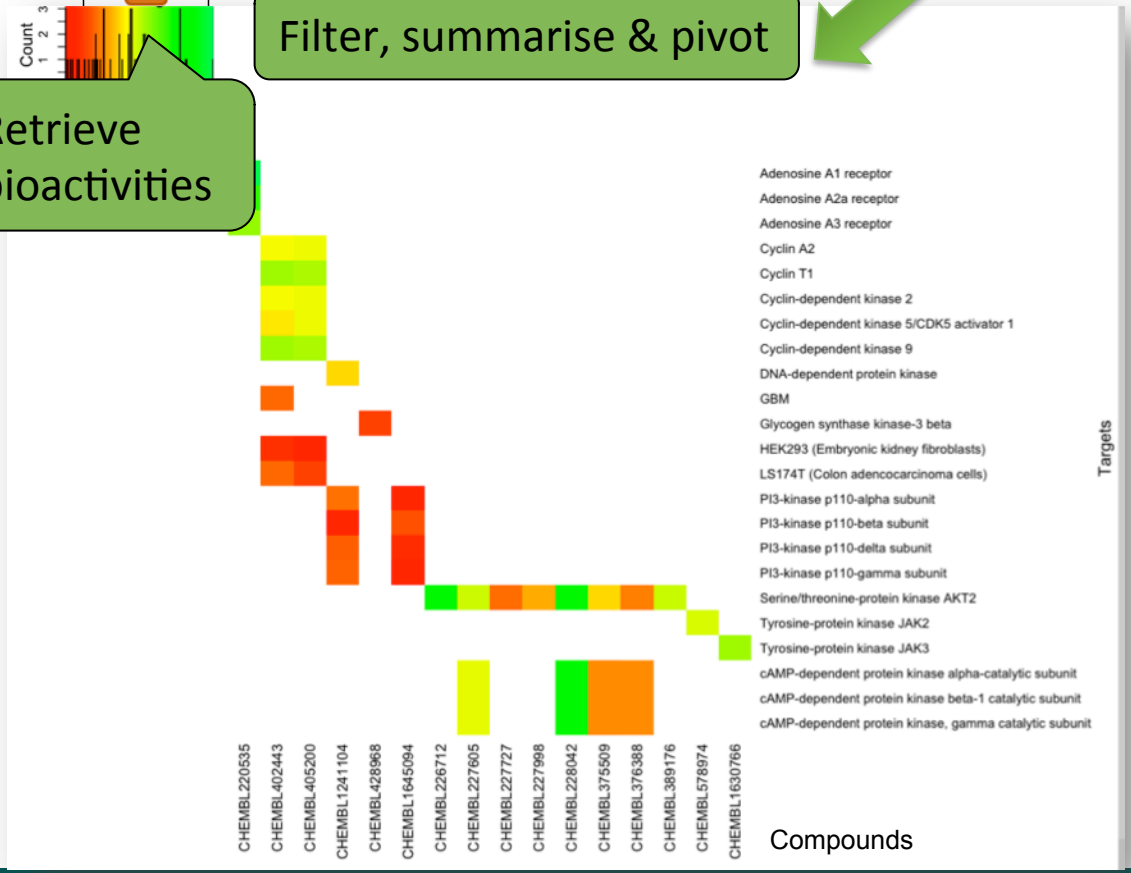
Find NNs

Retrieve bioactivities

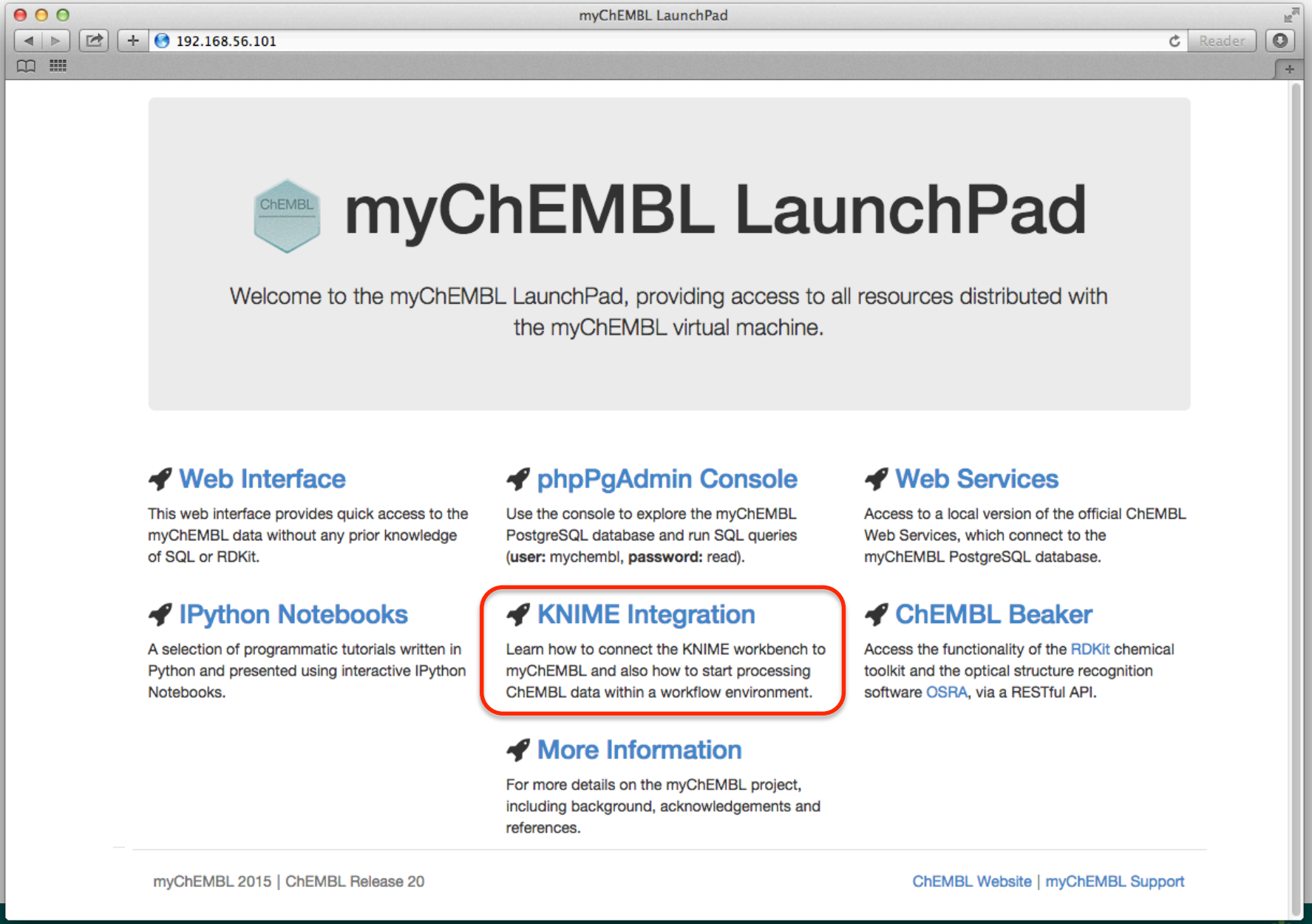
Filter, summarise & pivot



Query



Compounds


myChEMBL LaunchPad
192.168.56.101
Reader
ChEMBL
myChEMBL LaunchPad
Welcome to the myChEMBL LaunchPad, providing access to all resources distributed with the myChEMBL virtual machine.
Web Interface
This web interface provides quick access to the myChEMBL data without any prior knowledge of SQL or RDKit.
phpPgAdmin Console
Use the console to explore the myChEMBL PostgreSQL database and run SQL queries (user: mychembl, password: read).
Web Services
Access to a local version of the official ChEMBL Web Services, which connect to the myChEMBL PostgreSQL database.
IPython Notebooks
A selection of programmatic tutorials written in Python and presented using interactive IPython Notebooks.
KNIME Integration
Learn how to connect the KNIME workbench to myChEMBL and also how to start processing ChEMBL data within a workflow environment.
ChEMBL Beaker
Access the functionality of the RDKit chemical toolkit and the optical structure recognition software OSRA, via a RESTful API.
More Information
For more details on the myChEMBL project, including background, acknowledgements and references.
myChEMBL 2015 | ChEMBL Release 20
ChEMBL Website | myChEMBL Support

myChEMBL LaunchPad

192.168.56.101/knime

Reader

myChEMBL LaunchPad | Home | myChEMBL LaunchPad

myChEMBL KNIME Integration

Home

What is KNIME?

KNIME is an open source modular graphical workbench for data processing and analysis. KNIME users can build workflows by executing predefined components, called *nodes*, in a process known as *visual programming*. Among other fields, KNIME has several extensions for life sciences, including bio- and chemoinformatics.

How do I get KNIME?

Installing KNIME is really easy across all popular platforms (Windows, Mac OS, Linux). Just follow this link.

How do I find out more about KNIME?

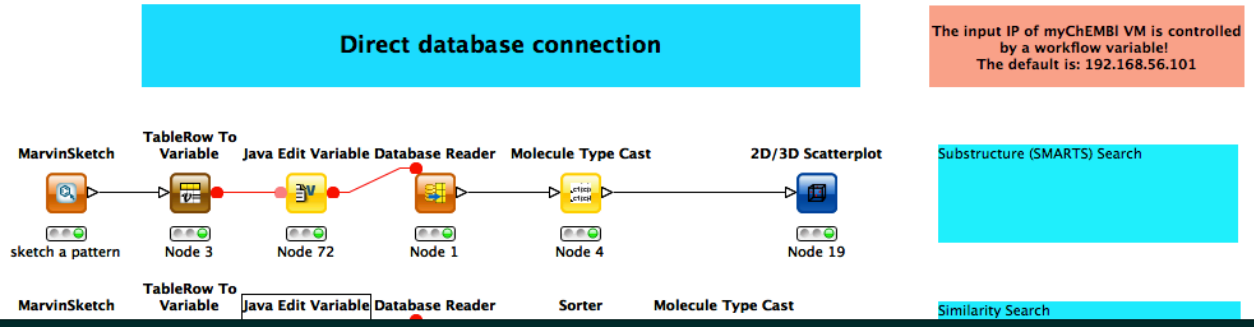
Here's a useful page. For a basic introduction to KNIME followed by a few chemoinformatics applciations, see this slideshow.

How to I use KNIME with the myChEMBL VM?

You can access the data in myChEMBL directly from a KNIME desktop running on your host machine. This can be achieved in two ways:

- Direct database connection with the Database Reader node. The node can execute arbitrary SQL SELECT commands and retrieve the results in a KNIME table.
- Calls to the RESTful web services that are hosted in myChEMBL. This is achieved with the KNIME REST (KREST) nodes, which are available as community node contribution. The KREST nodes send GET or POST requests to the server and format the response back to a KNIME table.

For an example workflow showcasing the KNIME/myChEMBL connectivity, download this file to your host machine. Then, open your KNIME desktop on your host machine and click File --> Import KNIME workflow... --> Select archive file --> Browse.. and then navigate to the zip file you've just downloaded and click Finish. If everything worked properly, you will be able to see something like this on your KNIME desktop:

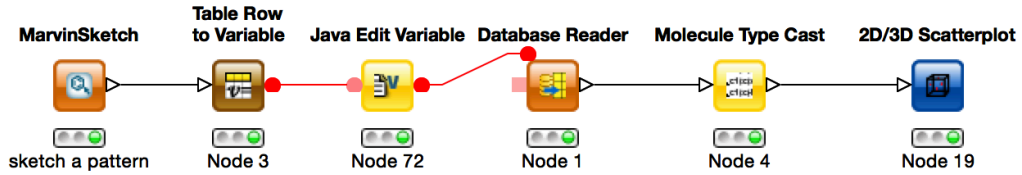


EMBL-EBI

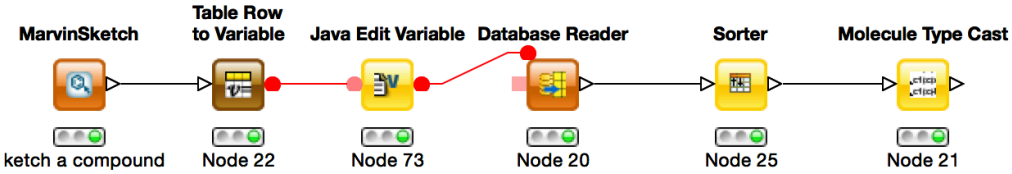
Accessing *local* data and services with myChEMBL

Direct database connection

The input IP of myChEMBl VM is controlled by a workflow variable!
The default is: 192.168.56.101



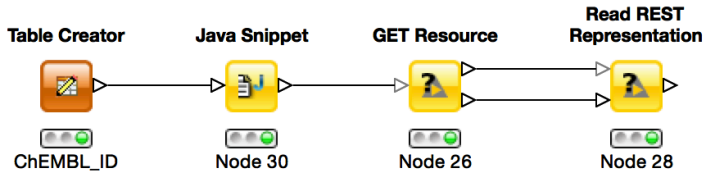
Substructure (SMARTS) Search



Similarity Search

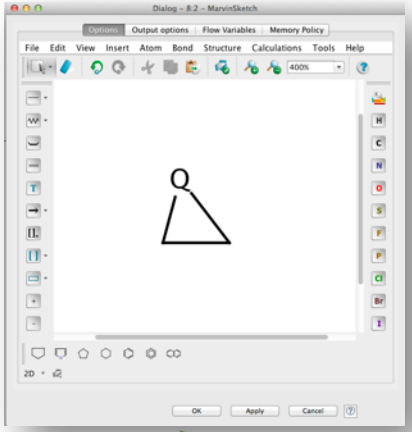
Web service connection

(requires the KREST community nodes
http://tech.knime.org/book/krest-rest-nodes-for-knime-trusted-extension



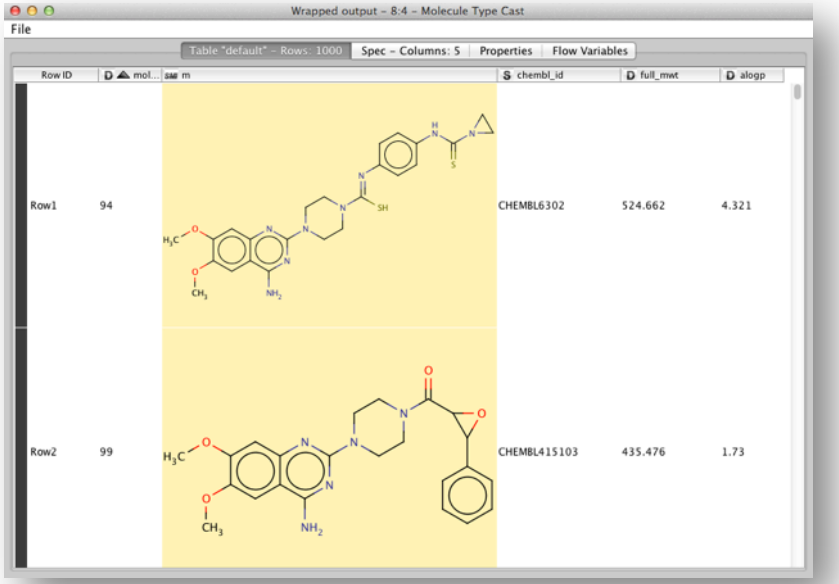
Get bioactivities for a compound

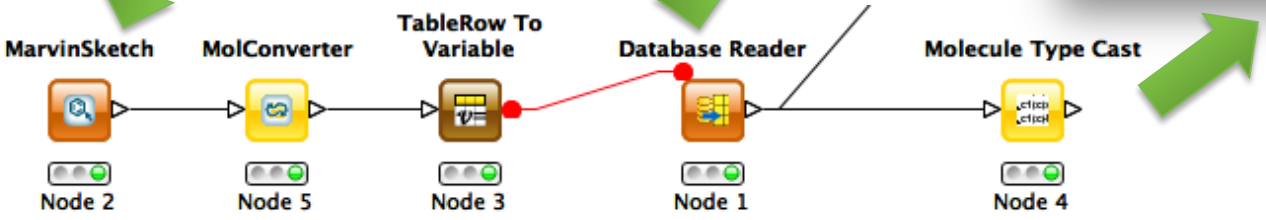
Using KNIME to connect to myChEMBL



```
SELECT mr.*, md.chembl_id,
cp.full_mwt, cp.alogp
from mols_rdkit mr,
molecule_dictionary md,
compound_properties cp
where
mr.m @> '$${SMolecule}$$'::qmol
and
mr.molregno = md.molregno
and
md.molregno = cp.molregno;
```

RDKit





EMBL-EBI

Cheminformatics utilities

ChEMBL Beaker API live documentation Explorer

Version: 0.5.24

- GET addHs/:CTAB
- GET atomIsInRing/:CTAB/:INDEX/:SIZE
- GET break_bonds/:CTAB
- GET ctab23D/:CTAB
- GET ctab2image/:CTAB
- GET ctab2inchi/:CTAB
- GET ctab2json/:CTAB
- GET ctab2smiles/:CTAB
- GET ctab2svg/:CTAB
- GET descriptors/:CTAB
- GET getNumAtoms/:CTAB
- GET image2ctab/:IMAGE
- GET inchi2ctab/:INCHI
- GET inchi2inchiKey/:INCHI

- Chemical format conversions
- Dynamic image generation
- Image processing (via OSRA)
- Descriptors and property calculations
- Chemical modifications and standardization

Description

Converts SMILES to PNG image. This method accepts urlsafe_base64 encoded string containing single or multiple SMILES optionally containing header line, specific to *.smi format. Size is the optional size of image in pixels (default value is 200 px). Legend is optional label in the bottom of image.

Requires

SMILES

Formats

text

Enter a value for SMILES and click GET to test the service!

GET /smiles2image/ ɔcnc5)c3)CC1

Request URI

smiles2image/Q04xQ0NOKENjMmNjYyhjYzIpQyg9TyIOYzNjY2MoQyIjKE5jNG5jY2MobjQpYzVjY2NuYzUpYzMpQ0MxIA==

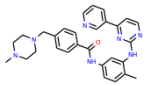
Response Code

200

Response

OK

Response Body Copy

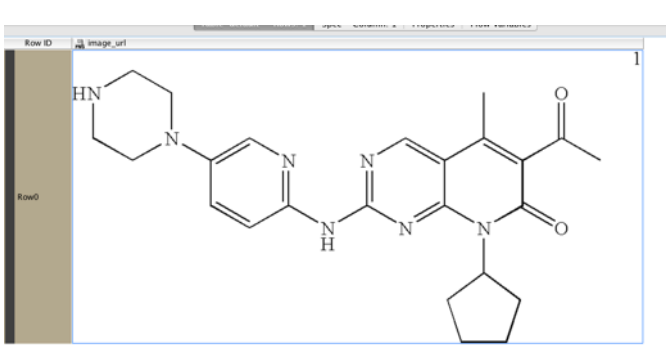


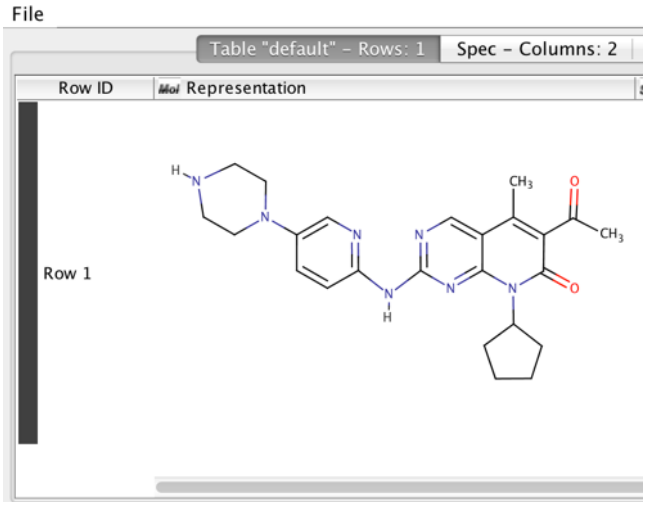
https://www.ebi.ac.uk/chembl/api/utils/docs

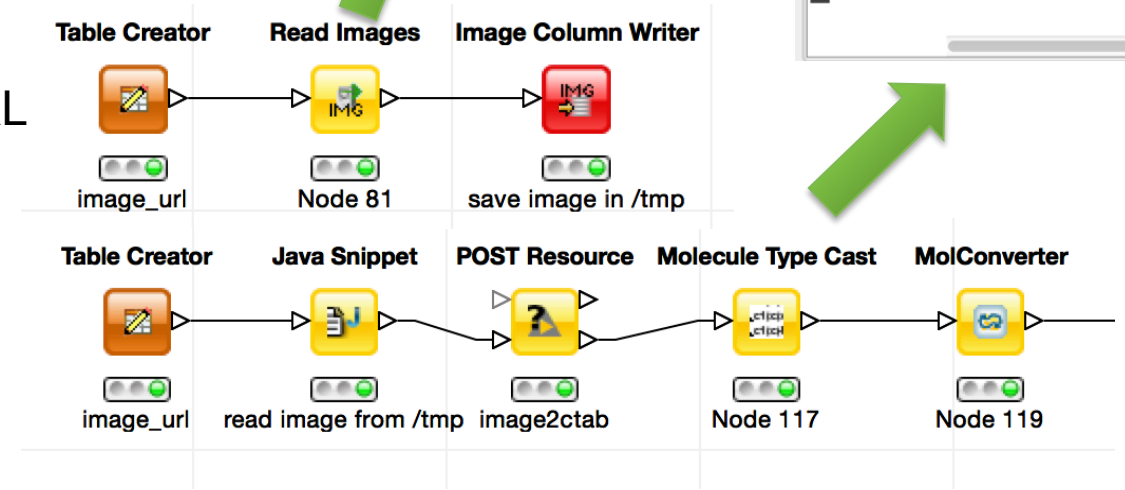
EMBL-EBI

Example: RESTful Image to Structure conversion

image URL

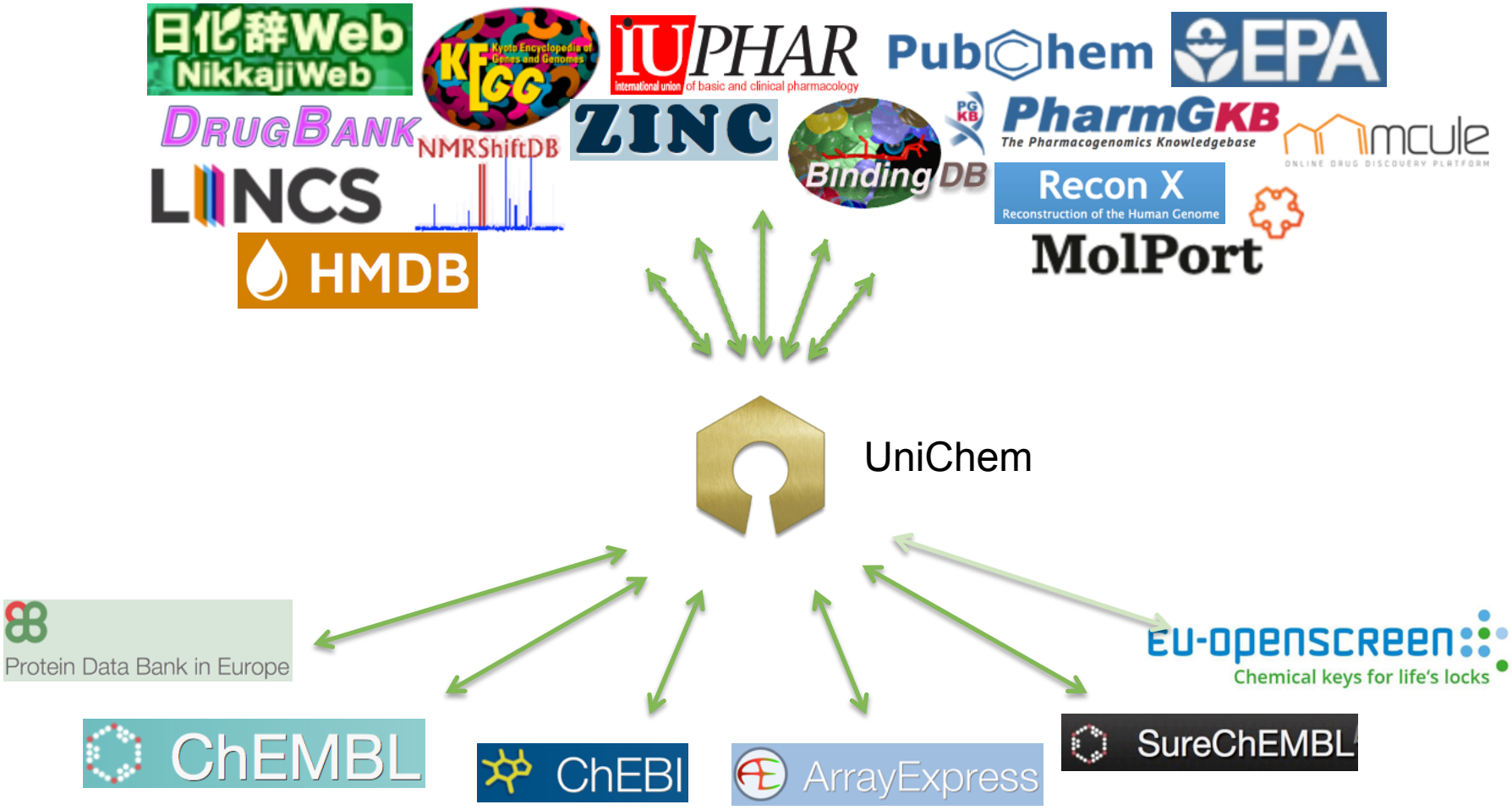






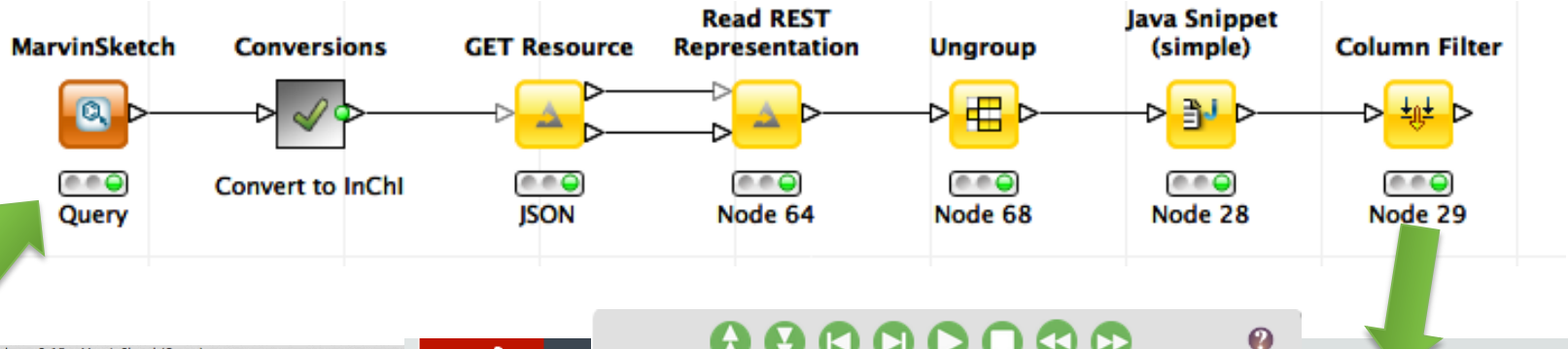
OSRA

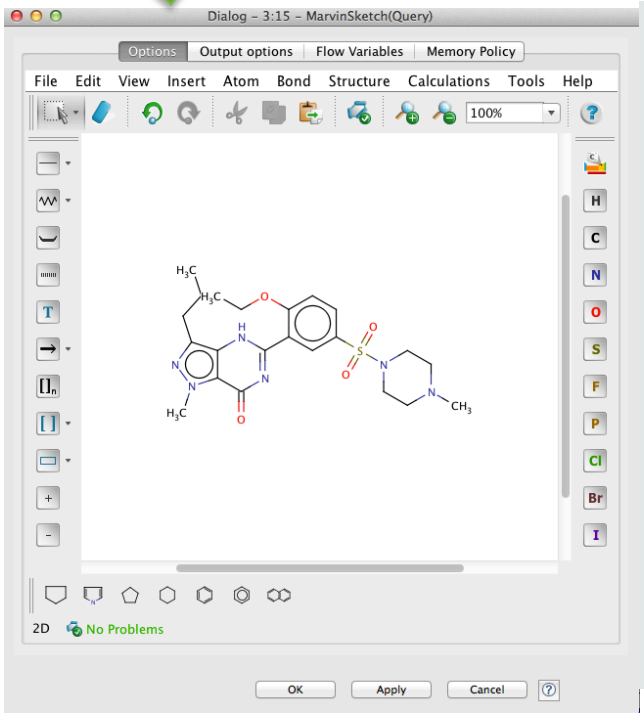
UniChem – Compound Mapping across Resources

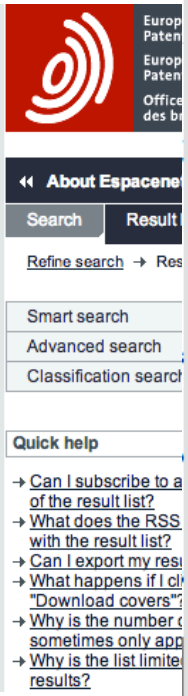


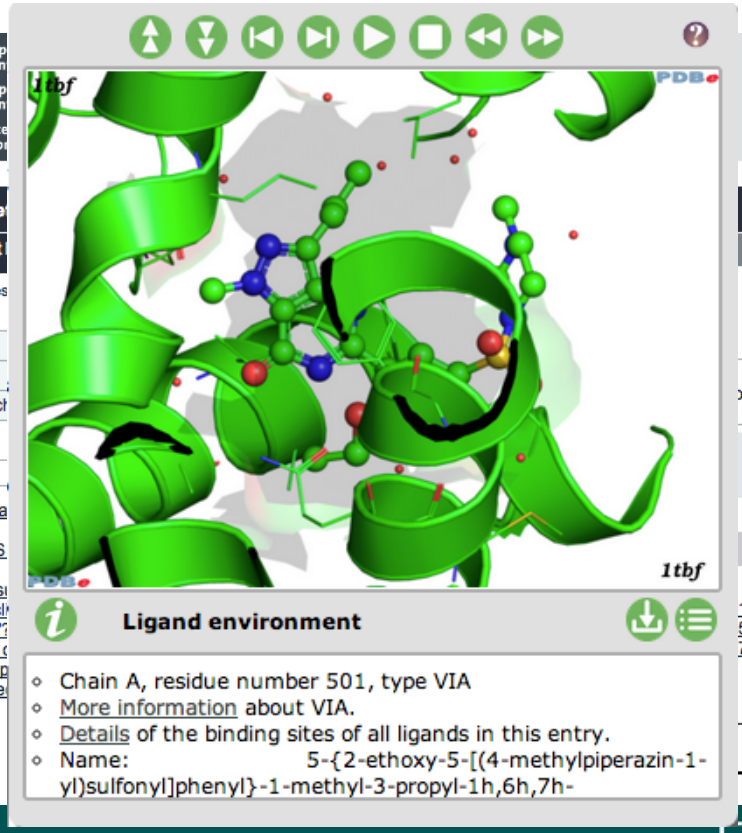
https://www.ebi.ac.uk/unichem/

Novelty checking with UniChem









EMBL-EBI