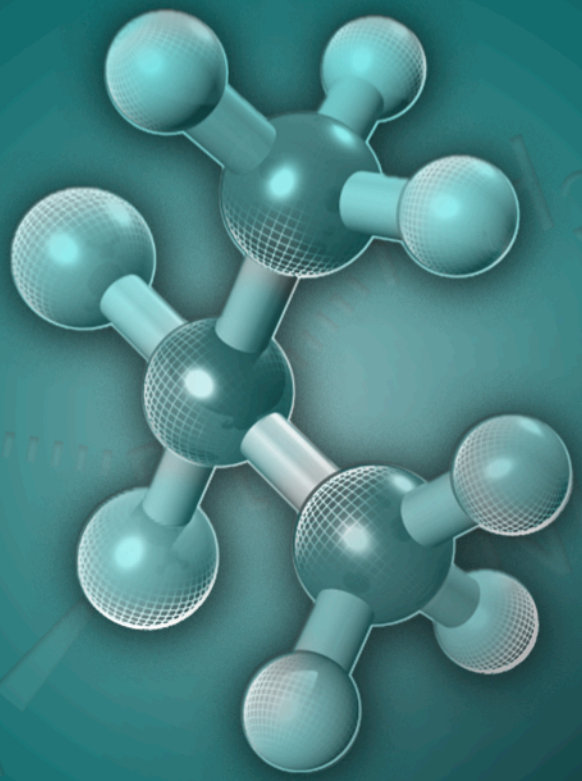
Mining Compounds, Targets and Indications from the Patent Corpus with **SureChEMBL** and **Open PHACTS**



George Papadatos, PhD

ChEMBL Group, EMBL-EBI

georgep@ebi.ac.uk

#ShefChem16

EMBL-EBI

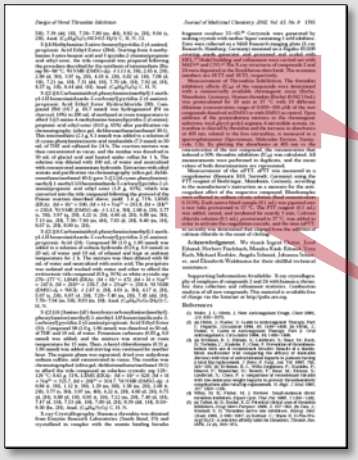
Outline

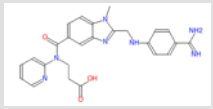
- Overview of chemical annotations in SureChEMBL
- Biological annotations for Open PHACTS
  - Relevance scoring
- Open PHACTS patent API
  - Use cases and examples

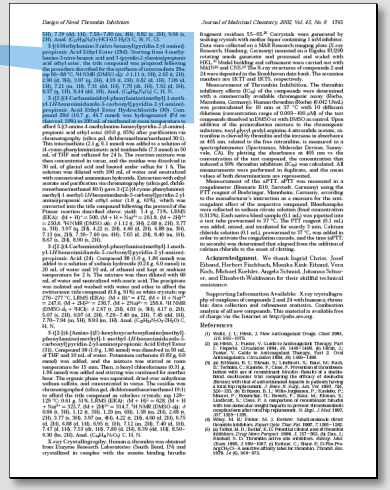
ChEMBL: Data for drug discovery

1. Scientific facts

3. Insight, tools and resources for translational drug discovery







>Thrombin
MAHVRGLQLPGCLALAALCSLVHSQHVFLAPQQARSLLQRVRRANTFLEEVRKGNLE
RECVEETCSYEEAFEALESSTATDVFWAKYTACETARTPRDKLAACLEGNCAEGLGT
NYRGHVNITRSGIECQLWRSRYPHKPEINSTTHPGADLQENFCRNPDSSTTGPWCYT
TDPTVRRQECSIPVCGQDQVTVAMTPRSEGSSVNLSPPLEQCVPDRGQQYQGRLAVT
THGLPCLAWASAQAKALSKHQDFNSAVQLVENFCRNPDGDEEGVWCYVAGKPGDFGY
CDLNYCEEAVEEETGDGLDEDSDRAIEGRTATSEYQTFFNPRTFGSGEADCGLRPLF
EKKSLEDKTERELLESYIDGRIVEGSDAEIGMSPWQVMLFRKSPQELLCGASLISDR
WVLTAAHCLLYPPWDKNFTENDLLVRIGKHSRTRYERNIEKISMLEKIYIHPRYNWR
ENLDRDIALMKLKKPVAFSDYIHPVCLPDRETAASLLQAGYKGRVTGWGNLKETWT
NVGKGQPSVLQVVNLPIVERPVCKDSTRIRITDNMFCAGYKPDEGKRGDACEGDSG
PFVMKSPFNNRWYQMGIVSWGEGCDRDGKYGFYTHVFRLLKWIQKVIDQFGE

Compound

K_i = 4.5nM

Assay/Target

Bioactivity data

APTT = 11 min.

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

What do people do with ChEMBL?

- Chemical space clustering & visualisation
- (Q)SAR analysis and Virtual Screening
  - Data modeling, activity cliffs, Free-Wilson, MMP analysis
  - Bioisosteric replacements mining
- *De novo* drug design
  - *In silico* evolutionary compound optimisation
- Target prediction (target fishing)
  - Off-target prediction and ADR analysis
- Polypharmacology networks
- Druggability / Drug-likeness
- Text mining
- Third-party application integration

Why looking at patent documents?

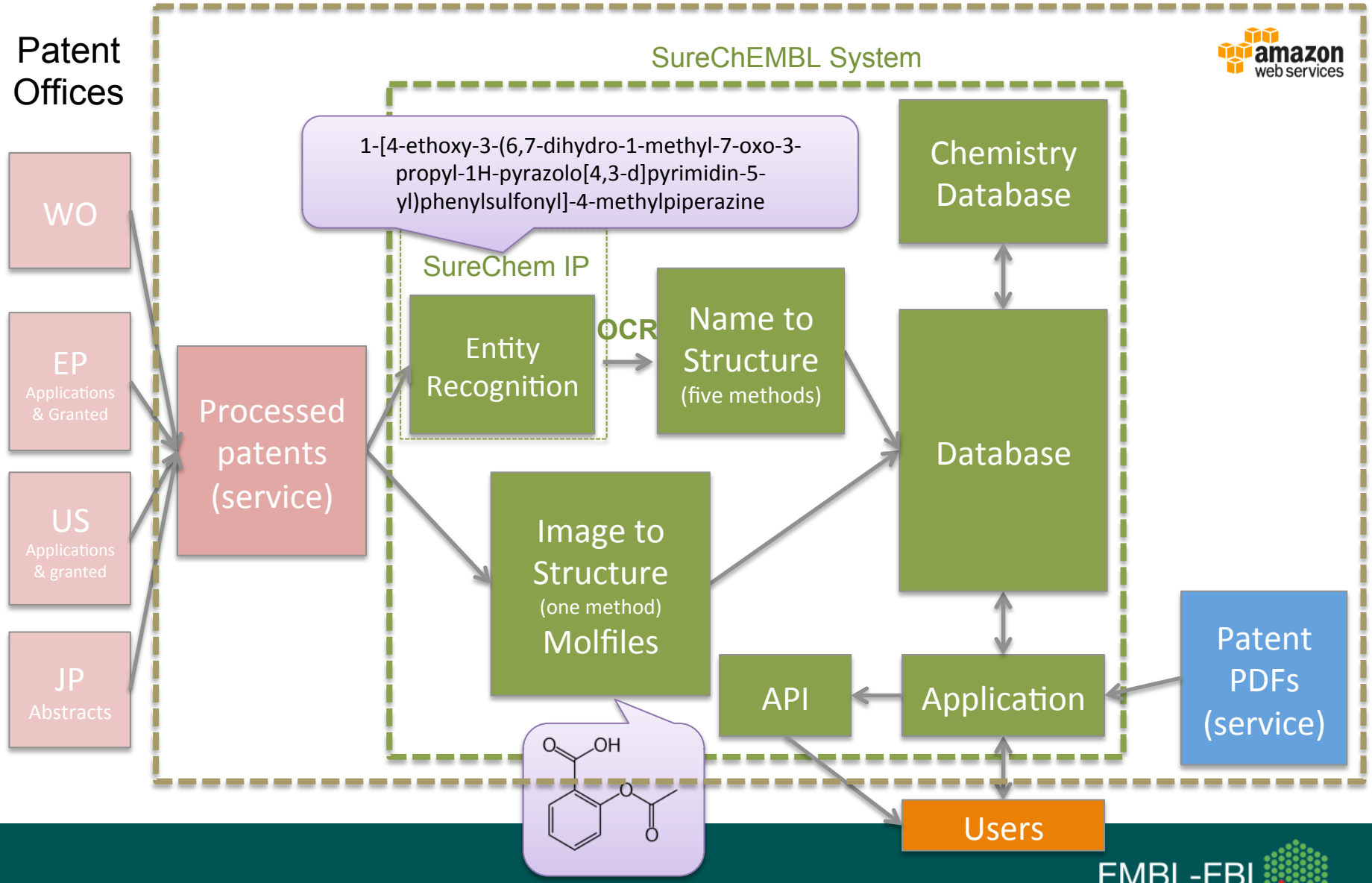
- Patent filing and searching
  - Legal, financial and commercial incentives & interests
  - Prior art, novelty, freedom to operate searches
  - Competitive intelligence
- Unprecedented wealth of knowledge
  - Most of the knowledge will never be disclosed anywhere else
    - Compounds, scaffolds, reactions,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

Rodriguez-Esteban & Bundschus. Text Mining Patents for Biomedical Knowledge. *Drug Discovery Today* **2016**.

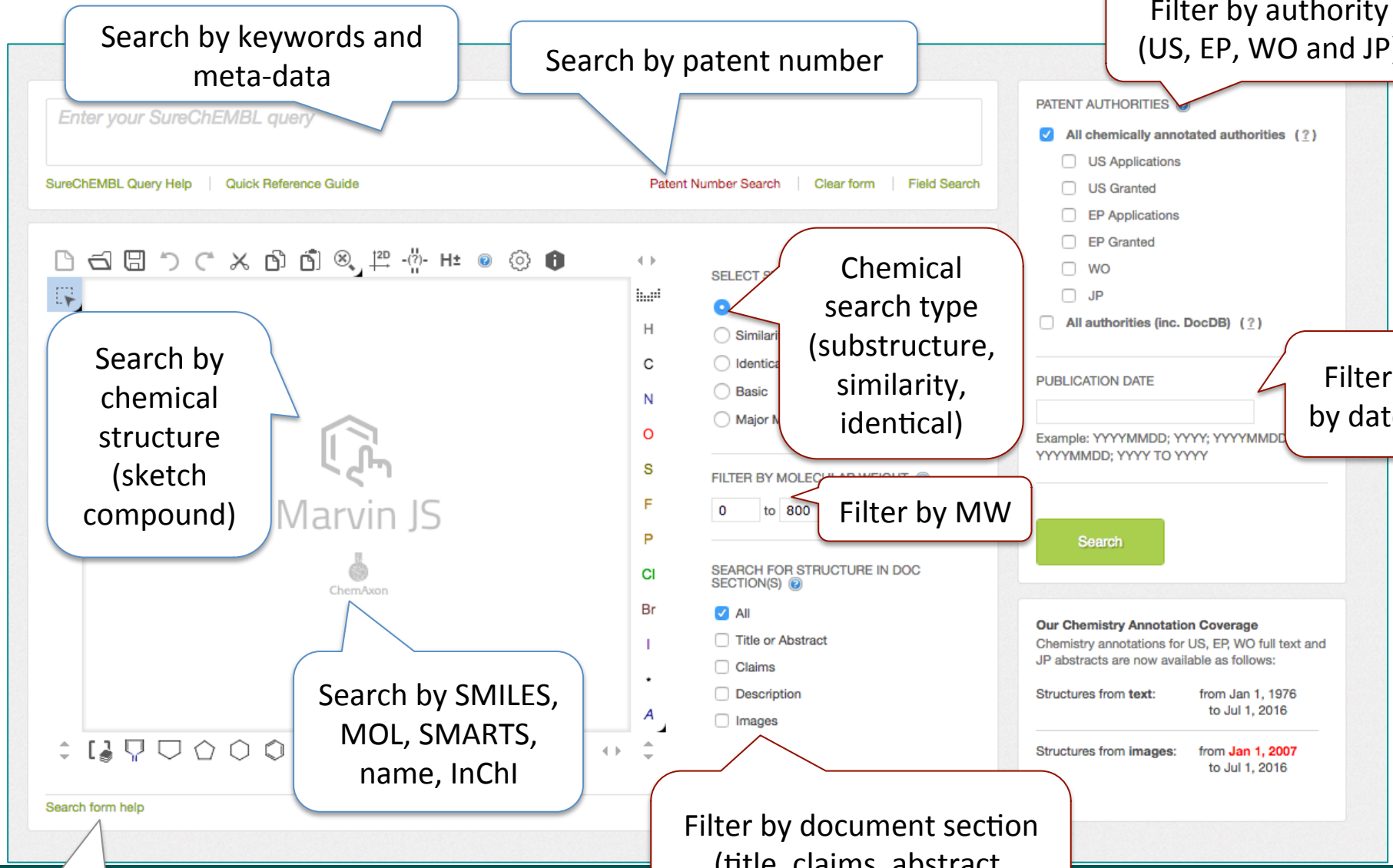
From SureChem to SureChEMBL

- 2007: SureChem is founded
- 2010: Acquired by Macmillan
- 2012: Revamped commercial product: SureChemDirect
  - Modular, cloud-based architecture
- **Dec 2013: Donated to EMBL-EBI**
- Sept 2014: SureChEMBL UI is launched
- Jan 2015: Data client and map file dumps
- April 2015: Bio-annotations are added
- May 2016: SureChEMBL data available via Open PHACTS API

SureChEMBL data processing

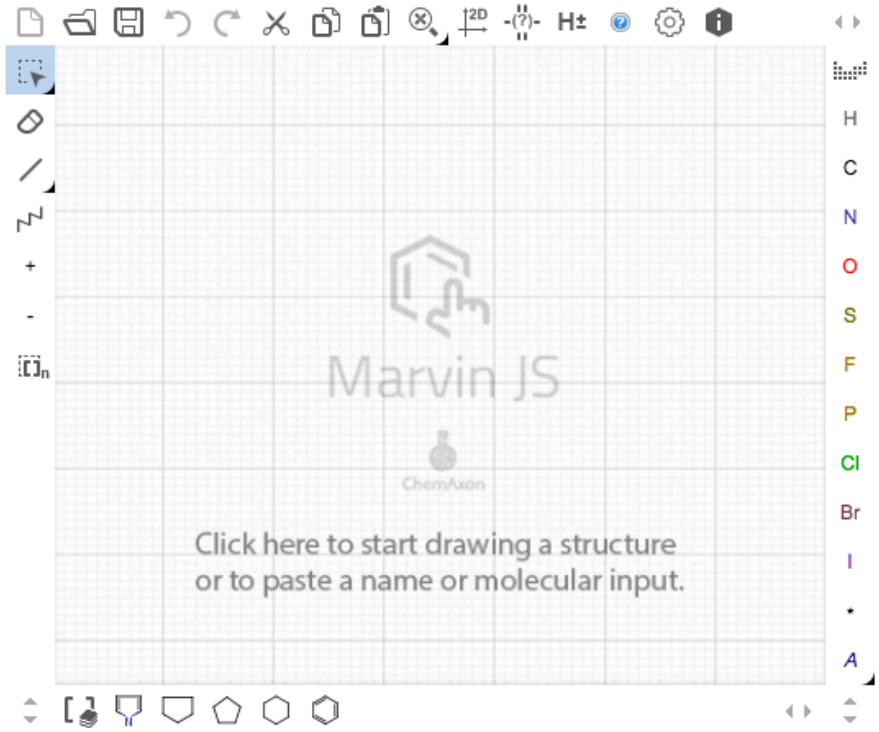


Homepage


Search by keywords and meta-data
Search by patent number
Filter by authority (US, EP, WO and JP)
Enter your SureChEMBL query
SureChEMBL Query Help
Quick Reference Guide
Patent Number Search
Clear form
Field Search
PATENT AUTHORITIES
All chemically annotated authorities (?)
US Applications
US Granted
EP Applications
EP Granted
WO
JP
All authorities (inc. DocDB) (?)
PUBLICATION DATE
Example: YYYYMMDD; YYYY; YYYYMMDD-YYYYMMDD; YYYY TO YYYY
Search by chemical structure (sketch compound)
Marvin JS
ChemAxon
Chemical search type (substructure, similarity, identical)
Similari
Identica
Basic
Major
Filter by date
FILTER BY MOLECULAR WEIGHT
0
to
800
Filter by MW
Search
SEARCH FOR STRUCTURE IN DOC SECTION(S)
All
Title or Abstract
Claims
Description
Images
Our Chemistry Annotation Coverage
Chemistry annotations for US, EP, WO full text and JP abstracts are now available as follows:
Structures from text: from Jan 1, 1976 to Jul 1, 2016
Structures from images: from Jan 1, 2007 to Jul 1, 2016
Search by SMILES, MOL, SMARTS, name, InChI
Search form help
Help
Filter by document section (title, claims, abstract, description and images)

pdyear:2015 AND ic:C07D AND (ttl:diabetes OR ab:diabetes) AND pa:Bristol*

SureChEMBL Query Help | Quick Reference Guide | Patent Number Search | Clear form | Field Search



SELECT STRUCTURE SEARCH

- (•) Substructure
- ( ) Similarity
- ( ) Identical
- ( ) Basic
- ( ) Major Match

FILTER BY MOLECULAR WEIGHT

0 to 800

SEARCH FOR STRUCTURE IN DOC SECTION(S)

- [x] All
- [ ] Title or Abstract
- [ ] Claims
- [ ] Description
- [ ] Images

Search form help

PATENT AUTHORITIES

- [ ] **All chemically annotated authorities** (?)
  - [x] US Applications
  - [x] US Granted
  - [ ] EP Applications
  - [ ] EP Granted
  - [ ] WO
  - [ ] JP
- [ ] **All authorities (inc. DocDB)** (?)

PUBLICATION DATE

Example: YYYYMMDD; YYYY; YYYYMMDD TO YYYYMMDD; YYYY TO YYYY

Search

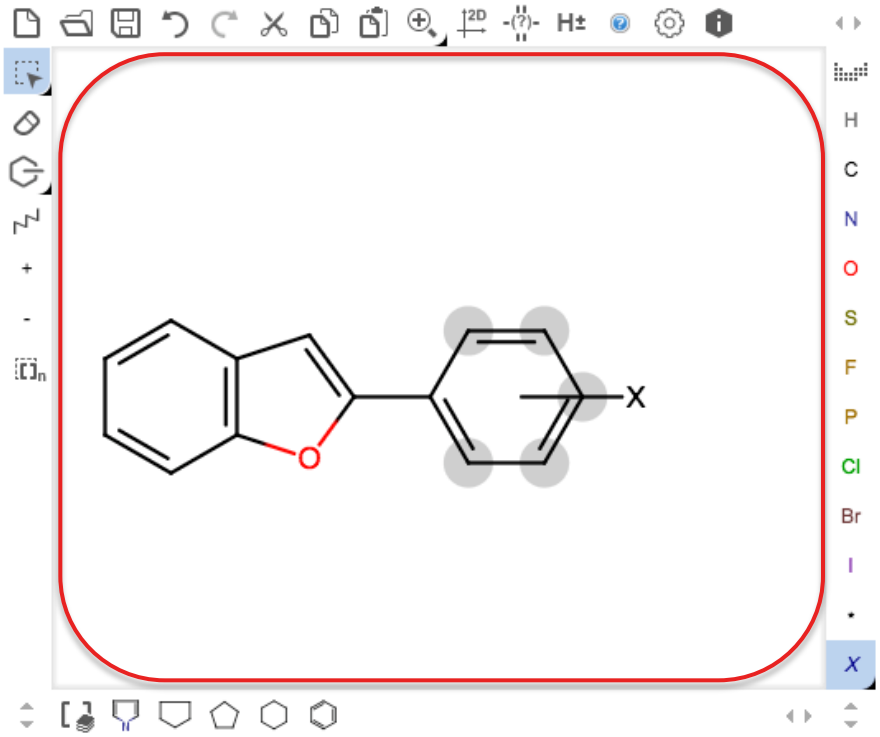
Our Chemistry Annotation Coverage

Chemistry annotations for US, EP, WO full text and JP abstracts are now available as follows:

Structures from **text**:	from Jan 1, 1976 to Jul 1, 2016
Structures from **images**:	from Jan 1, 2007 to Jul 1, 2016

pdyear:2015 AND ic:C07D AND (ttl:diabetes OR ab:diabetes) AND pa:Bristol*

SureChEMBL Query Help | Quick Reference Guide | Patent Number Search | Clear form | Field Search



SELECT STRUCTURE SEARCH
- (•) Substructure
- ( ) Similarity
- ( ) Identical
- ( ) Basic
- ( ) Major Match

FILTER BY MOLECULAR WEIGHT
200 to 500

SEARCH FOR STRUCTURE IN DOC SECTION(S)
- [ ] All
- [ ] Title or Abstract
- [x] Claims
- [x] Description
- [x] Images

Search form help

PATENT AUTHORITIES
- [ ] All chemically annotated authorities (?)
  - [x] US Applications
  - [x] US Granted
  - [ ] EP Applications
  - [ ] EP Granted
  - [ ] WO
  - [ ] JP
- [ ] All authorities (inc. DocDB) (?)

PUBLICATION DATE

Example: YYYYMMDD; YYYY; YYYYMMDD TO YYYYMMDD; YYYY TO YYYY

Search

Our Chemistry Annotation Coverage
Chemistry annotations for US, EP, WO full text and JP abstracts are now available as follows:

Structures from **text**: from Jan 1, 1976 to Jul 1, 2016

Structures from **images**: from Jan 1, 2007 to Jul 1, 2016

Back to search results

Front-page | Claims | Description

Your query

Highlight additional recognised chemical terms
Show select biological annotations
Powered by SciBite.com's Termite Engine and funded by Open PHACTS

Export document chemistry
Export family chemistry

Download patent PDF

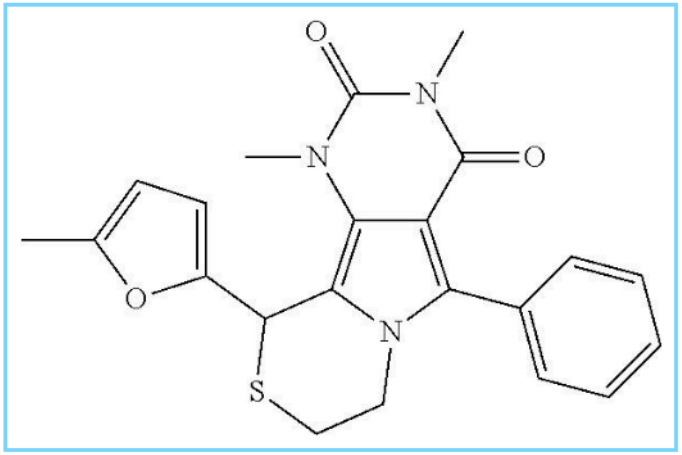
Flow rate: 1.0 mL/min

Gradient: 0.0 min 5% B, 0.2-1.8 min 5-98% B, 1.8-2.1 min 98% B

Preparation of Final Compounds

Unless otherwise indicated, the chemical name refers to the racemic compound or to the diastereomeric mixture. The term 'Enantiomer 1' refers to the more active of the separated enantiomers and does not indicate the relative position of the eluted peak within the chromatogram (where separation is achieved by SFC under the specified conditions). The term 'Diastereomer 1' refers to the more active of the separated diastereomers and does not indicate the relative position of the eluted peak within the chromatogram (where separation is achieved by SFC under the specified conditions).

Example 1.1 1,3-Dimethyl-10-(5-methylfuran-2-yl)-5-phenyl-7,8-dihydro-1H-pyrimido[4',5':3,4]pyrrolo[2,1-c][1,4]thiazine-2,4(3H,10H)-dione



To 6-(2-mercapto-ethyl)-1,3-dimethyl-5-phenyl-1,6-dihydro-pyrrolo[3,4-d]pyrimidine-2,4-dione (Intermediate A) (207 mg, 0.656 mmol) and bismuth triflate (43.1 mg, 0.066 mmol) in toluene (5 ml) was added 5-methylfuran-2-carbaldehyde (commercial) (0.065 ml, 0.656 mmol) the mixture heated to 100° C. for 1 hour. The reaction mixture was cooled to RT and the solvent was removed under reduced pressure. The residue was partitioned between EtOAc (20 mL) and water (20 mL) and the phases separated. The aqueous phase was extracted with EtOAc (2×20 ml) the combined organic phases were dried over magnesium sulfate and the solvent removed under vacuum. Purification by chromatography on silica, eluting with 0-50% EtOAc/hexane gave the title compound.

1H NMR (400 MHz, DMSO-d6) δ 7.45 (5H, s), 6.03 (1H, s), 5.99 (1H, dd), 5.88 (1H, d), 4.18-4.11 (1H, m), 4.02-3.97 (1H, m), 3.42 (3H, s), 3.11 (3H, s), 3.03-2.89 (2H, m), 2.25 (3H, s).

LC-MS: Rt=1.11 mins; MS m/z 408.4 [M+H]+ Method 2minLowpH

Chiral separation of the racemate by Supercritical Fluid Chromatography was carried out using the following conditions to afford the compounds listed hereinafter:

Back to search results

Front-page | Claims | Description

Your query

Highlight additional recognised chemical terms
Show select biological annotations
Powered by SciBite.com's Termite Engine and funded by Open PHACTS

Export document chemistry
Export family chemistry

Download patent PDF

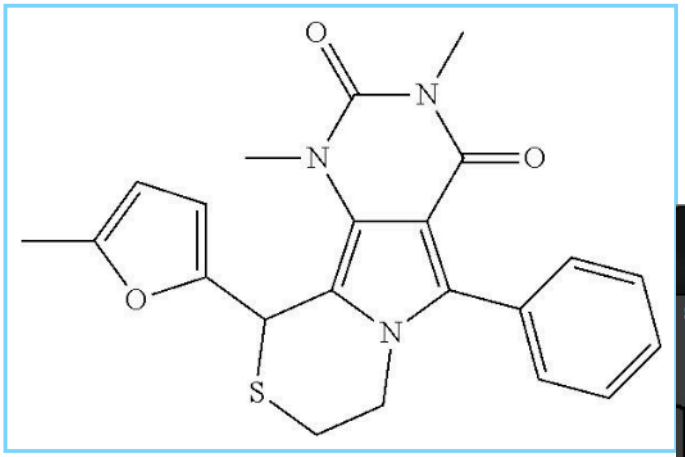
Flow rate: 1.0 mL/min

Gradient: 0.0 min 5% B, 0.2-1.8 min 5-98% B, 1.8-2.1 min 98% B

Preparation of Final Compounds

Unless otherwise indicated, the chemical name refers to the racemic compound or to the diastereomeric mixture. The term 'Enantiomer 1' refers to the more active of the separated enantiomers and does not indicate the relative position of the eluted peak within the chromatogram (where separation is achieved by SFC under the specified conditions). The term 'Diastereomer 1' refers to the more active of the separated diastereomers and does not indicate the relative position of the eluted peak within the chromatogram (where separation is achieved by SFC under the specified conditions).

Example 1.1 1,3-Dimethyl-10-(5-methylfuran-2-yl)-5-phenyl-7,8-dihydro-1H-pyrimido[4',5':3,4]pyrrolo[2,1-c][1,4]thiazine-2,4(3H,10H)-dione



To 6-(2-mercapto-ethyl)-1,3-dimethyl-5-phenyl-1,6-dihydro-pyrrolo[3,4-d]pyrimidine-2,4-dione (Intermediate A) (207 mg, 0.65
(43.1 mg, 0.066 mmol) in toluene (5 ml) was added 5-methylfuran-2-carbaldehyde (commercial) (0.065 ml, 0.656 mmol) the mi
hour. The reaction mixture was cooled to RT and the solvent was removed under reduced pressure. The residue was partitioned be
water (20 mL) and the phases separated. The aqueous phase was extracted with EtOAc (2×20 ml) the combined organic phases w
sulfate and the solvent removed under vacuum. Purification by chromatography on silica, eluting with 0-50% EtOAc/hexane gav

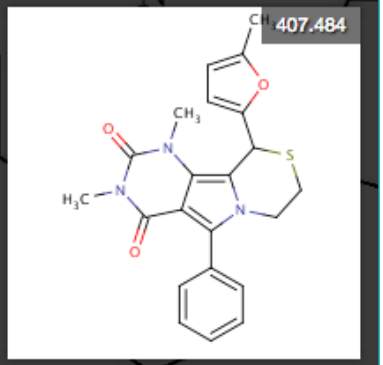
1H NMR (400 MHz, DMSO-d6) δ 7.45 (5H, s), 6.03 (1H, s), 5.99 (1H, dd), 5.88 (1H, d), 4.18-4.11 (1H, m), 4.02-3.97 (1H, m), 3.42
(2H, m), 2.25 (3H, s).

LC-MS: Rt=1.11 mins; MS m/z 408.4 [M+H]+ Method 2minLowpH

Chiral separation of the racemate by Supercritical Fluid Chromatography was carried out using the following conditions to afford t
hereinafter:

Chemical information

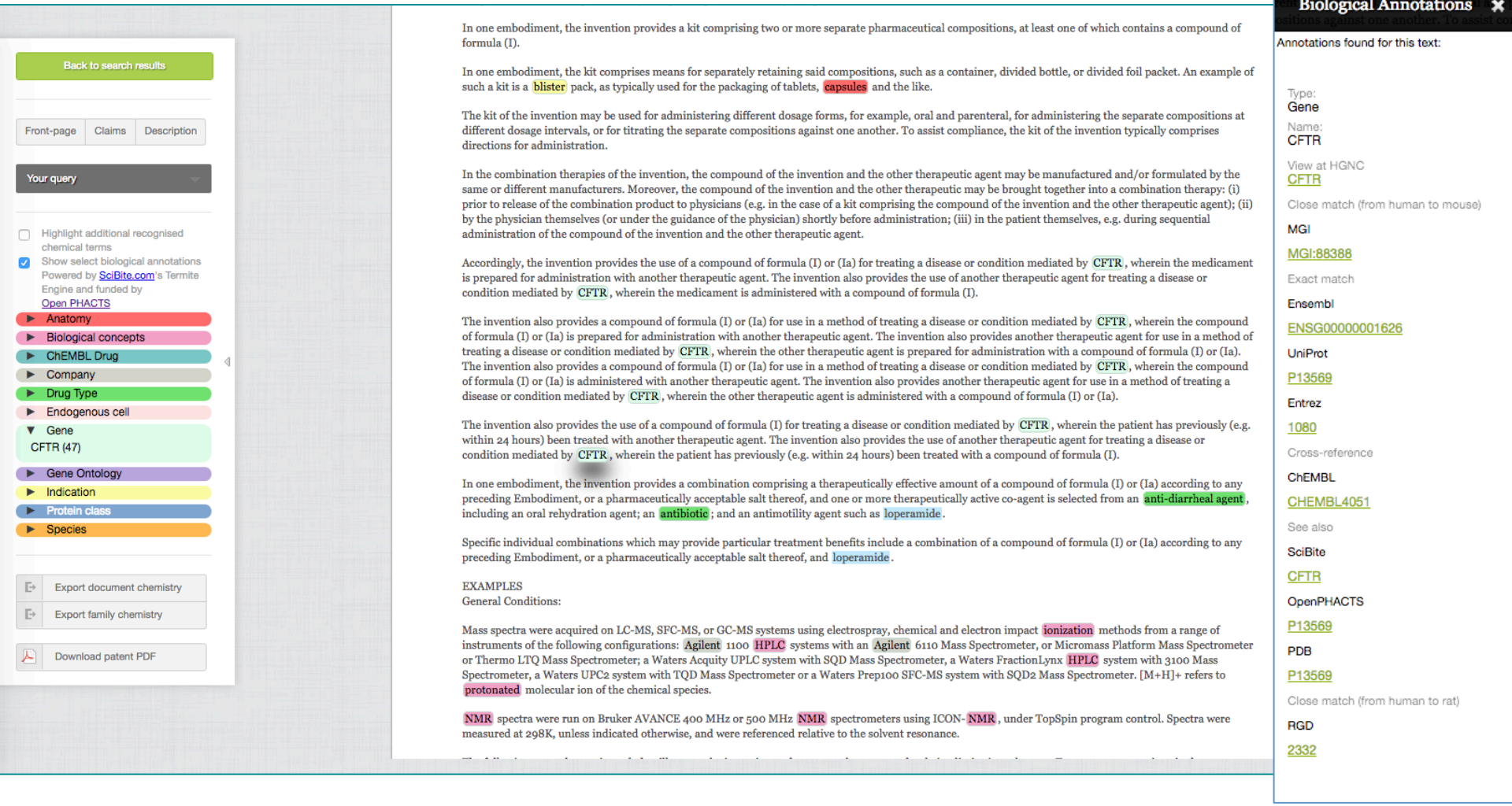
Structures generated for this name:



Name:
3,5-dimethyl-13-(5-methylfuran-2-yl)-8-phenyl-12-thia-3,5,9-triazatricyclo[7.4.0.0^{2,7}]trideca-1,7-diene-4,6-dione

View

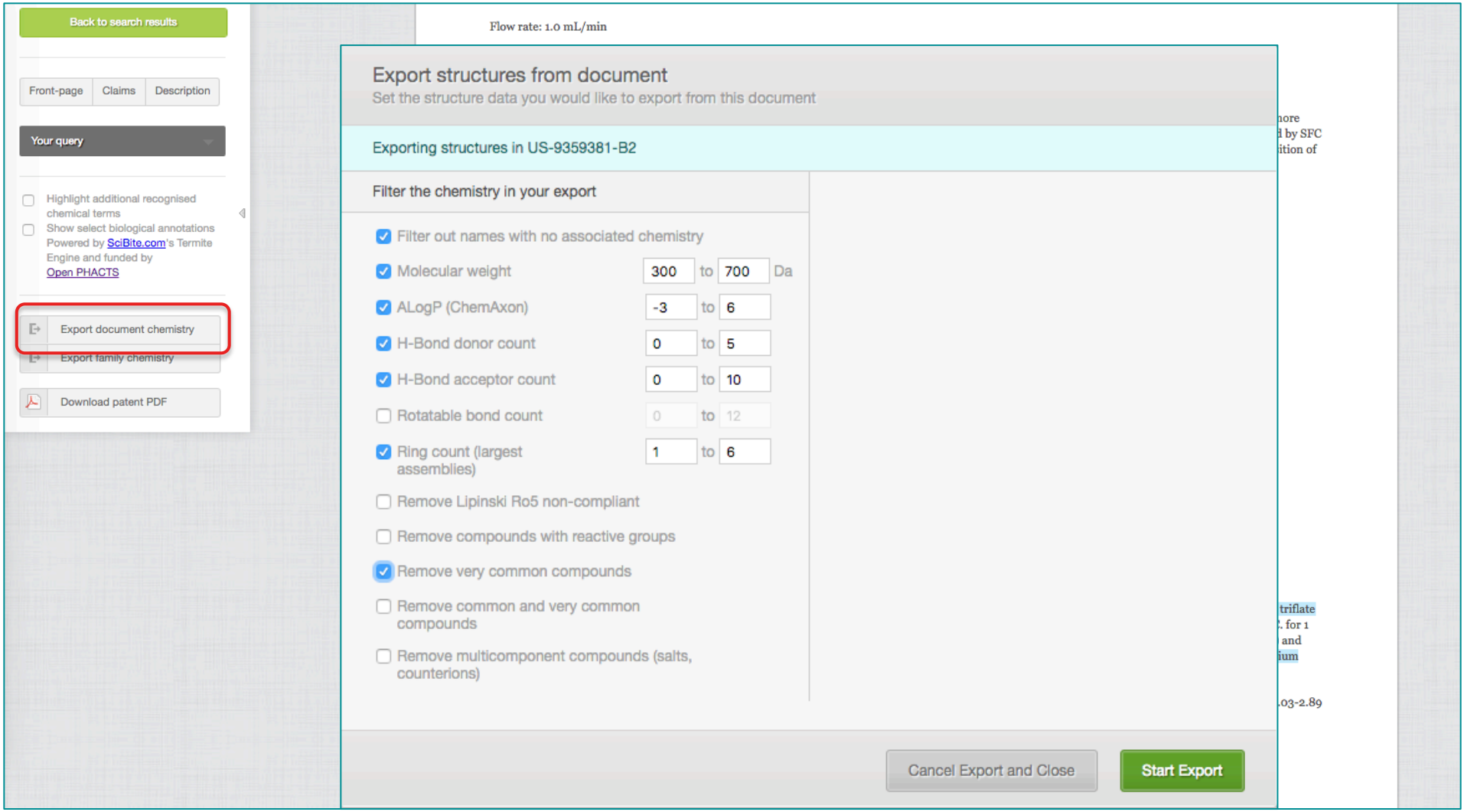
https://www.surechembl.org/document/US-9359381-B2



On-the-fly bio-annotation with mapping to DB IDs and ontologies

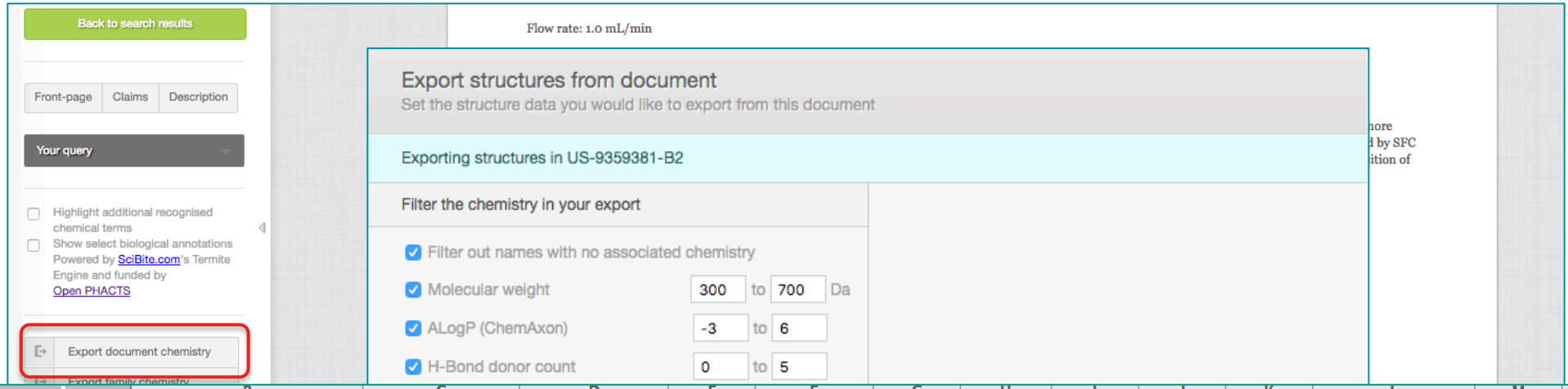
https://www.surechembl.org/document/US-9359381-B2

EMBL-EBI



https://www.surechembl.org/document/US-9359381-B2

EMBL-EBI



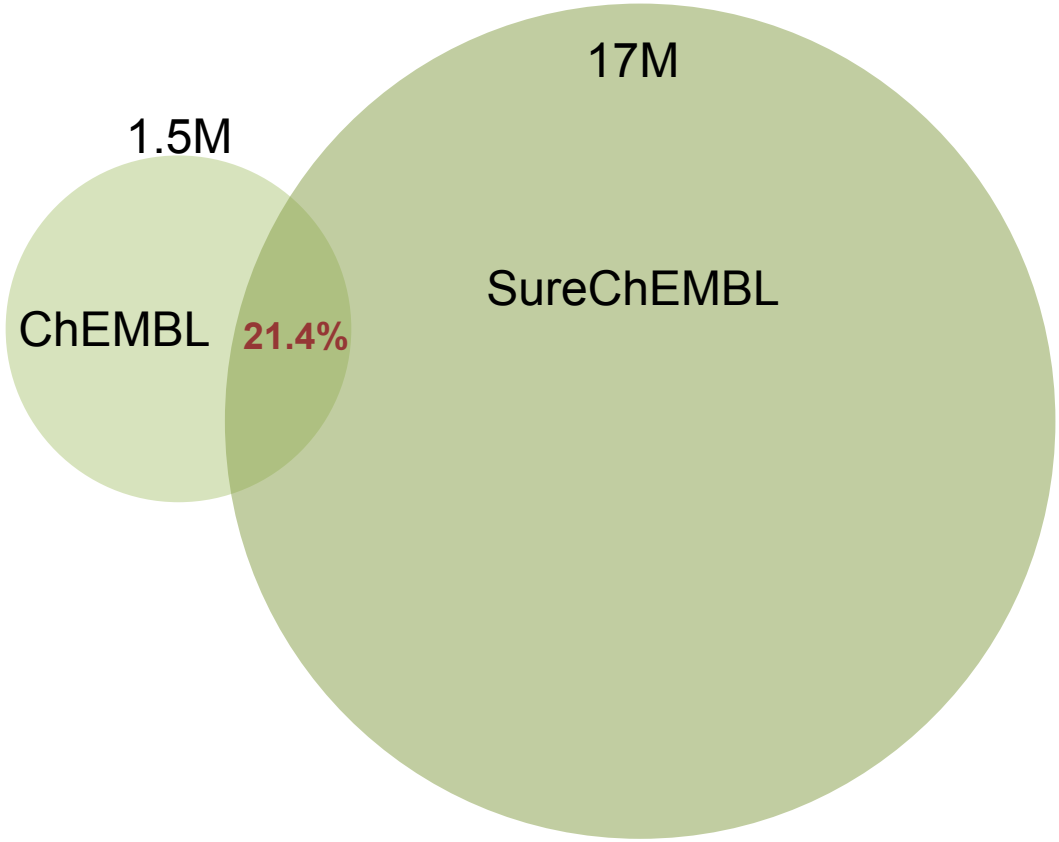
A	B	C	D	E	F	G	H	I	J	K	L	M
patent_id	annotation_reference	schembl_id	smiles	type	chemical_docum	annotation_	title_count	abstract_cou	claims_coun	description_	chemical_corpus_count	annotation_
US-9040690-B2	5-(3-chloro-phenyl)-3-(2-cyclohexy	SCHEMBL10188633	OC(=O)C1=C(C=C(S1)C	TEXT	1	1	0	0	0	1	1	1
US-9040690-B2	US09040690-20150526-C00237.M(	SCHEMBL10188633	OC(=O)C1=C(C=C(S1)C	MOLATTACH	1						1	
US-9040690-B2	US09040690-20150526-C00237.TIF	SCHEMBL10188633	OC(=O)C1=C(C=C(S1)C	IMAGE	1						1	
US-9040690-B2	3-(2-cyclohexyl-6-oxo-piperidin-1-y	SCHEMBL10188655	CC1=CC(=CC=C1)C1=C	TEXT	1	1	0	0	0	1	1	1
US-9040690-B2	US09040690-20150526-C00238.M(	SCHEMBL10188655	CC1=CC(=CC=C1)C1=C	MOLATTACH	1						1	
US-9040690-B2	US09040690-20150526-C00238.TIF	SCHEMBL10188655	CC1=CC(=CC=C1)C1=C	IMAGE	1						1	
US-9040690-B2	3-[4-(4-acetylamino-benzenesulfon	SCHEMBL10188842	CC(=O)NC1=CC=C(C=C	TEXT	1	1	0	0	0	1	1	1
US-9040690-B2	US09040690-20150526-C00154.TIF	SCHEMBL10188842	CC(=O)NC1=CC=C(C=C	IMAGE	1						1	
US-9040690-B2	US09040690-20150526-C00154.M(	SCHEMBL10188842	CC(=O)NC1=CC=C(C=C	MOLATTACH	1						1	
US-9040690-B2	3-(4-benzenesulfonyl-2-cyclohexyl-	SCHEMBL10188844	OC(=O)C1=C(C=C(S1)C	TEXT	1	1	0	0	0	1	1	1
US-9040690-B2	US09040690-20150526-C00156.M(	SCHEMBL10188844	OC(=O)C1=C(C=C(S1)C	MOLATTACH	1						1	
US-9040690-B2	US09040690-20150526-C00156.TIF	SCHEMBL10188844	OC(=O)C1=C(C=C(S1)C	IMAGE	1						1	
US-9040690-B2	3-[2-cyclohexyl-6-oxo-4-(2-trifluorc	SCHEMBL10189021	OC(=O)C1=C(C=C(S1)C	TEXT	1	1	0	0	0	1	1	1
US-9040690-B2	US09040690-20150526-C00172.TIF	SCHEMBL10189021	OC(=O)C1=C(C=C(S1)C	IMAGE	1						1	
US-9040690-B2	US09040690-20150526-C00172.M(	SCHEMBL10189021	OC(=O)C1=C(C=C(S1)C	MOLATTACH	1						1	
US-9040690-B2	3-[(r)-2-cyclohexyl-4-(1-ethyl-1h-py	SCHEMBL10189035	CCN1C=C(C=N1)S(=O)(	TEXT	1	1	0	0	0	1	1	1
US-9040690-B2	US09040690-20150526-C00400.M(	SCHEMBL10189035	CCN1C=C(C=N1)S(=O)(	MOLATTACH	1						1	
US-9040690-B2	US09040690-20150526-C00400.TIF	SCHEMBL10189035	CCN1C=C(C=N1)S(=O)(	IMAGE	1						1	
US-9040690-B2	3-[(r)-2-cyclohexyl-4-(1-methyl-1h-	SCHEMBL10189038	CN1C=C(C=N1)S(=O)(=	TEXT	1	1	0	0	0	1	1	1
US-9040690-B2	US09040690-20150526-C00399.TIF	SCHEMBL10189038	CN1C=C(C=N1)S(=O)(=	IMAGE	1						1	
US-9040690-B2	US09040690-20150526-C00399.M(	SCHEMBL10189038	CN1C=C(C=N1)S(=O)(=	MOLATTACH	1						1	
US-9040690-B2	3-(2-cyclohexyl-4,6-dioxo-piperidin	SCHEMBL10189076	OC(=O)C1=C(C=C(S1)C	TEXT	1	1	0	0	0	1	1	1
US-9040690-B2	US09040690-20150526-C00113.TIF	SCHEMBL10189076	OC(=O)C1=C(C=C(S1)C	IMAGE	1						1	
US-9040690-B2	US09040690-20150526-C00113.M(	SCHEMBL10189076	OC(=O)C1=C(C=C(S1)C	MOLATTACH	1						1	
US-9040690-B2	3-[2-cyclohexyl-6-oxo-4-(toluene-3	SCHEMBL10189092	CC1=CC=CC(=C1)S(=O)	TEXT	1	1	0	0	0	1	1	1

SureChEMBL Data Content and Growth

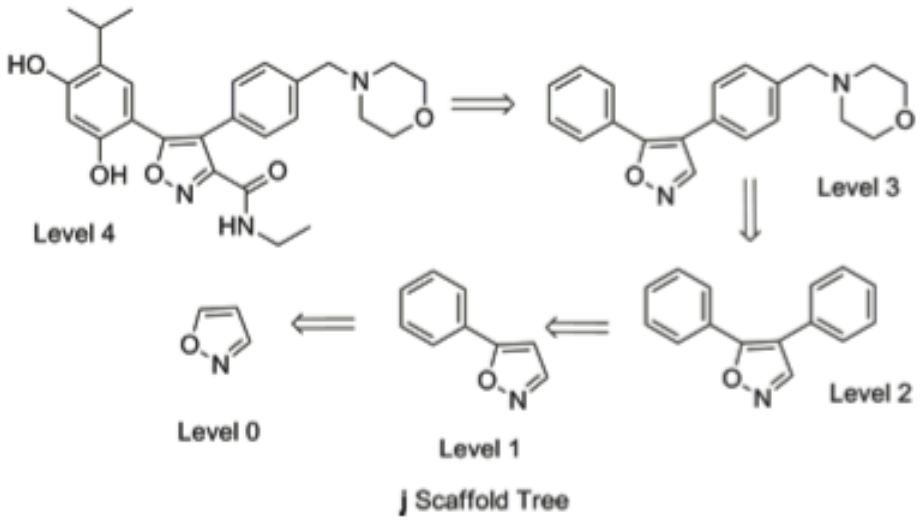
- > 18M unique structures
  - ~13M are in a drug-like phys-chem space
- > 15M chemically annotated patents
  - ~5M are life-science relevant (based on patent codes)
- ~80,000 *novel* compounds extracted from ~50,000 new patents *monthly*
- 1–7 days for a published patent to be chemically annotated, indexed and searchable in SureChEMBL
- SureChEMBL provides search access to all patents (not just chemically annotated ones) ~120M patents

ChEMBL-SureChEMBL overlap

- *Connectivity* match on single components - UniChem



Too granular? Try scaffolds instead



J. Chem. Inf. Model. **2007,** *47,* 47−58

The Scaffold Tree − Visualization of the Scaffold Universe by Hierarchical Scaffold Classification

Ansgar Schuffenhauer,*,† Peter Ertl,† Silvio Roggo,† Stefan Wetzel,‡ Marcus A. Koch,‡ and Herbert Waldmann‡

Novartis Institutes for BioMedical Research, CH-4002 Basel, Switzerland, and Max Planck Institute of Molecular Physiology and Fachbereich 3 − Chemical Biology, University of Dortmund, D-44227 Dortmund, Germany

Received August 2, 2006

A hierarchical classification of chemical scaffolds (molecular framework, which is obtained by pruning all terminal side chains) has been introduced. The molecular frameworks form the leaf nodes in the hierarchy trees. By an iterative removal of rings, scaffolds forming the higher levels in the hierarchy tree are obtained. Prioritization rules ensure that less characteristic, peripheral rings are removed first. All scaffolds in the hierarchy tree are well-defined chemical entities making the classification chemically intuitive. The classification is deterministic, data-set-independent, and scales linearly with the number of compounds included in the data set. The application of the classification is demonstrated on two data sets extracted from the PubChem database, namely, pyruvate kinase binders and a collection of pesticides. The examples shown demonstrate that the classification procedure handles robustly synthetic structures and natural products.

JOURNAL OF CHEMICAL INFORMATION AND MODELING

ARTICLE

pubs.acs.org/jcim

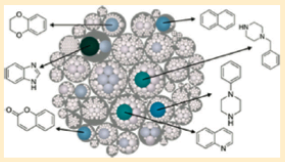
Scaffold Diversity of Exemplified Medicinal Chemistry Space

Sarah R. Langdon,† Nathan Brown,*,† and Julian Blagg*,†

†Cancer Research UK Cancer Therapeutics Unit, The Institute of Cancer Research, 15 Cotswold Road, Sutton, Surrey SM2 5NG, U.K.

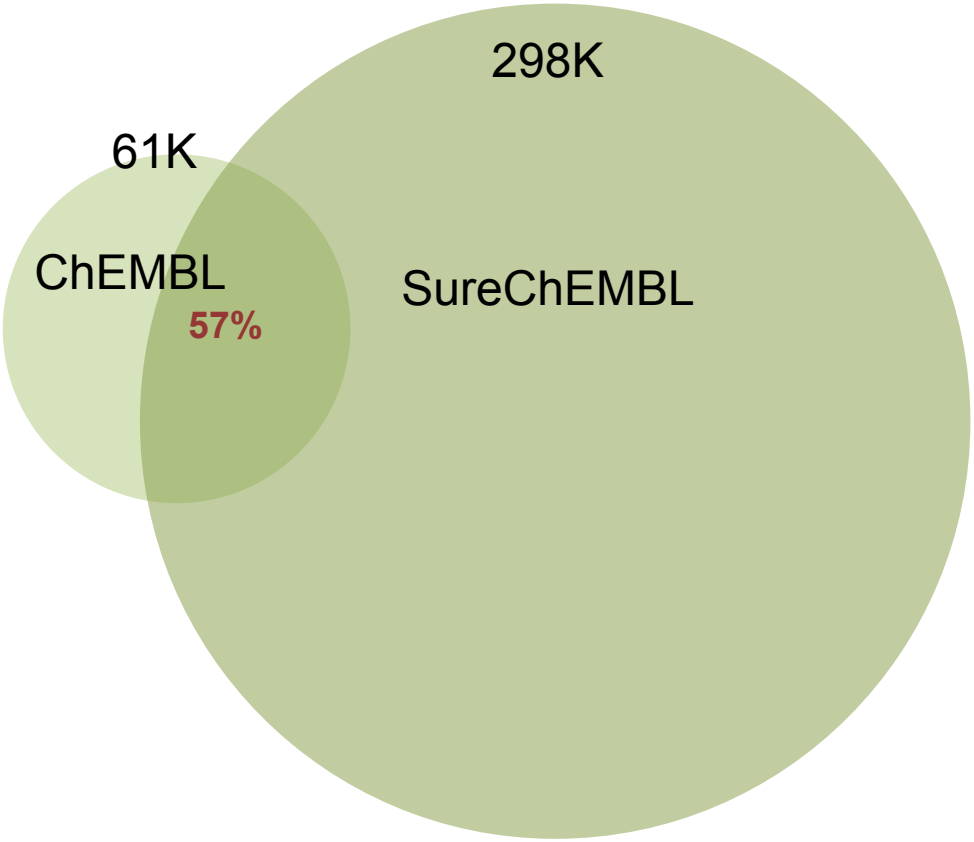
Supporting Information

ABSTRACT: The scaffold diversity of 7 representative commercial and proprietary compound libraries is explored for the first time using both Murcko frameworks and Scaffold Trees. We show that Level 1 of the Scaffold Tree is useful for the characterization of scaffold diversity in compound libraries and offers advantages over the use of Murcko frameworks. This analysis also demonstrates that the majority of compounds in the libraries we analyzed contain only a small number of well represented scaffolds and that a high percentage of singleton scaffolds represent the remaining compounds. We use Tree Maps to clearly visualize the scaffold space of representative compound libraries, for example, to display highly populated scaffolds and clusters of structurally similar scaffolds. This study further highlights the need for diversification of compound libraries used in hit discovery by focusing library enrichment on the synthesis of compounds with novel or underrepresented scaffolds.

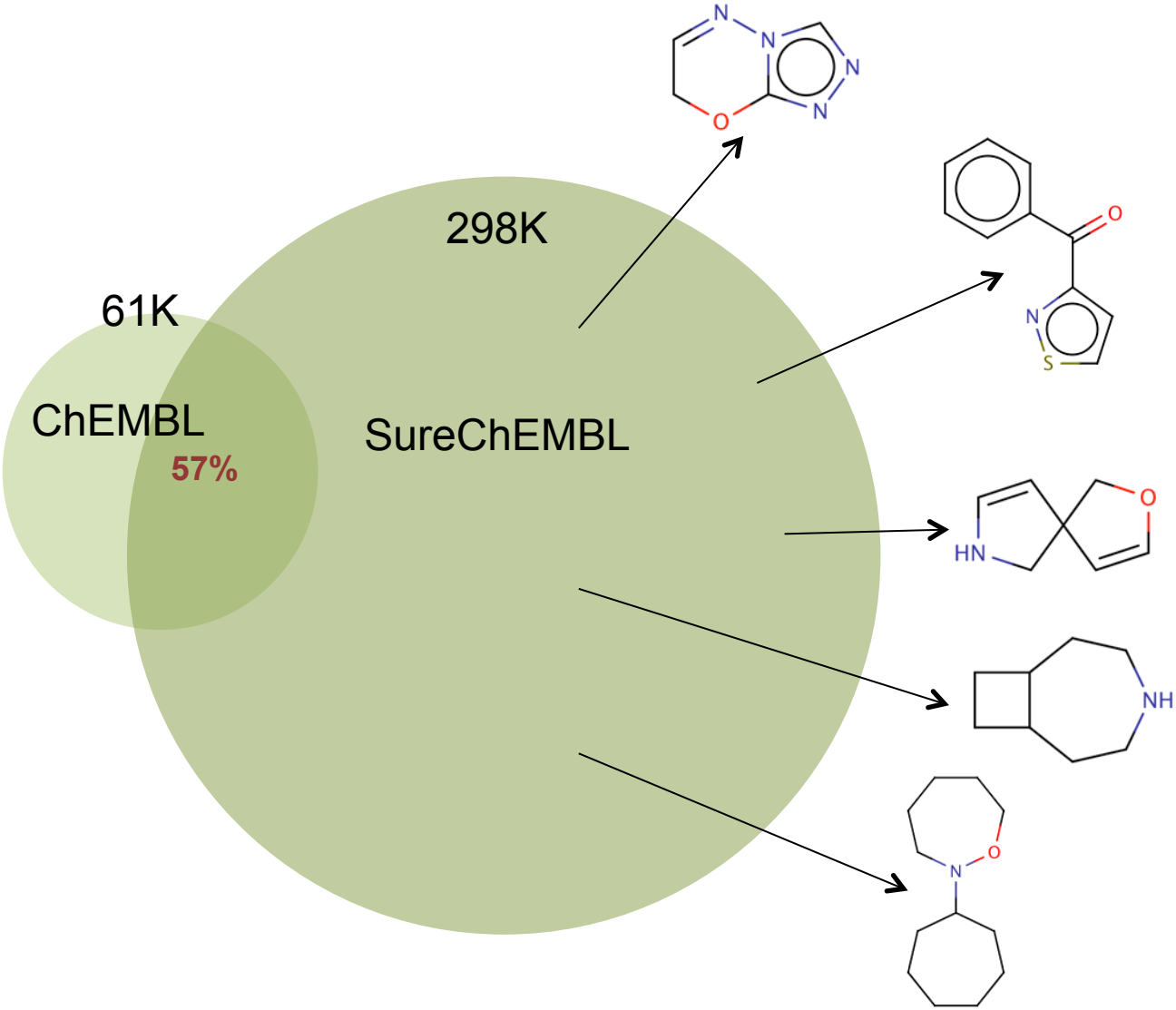


EMBL-EBI

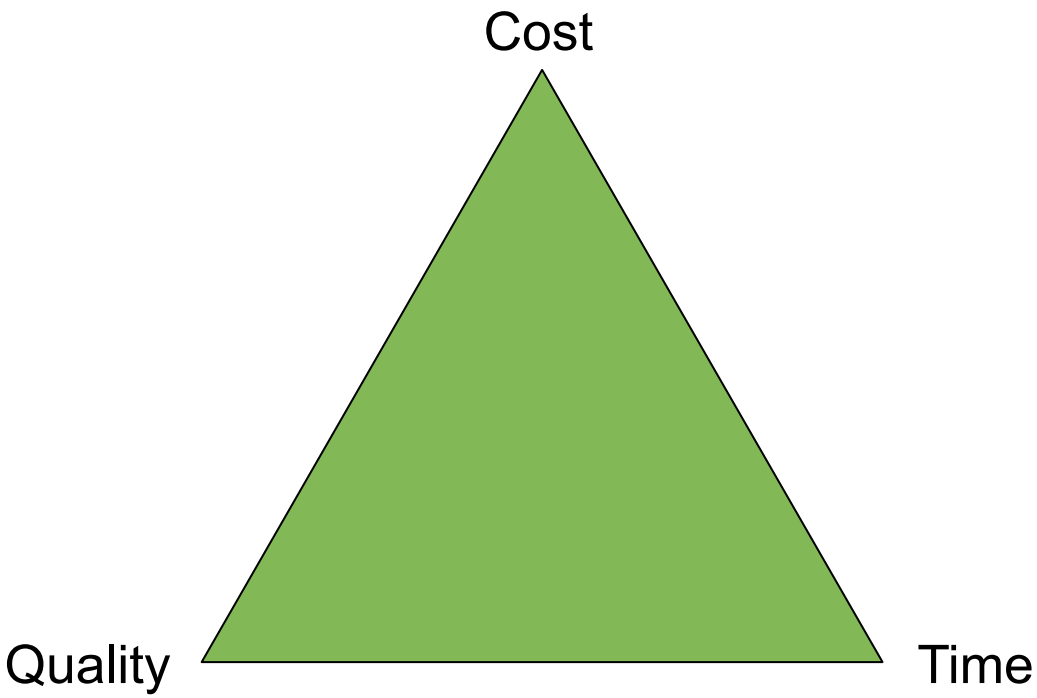
Level 1 scaffold overlap



Level 1 scaffold overlap

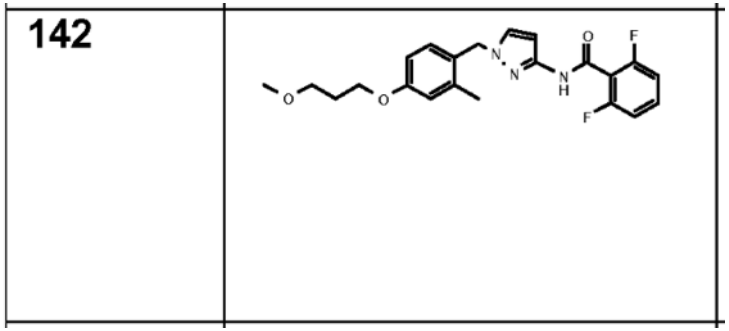


Can we have everything?



Common sources of errors

- Small, poor quality images



- OCR errors in names (OCR done by IFI). There is an OCR correction step, but cannot fix all errors

Example 77: 2,6-Difluoro-*N*-{1-[(4-iodo-2-methylphenyl)methyl]-1*H*-pyrazol-3-yl}benzamide

Becomes: '2,6-Difluoro- Λ/-{1 -r(4-iodo-2-methylphenyl)methvn-1 H-pyrazol-3- vDbenzamide'

- Reliability better for US patents due to inclusion of mol files

PDF

TABLE 2

ID.	Compound	IC50 (microM) (kinase assay)
51	(S)-8-(3-Aminopyrrolidin-1-yl)-2-(4-hydroxyphenyl)thieno[2,3-c]quinolin-4(5H)-one Dihydrochloride	0.078
61	8-Hydroxy-9-(1H-indazol-6-yl)thieno[2,3-c]quinolin-4(5H)-one	0.0035
65	4-(8-Hydroxy-4-oxo-4,5-dihydrothieno[2,3-c]quinolin-9-yl) benzenesulfonamide	0.0018

XML

<thead><row><entry namest=\"1\" nameend=\"3\" rowsep=\"1\">TABLE 2</entry></row><row><entry namest=\"1\" nameend=\"3\" align=\"center\" rowsep=\"1\"/></row><row><entry/><entry/><entry>IC50 (microM)</entry></row><row><entry>ID.</entry><entry>Compound</entry><entry>(kinase assay)</entry></row><row><entry namest=\"1\" nameend=\"3\" align=\"center\" rowsep=\"1\"/></row></thead><tbody valign=\"top\"><row><entry/></row></tbody></tgroup><tgroup align=\"left\" colsep=\"0\" rowsep=\"0\" cols=\"3\"><colspec colname=\"1\" colwidth=\"21pt\" align=\"char\" char=\".\"/><colspec colname=\"2\" colwidth=\"210pt\" align=\"left\"/><colspec colname=\"3\" colwidth=\"49pt\" align=\"char\" char=\".\"/><tbody valign=\"top\"><row><entry>51</entry><entry>**(S)-8-(3-Aminopyrrolidin-1-yl)-2-(4-hydroxyphenyl)thieno[2,3-**</entry><entry>**0.078**</entry></row><row><entry/><entry>**c]quinolin-4(5H)-one Dihydrochloride**</entry></row><row><entry>61</entry><entry>**8-Hydroxy-9-(1H-indazol-6-yl)thieno[2,3-c]quinolin-4(5H)-**</entry><entry>0.0035</entry></row><row><entry/><entry>**one**</entry></row><row><entry>65</entry><entry>4-(8-Hydroxy-4-oxo-4,5-dihydrothieno[2,3-c]quinolin-9-yl)</entry><entry>0.0018</entry></row><row><entry/><entry>benzenesulfonamide</entry></row><row><entry>72</entry><entry>9-[4-(Aminomethyl)phenyl]-8-hydroxythieno[2,3-c]quinolin-</entry><entry>0.00063</entry></row><row><entry/><entry>4(5H)-one</entry></row><row><entry>73</entry><entry>9-[4-(Aminomethyl)phenyl]-8-hydroxythieno[2,3-c]quinolin-</entry><entry>0.00038</entry></row><row><entry/><entry>4(5H)-one Hydrochloride</entry></row><row><entry>77</entry><entry>N-[4-(8-Hydroxy-4-oxo-4,5-dihydrothieno[2,3-c]quinolin-9-</entry><entry>0.0026</entry></row>

SC

TABLE 2

ID.	Compound	IC50 (microM) (kinase assay)
51	(S)-8-(3-Aminopyrrolidin-1-yl)-2-(4-hydroxyphenyl)thieno[2,3-c]quinolin-4(5H)-one Dihydrochloride	0.078
61	8-Hydroxy-9-(1H-indazol-6-yl)thieno[2,3-c]quinolin-4(5H)-one	0.0035
65	4-(8-Hydroxy-4-oxo-4,5-dihydrothieno[2,3-c]quinolin-9-yl) benzenesulfonamide	0.0018
72	9-[4-(Aminomethyl)phenyl]-8-hydroxythieno[2,3-c]quinolin-4(5H)-one	0.00063

Google patents

[0000]

TABLE 2

ID.	Compound	IC50 (microM) (kinase assay)
51	(S)-8-(3-Aminopyrrolidin-1-yl)-2-(4-hydroxyphenyl)thieno[2,3-c]quinolin-4(5H)-one Dihydrochloride	0.078
61	8-Hydroxy-9-(1H-indazol-6-yl)thieno[2,3-c]quinolin-4(5H)-one	0.0035
65	4-(8-Hydroxy-4-oxo-4,5-dihydrothieno[2,3-c]quinolin-9-yl) benzenesulfonamide	0.0018

https://www.surechembl.org/document/US-20150183799-A1 Table 2

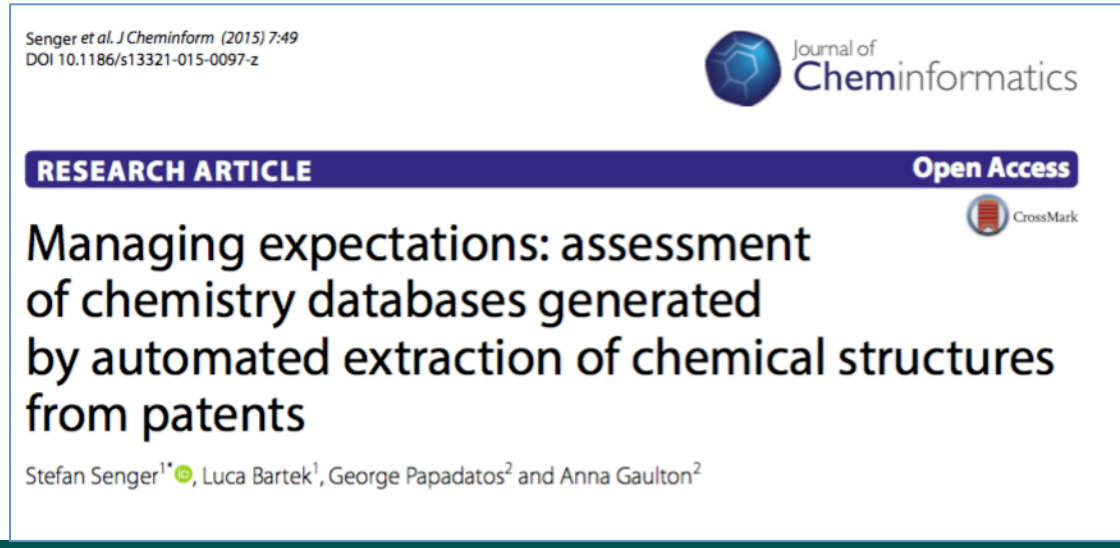
What about testing?

- Rigorous testing hampered by lack of good test sets
  - Akhondi corpus* annotates compounds, genes and diseases
  - → But not normalised to identifiers
  - → Does not convert to compound structures
  - → Identifies *all* entities, not *relevant* entities
  - BioCreative competition corpus covers only abstracts (21K)
- Comparison with 'gold standard'/'ground truth' data sets
  - SciFinder for compounds / Other sets for relevance
  - GVK BIO and BindingDB for targets
  - BUT not necessarily comprehensive, so can't assess precision

* Akhondi *et al* (2014)
http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0107477

Comparison of compound coverage

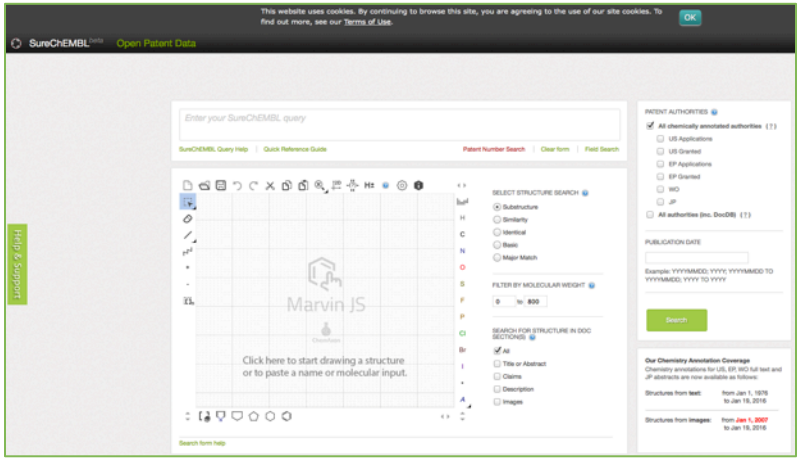
- Comparison of compounds with SciFinder
  - Set of 47 patents from Akhondi* corpus
  - SciFinder compounds for these extracted
  - **65%** of SciFinder compounds were found by SureChEMBL
  - 'Missed' compounds included some cases that were out of scope for SureChEMBL pipeline (e.g., Markush, tables)



* Akhondi *et al* (2014)
http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0107477

Accessing SureChEMBL data

Web Interface



Index of ftp://ftp.ebi.ac.uk/pub/databases/chembl/SureChEMBL/

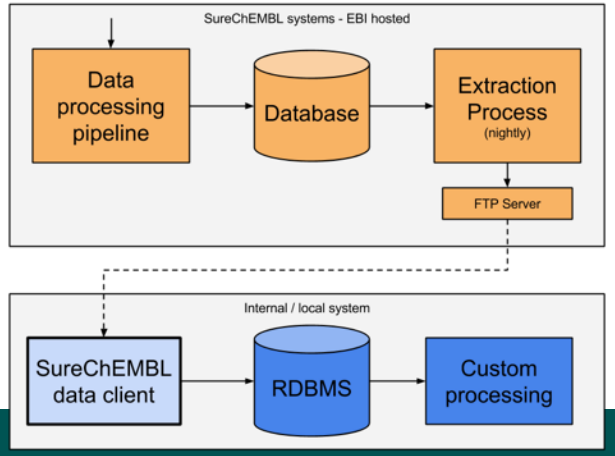
Up to higher level directory

Name	Size	Last Modified
data		14/10/2015 11:26:00

Compound-patent map file
Quarterly updates

UniChem
Weekly updates
Access to 135M structures

Data client feed
Local compound-patent db schema
Daily updates



http://chembl.blogspot.co.uk/2015/08/accessing-surechembl-data-in-bulk.html

EMBL-EBI

Identify key compounds in patent docs

- What are key compounds in med. chem. patents?
  - ‘Best’, most promising/important compound
    - Drug candidate
    - Optimal property profile
    - Most active
- Automated identification of key compounds in patents
  - Using compound information only
  - Main assumption: busy chemical space around key compounds
  - Number of NNs, MCS extraction and R-group frequency

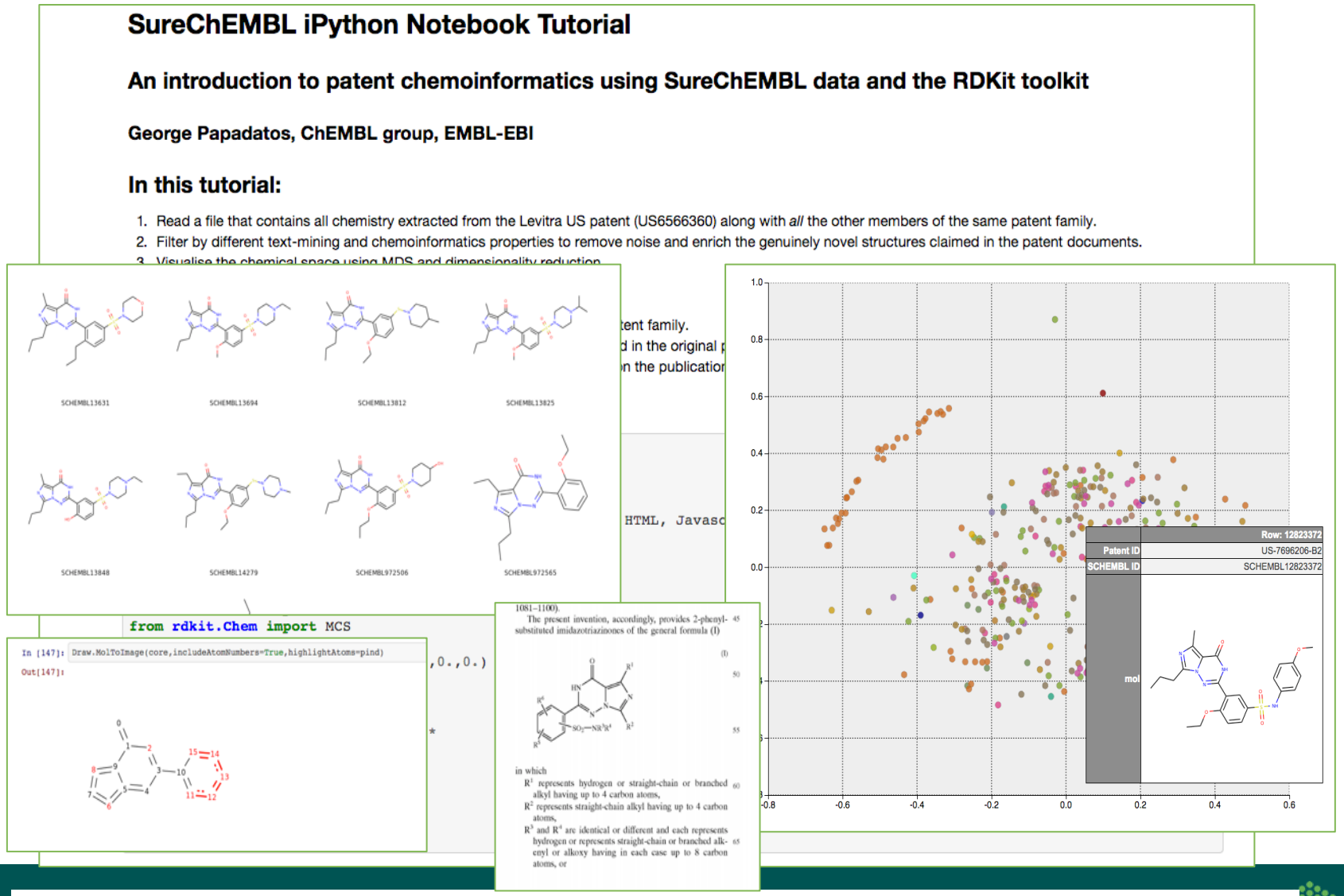
Hattori, K., Wakabayashi, H., & Tamaki, K. (2008). *JCIM*, *48*(1), 135–142. doi:10.1021/ci7002686
Tyrchan, C., Boström, J., Giordanetto, F., Winter, J., & Muresan, S. (2012). *JCIM*, *52*(6), 1480–1489. doi:10.1021/ci3001293

Workflow

1. Read a file with chemistry extracted from the Levitra family of patents
   - US6566360[1]
2. Filter by different text-mining and chemoinformatics properties to remove noise and enrich the genuinely novel structures
3. Visualise the chemical space using MDS and dimensionality reduction
   - Identify and fix outliers in the chemistry space.
4. Find the Murcko and MCS scaffolds
5. Compare the derived MCS core with the actual Markush structure
6. Identify key compounds using structural information only
   - Count number of NNs per compound
   - R-Group decomposition and frequency of R-Groups

[1]Sayle, R. *et al*. (2012). *JCIM*, *52*(1), 51–62. doi:10.1021/ci200463r

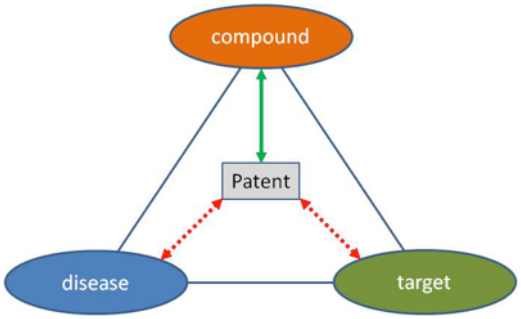
Scaffold and chemical space analysis

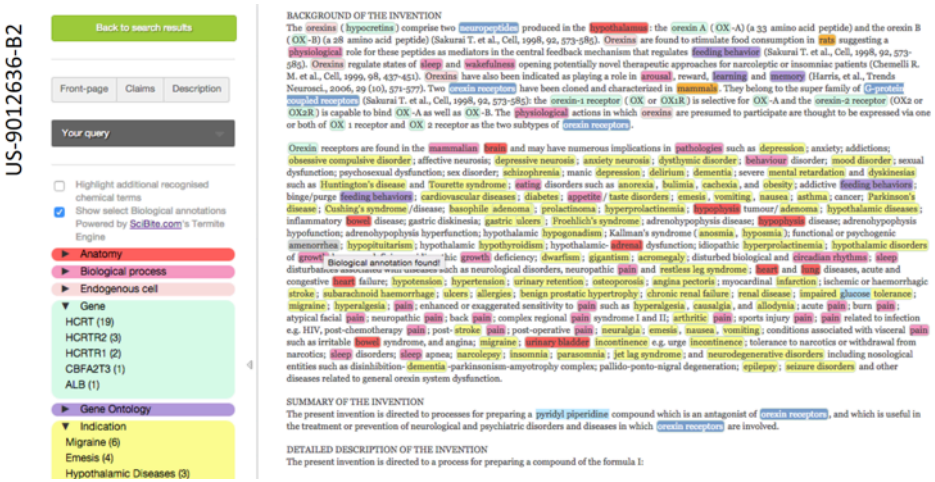


http://nbviewer.ipython.org/github/rdkit/UGM_2014/blob/master/Notebooks/Vardenafil.ipynb

Bio-annotations and Open PHACTS

SureChEMBL bio-annotation

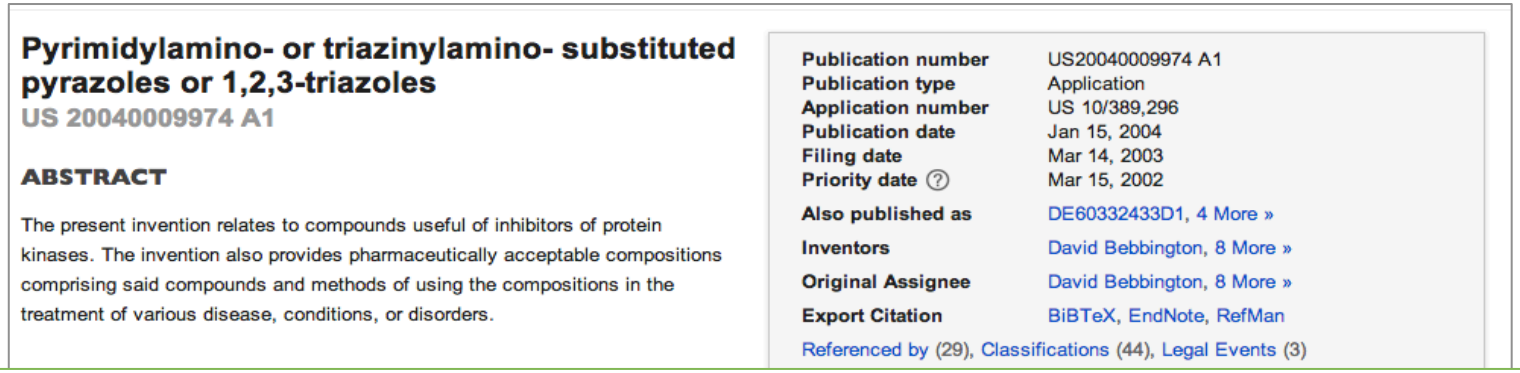




- SciBite's Termite text-mining engine run on 4M life-science patents from SureChEMBL corpus
- Genes (identified by HGNC symbols, e.g. FDFT1) and diseases (identified by MeSH IDs, e.g. D009765) annotated
- Section/frequency information annotated (e.g., in title, abstract, claims, total frequency)
- **Relevance score (0-3)** to flag important chemical and biological entities and remove noise

Relevance scoring

- Patents designed to obfuscate – **not** all entities are equal



Pyrimidylamino- or triazinylamino- substituted pyrazoles or 1,2,3-triazoles

US 20040009974 A1

ABSTRACT

The present invention relates to compounds useful of inhibitors of protein kinases. The invention also provides pharmaceutically acceptable compositions comprising said compounds and methods of using the compositions in the treatment of various disease, conditions, or disorders.

Publication number	US20040009974 A1
Publication type	Application
Application number	US 10/389,296
Publication date	Jan 15, 2004
Filing date	Mar 14, 2003
Priority date ⓘ	Mar 15, 2002
Also published as	DE60332433D1, 4 More »
Inventors	David Bebbington, 8 More »
Original Assignee	David Bebbington, 8 More »
Export Citation	BiBTeX, EndNote, RefMan

Referenced by (29), Classifications (44), Legal Events (3)

Over 50 proteins mentioned!

SRC proto-oncogene, non-receptor tyrosine kinase ; aurora kinase A ; insulin ; microtubule-associated protein tau ; catenin (cadherin-associated protein), beta 1, 88kDa ; interleukin 1, alpha ; c-src tyrosine kinase ; phosducin-like 2 ; colony stimulating factor 2 (granulocyte-macrophage) ; tumor necrosis factor ; glycogen synthase kinase 3 beta ; amyloid beta (A4) precursor protein ; asparaginase ; interferon, beta 1, fibroblast ; jun proto-oncogene ; coagulation factor II (thrombin) ; acetylcholinesterase ; FGR proto-oncogene, Src family tyrosine kinase ; BLK proto-oncogene, Src family tyrosine kinase ; angiotensin I converting enzyme ; albumin ; axin 1 ; heat shock transcription factor 1 ; v-myb avian myeloblastosis viral oncogene homolog ; NADH dehydrogenase, subunit 1 (complex I) ; LYN proto-oncogene, Src family tyrosine kinase ; LCK proto-oncogene, Src family tyrosine kinase ; CCAAT/enhancer binding protein (C/EBP), alpha ; HCK proto-oncogene, Src family tyrosine kinase ; ATP citrate lyase ; Glycogen synthase kinase 3 ; beta-catenin ; interferon ; tnf superfamily ; lactate dehydrogenase ; neurotrophic factor ; pyruvate kinase ; fibroblast growth factor ; glycogen synthase ; tumor necrosis factor alpha ; serine threonine kinase ; tau protein ; interleukin ; albumin ; thrombin ; eif2b ; ion channel ; cytokine ; heat shock factor ; axin ; cAMP Response element binding protein ; amyloid beta ; Microtubule Interacting Protein ; c-Jun N-terminal kinases ; src homology ; calcium channel ; atpase ; acetylcholinesterase

Some more interesting than others…

The method of claim 27, wherein the method comprises inhibiting Aurora-2, GSK-3, or Src activity

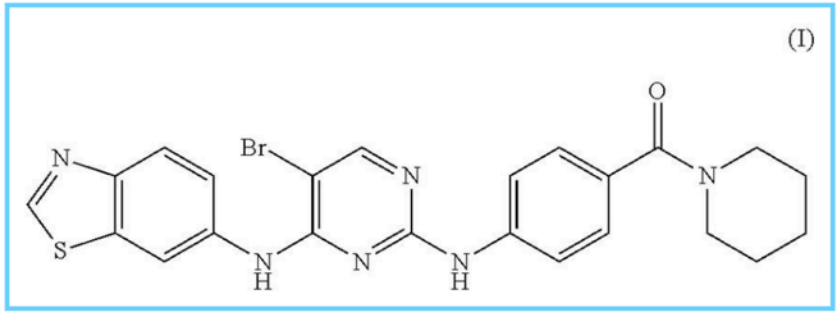
Same for compounds!

US-20160031890-A1 / 2016-02-04

EN NOVEL METHODS, COMPOUNDS, AND COMPOSITIONS FOR INHIBITION OF ROS

EN ABSTRACT

The present invention relates to a method using some novel compounds and compositions for the inhibition of ROS tyrosine kinase. In particular, the present invention covers a method to treat abnormal cell growth, such as cancer, with ROS 10 inhibitors and compositions having ROS inhibitors. An illustrative compound of the invention is shown below Formula (I).



Real scope of patent

Same for compounds!

US-20160031890-A1 / 2016-02-04

EN NOVEL METHODS, COMPOUNDS, AND COMPOSITIONS FOR INHIBITION OF ROS

EN ABSTRACT

The present invention relates to a method using some novel compounds and compositions for the inhibition of ROS tyrosine kinase. In particular, the present invention covers a method to treat abnormal cell growth, such as cancer, with ROS 10 inhibitors and compositions having ROS inhibitors. An illustrative compound of the invention is shown below Formula (I).

Real scope of patent

31. The method of claim 30, wherein said one or more additional therapeutic agents are selected from the group consisting of cytotoxic agent, cisplatin, doxorubicin, taxotere, taxol, etoposide, irinotecan, camptostar, topotecan, paclitaxel, docetaxel, the epothilones, tamoxifen, 5-fluorouracil, methoxtrexate, temozolomide, cyclophosphamide, SCH 66336, tipifarnib (Zarnestra®), R115777, L778,123, BMS 214662, Iressa®, Tarceva®, C225, GLEEVEC®, Intron®, Peg-Intron®, aromatase combinations, ara-C, adriamycin, cytoxan, gemcitabine, Uracil mustard, Chlormethine, Ifosfamide, Melphalan, Chlorambucil, Pipobroman, Triethylenemelamine, Triethylenethiophosphoramine, Busulfan, Carmustine, Lomustine, Streptozocin, Dacarbazine, Floxuridine, Cytarabine, 6-Mercaptopurine, 6-Thioguanine, Fludarabine phosphate, oxaliplatin, leucovirin, oxaliplatin (ELOXATIN®), Pentostatine, Vinblastine, Vincristine, Vindesine, Bleomycin, Dactinomycin, Daunorubicin, Doxorubicin, Epirubicin, Idarubicin, Mithramycin™ Deoxycoformycin, Mitomycin-C, L-Asparaginase, Teniposide 17α-Ethinylestradiol, Diethylstilbestrol, Testosterone, Prednisone, Fluoxymesterone, Dromostanolone propionate, Testolactone, Megestrol acetate, Methylprednisolone, Methyltestosterone, Prednisolone, Triamcinolone, Chlorotrianisene, Hydroxyprogesterone, Aminoglutethimide, Estramustine, Medroxyprogesteroneacetate, Leuprolide, Flutamide, Toremifene, goserelin, Carboplatin, Hydroxyurea, Amsacrine, Procarbazine, Mitotane, Mitoxantrone, Levamisole, Navelbene, Anastrazole, Letrazole, Capecitabine, Reloxafine, Droloxafine, Hexamethylmelamine, Avastin, herceptin, Bexxar, Velcade, Zevalin, Trisenox, Xeloda, Vinorelbine, Porfimer, Erbitux, Liposomal, Thiotepa, Altretamine, Melphalan, Trastuzumab, Lerozole, Fulvestrant, Exemestane, Ifosfomide, Rituximab, C225, Campath, leucovorin, and dexamethasone, bicalutamide, carboplatin, chlorambucil, cisplatin, letrozole, megestrol, valrubicin, and vinvlastin.

Tens of drugs trivially mentioned

Same for compounds!

US-20160031890-A1 / 2016-02-04

EN NOVEL METHODS, COMPOUNDS, AND COMPOSITIONS FOR INHIBITION OF ROS

EN ABSTRACT

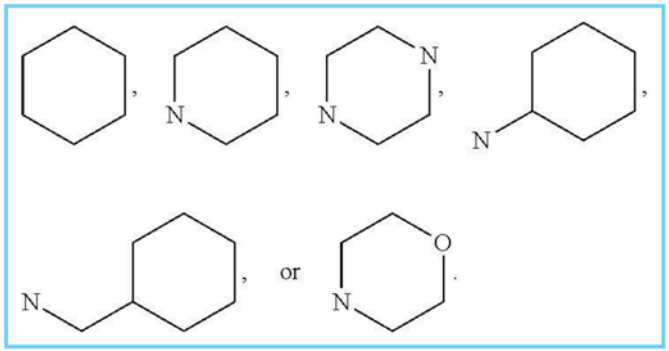
The present invention relates to a method using some novel compounds and compositions for the inhibition of ROS tyrosine kinase. In particular, the present invention covers a method to treat abnormal cell growth, such as cancer, with ROS 10 inhibitors and compositions having ROS inhibitors. An illustrative compound of the invention is shown below Formula (I).

Real scope of patent

31. The method of claim 30, wherein said one or more additional therapeutic agents are selected from the group consisting of cytotoxic agent, cisplatin, doxorubicin, taxotere, taxol, etoposide, irinotecan, camptostar, topotecan, paclitaxel, docetaxel, the epothilones, tamoxifen, 5-fluorouracil, methoxtrexate, temozolomide, cyclophosphamide, SCH 66336, tipifarnib (Zarnestra®), R115777, L778,123, BMS 214662, Iressa®, Tarceva®, C225, GLEEVEC®, Intron®, Peg-Intron®, aromatase combinations, ara-C, adriamycin, cytoxan, gemcitabine, Uracil mustard, Chlormethine, Ifosfamide, Melphalan, Chlorambucil, Pipobroman, Triethylenemelamine, Triethylenethiophosphoramine, Busulfan, Carmustine, Lomustine, Streptozocin, Dacarbazine, Floxuridine, Cytarabine, 6-Mercaptopurine, 6-Thioguanine, Fludarabine phosphate, oxaliplatin, leucovirin, oxaliplatin (ELOXATIN®), Pentostatine, Vinblastine, Vincristine, Vindesine, Bleomycin, Dactinomycin, Daunorubicin, Doxorubicin, Epirubicin, Idarubicin, Mithramycin™ Deoxycoformycin, Mitomycin-C, L-Asparaginase, Teniposide 17α-Ethinylestradiol, Diethylstilbestrol, Testosterone, Prednisone, Fluoxymesterone, Dromostanolone propionate, Testolactone, Megestrol acetate, Methylprednisolone, Methyltestosterone, Prednisolone, Triamcinolone, Chlorotrianisene, Hydroxyprogesterone, Aminoglutethimide, Estramustine, Medroxyprogesteroneacetate, Leuprolide, Flutamide, Toremifene, goserelin, Carboplatin, Hydroxyurea, Amsacrine, Procarbazine, Mitotane, Mitoxantrone, Levamisole, Navelbene, Anastrazole, Letrazole, Capecitabine, Reloxafine, Droloxafine, Hexamethylmelamine, Avastin, herceptin, Bexxar, Velcade, Zevalin, Trisenox, Xeloda, Vinorelbine, Porfimer, Erbitux, Liposomal, Thiotepa, Altretamine, Melphalan, Trastuzumab, Lerozole, Fulvestrant, Exemestane, Ifosfomide, Rituxi
carboplatin, chlorambucil, cisplatin, letrozole, megestrol, valrubicin, and vinvl

Tens of drugs trivially mentioned

11. The method of claim 10, wherein R1 is



Substituents

Relevance scoring – genes/diseases

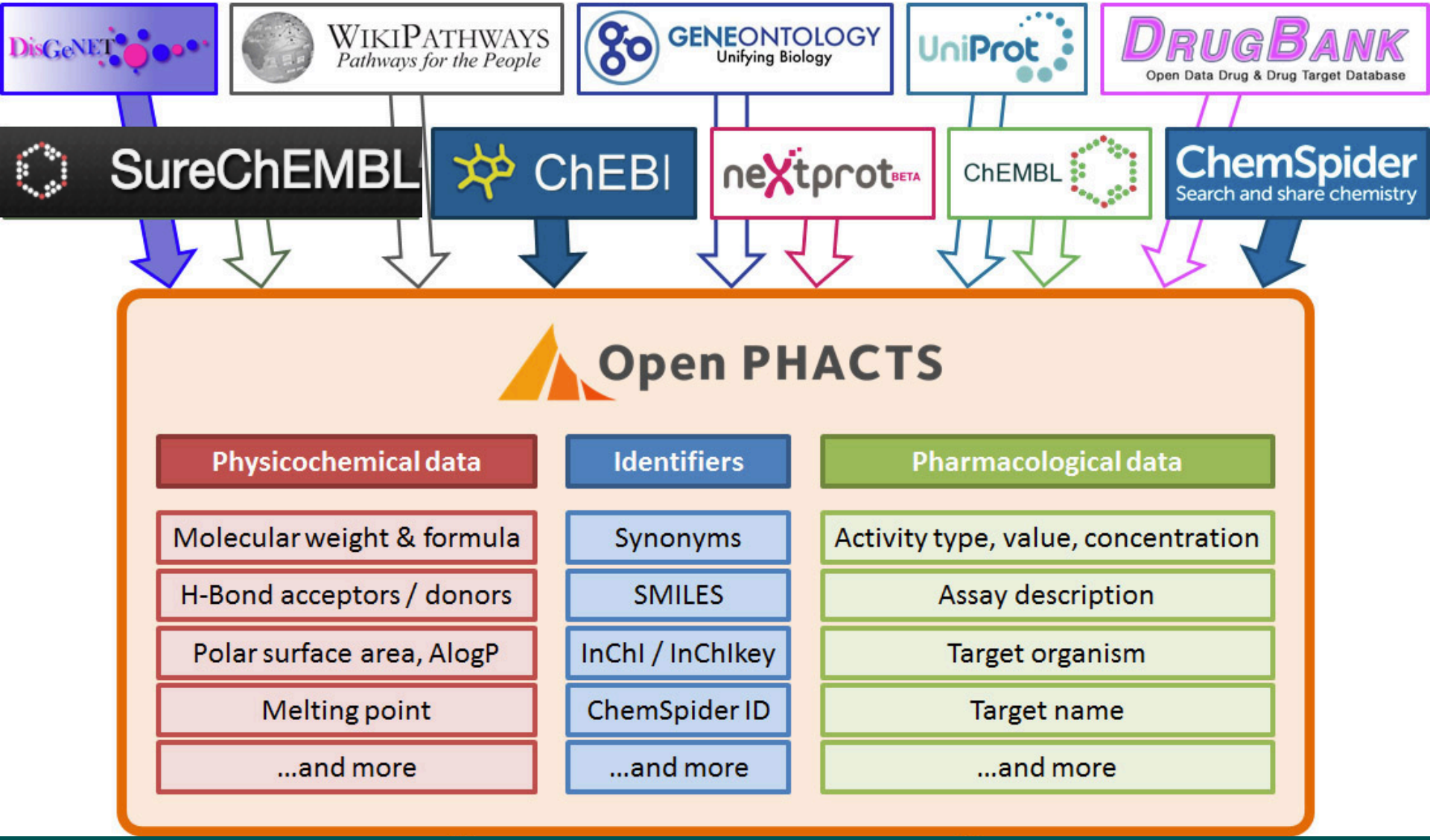
- Various features used:
  - Term frequency
  - Position (title, abstract, figure, caption, table)
  - Frequency distribution
- Scores range from 0 – 3
  - 3 – most important entities in the patent
  - 2 – important entities in the patent
  - 1 – mentioned entities in the patent
  - 0 – ambiguous entity/likely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

Relevance scoring - compounds

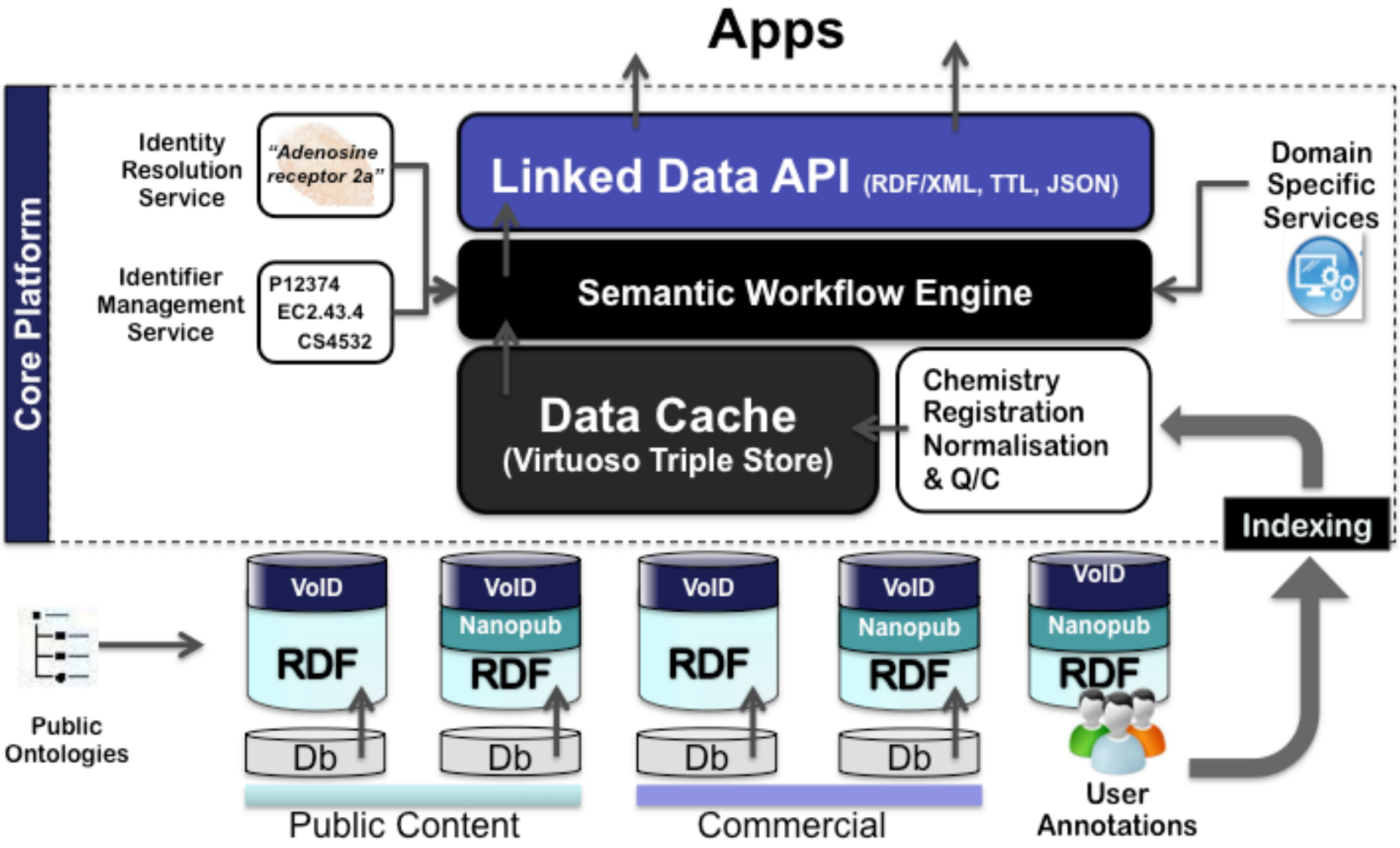
- Main assumptions for relevance:
  1. Very frequent compounds are usually irrelevant
  2. Compounds with busy chemical space around them are **interesting**
     - Use distribution of close analogues (NNs) among compounds found in the **same patent family** – FCFP_4 similarity
- Scores range from 0 – 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

Open PHACTS Sources



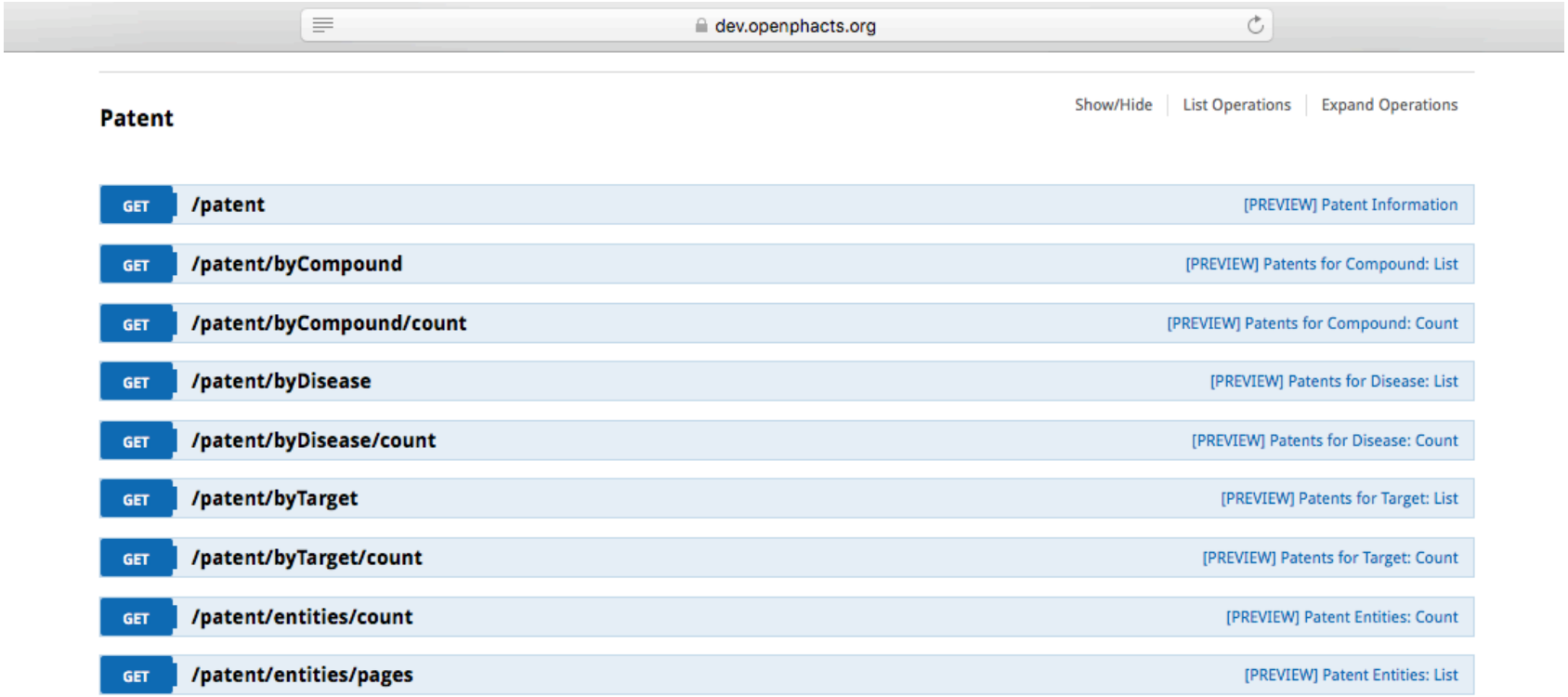
Open PHACTS Architecture



Drug Discovery Today 2012, 17:21 (doi:10.1016/j.drudis.2012.05.016)

EMBL-EBI

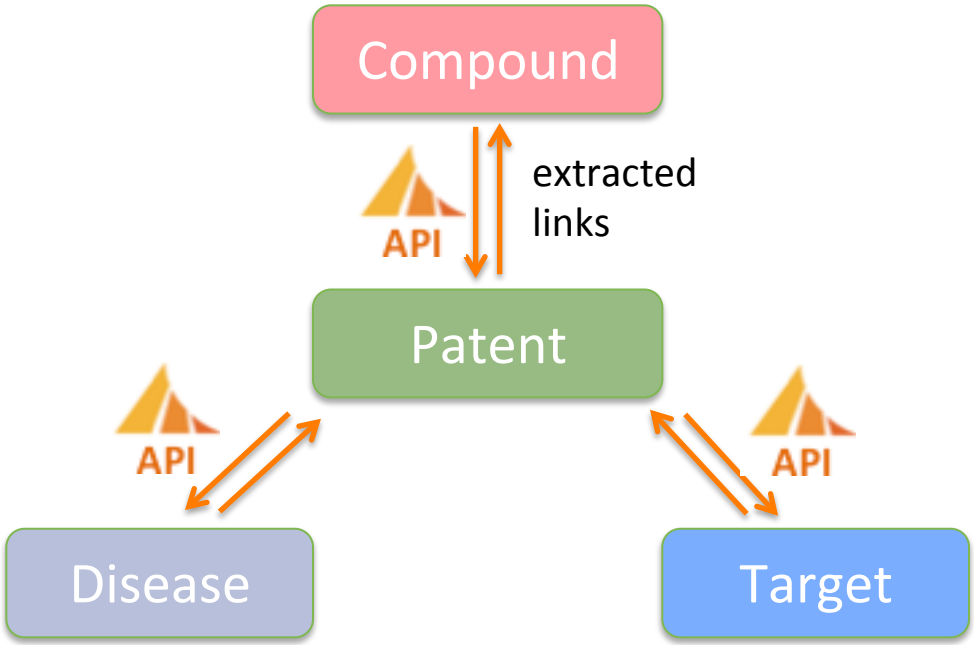
Open PHACTS Patent API



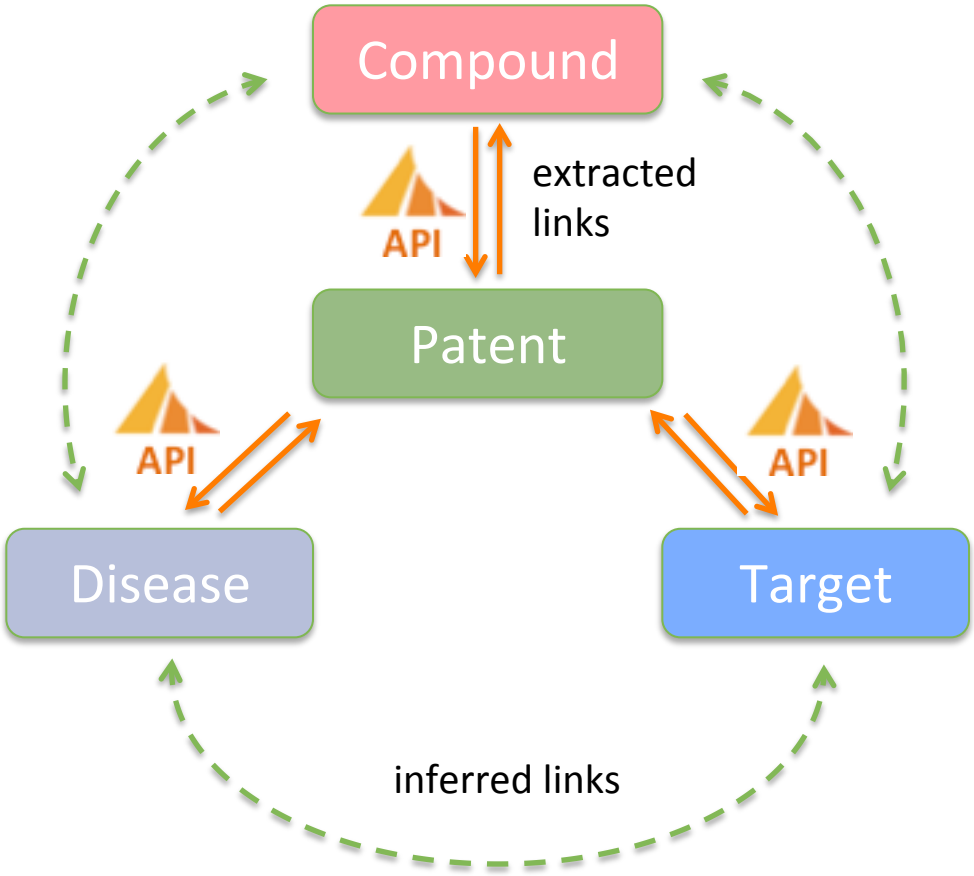
https://dev.openphacts.org/docs/2.1

EMBL-EBI

Open PHACTS Patent API

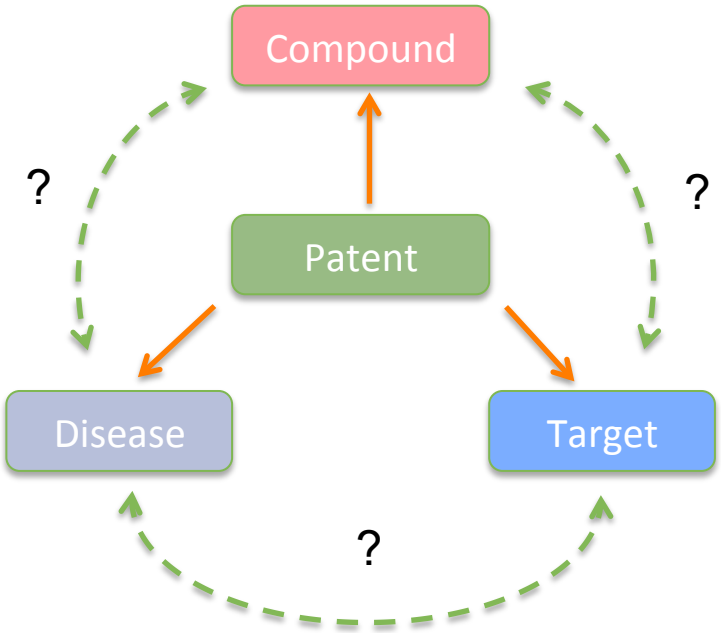


Open PHACTS Patent API



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

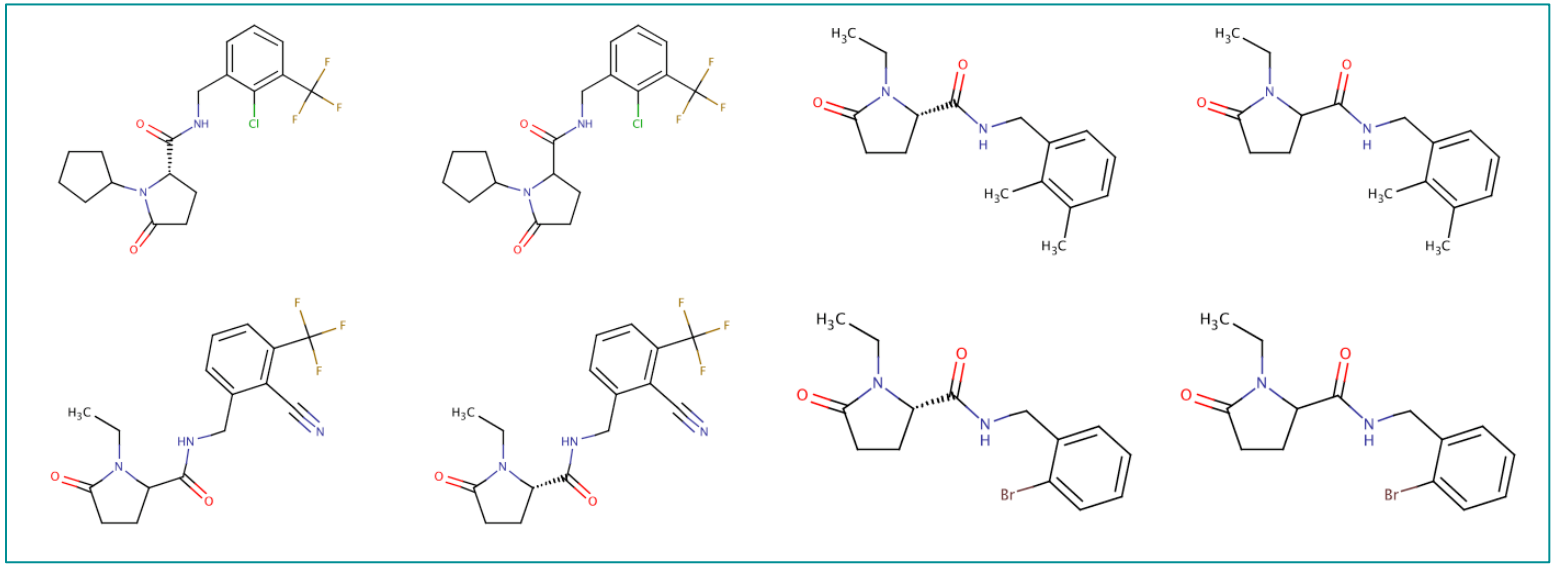
Angier et al.; "Some N-Substituted-5-oxo-2-pyrrolidinecarboxamides"; Journal of Organic Chemistry; 1956; vol. 21, No. 12; pp. 1540-1543.
Zhao et al., *Bioorganic & Medicinal Chemistry*, 7(8), pp. 1647-1654 (1999).
Møss et al., *International Journal of Pharmaceutics*, 52(3), pp. 255-263 (1989).
Tye et 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

23 relevant targets and diseases:

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

201 relevant structures:



https://www.surechembl.org/document/US-7718693-B2/

EMBL-EBI

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**

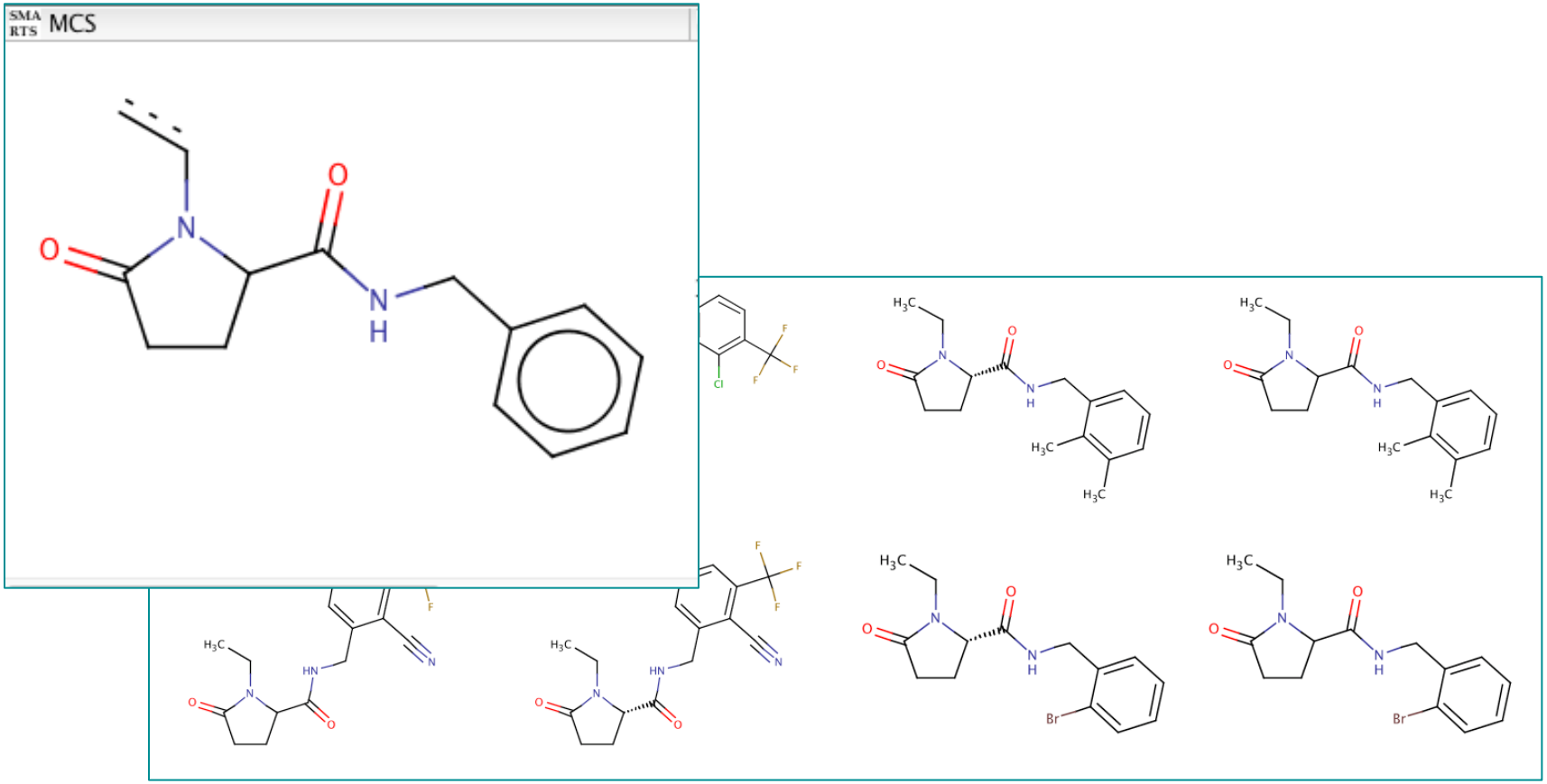
The present invention relates to novel oxo-prolinamide derivatives of formula (I) which modulate P2X7 receptor function and are capable of antagonizing the effects of ATP at the P2X7 receptor

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

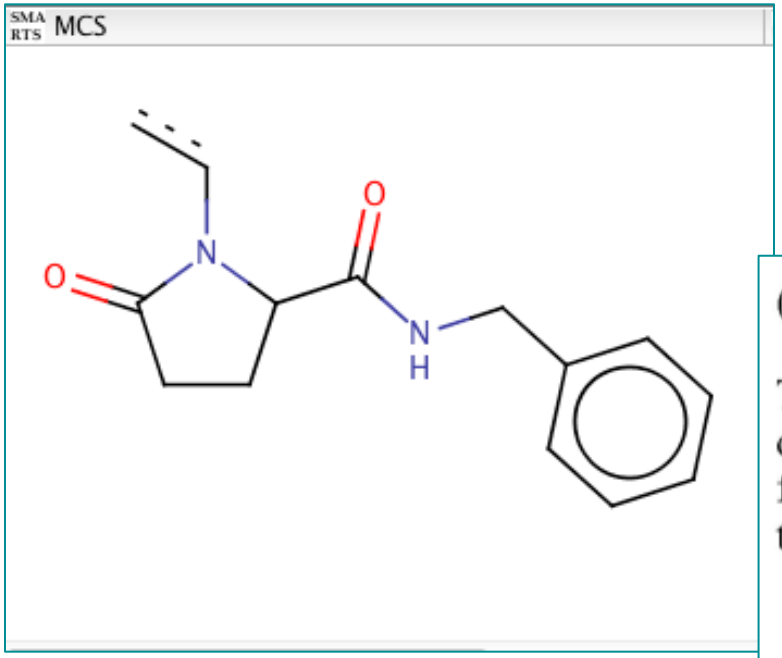
Does it make sense?

- Calculate MCS



Does it make sense?

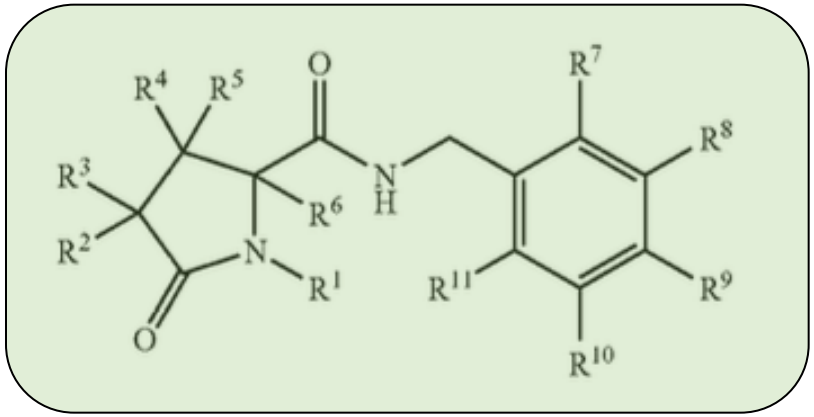
- Calculate MCS



US-7718693-B2

(57) **ABSTRACT**

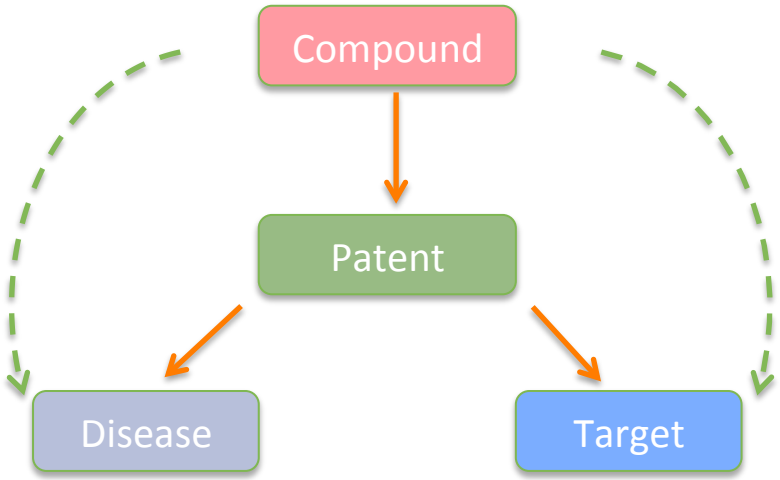
The present invention relates to novel oxo-prolinamide derivatives of formula (I) which modulate P2X7 receptor function and are capable of antagonizing the effects of ATP at the P2X7 receptor



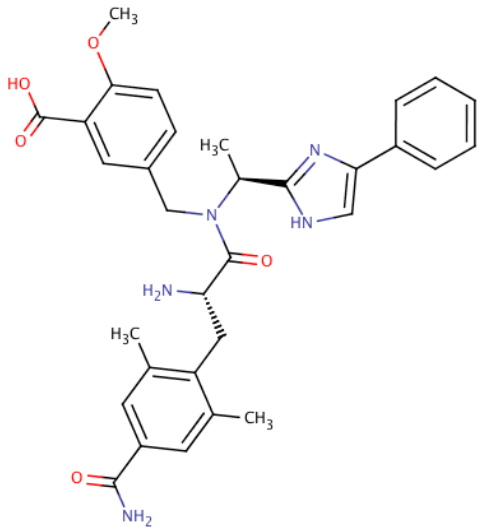
(I)

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



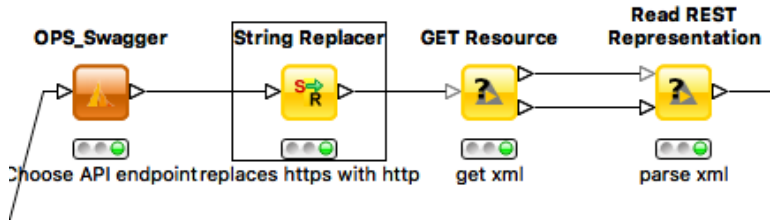
Eluxadoline (JNJ-27018966, VIBERZI)



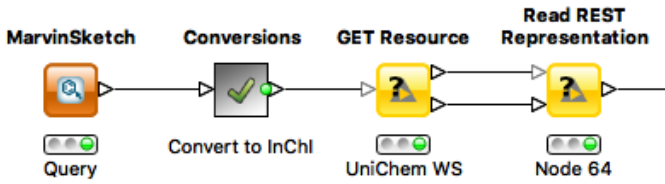
CHEMBL2159122
SCHEMBL12971682
FDA Approval: 2015

1) Get patents for Eluxadoline

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



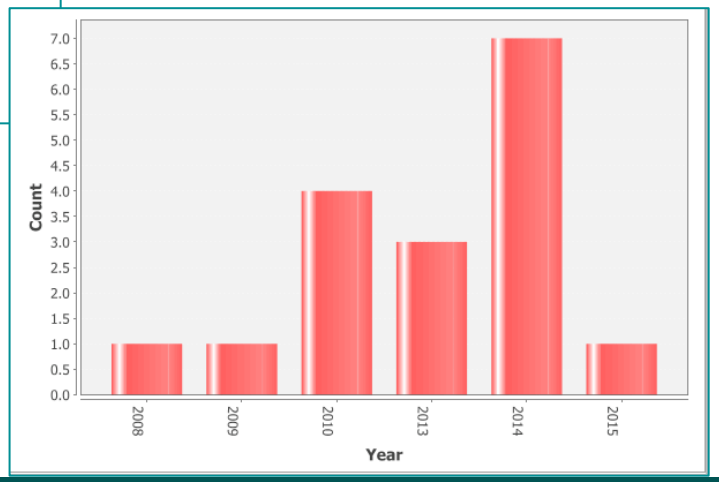
  - Relevance score >=1 → 17 patents (**patentome**):



1) Get patents for Eluxadoline

- 17 relevant 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



Results

Relevant targets:

Relevant diseases:

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

New PARP inhibitors

IN THE PIPELINE

Derek Lowe's commentary on drug discovery and the pharma industry. An editorially independent blog from the publishers of *Science Translational Medicine*.

By Derek Lowe

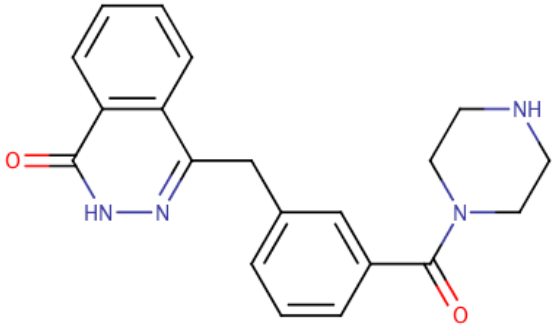
PARP Inhibitors Come Through

By Derek Lowe | June 29, 2016

PARP (poly ADP-ribose polymerase) inhibitors have had a rough time of it in the clinic. Sanofi had one (**iniparib**) fail, but it later turned out, **embarrassingly**, that it **wasn't** really a PARP inhibitor at all (*update: more on this **here***). AstraZeneca had a legitimate one fail as well (**olaparib**). Merck got out of the area and **outlicensed** their **compound** to a small company called Tesaro, and Pfizer **spun** their **molecule** off to another small outfit, Clovis (who have since had a **wild ride** of their own with other compounds). Meanwhile, Medivation **recently acquired** Biomarin's PARP inhibitor, **talazoparib**.

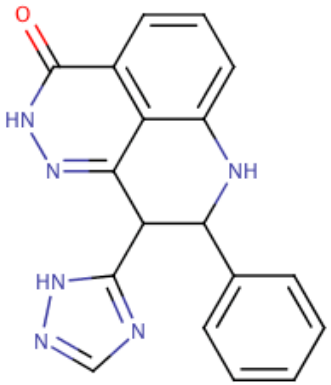
From target to scaffolds

PARP1



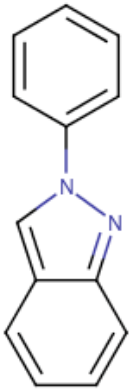
Olaparib, FDA-approved

WO-2008020180-A2



Talazoparib, Phase 3

US-20100035883-A1

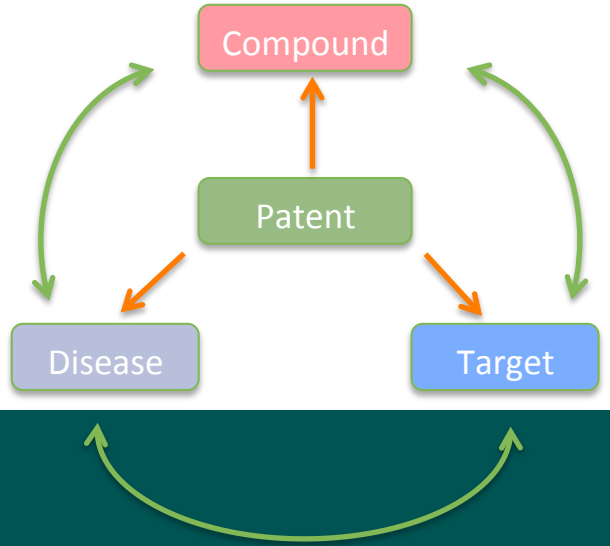


Niraparib, Phase 3

US-20080167345-A1

Take home message

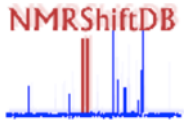
- It is now possible to search and interlink the *key* structures, scaffolds, targets and diseases from med. chem. patent corpus *automatically*
  - By high-throughput text-mining only
  - Thanks to simple heuristics (relevance scores and frequency)
  - Using the Open PHACTS API
  - For the first time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

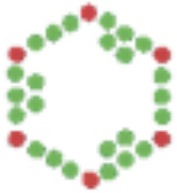


What next: Other ideas and use cases

- Target validation / Druggability
  - For a target, get me all related and relevant diseases
    - Compare with DisGeNET / Open Targets, etc.
  - Any known patented scaffolds for my target?
    - Start a pharmacophore hypothesis for patent busting
    - Find a tool compound for a new assay
- Novelty checking / Due diligence
  - What do we know about this scaffold / compound?
- Add ChEMBL pharmacology and pathway information

The Data Are Out There





What’s cooking in ChEMBL22

BindingDB patent set

- BindingDB manually extracts binding activities from tables in *granted* US med. chem. patents
  - Coverage 2013-2016, currently ~1,100 patents
    - On-going effort
  - Includes exact and binned bioactivities
    - Ki, IC50, Kd, EC50 activity types
  - Compounds mapped to example no in patent tables
  - Detailed assay description
  - http://www.bindingdb.org/bind/ByPatent.jsp

BindingDB patent bioactivity integration

- 1,017 US patents (mapped to SureChEMBL IDs)
- 68.6K compounds
  - ~70 unique cmpds per patent
  - Only **5%** exist in ChEMBL (exact match) → complementary
- 450 single protein targets
  - ~10% not in ChEMBL, e.g. GPR4 in US-8748435-B2
- 100,329 bioactivity values
  - ~100 activities per patent
- Compounds and targets curated by ChEMBL
- **Available in ChEMBL22**

Example: BI patent

https://www.surechembl.org/document/US-9120769-B2/

US-9120769-B2 / 2015-09-01

Aryl-and heteroarylcarbonyl derivatives of hexahydroindenopyridine and octahydrobenzoquinoline

TABLE 2

174 activities in Table 2

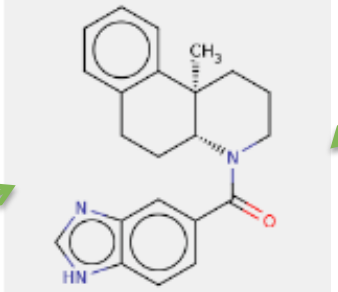
Inhibitory activity on 11β-HSD of
Examples compiled in the experimental part

Example	IC50 [nM]
1	106
2	65
3	1380
4	573
5	861
6	296
7	410
8	342
9	481
10	321
11	94
12	367
13	257
14	405
15	1617
16	1412
17	2107
18	1027
19	1902
20	227

Example 11 (1H-Benzoimidazol-5-yl)-(trans-10b-methyl-2,3,4a,5,6,10b-hexahydro-1H-benzo[f]quinolin-4-yl)-methanone

Manual example number dereference

BDBM177587



(US9120769, 11)

IC50=**94nM** against **HSD11B1** (P28845)

Assay Name: Homogeneous Time-Resolved Fluorescence (HTRF) Assay

Description: n vitro inhibition of 11β-HSD1 by test compounds is determined with HTRF (Homogeneous Time-Resolved Fluorescence) technology (cisbio international, France) detecting cortisol generated from cortisterone by human liver microsomes. Briefly, compounds are incubated for 1 hour at 37° C. in Tris buffer (20 mM tris, 5 mM EDTA, pH 6.0) containing NADPH (200 μM) and cortisone (80 nM). Cortisol generated in the reaction is then detected with a competitive immunoassay, involving two HTRF conjugates: cortisol linked to XL665 and anti-cortisol antibody labeled with Europium cryptate. The incubation period for detection reaction is typically 2 hours. The amount of cortisol is determined by reading the time-resolved fluorescence of the wells (Ex 320/75 nm; Em 615/8.5 nm and 665/7.5 nm).

Conclusion: current patent data landscape

- ChEMBL
  - Literature SAR data, drug annotations, *manually* curated
  - Coming soon: bioactivities from patents
- SureChEMBL
  - Chemical annotations *automatically* mined from patents
- Open PHACTS
  - The above plus biological annotations *automatically* mined from patents, with relevance scoring
  - Public API

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.

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 *

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

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
- surechembl-help@ebi.ac.uk

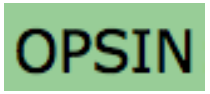
innovative medicines initiative

wellcome trust Strategic Award

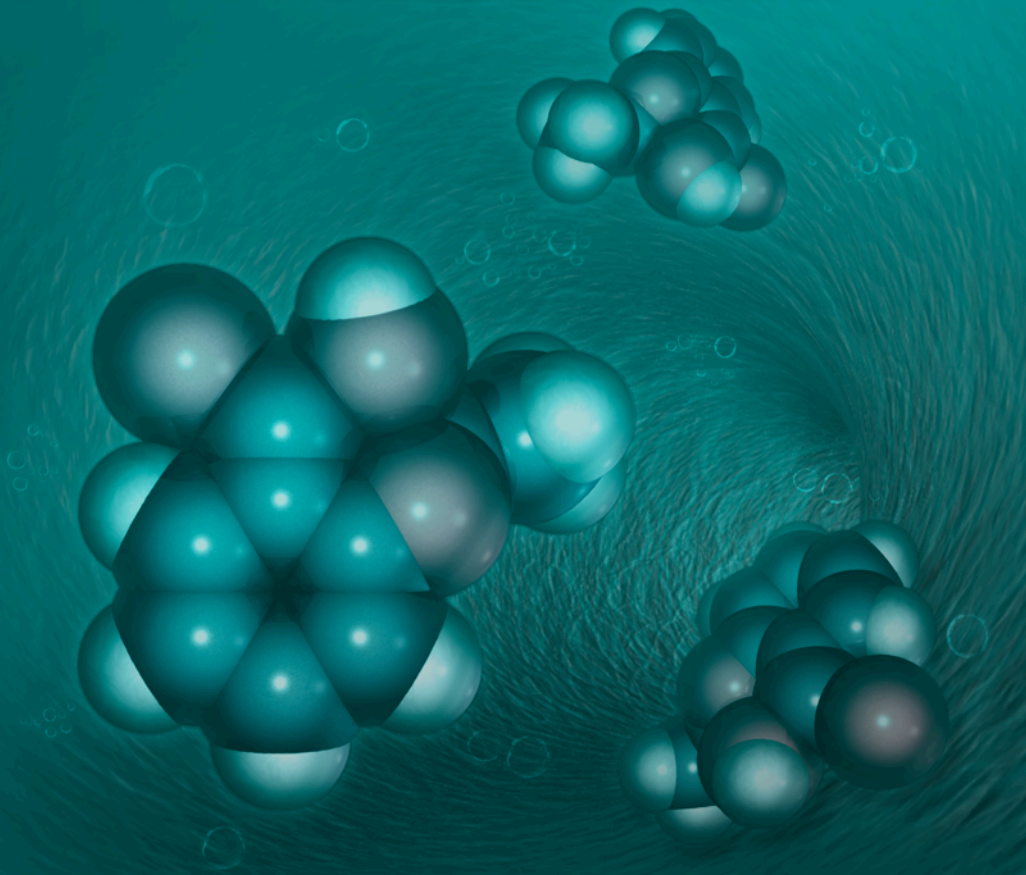
EMBL-EBI

The Open PHACTS Foundation

Technology partners



Mining Compounds, Targets and Indications from the Patent Corpus with **SureChEMBL** and **Open PHACTS**



George Papadatos, PhD

ChEMBL Group, EMBL-EBI

georgep@ebi.ac.uk

#ShefChem16

EMBL-EBI