SureChEMBL: An open patent chemistry resource

George Papadatos, PhD

ChEMBL Group, EMBL-EBI

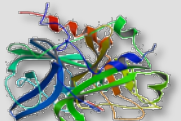
georgep@ebi.ac.uk

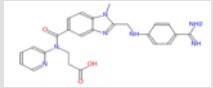
EMBL-EBI

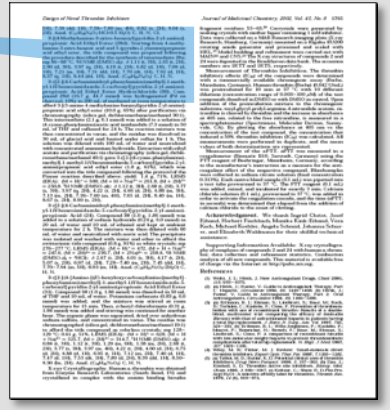
ChEMBL: Data for drug discovery

1. Scientific facts

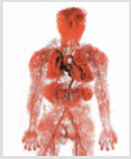
3. Insight, tools and resources for translational drug discovery







>Thrombin
MAHVRGLQLPGCLALAALCSLVHSQHVFLAPQQARSLLQRVRRANTFLEEVRKGNLE
RECVEETCSYEEAFEALESSTATDVFWAKYTACETARTPRDKLAACLEGNCAESLGT
NYRGHVNITRSGIECQLWRSRYPHKPEINSTTHPGADLQENFCRNPDSSTTGPWCYT
TDPTVRRQECSIPVCGQDQVTVAMTPRSEGSSVNLSPPLEQCVPDRGQQYQGRL
THGLPCLAWASAQAKALSKHQDFNSAVQLVENFCRNPDGDEEGVWCYVAGKPGDFGY
CDLNYCEEAVEEETGDGLDEDSDRAIEGRTATSEYQTFFNPRTFGSGEADCGLRP
EKKSLEDKTERELLESYIDGRIVEGSDAEIGMSPWQVMLFRKSPQELLCGASLIS
WVLTAAHCLLYPPWDKNFTENDLLVRIGKHSRTRYERNIEKISMLEKIYIHPRYN
ENLDRDIALMKLKKPVAFSDYIHPVCLPDRETAASLLQAGYKGRVTGWGNLKET
NVGKGQPSVLQVVNLPIVERPVCKDSTRIRITDNMFCAGYKPDEGKRGDACEGD
PFVMKSPFNNRWYQMGIVSWGEGCDRDGKYGFYTHVFRLKKWIQKVIDQFGE



Assay/Target

Compound

K_i = 4.5nM

Bioactivity data

APTT = 11 min.

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

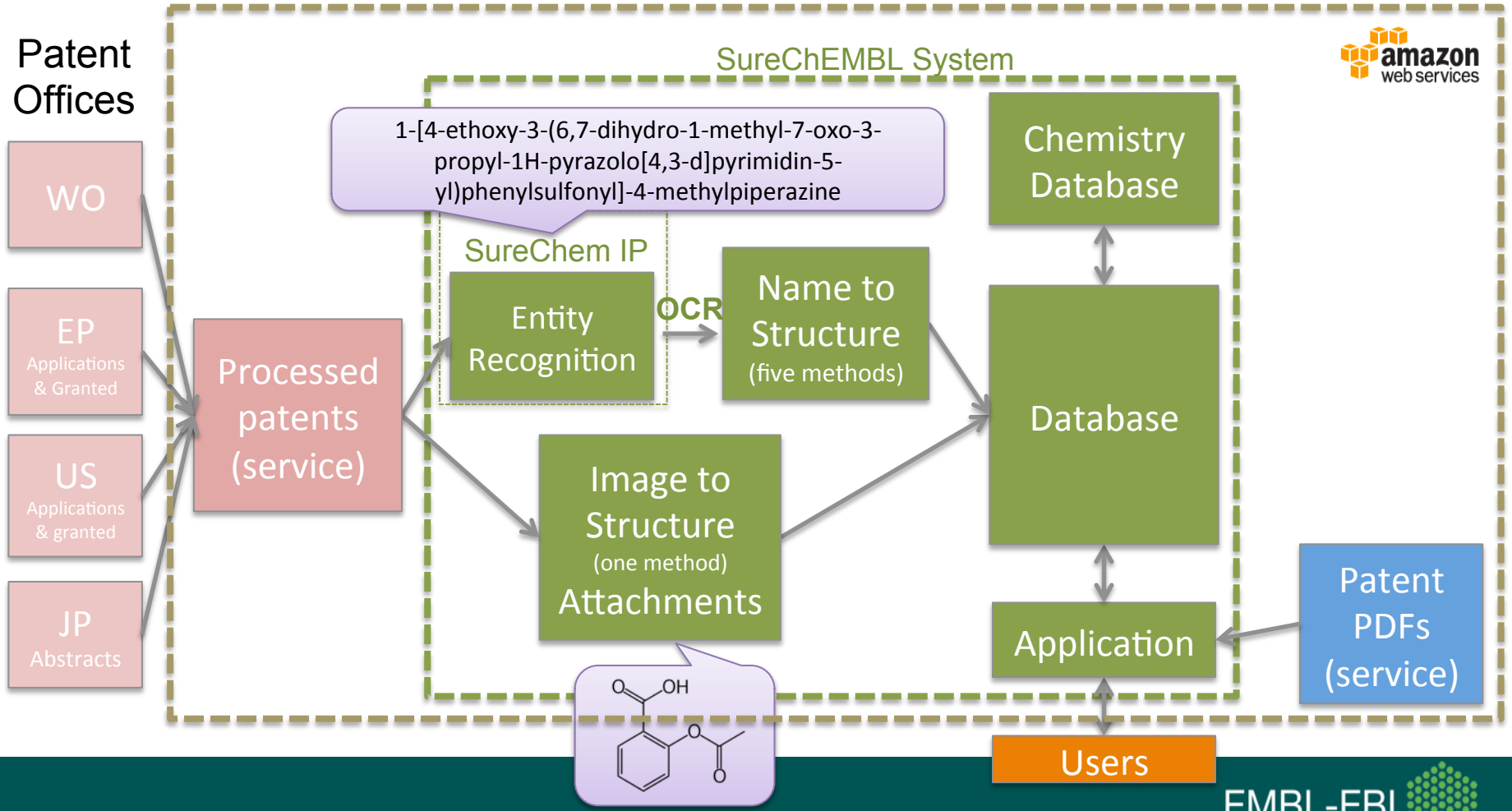
Why looking at patent documents?

- 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
  - Average lag of 1-2 years between patent document and journal publication disclosure for chemistry
    - Compounds, scaffolds, reactions
    - Biological targets, diseases, indications

From SureChem to SureChEMBL

- Digital Science/Macmillan donated SureChem to EMBL-EBI
  - SureChem: commercial patent chemistry mining product
- Wellcome Trust funds further development
- EMBL-EBI provides an on-going, live service
  - Full functionality *freely* available to everyone
  - Query, view and export chemistry from patents
  - Complemented with biological annotations *via* Open PHACTS

SureChEMBL data processing



www.surechembl.org

Homepage

Search by keyword 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 | Fielded Search

PATENT AUTHORITIES

All chemically annotated authorities (?)

US Applications

US Granted

EP Applications

EP Granted

WO

JP

ll authorities (inc. DocDB) (?)

hEMBL Patent Number Search Format

Search by chemical structure (sketch compound)

Click here to draw a structure

Manual structure input

SELECT STRUCTURE SEAR

Substructure

Similarity

Identical

Basic

Major Match

Chemical search type (substructure, similarity, identical)

FILTER BY MOLECULAR WEIG

0 to 800

Filter by MW

CATION DATE

YYYYMMDD; YYYY; YYYYMMDD TO DD; YYYY TO YYYY

Filter by date

Search

SEARCH FOR STRUCTURE IN DOC SECTION(S)

All

Title or Abstract

Claims

Description

Images

Search for

Our Chemistry Annotation Coverage NEW!

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

Jan 1, 1976 to

Jan 1, 2007 to

Search by SMILES, MOL, SMARTS, name

Filter by document section (title, claims, abstract, description and images)

Help

EMBL-EBI

pdyear:2014 AND ic:C07D AND (ttl:hepatitis OR ab:hepatitis) AND pa:Bristol*

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

Manual structure input

SELECT STRUCTURE SEARCH

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

FILTER BY MOLECULAR WEIGHT

0 to 800

SEARCH FOR STRUCTURE IN DOC SECTION(S)

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

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

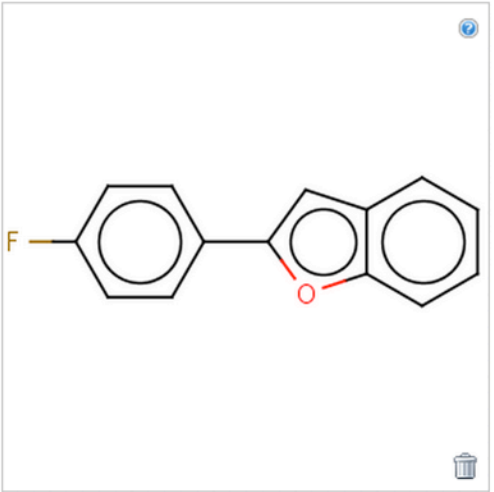
Search

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

Structures from **text** annotations:	from Jan 1, 1976 to date
Structures from **images**:	from **Jan 1, 2007** to date

pdyear:2014 AND ic:C07D AND (ttl:hepatitis OR ab:hepatitis) AND pa:Bristol*

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



Manual structure input

SELECT STRUCTURE SEARCH

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

FILTER BY MOLECULAR WEIGHT

0 to 800

SEARCH FOR STRUCTURE IN DOC SECTION(S)

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

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

Structures from **text** annotations:	from Jan 1, 1976 to date
Structures from **images**:	from **Jan 1, 2007** to date

Back to search results

Front-page | Claims | Description

Your query

Highlight additional recognised chemical terms

Export document chemistry

Export family chemistry

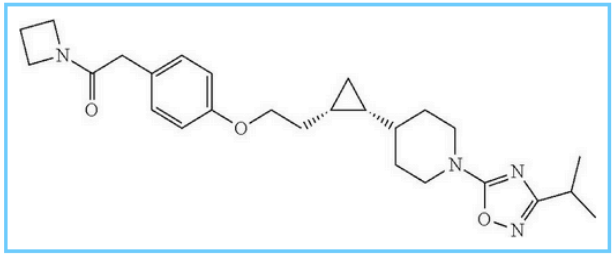
Download patent PDF

water (1:1). The solution was refluxed for 1 hour, cooled to RT, and concentrated to dryness under reduced pressure. The residue was dissolved in methanol (5 mL), neutralized by the addition of excess potassium carbonate, mixed with silica gel (2 g), and concentrated to dryness under reduced pressure. The residue was loaded on a silica gel column (15 g of silica gel) and eluted with dichloromethane / methanol (199:1, 1 L) to provide impure product (72 mg).

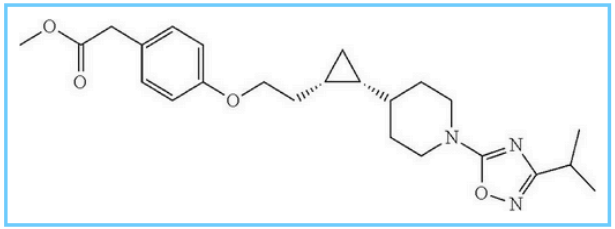
The compound was further purified by prep-HPLC (SunFire C18 OBD, 10 50×150 mm, 118 mL/min, acetonitrile/water 10:90 to 90:10 at 25 min, total run 30 min) to provide the title compound as an off-white solid. 1H NMR (300 MHz, $CDCl_3$) δ 8.89 (s, 1H), 7.60 (d, J=9.0 Hz, 2H), 7.07 (d, J=9.0 Hz, 2H), 4.20-4.05 (m, 4H), 3.12-2.95 (m, 2H), 2.95-2.80 (m, 1H), 2.30-2.10 (m, 1H), 1.95-1.80 (m, 2H), 1.65-1.40 (m, 3H), 1.29 (d, J=6.9 Hz, 6H), 1.15-0.90 (m, 2H), 0.80-0.55 (m, 2H), −0.06 (q, J=4.5 Hz, 1H). MS (ESI) m/z 424 [M+H]+.

GPR119 Human EC_{50}: 11.8 nM

Example 99Preparation of 4-[(1R,2S)-2-{2-[4-(2-azetidin-1-yl-2-oxoethyl)phenoxy]ethyl}cyclopropyl]-1-[3-(1-methylethyl)-1,2,4-oxadiazol-5-yl]piperidine



Step A: Methyl 2-(4-(2-((1S,2R)-2-(1-(3-isopropyl-1,2,4-oxadiazol-5-yl)piperidin-4-yl)cyclopropyl)ethoxy)phenyl)acetate



2-((1S,2R)-2-(1-(3-isopropyl-1,2,4-oxadiazol-5-yl)piperidin-4-yl)cyclopropyl)ethanol (1.7 g, 6.08 mmol), methyl 2-(4-hydroxyphenyl)acetate (1.5 g, 9.1 mmol) and triphenylphosphine (2.4 g, 9.1 mmol) were dissolved in THF (30 ml). The mixture was stirred at RT under N2 for 5 min and disisopropyl azodicarboxylate (1.78 ml, 9.1 mmol) was added. The mixture was stirred at RT overnight. The mixture was diluted with DCM (50 ml), washed with water, dried and evaporated. The crude material was purified by silica gel column (100 g SNAP, 5-25% EtOAc in hexane) to afford the desired product. LC/MS (m/z): 428 (M+H)+.

Step B: 2-(4-(2-((1S,2R)-2-(1-(3-isopropyl-1,2,4-oxadiazol-5-yl)piperidin-4-yl)cyclopropyl)ethoxy)phenyl)acetic acid

Back to search results

Front-page | Claims | Description

Your query

Highlight additional recognised chemical terms

Export document chemistry

Export family chemistry

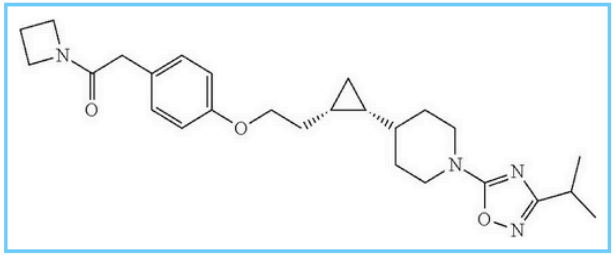
Download patent PDF

water (1:1). The solution was refluxed for 1 hour, cooled to RT, and concentrated to dryness under reduced pressure. The residue was dissolved in methanol (5 mL), neutralized by the addition of excess potassium carbonate, mixed with silica gel (2 g), and concentrated to dryness under reduced pressure. The residue was loaded on a silica gel column (15 g of silica gel) and eluted with dichloromethane / methanol (199:1, 1 L) to provide impure product (72 mg).

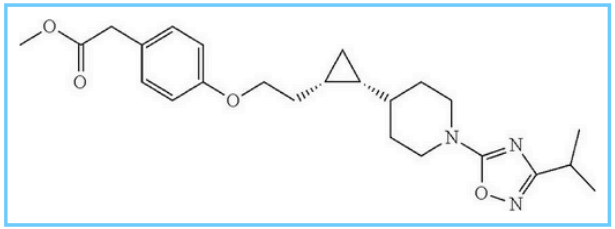
The compound was further purified by prep-HPLC (SunFire C18 OBD, 10 50×150 mm, 118 mL/min, acetonitrile/water 10:90 to 90:10 at 25 min, total run 30 min) to provide the title compound as an off-white solid. 1H NMR (300 MHz, $CDCl_3$) δ 8.89 (s, 1H), 7.60 (d, J=9.0 Hz, 2H), 7.07 (d, J=9.0 Hz, 2H), 4.20-4.05 (m, 4H), 3.12-2.95 (m, 2H), 2.95-2.80 (m, 1H), 2.30-2.10 (m, 1H), 1.95-1.80 (m, 2H), 1.65-1.40 (m, 3H), 1.29 (d, J=6.9 Hz, 6H), 1.15-0.90 (m, 2H), 0.80-0.55 (m, 2H), −0.06 (q, J=4.5 Hz, 1H). MS (ESI) m/z 424 [M+H]+.

GPR119 Human EC_{50}: 11.8 nM

Example 99Preparation of 4-[(1R,2S)-2-{2-[4-(2-azetidin-1-yl-2-oxoethyl)phenoxy]ethyl}cyclopropyl]-1-[3-(1-methylethyl)-1,2,4-oxa



Step A: Methyl 2-(4-(2-((1S,2R)-2-(1-(3-isopropyl-1,2,4-oxadiazol-5-yl)piperidin-4-yl)cyclopropyl)ethoxy)phenyl)acetate

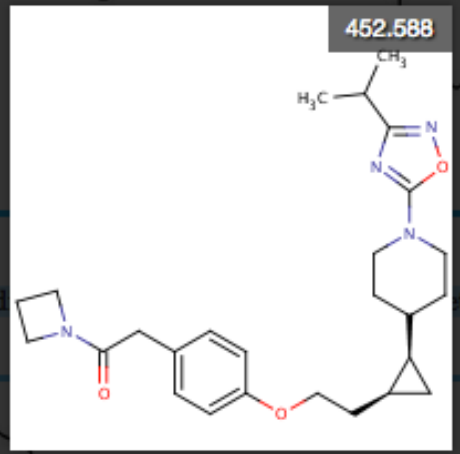


2-((1S,2R)-2-(1-(3-isopropyl-1,2,4-oxadiazol-5-yl)piperidin-4-yl)cyclopropyl)ethanol (1.7 g, 6.08 mmol), methyl 2-(4-hydroxyphenyl and triphenylphosphine (2.4 g, 9.1 mmol) were dissolved in THF (30 ml). The mixture was stirred at RT under N2 for 5 min and disis (1.78 ml, 9.1 mmol) was added. The mixture was stirred at RT overnight. The mixture was diluted with DCM (50 ml), washed with wat The crude material was purified by silica gel column (100 g SNAP, 5-25% EtOAc in hexane) to afford the desired product. LC/MS (m

Step B: 2-(4-(2-((1S,2R)-2-(1-(3-isopropyl-1,2,4-oxadiazol-5-yl)piperidin-4-yl)cyclopropyl)ethoxy)phenyl)acetic acid

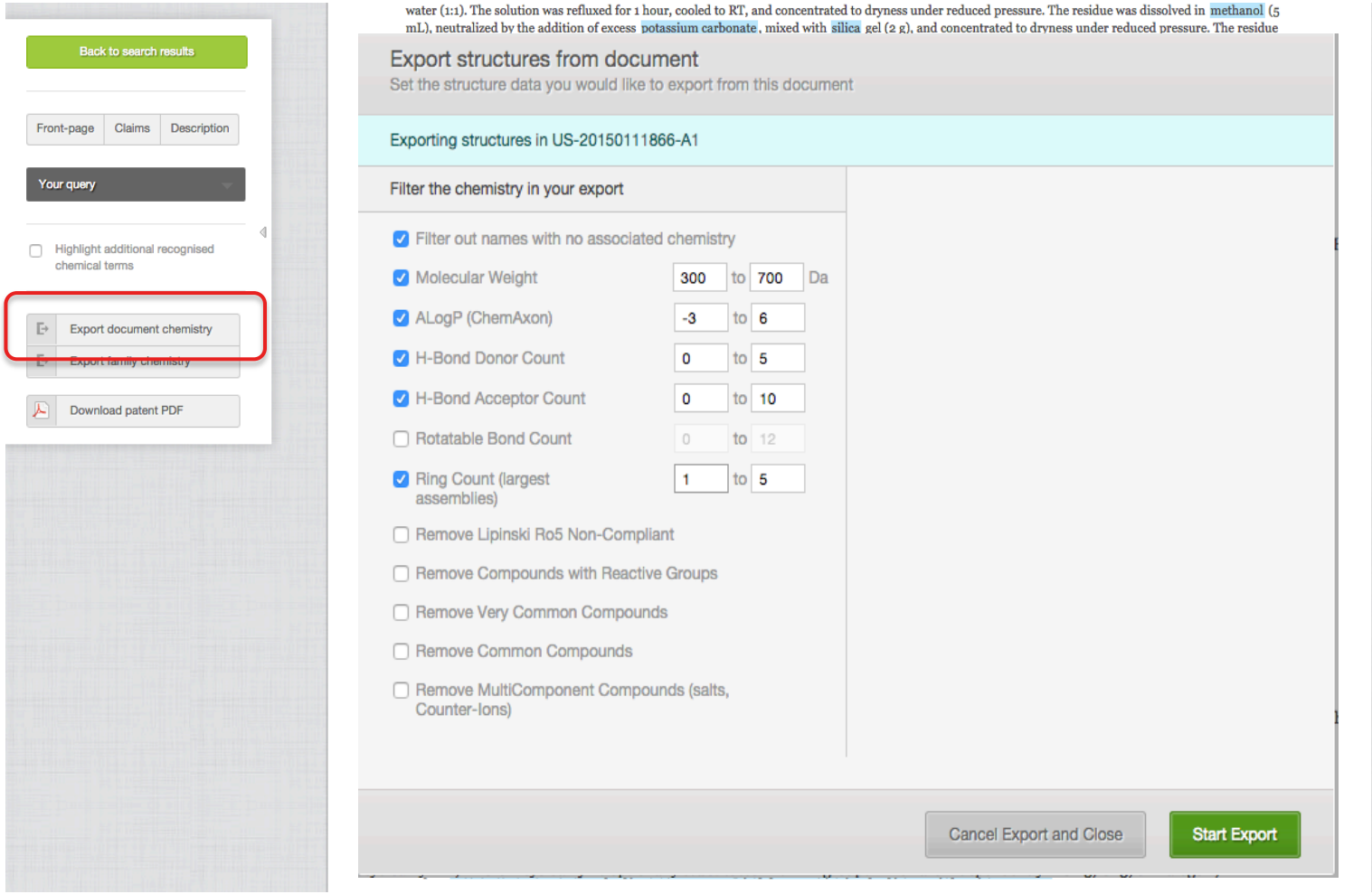
Chemical information

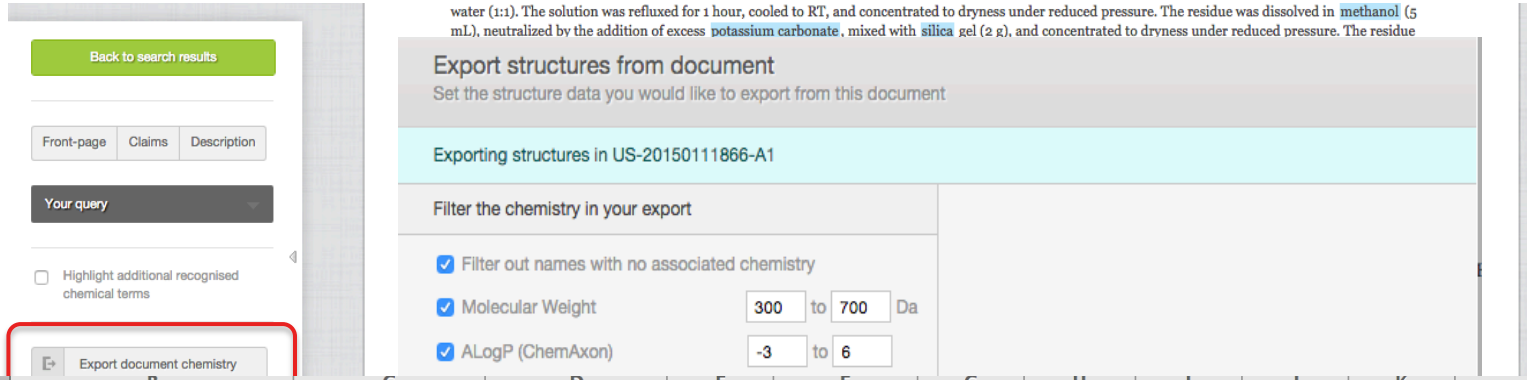
Structures generated for this name:



Name:
1-(azetidin-1-yl)-2-(4-{2-[(1S,2R)-2-{1-[3-(propan-2-yl)-1,2,4-oxadiazol-5-yl]piperidin-4-yl}cyclopropyl]ethoxy}phenyl)ethan-1-one

View


water (1:1). The solution was refluxed for 1 hour, cooled to RT, and concentrated to dryness under reduced pressure. The residue was dissolved in methanol (5 mL), neutralized by the addition of excess potassium carbonate, mixed with silica gel (2 g), and concentrated to dryness under reduced pressure. The residue
Back to search results
Front-page
Claims
Description
Your query
Highlight additional recognised chemical terms
Export document chemistry
Export family chemistry
Download patent PDF
Export structures from document
Set the structure data you would like to export from this document
Exporting structures in US-20150111866-A1
Filter the chemistry in your export
Filter out names with no associated chemistry
Molecular Weight 300 to 700 Da
ALogP (ChemAxon) -3 to 6
H-Bond Donor Count 0 to 5
H-Bond Acceptor Count 0 to 10
Rotatable Bond Count 0 to 12
Ring Count (largest assemblies) 1 to 5
Remove Lipinski Ro5 Non-Compliant
Remove Compounds with Reactive Groups
Remove Very Common Compounds
Remove Common Compounds
Remove MultiComponent Compounds (salts, Counter-Ions)
Cancel Export and Close
Start Export



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

Data contents and growth

- 16.5M unique compounds
- 13.5M annotated patent documents
- 3.5M life-sciences relevant patent documents
- 120M patent documents in total
- ~80K *novel* compounds every month
- ~1M *novel* compounds since EBI took over
- 1–4 days for a published patent to be chemically annotated and searchable in SureChEMBL

UniChem and SureChEMBL

RDF and REST API interfaces

Atlas	PDBe	ChEBI	ChEMBL	SureChEMBL	3rd Party Data
Ligand induced transcript response	Ligand structures from protein complexes	Nomenclature of primary and secondary metabolites. Chemical Ontology	Bioactivity data from literature and depositions	Chemical structures from patent literature	ZINC, PubChem, ThomsonPharma DOTF, IUPHAR, DrugBank, KEGG, NIH NCC, eMolecules, FDA SRS, PharmGKB, Selleck, ...
750	15K	24K	1.5M	**16.5M**	~65M

UniChem – InChI-based chemical resolver (full + relaxed 'lenses') >90M

REST API Interface - https://www.ebi.ac.uk/unichem/

Data access & exports

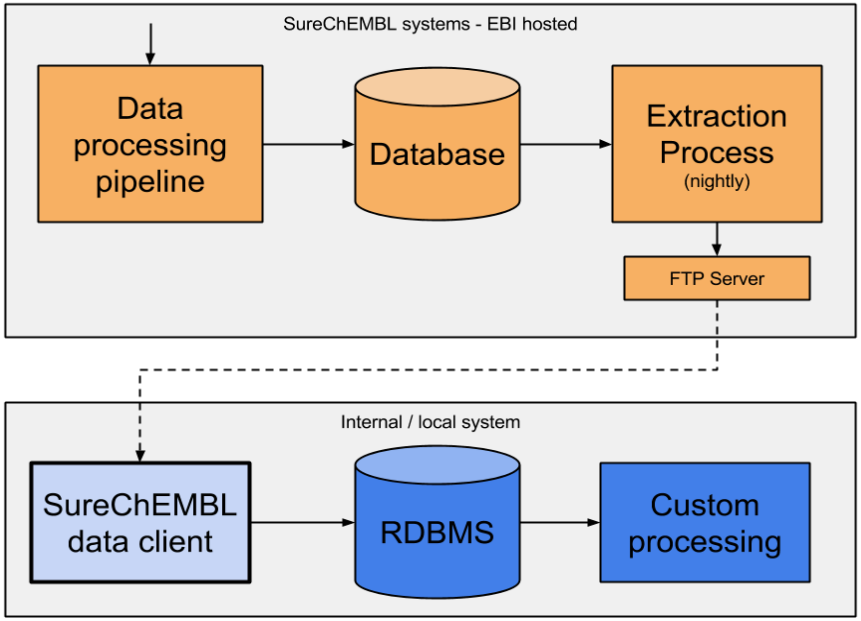
- Full compound repository
  - FTP download, SDF and CSV format
  - Updates quarterly
- Full compound-patent map
  - FTP download, flat file
  - Updates quarterly
- Data feed client
  - Creates a *local* replica database of SureChEMBL
  - Updates daily

Compound-patent map

- Flat file with
  - Compound, global frequency, document, section, section frequency, publication date
  - Back file
    - 18M unique patent-compound pairs
    - 14M unique compound IDs
    - 3.5M EP, JP, WO and US patent docs
    - 1960-2014
  - Quarterly incremental updates
  - Q1 & Q2 2015 are also now available on the FTP

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

Data feed client



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

Use cases with SureChEMBL

- Chemoinformatics
  - Chemistry landscape for a particular biological target/disease
    - Novel chemistry & scaffolds
  - Scaffold/chemical space analysis for a particular patent family claimed chemistry
  - (Negative) novelty checking with UniChem
- Competitive intelligence
  - Reporting
  - Patent alerts

Scaffold and chemical space analysis

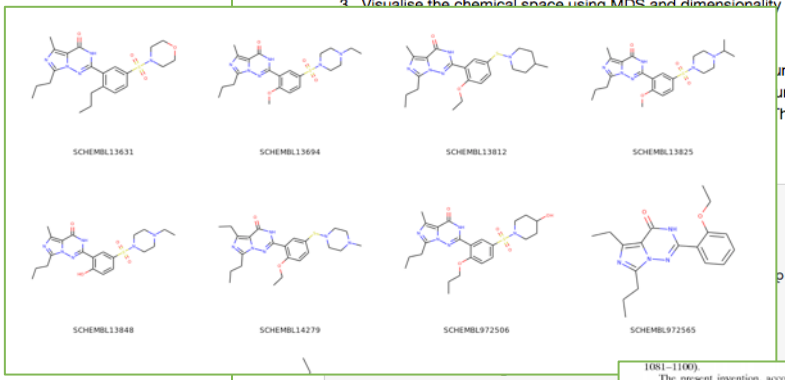
SureChEMBL iPython Notebook Tutorial

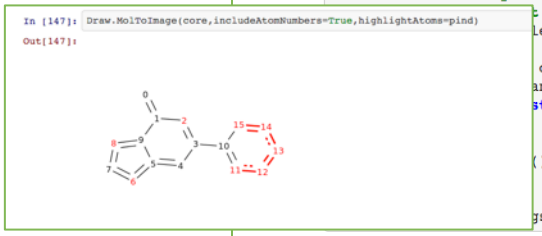
An introduction to patent chemoinformatics using SureChEMBL data and the RDKit toolkit

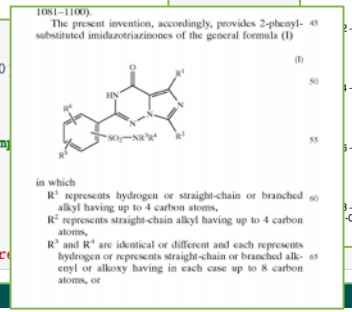
George Papadatos, ChEMBL group, EMBL-EBI

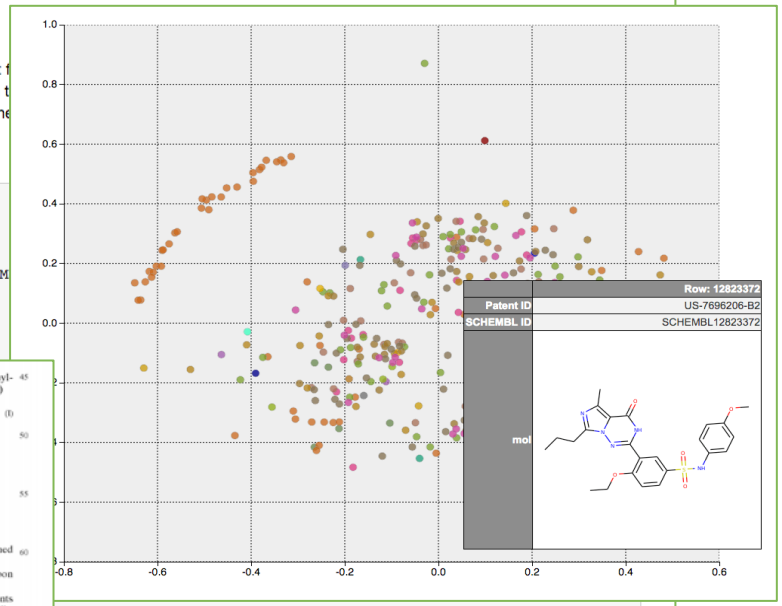
In this tutorial:

1. Read a file that contains all chemistry extracted from the Levitra US patent (US6566360) along with *all* the other members of the same patent family.
2. Filter by different text-mining and chemoinformatics properties to remove noise and enrich the genuinely novel structures claimed in the patent documents.
3. Visualise the chemical space using MDS and dimensionality reduction.









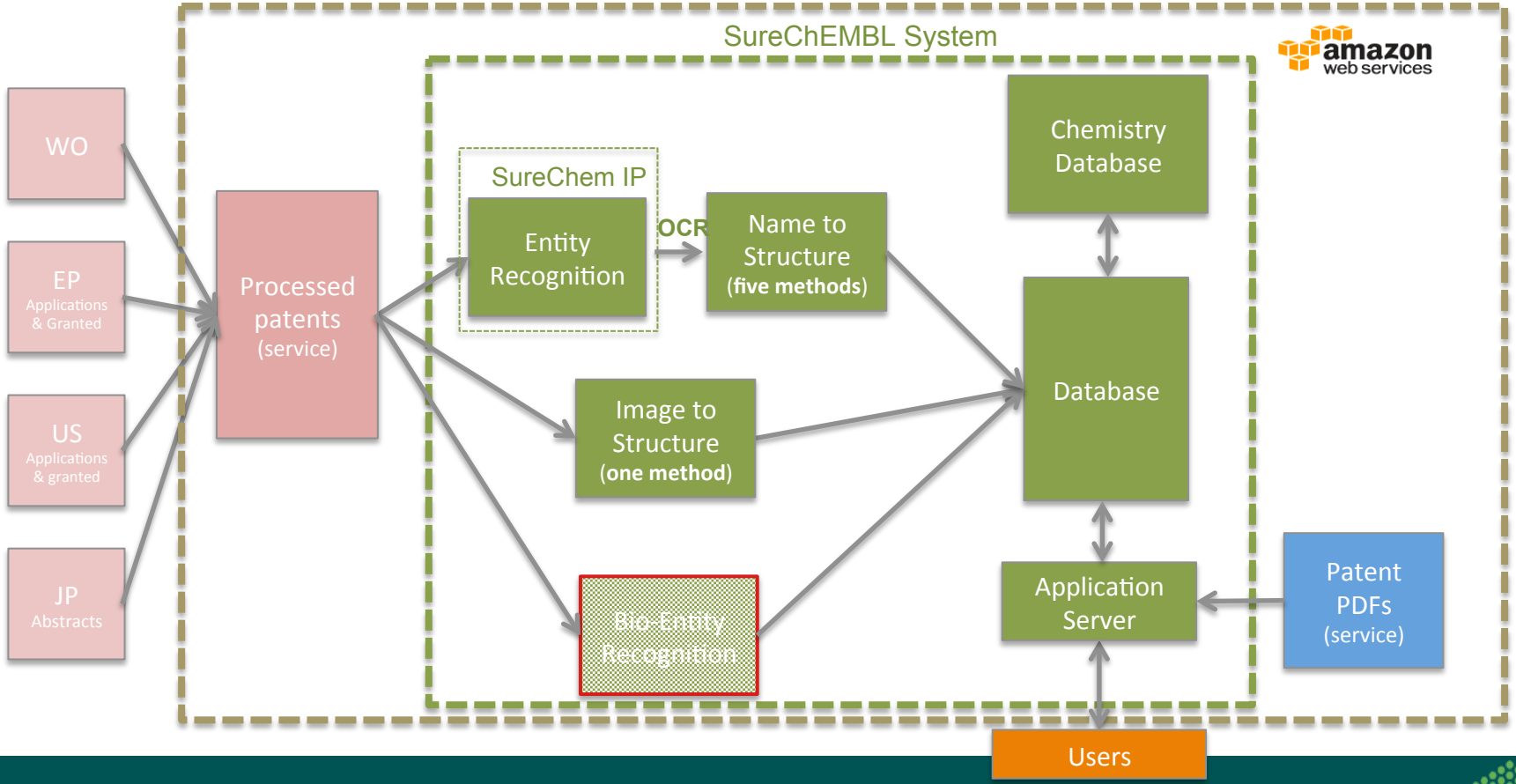
http://nbviewer.ipython.org/github/rdkit/UGM_2014/blob/master/Notebooks/Vardenafil.ipynb

Challenges / Opportunities

- Extraction of Markush structures
- Extraction of bioactivities from tables
  - Mapping to targets
- Increase precision and recall
  - Reduce OCR errors and false negatives

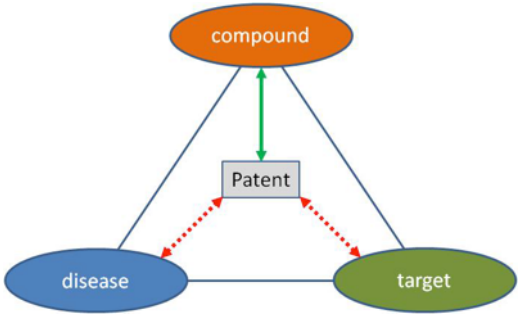
The (near) Future

SureChEMBL data processing v2



Future steps and plans

- Open PHACTS ENSO
  - Biological tagging of genes and indications (SciBite)
  - Development of integrated use-cases
    - Search patents for a gene or indication
    - Relevance score to filter out noise
    - Combine chemistry & biology from patents, literature, pathways, etc.
    - Part of the Open PHACTS API in late Autumn
      - KNIME and PP clients



US-9012636-B2

Back to search results

Front-page | Claims | Description

Your query

- [ ] Highlight additional recognised chemical terms
- [x] Show select Biological annotations Powered by SciBite.com's Termite Engine

▶ Anatomy
▶ Biological process
▶ Endogenous cell
▼ Gene
HCRT (19)
HCRTR2 (3)
HCRTR1 (2)
CBFA2T3 (1)
ALB (1)
▶ Gene Ontology
▼ Indication
Migraine (6)
Emesis (4)
Hypothalamic Diseases (3)
Hyperalgesia (3)
Hypothyroidism (3)
Depressive Neurosis (3)
Diabetes (3)
Ulcers (2)
Cachexia (2)
Renal Disease (2)

Export document chemistry
Export family chemistry

Download patent PDF

BACKGROUND OF THE INVENTION

The orexins (hypocretins) comprise two neuropeptides produced in the hypothalamus: the orexin A (OX-A) (a 33 amino acid peptide) and the orexin B (OX-B) (a 28 amino acid peptide) (Sakurai T. et al., Cell, 1998, 92, 573-585). Orexins are found to stimulate food consumption in rats suggesting a physiological role for these peptides as mediators in the central feedback mechanism that regulates feeding behavior (Sakurai T. et al., Cell, 1998, 92, 573-585). Orexins regulate states of sleep and wakefulness opening potentially novel therapeutic approaches for narcoleptic or insomniac patients (Chemelli R. M. et al., Cell, 1999, 98, 437-451). Orexins have also been indicated as playing a role in arousal, reward, learning and memory (Harris, et al., Trends Neurosci., 2006, 29 (10), 571-577). Two orexin receptors have been cloned and characterized in mammals. They belong to the super family of G-protein coupled receptors (Sakurai T. et al., Cell, 1998, 92, 573-585): the orexin-1 receptor (OX or OX1R) is selective for OX-A and the orexin-2 receptor (OX2 or OX2R) is capable to bind OX-A as well as OX-B. The physiological actions in which orexins are presumed to participate are thought to be expressed via one or both of OX 1 receptor and OX 2 receptor as the two subtypes of orexin receptors.

Orexin receptors are found in the mammalian brain and may have numerous implications in pathologies such as depression; anxiety; addictions; obsessive compulsive disorder; affective neurosis; depressive neurosis; anxiety neurosis; dysthymic disorder; behaviour disorder; mood disorder; sexual dysfunction; psychosexual dysfunction; sex disorder; schizophrenia; manic depression; delirium; dementia; severe mental retardation and dyskinesias such as Huntington's disease and Tourette syndrome; eating disorders such as anorexia, bulimia, cachexia, and obesity; addictive feeding behaviors; binge/purge feeding behaviors; cardiovascular diseases; diabetes; appetite/taste disorders; emesis, vomiting, nausea; asthma; cancer; Parkinson's disease; Cushing's syndrome/disease; basophile adenoma; prolactinoma; hyperprolactinemia; hypophysis tumour/adenoma; hypothalamic diseases; inflammatory bowel disease; gastric diskinesia; gastric ulcers; Froehlich's syndrome; adrenohypophysis disease; hypophysis disease; adrenohypophysis hypofunction; adrenohypophysis hyperfunction; hypothalamic hypogonadism; Kallman's syndrome (anosmia, hyposmia); functional or psychogenic amenorrhea; hypopituitarism; hypothalamic hypothyroidism; hypothalamic-adrenal dysfunction; idiopathic hyperprolactinemia; hypothalamic disorders of growth [illegible] hic growth deficiency; dwarfism; gigantism; acromegaly; disturbed biological and circadian rhythms; sleep disturbances associated with diseases such as neurological disorders, neuropathic pain and restless leg syndrome; heart and lung diseases, acute and congestive heart failure; hypotension; hypertension; urinary retention; osteoporosis; angina pectoris; myocardinal infarction; ischemic or haemorrhagic stroke; subarachnoid haemorrhage; ulcers; allergies; benign prostatic hypertrophy; chronic renal failure; renal disease; impaired glucose tolerance; migraine; hyperalgesia; pain; enhanced or exaggerated sensitivity to pain such as hyperalgesia, causalgia, and allodynia; acute pain; burn pain; atypical facial pain; neuropathic pain; back pain; complex regional pain syndrome I and II; arthritic pain; sports injury pain; pain related to infection e.g. HIV, post-chemotherapy pain; post-stroke pain; post-operative pain; neuralgia; emesis, nausea, vomiting; conditions associated with visceral pain such as irritable bowel syndrome, and angina; migraine; urinary bladder incontinence e.g. urge incontinence; tolerance to narcotics or withdrawal from narcotics; sleep disorders; sleep apnea; narcolepsy; insomnia; parasomnia; jet lag syndrome; and neurodegenerative disorders including nosological entities such as disinhibition-dementia-parkinsonism-amyotrophy complex; pallido-ponto-nigral degeneration; epilepsy; seizure disorders and other diseases related to general orexin system dysfunction.

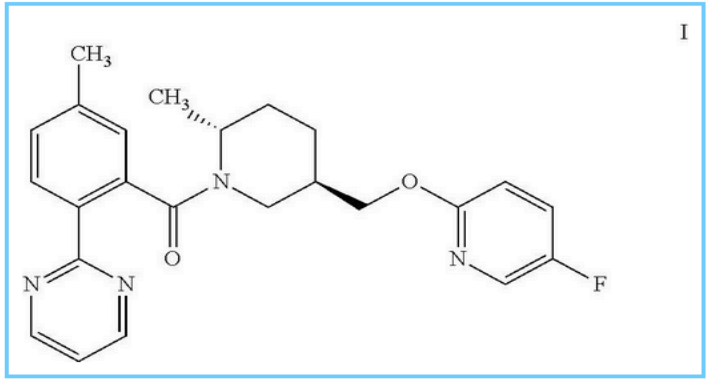
Biological annotation found!

SUMMARY OF THE INVENTION

The present invention is directed to processes for preparing a pyridyl piperidine compound which is an antagonist of orexin receptors, and which is useful in the treatment or prevention of neurological and psychiatric disorders and diseases in which orexin receptors are involved.

DETAILED DESCRIPTION OF THE INVENTION

The present invention is directed to a process for preparing a compound of the formula I:



Acknowledgements

- ChEMBL team
  - John Overington
  - Anna Gaulton
  - Jon Chambers
  - Mark Davies
- Digital Science
  - Nicko Goncharoff
  - James Siddle
  - Richard Koks
- SciBite
  - Lee Harland

Open PHACTS consortium

http://www.openphacts.org

Funding

The research leading to these results has received support from the Innovative Medicines Initiative Joint Undertaking under grant agreement n° 115191, resources of which are composed of financial contributions from the EU's Seventh Framework Programme (FP7/2007-2013) and EFPIA companies' in-kind contribution. (Open PHACTS)

Wellcome Trust Strategic Award for Chemogenomics, WT086151/Z/08/Z

European Molecular Biology Laboratory

European Commission FP7 Capacities Specific Programme, grant agreement no. 284209 (BioMedBridges)

wellcome trust EMBL imi Innovative Medicines Initiative Open PHACTS SEVENTH FRAMEWORK PROGRAMME

Technology partners

ChemAxon

SciBite

IFI CLAIMS® Patent Services
a division of Fairview Research

PerkinElmer
For the Better

amazon web services

Twin Dolphin Software

NextMove Software

OPSIN

ACD/Labs

keymodule

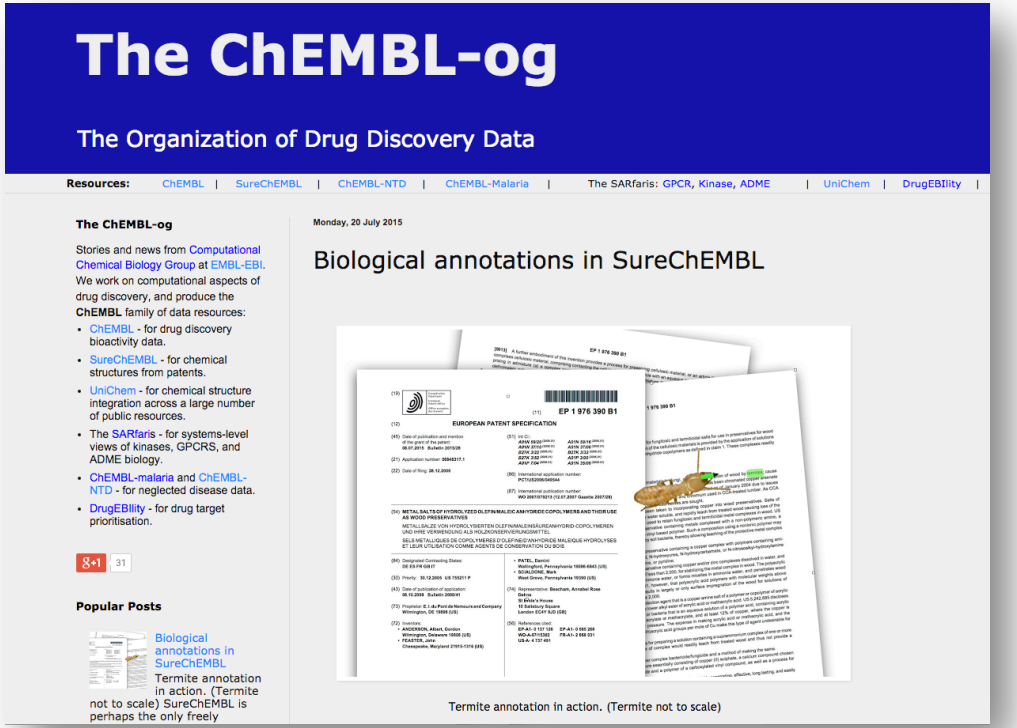
OpenEye Scientific Software
Shaping the Industry

BIOVIA

Open for Innovation KNIME

RDKit
Open-Source Cheminformatics and Machine Learning

ChEMBL-og



Support helpdesk:
surechembl-help@ebi.ac.uk

Webinar:
http://www.ebi.ac.uk/training/online/course/surechembl-accessing-chemical-patent-data-webinar

SureChEMBL: An open patent chemistry resource

George Papadatos, PhD

ChEMBL Group, EMBL-EBI

georgep@ebi.ac.uk

EMBL-EBI

Bioactivity data extraction?

Compounds

Target/Assay

Bioactivity

TABLE 1-continued

Ex. No.	Compound Name	Method	Observed MS	BACE 1 FRET assay (uM	HEK cell assay (uM)
39	3-(2-amino-6-o-tolylquinolin-3-yl)-N-(1-methylpiperidin-3-yl)propanamide		403	++	++
40	3-(2-amino-6-o-tolylquinolin-3-yl)-N-((tetrahydro-2H-pyran-2-yl)methyl)propanamide		404	++++	++
3	3-(2-amino-6-(2-chlorophenyl)quinolin-3-yl)-N-(cyclohexylmethyl)propanamide		422	++++	++
41	3-(2-amino-6-(2-cyanophenyl)quinolin-3-yl)-N-(cyclohexylmethyl)propanamide		413	++++	++
42	3-(2-amino-6-(2-fluorophenyl)quinolin-3-yl)-N-(cyclohexylmethyl)propanamide		406	++++	+
43	3-(2-amino-6-(3-fluorophenyl)quinolin-3-yl)-N-(cyclohexylmethyl)propanamide		406	+++	++
44	3-(2-amino-6-(2-methoxyphenyl)quinolin-3-yl)-N-(cyclohexylmethyl)propanamide		418	++++	++
45	3-(2-amino-6-phenylquinolin-3-yl)-N-(cyclohexylmethyl)propanamide		388	+++	+
46	3-(2-amino-6-(4-methylpyridin-3-yl)quinolin-3-yl)-N-(cyclohexylmethyl)propanamide		403	++	+
47	3-(2-amino-6-o-tolylquinolin-3-yl)-N-((R)-1-cyclohexylethyl)-2-methylpropanamide		430	++++	++
48	N-(3-(2-amino-6-o-tolylquinolin-3-yl)propyl)cyclohexanecarboxamide				
49	3-(2-amino-6-(2-ethylphenyl)quinolin-3-yl)-N-(cyclohexylmethyl)propanamide				
50	2-((2-amino-6-o-tolylquinolin-3-yl)methyl)-N-(2-fluorophenyl)butanamide				
51	3-(2-amino-6-o-tolylquinolin-3-yl)-2-methyl-N-phenylpropanamide				
52	3-(2-amino-6-o-tolylquinolin-3-yl)-N-benzyl-2-methylpropanamide				
53	3-(2-amino-6-o-tolylquinolin-3-yl)-N-(2-fluoro-4-methylphenyl)propanamide				
54	3-(2-amino-6-o-tolylquinolin-3-yl)-N-((1-methyl-1H-pyrazol-3-yl)methyl)propanamide				

[0474] Compounds of the present invention that contain the aforementioned isotopes and/or other isotopes of other atoms are within the scope of this invention. Certain isotopically-labelled compounds of the present invention, for example those into which radioactive isotopes such as ^{3}H and ^{14}C are incorporated, are useful in drug and/or substrate tissue distribution assays. Tritiated, i.e., ^{3}H, and carbon-14, i.e., ^{14}C, isotopes are particularly preferred for their ease of preparation and detection. Further, substitution with heavier isotopes such as deuterium, i.e., ^{2}H, can afford certain therapeutic advantages resulting from greater metabolic stability, for example increased in vivo half-life or reduced dosage requirements and, hence, may be preferred in some circumstances. Isotopically labelled compounds of this invention can generally be prepared by substituting a readily available isotopically labelled reagent for a non-isotopically labelled reagent.

Biological Evaluation

[0475] The compounds of the invention may be modified by appending appropriate functionalities to enhance selective biological properties. Surprisingly, the compounds of the present invention exhibit improved pharmacokinetics and

[0478] Of the compounds tested, the in-vitro BACE FRET enzyme data for each of Examples 1-171, where available at the time of filing this applicacation, is provided in Table 1. Data key for the in-vitro BACE FRET assay is as follows:

[0479] "+" means the compound example has an IC_{50} value of >5.0 uM;

[0480] "++" means the compound example has an IC_{50} value in the range from 1.0 uM-5.0 uM;

[0481] "+++" means the compound example has an IC_{50} value in the range from 500 nM-1.0 uM;

[0482] "++++" means the compound example has an IC_{50} value in the range less than 500 nM.

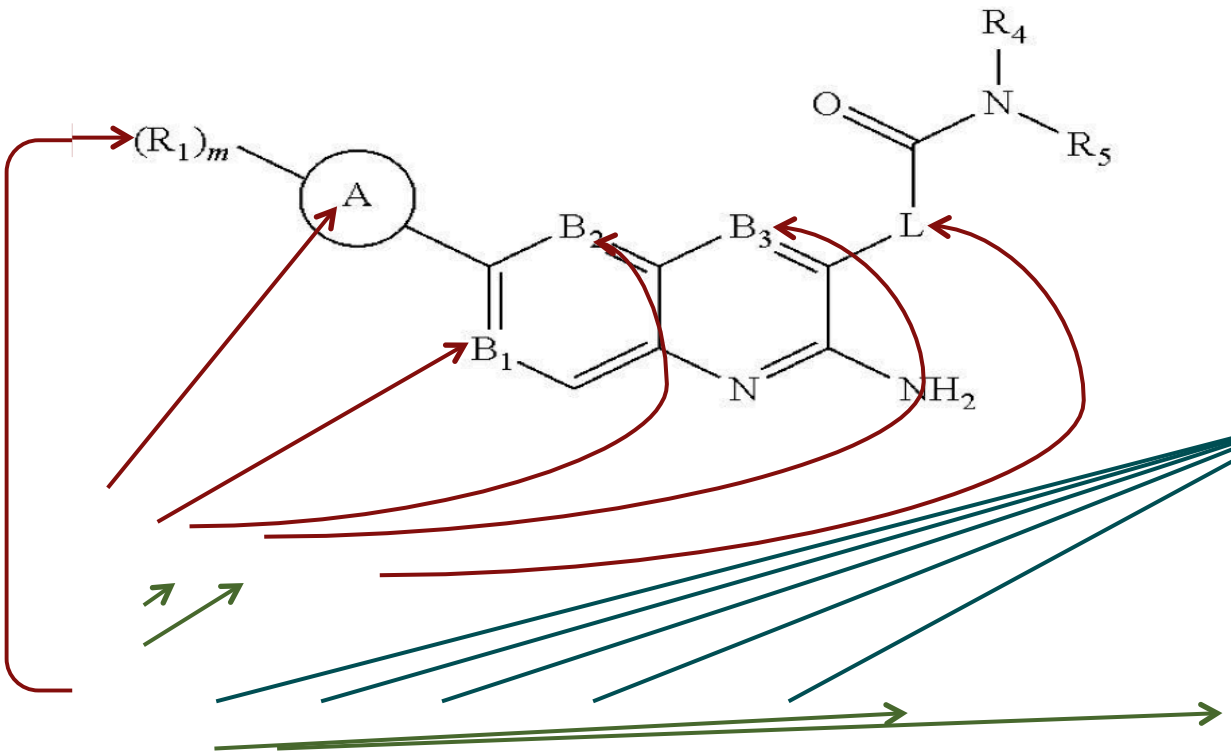
[0483] A majority of the exemplary compounds tested had IC_{50}'s for the enzyme BACE of less than 50 nM. For instance, example numbers 81, 84, 87, 90-93, 96, 98, 103, 106, 126, 128-130, 135, 136, 138, 12, 145-151, 153, 9, 156, 10, 158-164, 11 and 169-171 each exhibited an IC_{50} value of less than 50 nM in the FRET BACE enzyme assay.

[0484] In Vitro BACE Cell-Based Assay:

[0485] The cell-based assay measures inhibition or reduction of Aβ40 in conditioned medium of test compound treated cells expressing amyloid precursor protein.

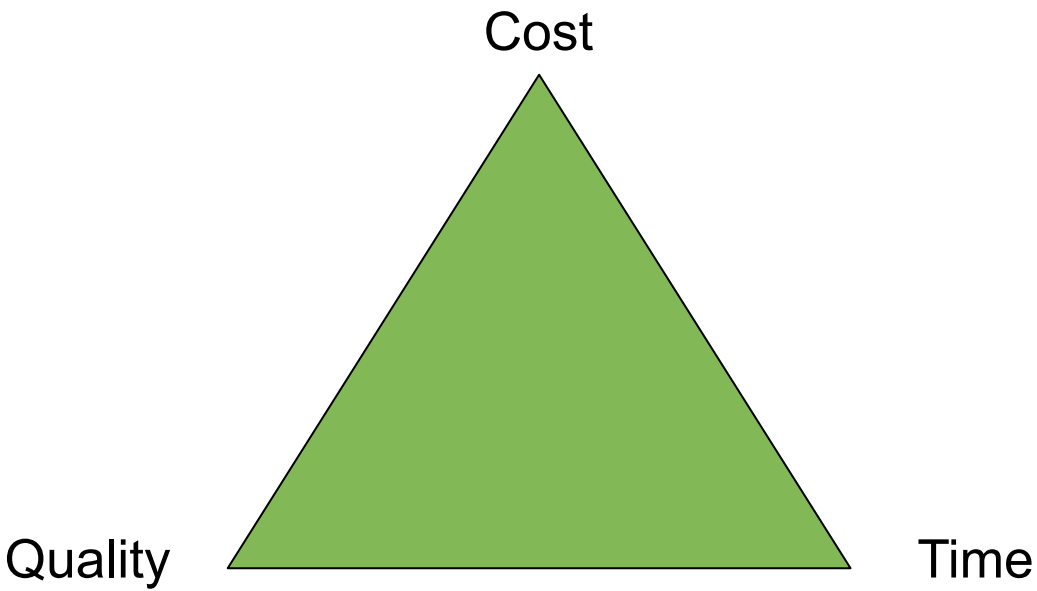
[0486] Cells stably expressing Amyloid Precursor Protein

Markush structure extraction?



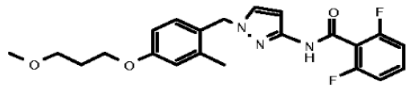
-alkyl
-aryl
-heteroaryl
-heterocyclyl
-cycloalkyl
….

Can we have everything?



Common sources of errors

- Small, poor quality images

142	

- OCR errors in names (OCR done by IFI). There is an OCR correction

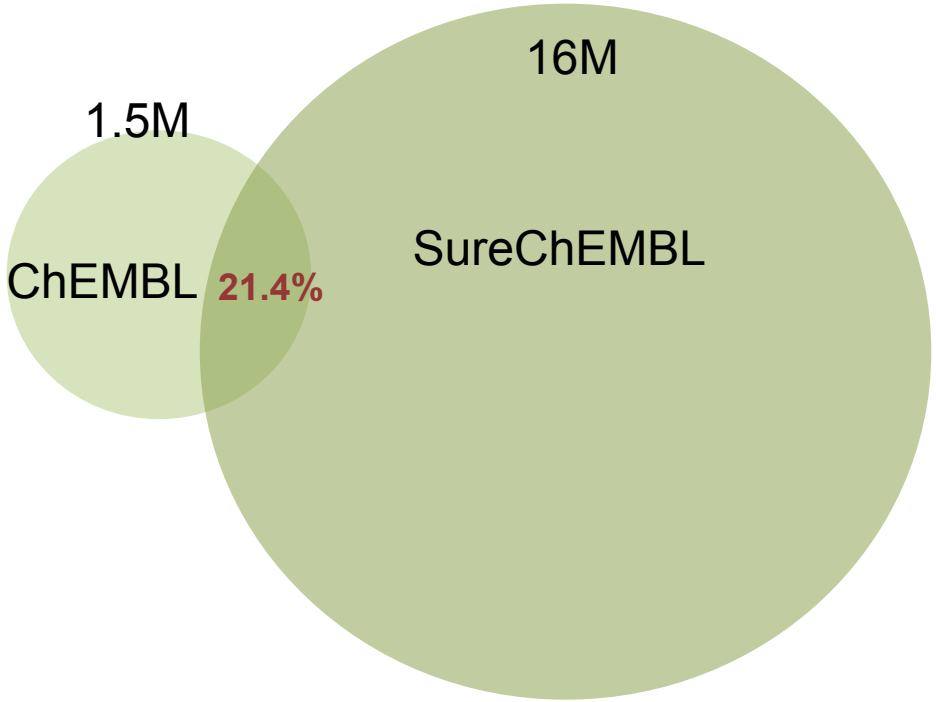
Example 77: 2,6-Difluoro-*N*-{1-[(4-iodo-2-methylphenyl)methyl]-1*H*-pyrazol-3-yl}benzamide

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

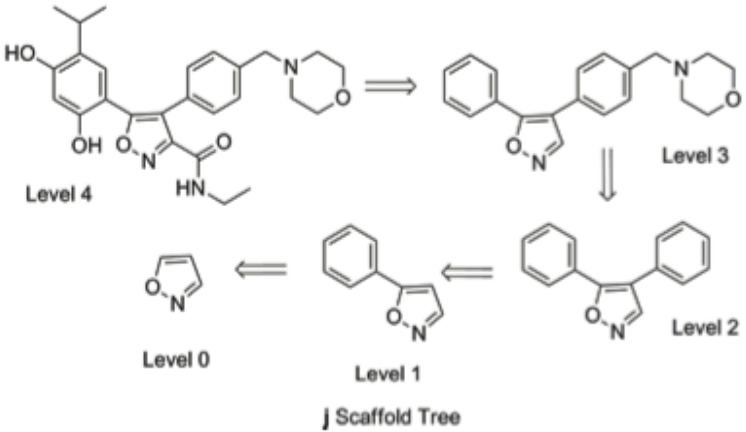
- Reliability better for US patents due to inclusion of mol files

ChEMBL-SureChEMBL compound overlap

- *Connectivity* match on single components - UniChem



Too granular? Try scaffolds instead



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

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

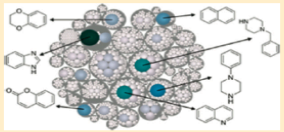
Scaffold Diversity of Exemplified Medicinal Chemistry Space

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.

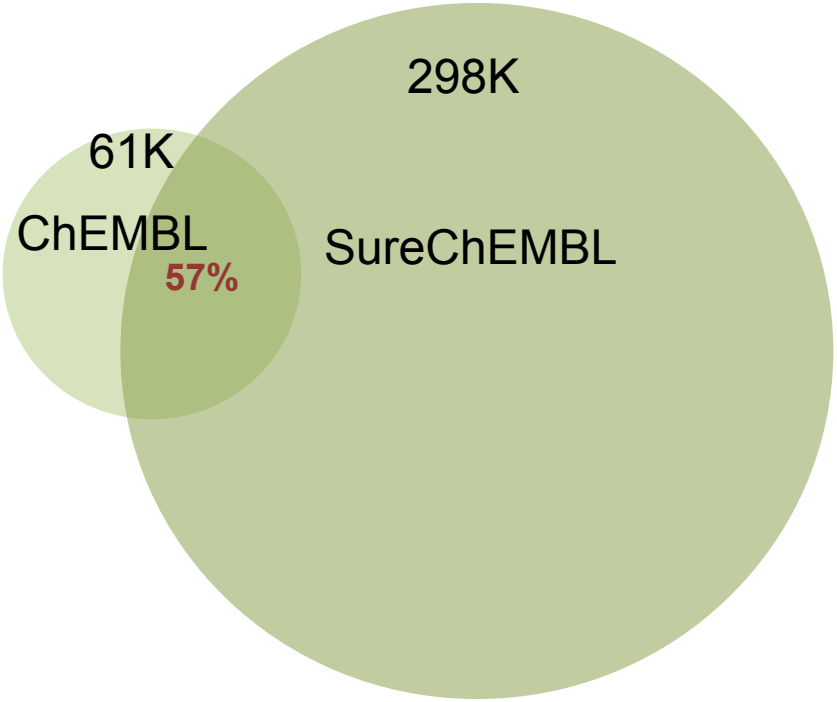
Supporting Information

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