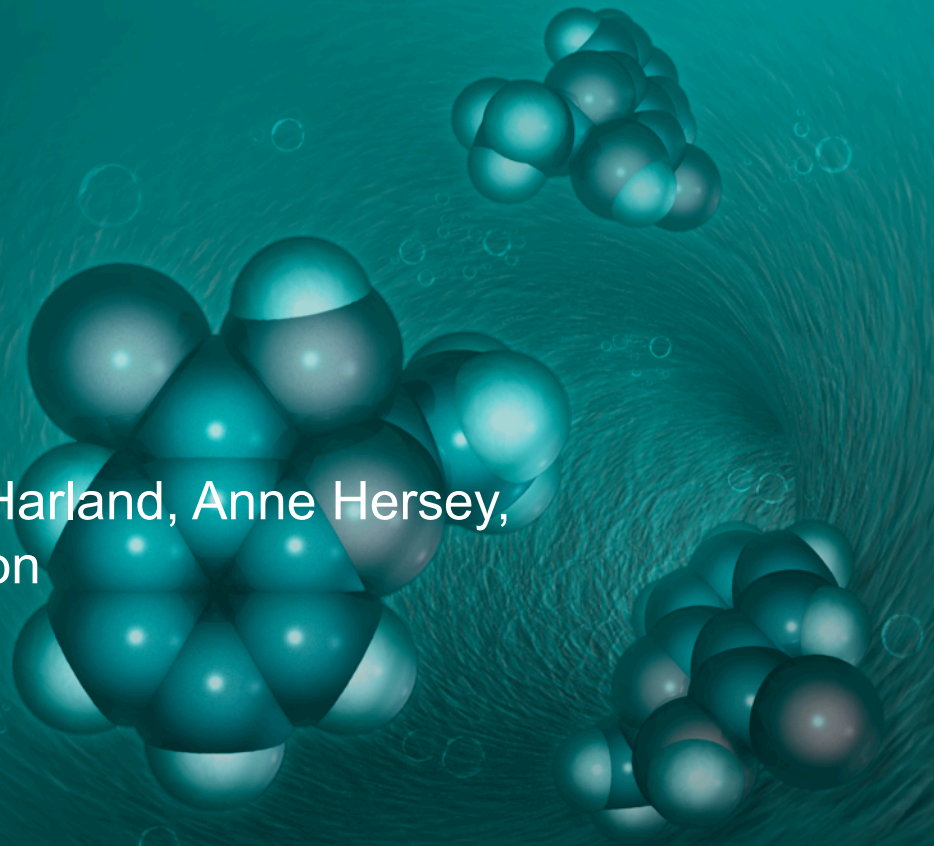
SureChEMBL – Open Patent Data



Anna Gaulton
Jon Chambers, Mark Davies, Lee Harland, Anne Hersey,
George Papadatos, John Overington

EMBL-EBI

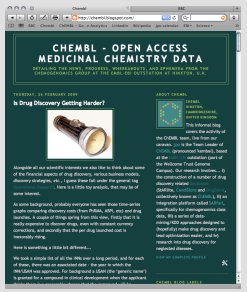
ChEMBL

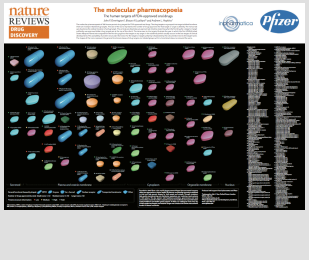
Nucleic Acids Research 2014, 42:D1 (doi:10.1093/nar/gkt1031)

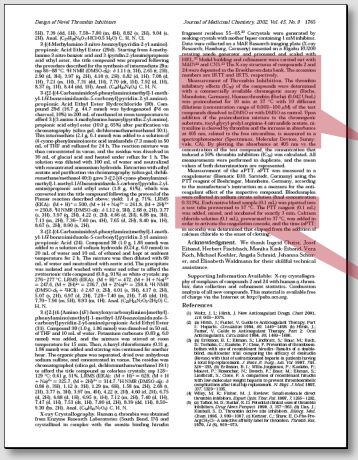
1. Scientific facts

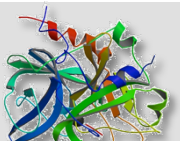
Design of Selective Thrombin Inhibitors Based on the (R)-Phe-Pro-Arg Sequence

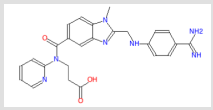
3. Insight, tools and resources for translational drug discovery

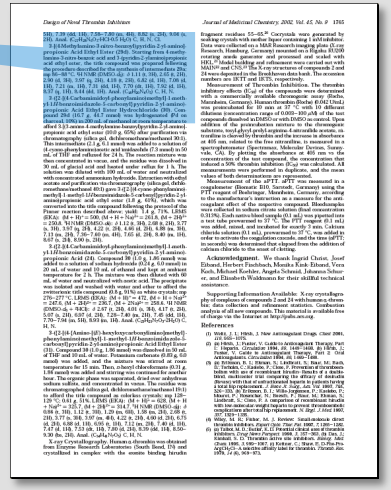












>Thrombin

MAHVRGLQLPGCLALAALCSLVHSQHVFLAPQQARSLLQRVRRANTFLEEVRKGNLE
RECVEETCSYEEAFEALESSTATDVFWAKYTACETARTPRDKLAACLEGNCAEGLGT
NYRGHVNITRSGIECQLWRSRYPHKPEINSTTHPGADLQENFCRNPDSSTTGPWCYT
TDPTVRRQECSIPVCGQDQVTVAMTPRSEGSSVNLSPPLEQCVPDRGQQYQGRLAVT
THGLPCLAWASAQAKALSKHQDFNSAVQLVENFCRNPDGDEEGVWCYVAGKPGDFG
CDLNYCEEAVEEETGDGLDEDSDRAIEGRTATSEYQTFFNPRTFGSGEADCGLRPLF
EKKSLEDKTERELLESYIDGRIVEGSDAEIGMSPWQVMLFRKSPQELLCGASLISDR
WVLTAAHCLLYPPWDKNFTENDLLVRIGKHSRTRYERNIEKISMLEKIYIHPRYNWR
ENLDRDIALMKLKKPVAFSDYIHPVCLPDRETAASLLQAGYKGRVTGWGNLKETWT
NVGKGQPSVLQVVNLPIVERPVCKDSTRIRITDNMFCAGYKPDEGKRGDACEGDSG
PFVMKSPFNNRWYQMGIVSWGEGCDRDGKYGFYTHVFRLKKWIQKVIDQFGE

Compound

K_i = 4.5nM

Assay/Target

Bioactivity data

APTT = 11 min.

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

Patent Data

- Do we include patent data in the ChEMBL database?
  - One of most common questions asked about ChEMBL during training and outreach
  - We do provide limited cross-references (via UniChem), but not the underlying chemical data
- Why is this important to Drug Discovery researchers?
  - Patent literature 2-3 years ahead of published literature
  - Prior art and freedom to operate, competitor intelligence
  - Lots more data – *but high cost to extract* + *lots of noisy data*

SureChEMBL

- December 2013 EMBL-EBI acquired SureChem – a leading ‘chemistry patent mining’ product from Digital Science, Macmillan Group
- SureChem provides a live (updated daily) view of chemical patent space, with ~50K new documents added per month
- EMBL-EBI will provide an ongoing, free and open resource to entire community
- Rebranded SureChEMBL

SureChEMBL Data Coverage

IFI CLAIMS® Patent Services
a division of Fairview Research

	Data	Description & Languages	Years
EP applications	Bib. data Full text	DocDB + Original Original (EN, DE, FR)	from 1978
EP granted	Bib. data Full text	DocDB + Original Original (EN, DE, FR)	From 1980
WO applications	Bib. data Full text	DocDB + Original Original (EN, DE, FR, ES, RU)	From 1978 From 1978
US applications	Bib. data Full text	DocDB + Original Original (EN)	From 2001 From 2001
US granted	Bib. data Full text	DocDB + Original Original (EN)	From 1920 From 1976
JP applications	Bib. data Full text	DocDB PAJ - English abstracts/titles	From 1973 From 1976
JP granted	Bib. data	DocDB	From 1994
90+ countries	Bib. data	DocDB	From 1920

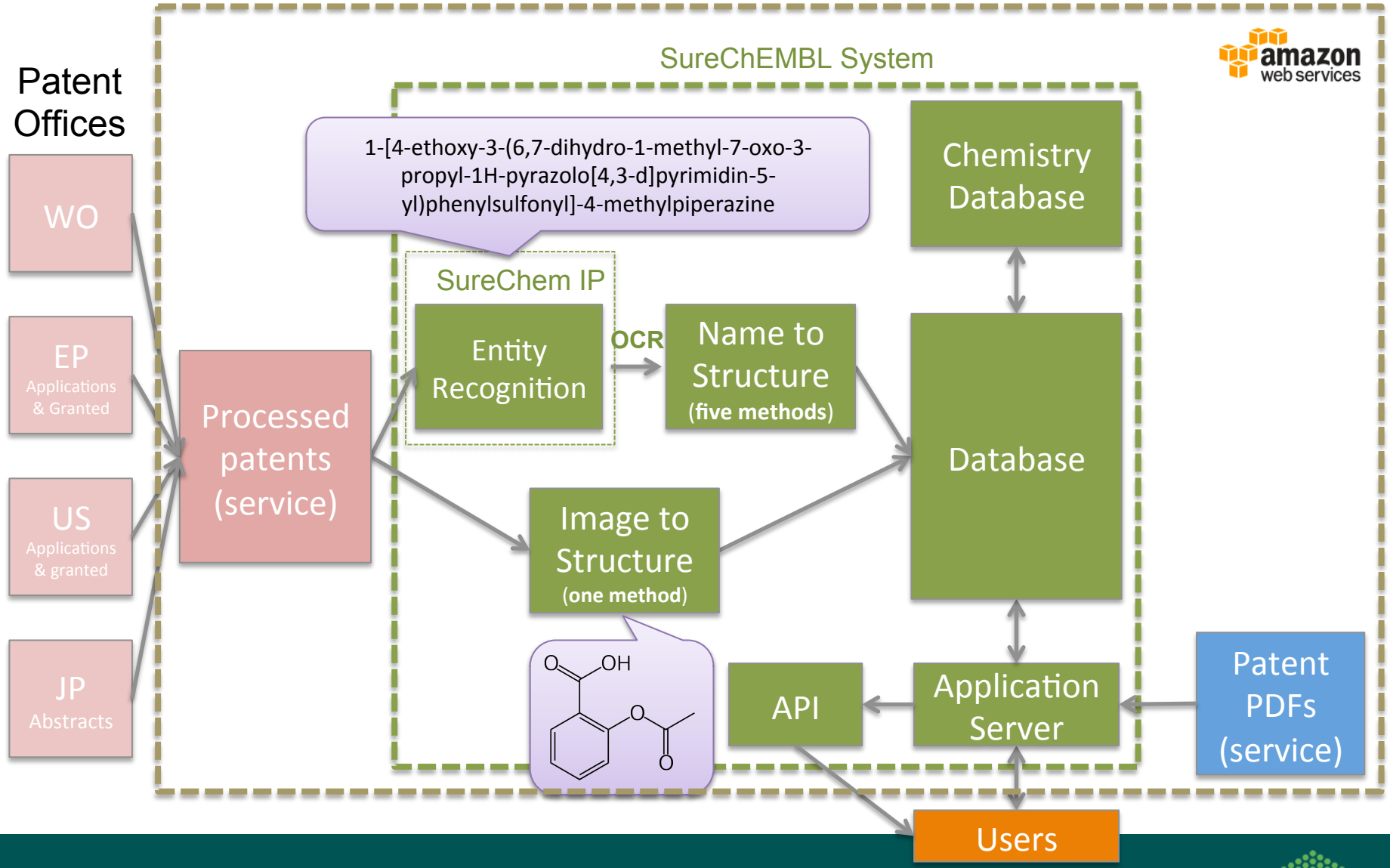
All patents from above data sources are searchable via SureChEMBL

EMBL-EBI

SureChEMBL Chemistry Data Coverage

- Exemplified structures from patent title, description, abstract and claims
- Structures from text 1976 onwards
  - SureChem chemical entity recognition system
  - ACD/Labs, ChemAxon, OpenEye, Opsin + PerkinElmer name-structure conversion
- Structures from images 2007 onwards
  - CLiDE image-structure conversion
- USPTO have provided ‘Complex Work Units’ since 2001
  - CWU file types include MOL and CDX
  - CWUs processed as part of pipeline (from 2007 onwards)

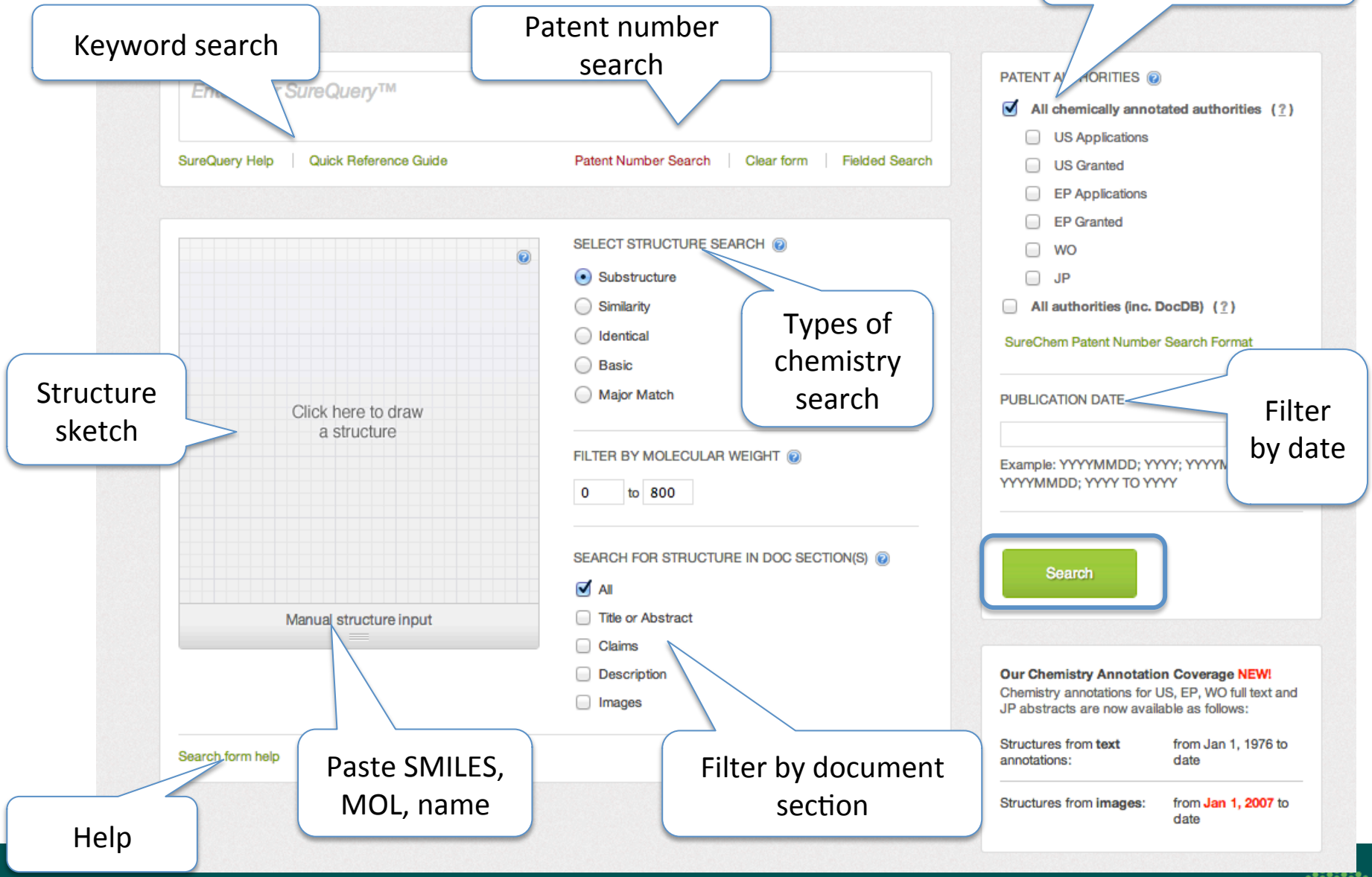
SureChEMBL Data Processing



SureChEMBL System

- Searching capabilities
  - Free text keywords and Lucene fields
  - Patent IDs & bibliographic information
  - Patent authority & date
  - Chemical structure
- Retrieving capabilities
  - Retrieve chemistry (with additional filters)
  - Retrieve patent family information
  - Retrieve *annotated* full patent text
- Accessible via Web Interface and API

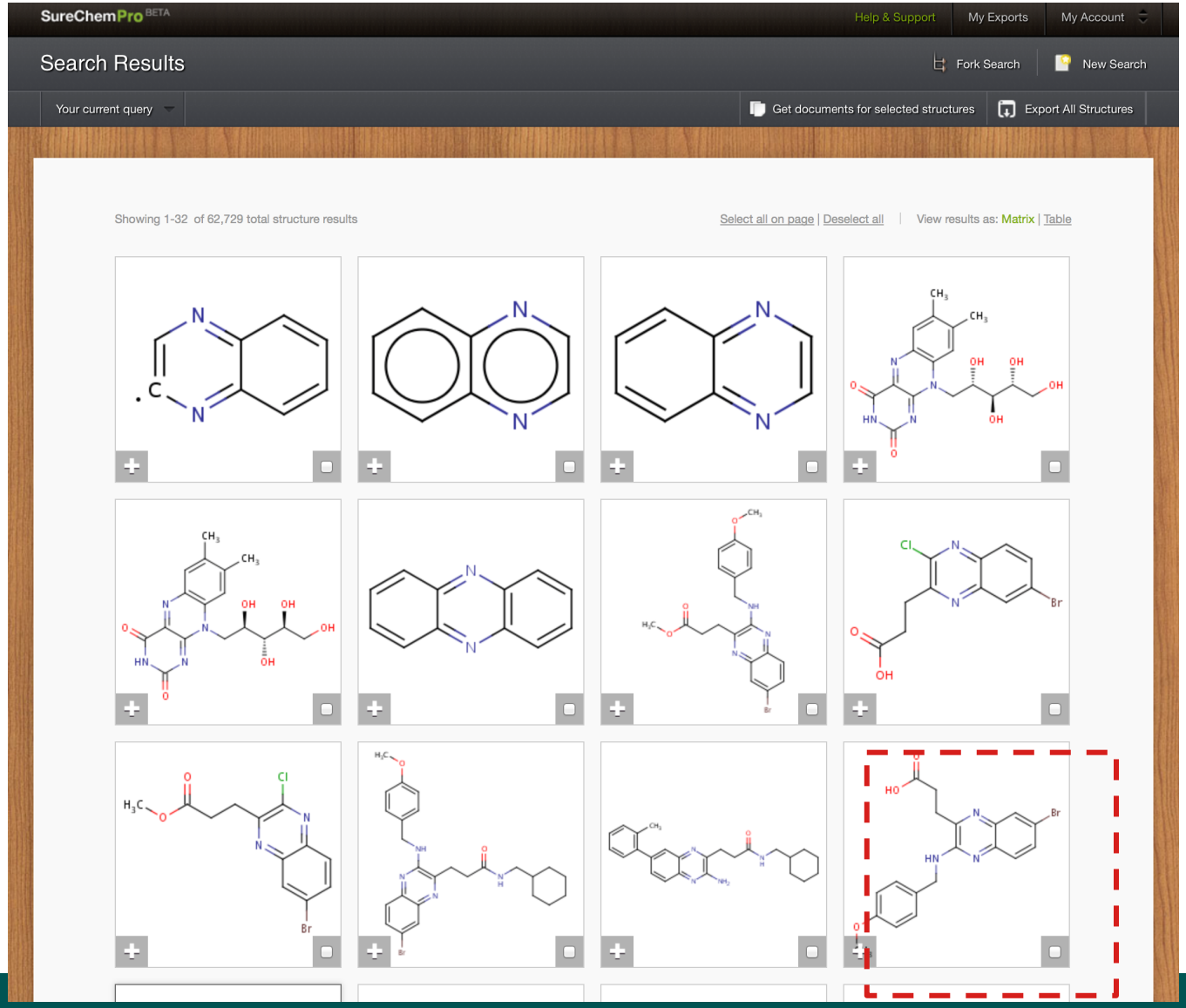
SureChEMBL System



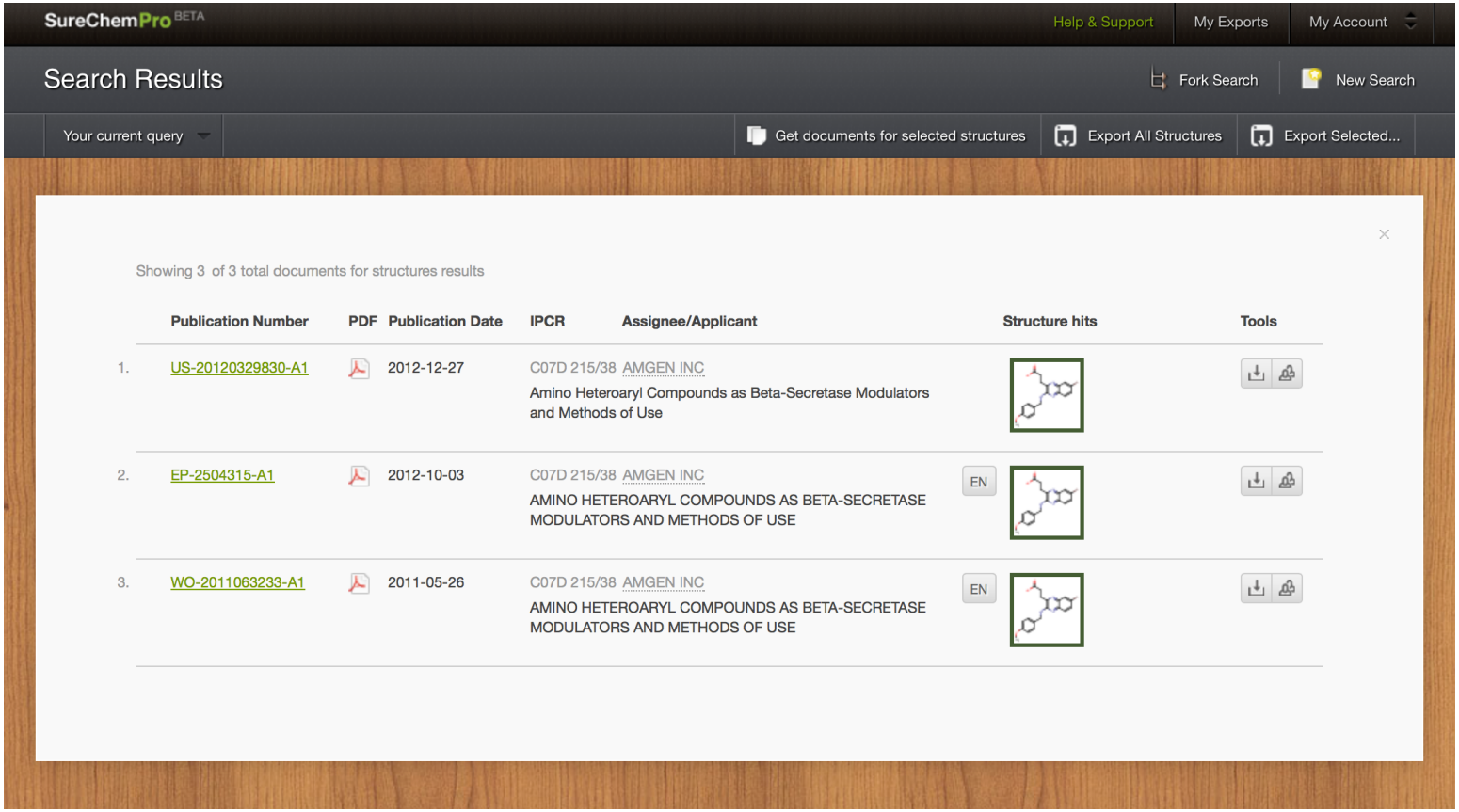
https://www.surechembl.org/

EMBL-EBI

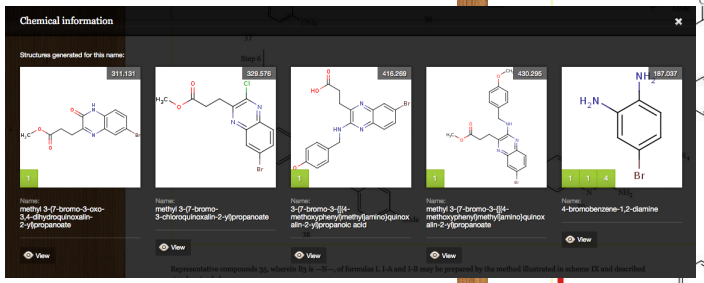
SureChEMBL System

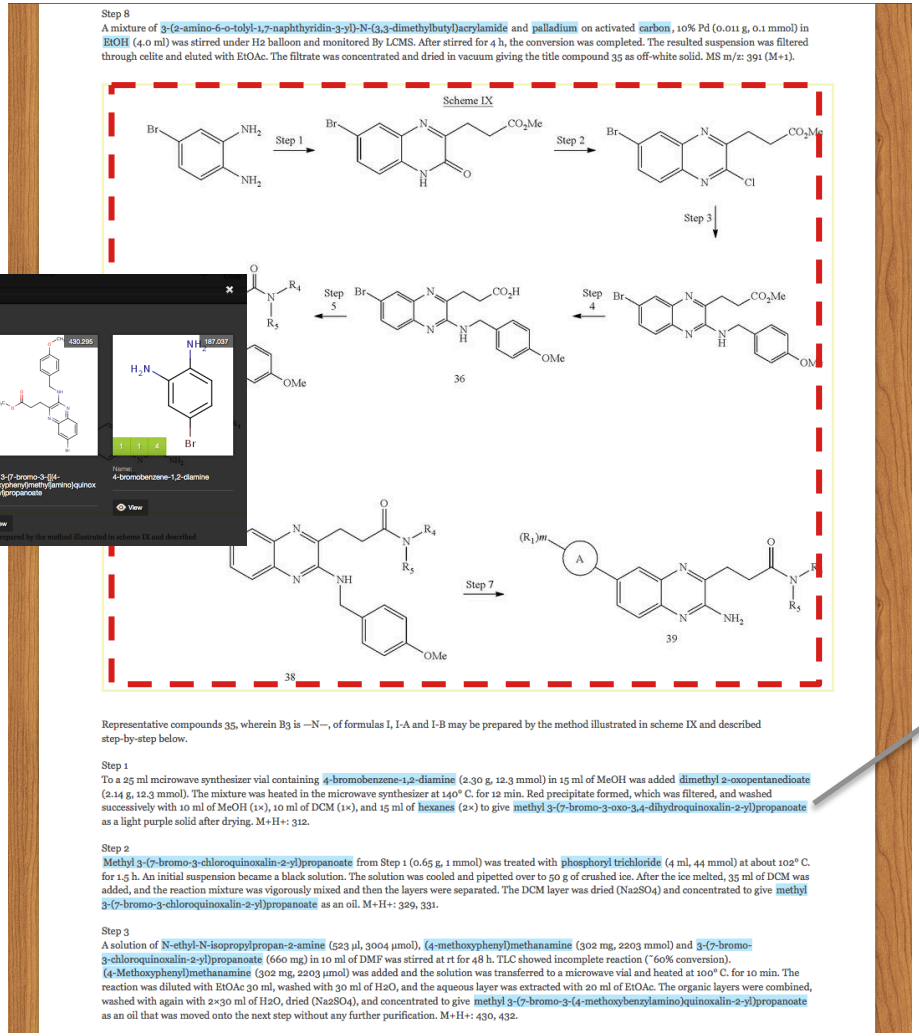


SureChEMBL System



SureChEMBL System





Step 8
A mixture of 3-(2-amino-6-o-tolyl-1,7-naphthyridin-3-yl)-N-(3,3-dimethylbutyl)acrylamide and palladium on activated carbon, 10% Pd (0.011 g, 0.1 mmol) in EtOH (4.0 ml) was stirred under H2 balloon and monitored By LCMS. After stirred for 4 h, the conversion was completed. The resulted suspension was filtered through celite and eluted with EtOAc. The filtrate was concentrated and dried in vacuum giving the title compound 35 as off-white solid. MS m/z: 391 (M+1).

Scheme IX

Step 1 Step 2 Step 3 Step 4 Step 5 Step 7

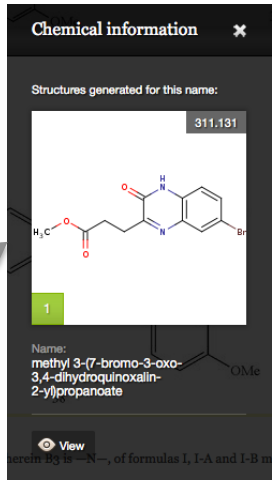
36 38 39

Representative compounds 35, wherein B3 is —N—, of formulas I, I-A and I-B may be prepared by the method illustrated in scheme IX and described step-by-step below.

Step 1
To a 25 ml mcirowave synthesizer vial containing 4-bromobenzene-1,2-diamine (2.30 g, 12.3 mmol) in 15 ml of MeOH was added dimethyl 2-oxopentanedioate (2.14 g, 12.3 mmol). The mixture was heated in the microwave synthesizer at 140° C. for 12 min. Red precipitate formed, which was filtered, and washed successively with 10 ml of MeOH (1×), 10 ml of DCM (1×), and 15 ml of hexanes (2×) to give methyl 3-(7-bromo-3-oxo-3,4-dihydroquinoxalin-2-yl)propanoate as a light purple solid after drying. M+H+: 312.

Step 2
Methyl 3-(7-bromo-3-chloroquinoxalin-2-yl)propanoate from Step 1 (0.65 g, 1 mmol) was treated with phosphoryl trichloride (4 ml, 44 mmol) at about 102° C. for 1.5 h. An initial suspension became a black solution. The solution was cooled and pipetted over to 50 g of crushed ice. After the ice melted, 35 ml of DCM was added, and the reaction mixture was vigorously mixed and then the layers were separated. The DCM layer was dried (Na2SO4) and concentrated to give methyl 3-(7-bromo-3-chloroquinoxalin-2-yl)propanoate as an oil. M+H+: 329, 331.

Step 3
A solution of N-ethyl-N-isopropylpropan-2-amine (523 µl, 3004 µmol), (4-methoxyphenyl)methanamine (302 mg, 2203 mmol) and 3-(7-bromo-3-chloroquinoxalin-2-yl)propanoate (660 mg) in 10 ml of DMF was stirred at rt for 48 h. TLC showed incomplete reaction (~60% conversion). (4-Methoxyphenyl)methanamine (302 mg, 2203 µmol) was added and the solution was transferred to a microwave vial and heated at 100° C. for 10 min. The reaction was diluted with EtOAc 30 ml, washed with 30 ml of H2O, and the aqueous layer was extracted with 20 ml of EtOAc. The organic layers were combined, washed with again with 2×30 ml of H2O, dried (Na2SO4), and concentrated to give methyl 3-(7-bromo-3-(4-methoxybenzylamino)quinoxalin-2-yl)propanoate as an oil that was moved onto the next step without any further purification. M+H+: 430, 432.



EMBL-EBI

UniChem

- Chemical cross-referencing system developed within EMBL-EBI
- Based on Standard InChI
- Provides fast, scalable mapping between chemistry resources, accessible via web services
- Currently provides mappings between 22 different resources (e.g., ChEMBL, ChEBI, PubChem, Zinc, eMolecules, DrugBank)
- Contains more than 65 million distinct structures (InChIs) and 100 million database identifiers

Integration of SureChEMBL with ChEMBL

- Compounds extracted from patents in SureChEMBL loaded as a source in UniChem
- Then possible to dynamically link other resources such as ChEMBL to patent documents
- Lookup can be carried out with:
  - A database identifier (e.g., does ZINC19796168 exist in SureChEMBL?)
  - A standard InChI key (e.g., does BNRNXUUZRGQAQC-UHFFFAOYSA-N exist in SureChEMBL?)
- https://www.ebi.ac.uk/unichem/rest/verbose_inchikey/BNRNXUUZRGQAQC-UHFFFAOYSA-N

ChEMBL

ChEMBL
Downloads
UniChem New
Malaria Data
ChEMBL-NTD
ADME SARfari New
Kinase SARfari
GPCR SARfari
DrugEBIlity
Web Services
EBI RDF Platform New
FAQ

ChEMBL Statistics

- DB: ChEMBL_18
- Targets: 9,414
- Compound records: 1,566,998
- Distinct compounds: 1,359,508
- Activities: 12,419,715
- Publications: 53,298
- Release Notes

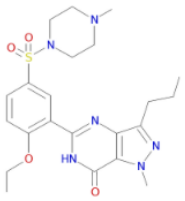
ChEMBL Blog

- Paper: Towards predictive resistance models for agrochemicals by combining chemical and protein similarity via proteochemometric modelling
- New Drug Approvals 2014 - Pt. VIII - Siltuximab (Sylvant™)

Compound Report Card

Compound Name and Classification

Compound ID	CHEMBL192
Compound Name	SILDENAFIL
ChEMBL Synonyms	Viagra | UK-92480 | UK-9248010 | SID26748898 | Sildenafil | UK-92480-10 | SID144205270 | Sildenafil Citrate | SID50085897
Max Phase	4 (Approved)
Trade Names	Viagra | Revatio | Sildenafil Citrate
Molecular Formula	C22H30N6O4S



CHEMBL192

Additional synonyms for CHEMBL192 found using NCI Chemical Identifier Resolver

Compound Representations

Molfile	Download MolFile
Canonical SMILES	CCCc1nn(C)c2C(=O)NC(=Nc12)c3cc(ccc3OCC)S(=O)(=O)N4CCN(C)CC4
Standard InChI	InChI=1S/C22H30N6O4S/c1-5-7-17-19-20(27(4)25-17)22(29)24-21(... Download InChI
Standard InChI Key	BNRNXUUZRGQAQC-UHFFFAOYSA-N

Mechanism of Action

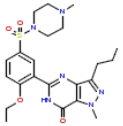
Mechanism of Action	ChEMBL Target	References
Phosphodiesterase 5A inhibitor	Phosphodiesterase 5A	DailyMed

Molecule Features

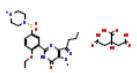
Clinical Trials for Compound

Number of clinical trials registered at clinicaltrials.gov	268

Alternate Forms of Compound in ChEMBL



CHEMBL192



CHEMBL1737

ChEMBL

ChEMBL
Downloads
UniChem New
Malaria Data
ChEMBL-NTD
ADME SARfari New
Kinase SARfari
GPCR SARfari
DrugEBIlity
Web Services
EBI RDF Platform New
FAQ

ChEMBL Statistics

- DB: ChEMBL_18
- Targets: 9,414
- Compound records: 1,566,998
- Distinct compounds: 1,359,508
- Activities: 12,419,715
- Publications: 53,298
- Release Notes

ChEMBL Blog

- Paper: Towards predictive resistance models for agrochemicals by combining chemical and protein similarity via proteochemometric modelling
- New Drug Approvals 2014 - Pt. VIII - Siltuximab (Sylvant™)

EBI › Databases › Small Molecules › ChEMBL Database › Compound Search › CHEMBL192

Compound Report Card

Compound Name and Classification

Compound ID	CHEMBL192
Compound Name	SILDENAFIL
ChEMBL Synonyms	Viagra | UK-92480 | UK-9248010 | SID26748898 | Sildenafil | UK-92480-1 SID144205270 | Sildenafil Citrate | SID50085897
Max Phase	4 (Approved)
Trade Names	Viagra | Revatio | Sildenafil Citrate
Molecular Formula	C22H30N6O4S

Additional synonyms for CHEMBL192 found using NCI Chemic

Compound Representations

Molfile	Download MolFile
Canonical SMILES	CCCc1nn(C)c2C(=O)NC(=Nc12)c3cc
Standard InChI	InChI=1S/C22H30N6O4S/c1-5-7-17- Download InChI
Standard InChI Key	BNRNXUUZRGQAQC-UHFFFAOYS

Mechanism of Action

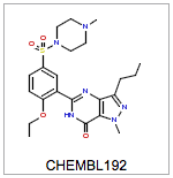
Mechanism of Action	Ch
Phosphodiesterase 5A inhibitor	Phos

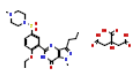
Molecule Features

Clinical Trials for Compound

Number of clinical trials registered at clinicaltrials.

Alternate Forms of Compound in ChEMBL


CHEMBL192


CHEMBL1737

UniChem Cross References

New View the UniChem Connectivity matches for CHEMBL192

DrugBank	DB00203
PDBe	VIA
KEGG Ligand	C07259
ChEBI	9139
ZINC	ZINC19796168
eMolecules	902463
IBM Patent System	D814FFE26EDA163F0CDC1115AD9C7CC3
Patent	WO2000007597A1 WO1999066933A1 EP1020190A3 US6066735 EP1027887A2 US6087362 WO2000054773A1 EP0812845A1 WO2000000199A1 WO2000038655A1 WO2000054777A1 US5250534 EP0995441A3 WO2000054774A1 WO1999066924A1 WO2000004875A2 WO2000072827A2 WO1999059584A1 EP1027888A3 WO2000059475A1 WO2000012076A1 WO2000010542A2 WO2000003721A1 WO1999060985A2 WO1999064033A1 WO2000067735A2 WO1999067231A1 WO1999051252A1 WO1999030697A2 EP1037616A2 WO1999066870A1 WO1999039763A1 WO2000044363A2 WO2000006121A1 WO2000007596A1 WO2000078760A1 WO1999021558A2 WO2000000212A1 US6037346 WO1998055176A1 WO2000040226A2 EP1027054A1 EP1027887A3 WO2000045795A2 EP0967214A1 WO2000043012A1 WO2000051978A1 WO2000057857A1 WO1999027905A1 WO2000050007A1 EP0960621A2 WO2000074652A1 WO1999020251A1 US6075028 WO2000015233A1 WO1999030688A1 EP0812845B1 WO2000053148A2 EP0995441A2 WO1999038507A1 US5955611 EP0951908A2 WO2000066084A1 WO2000012110A2 EP1027888A2 WO2000042992A2 EP0916675A2 WO1999002161A1
FDA SRS	3M7OB98Y7H
SureChem	SureCN1895
PharmGKB	PA451346
Human Metabolome Database	HMDB05039
PubChem: Thomson Pharma	14883184 14834479
PubChem	5212
Mcule	MCULE-2481702478

UniChem REST Service Call: https://www.ebi.ac.uk/unichem/rest/verbose_inchikey/BNRNXUUZRGQAQC-UHFFFAOYSA-N

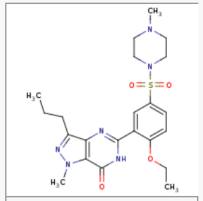
EMBL-EBI

5-[2-ethoxy-5-(4-methylpiperazine-1-sulfonyl)phenyl]-1-methyl-3-propyl-1H,6H,7H-pyrazolo[4,3-d]pyrimidin-7-one

Inchi key: BNRNXUUZRGQAQC-UHFFFAOYSA-N

Occurences across our corpus

53,597



Enlarge image

SMILES

InChi

Mol Weight	Lipinski ROF
474.575	Yes

Log P	Donor Cnt	Ring Cnt
1.35	1	4

Accept Cnt
7

Rotable Bond Cnt
6

ChemSpider database results 1

View ChemSpider hits

RSC journal results 28

View RSC hits

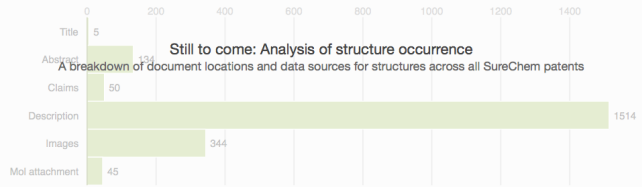
PubChem database results 1

View PubChem hits

ChEMBL database results 1

View ChEMBL hits

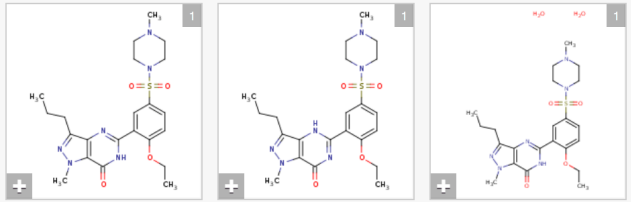
Found	Title	Abstract	Claims	Description	Images	CWU's (?)
2092 times	5 times	134 times	50 times	1514 times	344 times	45 times



Related structures | Patent hits

Related structures are those within our corpus with a similarity equal to or greater than 0.95

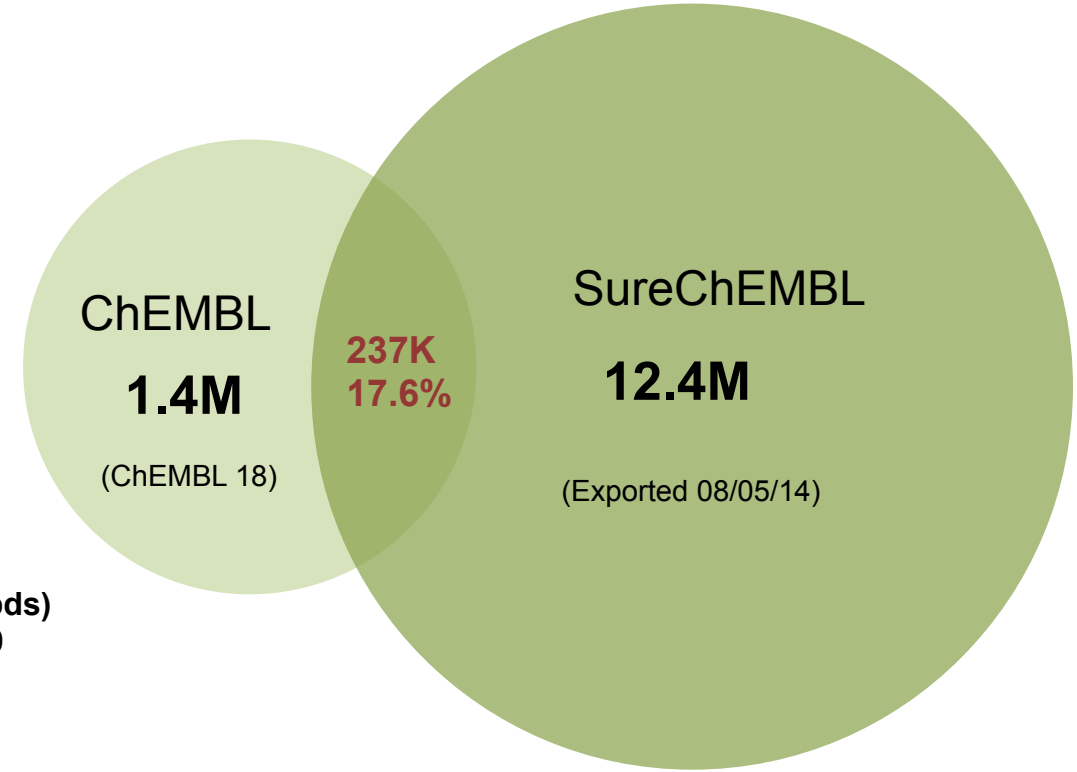
Showing 1-3 of 59 total structure results Select all on page |



View all results

SureChEMBL-ChEMBL Overlap

- InChI based comparison using UniChem



Filters (removes ~3 million cpds)

- MW between 100 and 1200
- #Atoms between 6 and 70
- ALogP between -10 and 10
- #Rings > 0
- #C > 0
- #C != #Atoms
- RTB <= 20
- Not frequently occurring (<100K)

SureChEMBL Overlap with Other Sources

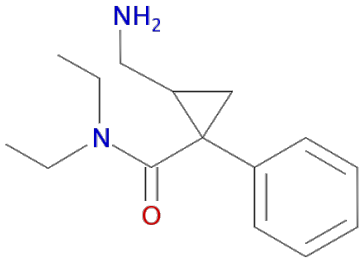
Data Source	Total Structures	% in SureChEMBL
PubChem	49,224,050	19.2
Zinc	21,702,928	0.9
mcule	5,999,424	2.2
eMolecules	5,168,336	4.4
IBM patents	2,473,730	62.7
FDA SRS	33,780	51.1
DrugBank	6,352	50.3

2.74 million SureChEMBL compounds new to UniChem (not in any of other 21 sources)

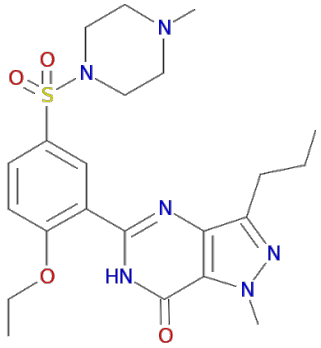
Limitations of InChI Key-Based Matching

- Matching on standard InChI keys doesn't consider different stereochemistry, isotopes, salts and mixtures

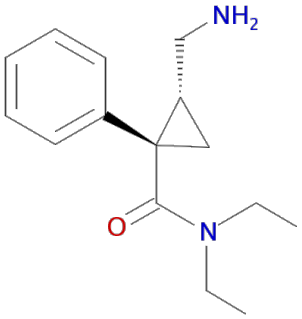
Milnacipran
GJJFMKBJSRMPLA-UHFFFAOYSA-N



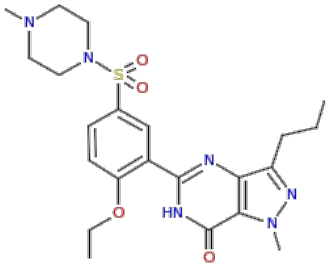
Sildenafil
BNRNXUUZRGQAQC-UHFFFAOYSA-N

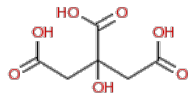


Levomilnacipran
GJJFMKBJSRMPLA-DZGCQCFKSA-N



Sildenafil citrate
DEIYFTQMQPDXOT-UHFFFAOYSA-N





Connectivity Matching with UniChem

- Layered construction of InChI allows matching criteria to be varied and mixtures to be separated

InChI=1S/C22H31NO.C4H6O6/
c1-16(2)23(17(3)4)14-13-20(19-9-7-6-8-10-19)21-15-18(5)11-12-22(21)
24;5-1(3(7)8)2(6)4(9)10/h6-12,15-17,20,24H,13-14H2,1-5H3;1-2,5-6H,
(H,7,8)(H,9,10)/t20-;/m1./s1

- Allows matching of ‘related’ compounds (i.e., same connectivity or components of mixtures)
- Allows identification of relationships between molecules (e.g., ‘isotopic variant of’, ‘stereoisomer of’ etc)

ChEMBL Statistics

- DB: ChEMBL_18
- Targets: 9,414
- Compound records: 1,566,998
- Distinct compounds: 1,359,508
- Activities: 12,419,715
- Publications: 53,298
- Release Notes

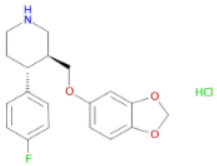
ChEMBL Blog

- Paper: An atlas of genetic influences on human blood metabolites
- Paper: Towards predictive resistance models for agrochemicals by combining chemical and protein similarity via proteochemometric modelling

Compound Report Card

Compound Name and Classification

Compound ID	CHEMBL1708
Compound Name	PAROXETINE HYDROCHLORIDE
ChEMBL Synonyms	Paxil | SID26753519 | SID85273719 | Paroxetine HCl
Max Phase	4 (Approved)
Trade Names	Paxil | Paxil Cr | Paroxetine HCl
Molecular Formula	C19H21ClFNO3



CHEMBL1708

Additional synonyms for CHEMBL1708 found using NCI Chemical Identifier Resolver

Compound Representations

Molfile	Download MolFile
Canonical SMILES	Cl.Fc1ccc(cc1)[C@@H]2CCNC[C@H]2COc3ccc4OCOc4c3
Standard InChI	InChI=1S/C19H20FNO3.ClH/c20-15-3-1-13(2-4-15)17-7-8-21-10-14 ... Download InChI
Standard InChI Key	GELRVIPPMNMYGS-RVXRQPKJSA-N

Mechanism of Action

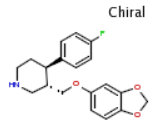
Mechanism of Action	ChEMBL Target	References
Serotonin transporter inhibitor	Serotonin transporter	DailyMed

Molecule Features

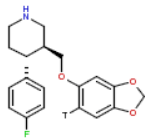
Clinical Trials for Compound

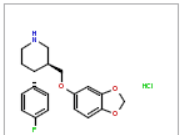
Number of clinical trials registered at clinicaltrials.gov	201

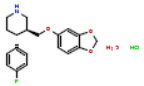
Alternate Forms of Compound in ChEMBL

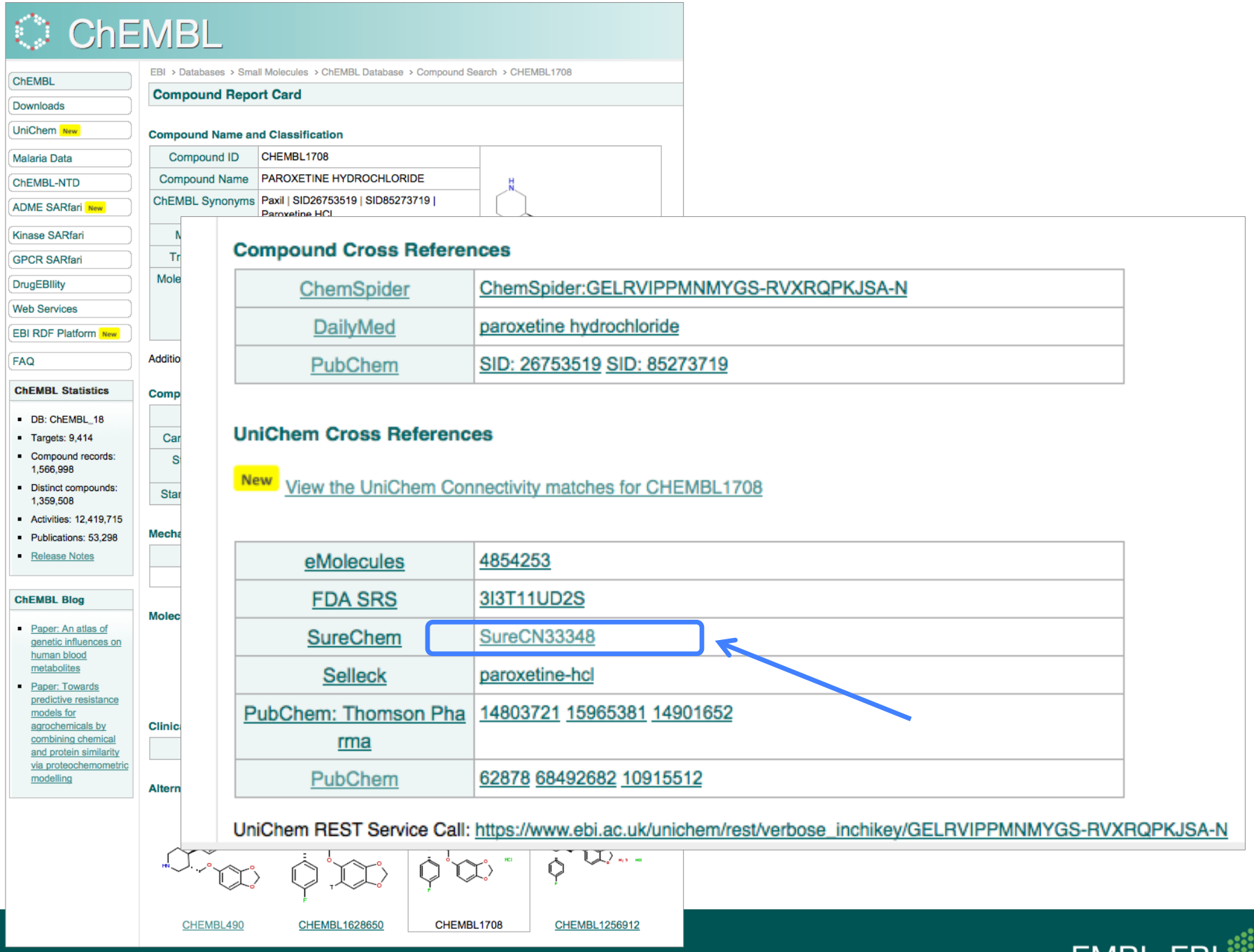


CHEMBL490


CHEMBL1628650


CHEMBL1708


CHEMBL1256912


ChEMBL
ChEMBL
Downloads
UniChem New
Malaria Data
ChEMBL-NTD
ADME SARfari New
Kinase SARfari
GPCR SARfari
DrugEBIlity
Web Services
EBI RDF Platform New
FAQ
ChEMBL Statistics
DB: ChEMBL_18
Targets: 9,414
Compound records: 1,566,998
Distinct compounds: 1,359,508
Activities: 12,419,715
Publications: 53,298
Release Notes
ChEMBL Blog
Paper: An atlas of genetic influences on human blood metabolites
Paper: Towards predictive resistance models for agrochemicals by combining chemical and protein similarity via proteochemometric modelling
EBI > Databases > Small Molecules > ChEMBL Database > Compound Search > CHEMBL1708
Compound Report Card
Compound Name and Classification
Compound ID CHEMBL1708
Compound Name PAROXETINE HYDROCHLORIDE
ChEMBL Synonyms Paxil | SID26753519 | SID85273719 | Paroxetine HCl
Compound Cross References
ChemSpider ChemSpider:GELRVIPPMNMYGS-RVXRQPKJSA-N
DailyMed paroxetine hydrochloride
PubChem SID: 26753519 SID: 85273719
UniChem Cross References
New View the UniChem Connectivity matches for CHEMBL1708
eMolecules 4854253
FDA SRS 3I3T11UD2S
SureChem SureCN33348
Selleck paroxetine-hcl
PubChem: Thomson Pharma 14803721 15965381 14901652
PubChem 62878 68492682 10915512
UniChem REST Service Call: https://www.ebi.ac.uk/unichem/rest/verbose_inchikey/GELRVIPPMNMYGS-RVXRQPKJSA-N
CHEMBL490
CHEMBL1628650
CHEMBL1708
CHEMBL1256912

EMBL-EBI

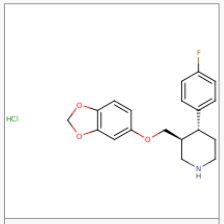
SureChemOpen BETA

Help & Support | My Account

(3S,4R)-3-[(2H-1,3-benzodioxol-5-yloxy)methyl]-4-(4-fluorophenyl)piperidine hydrochloride

Occurences across our corpus
11,865

Inchi key: GELRVIPPMNMYGS-RVXRQPKJSA-N



Enlarge image

SMILES

InChi

Mol Weight	Lipinski ROF
365.825	Yes

Log P	Donor Cnt	Ring Cnt
3.14	1	4

Accept Cnt	Rotable Bond Cnt
4	4

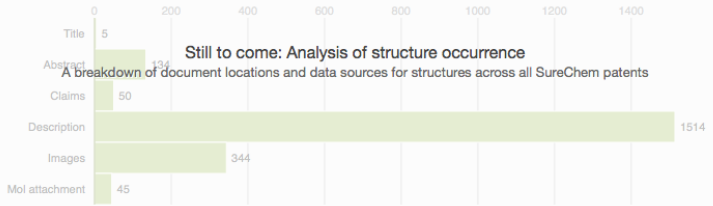
ChemSpider database results
1
View ChemSpider hits

RSC journal results
2
View RSC hits

PubChem database results
1
View PubChem hits

ChEMBL database results
1
View ChEMBL hits

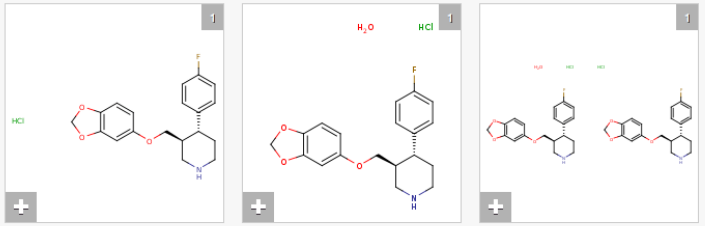
Found	Title	Abstract	Claims	Description	Images	CWU's (?)
2092 times	5 times	134 times	50 times	1514 times	344 times	45 times



Related structures | Patent hits

Related structures are those within our corpus with a similarity equal to or greater than 0.95

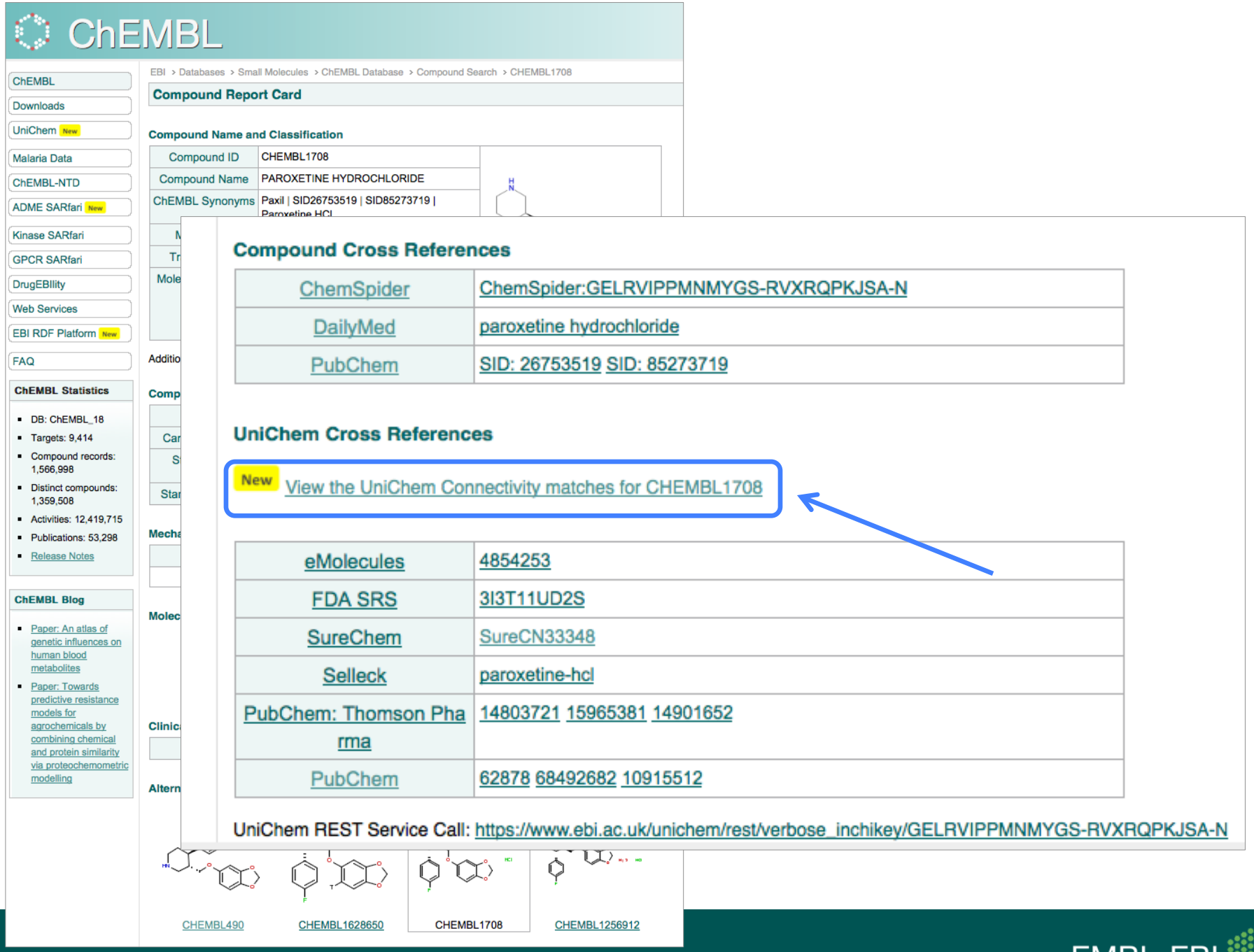
Showing 1-3 of 364 total structure results Select all on page |



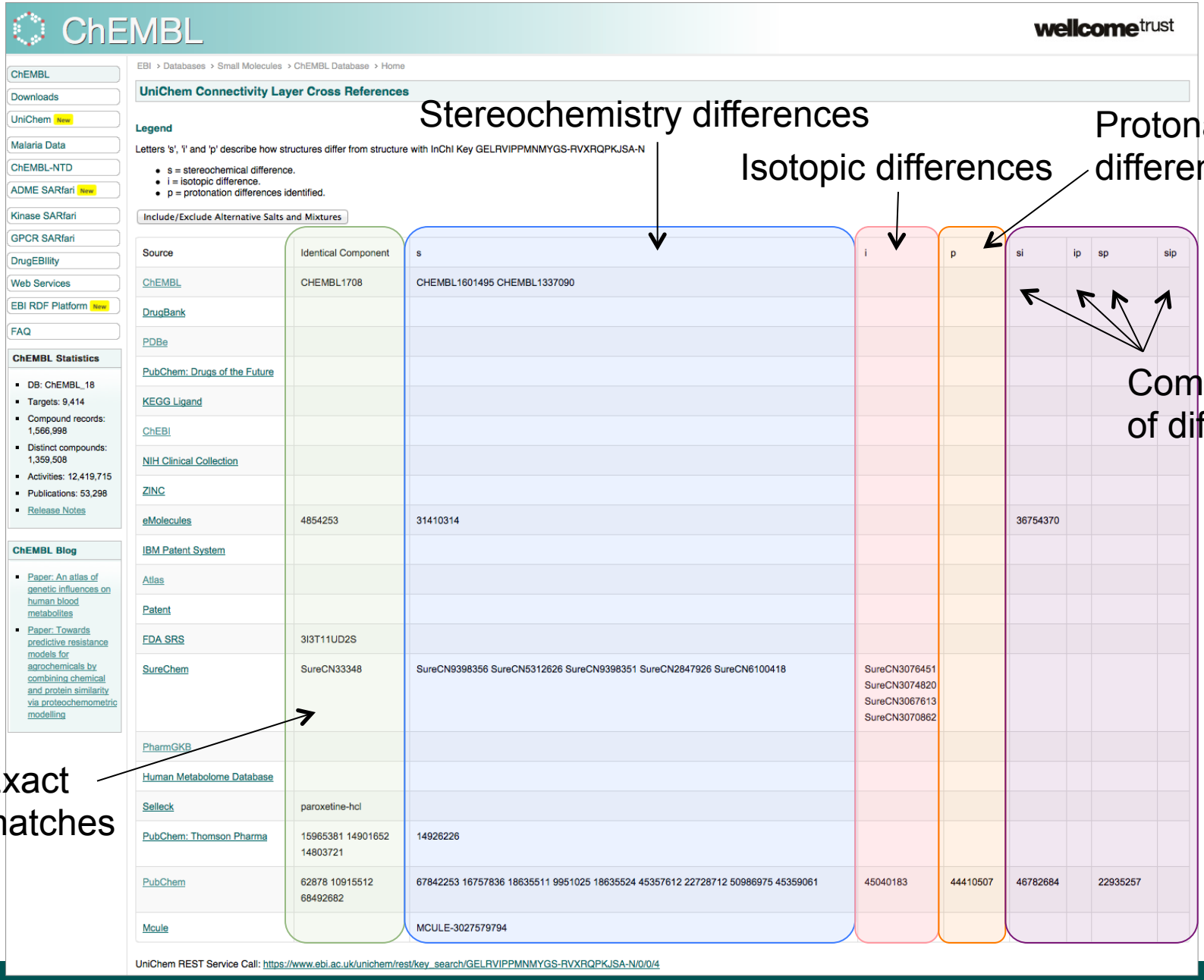
View all results

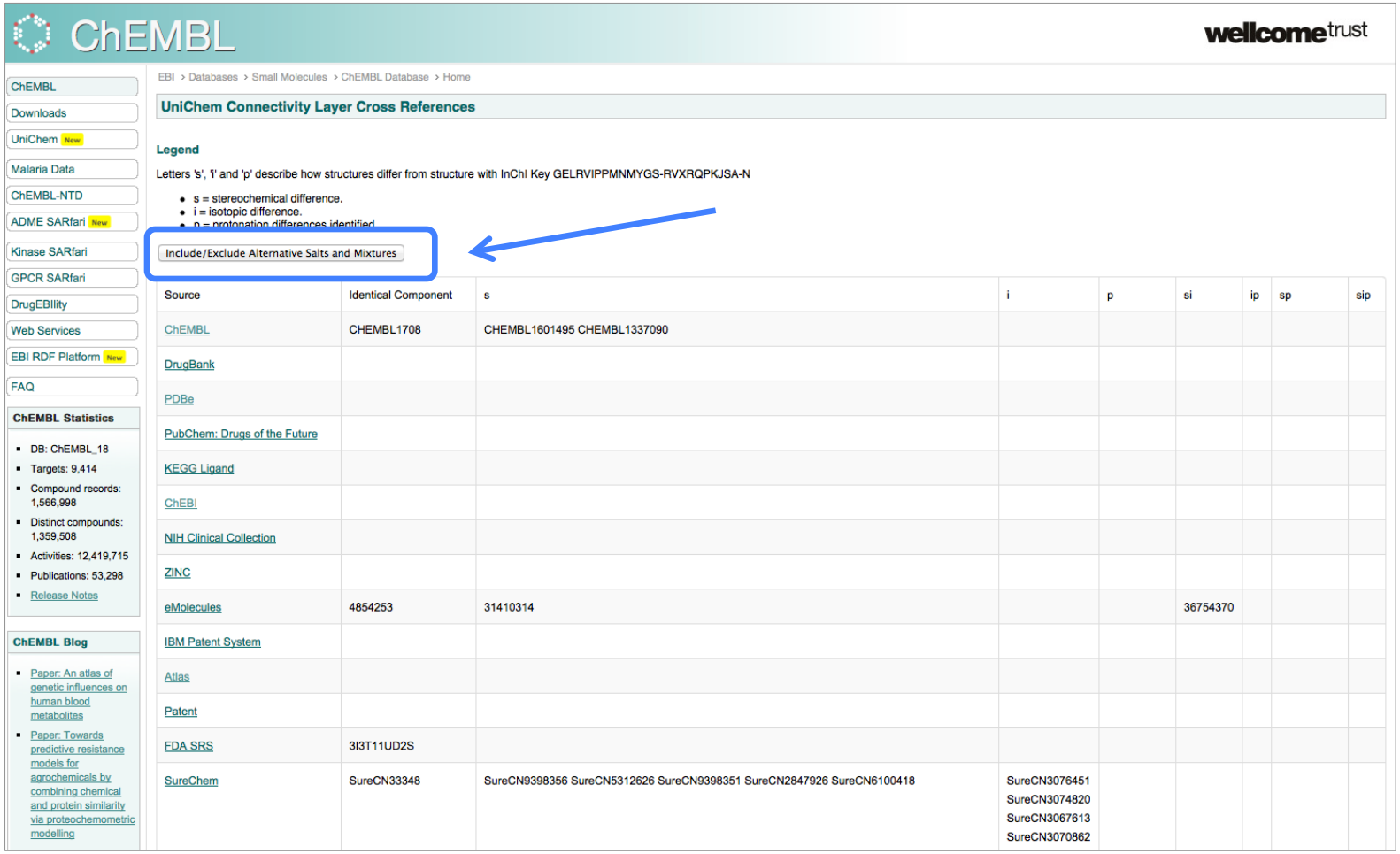
Feedback & support

EMBL-EBI


ChEMBL
ChEMBL
Downloads
UniChem New
Malaria Data
ChEMBL-NTD
ADME SARfari New
Kinase SARfari
GPCR SARfari
DrugEBIlity
Web Services
EBI RDF Platform New
FAQ
ChEMBL Statistics
DB: ChEMBL_18
Targets: 9,414
Compound records: 1,566,998
Distinct compounds: 1,359,508
Activities: 12,419,715
Publications: 53,298
Release Notes
ChEMBL Blog
Paper: An atlas of genetic influences on human blood metabolites
Paper: Towards predictive resistance models for agrochemicals by combining chemical and protein similarity via proteochemometric modelling
EBI > Databases > Small Molecules > ChEMBL Database > Compound Search > CHEMBL1708
Compound Report Card
Compound Name and Classification
Compound ID CHEMBL1708
Compound Name PAROXETINE HYDROCHLORIDE
ChEMBL Synonyms Paxil | SID26753519 | SID85273719 | Paroxetine HCl
Compound Cross References
ChemSpider ChemSpider:GELRVIPPMNMYGS-RVXRQPKJSA-N
DailyMed paroxetine hydrochloride
PubChem SID: 26753519 SID: 85273719
UniChem Cross References
New View the UniChem Connectivity matches for CHEMBL1708
eMolecules 4854253
FDA SRS 3I3T11UD2S
SureChem SureCN33348
Selleck paroxetine-hcl
PubChem: Thomson Pharma 14803721 15965381 14901652
PubChem 62878 68492682 10915512
UniChem REST Service Call: https://www.ebi.ac.uk/unichem/rest/verbose_inchikey/GELRVIPPMNMYGS-RVXRQPKJSA-N
CHEMBL490
CHEMBL1628650
CHEMBL1708
CHEMBL1256912

EMBL-EBI


ChEMBL
wellcome trust
ChEMBL
Downloads
UniChem New
Malaria Data
ChEMBL-NTD
ADME SARfari New
Kinase SARfari
GPCR SARfari
DrugEBIlity
Web Services
EBI RDF Platform New
FAQ
ChEMBL Statistics
DB: ChEMBL_18
Targets: 9,414
Compound records: 1,566,998
Distinct compounds: 1,359,508
Activities: 12,419,715
Publications: 53,298
Release Notes
ChEMBL Blog
Paper: An atlas of genetic influences on human blood metabolites
Paper: Towards predictive resistance models for agrochemicals by combining chemical and protein similarity via proteochemometric modelling
EBI › Databases › Small Molecules › ChEMBL Database › Home
UniChem Connectivity Layer Cross References
Legend
Letters 's', 'i' and 'p' describe how structures differ from structure with InChI Key GELRVIPPMNMYGS-RVXRQPKJSA-N
s = stereochemical difference.
i = isotopic difference.
p = protonation differences identified.
Include/Exclude Alternative Salts and Mixtures
Source
Identical Component
s
i
p
si
ip
sp
sip
ChEMBL
CHEMBL1708
CHEMBL1601495 CHEMBL1337090
DrugBank
PDBe
PubChem: Drugs of the Future
KEGG Ligand
ChEBI
NIH Clinical Collection
ZINC
eMolecules
4854253
31410314
36754370
IBM Patent System
Atlas
Patent
FDA SRS
3I3T11UD2S
SureChem
SureCN33348
SureCN9398356 SureCN5312626 SureCN9398351 SureCN2847926 SureCN6100418
SureCN3076451 SureCN3074820 SureCN3067613 SureCN3070862
PharmGKB
Human Metabolome Database
Selleck
paroxetine-hcl
PubChem: Thomson Pharma
15965381 14901652 14803721
14926226
PubChem
62878 10915512 68492682
67842253 16757836 18635511 9951025 18635524 45357612 22728712 50986975 45359061
45040183
44410507
46782684
22935257
Mcule
MCULE-3027579794
UniChem REST Service Call: https://www.ebi.ac.uk/unichem/rest/key_search/GELRVIPPMNMYGS-RVXRQPKJSA-N/0/0/4
Stereochemistry differences
Isotopic differences
Protonation differences
Combinations of differences
Exact matches


ChEMBL
wellcome trust
ChEMBL
Downloads
UniChem New
Malaria Data
ChEMBL-NTD
ADME SARfari New
Kinase SARfari
GPCR SARfari
DrugEBIlity
Web Services
EBI RDF Platform New
FAQ
ChEMBL Statistics
DB: ChEMBL_18
Targets: 9,414
Compound records: 1,566,998
Distinct compounds: 1,359,508
Activities: 12,419,715
Publications: 53,298
Release Notes
ChEMBL Blog
Paper: An atlas of genetic influences on human blood metabolites
Paper: Towards predictive resistance models for agrochemicals by combining chemical and protein similarity via proteochemometric modelling
EBI › Databases › Small Molecules › ChEMBL Database › Home
UniChem Connectivity Layer Cross References
Legend
Letters 's', 'i' and 'p' describe how structures differ from structure with InChI Key GELRVIPPMNMYGS-RVXRQPKJSA-N
s = stereochemical difference.
i = isotopic difference.
Include/Exclude Alternative Salts and Mixtures
Source
Identical Component
s
i
p
si
ip
sp
sip
ChEMBL
CHEMBL1708
CHEMBL1601495 CHEMBL1337090
DrugBank
PDBe
PubChem: Drugs of the Future
KEGG Ligand
ChEBI
NIH Clinical Collection
ZINC
eMolecules
4854253
31410314
36754370
IBM Patent System
Atlas
Patent
FDA SRS
3I3T11UD2S
SureChem
SureCN33348
SureCN9398356 SureCN5312626 SureCN9398351 SureCN2847926 SureCN6100418
SureCN3076451
SureCN3074820
SureCN3067613
SureCN3070862

UniChem Connectivity Layer Cross References

Legend

Letters 's', 'i' and 'p' describe how structures differ from structure with InChI Key GELRVIPPMNMYGS-RVXRQPKJSA-N

- s = stereochemical difference.
- i = isotopic difference.
- p = protonation differences identified.

Include/Exclude Alternative Salts and Mixtures

Matching component of mixture (red = paroxetine, blue = HCl)

Source	Identical Component	s	i	p	si	ip	sp	sip
ChEMBL	CHEMBL1708 CHEMBL490 CHEMBL1256912 CHEMBL1200609 CHEMBL1449490 CHEMBL19429 CHEMBL1601495 CHEMBL1256912 CHEMBL1337090 CHEMBL1231821	CHEMBL1337090 CHEMBL1601495 CHEMBL169007 CHEMBL1354777 CHEMBL1616103 CHEMBL1601495 CHEMBL1337090 CHEMBL104677	CHEMBL1628650					

...

SureChem	SureCN33348 SureCN9937114 SureCN6978081 SureCN7070794 SureCN8419494 SureCN6599797 SureCN4188681 SureCN3197349 SureCN3888530 SureCN8126670 SureCN7074912 SureCN6586019 SureCN7827325 SureCN143905 SureCN14778455 SureCN297376 SureCN7836920 SureCN7975085 SureCN505869 SureCN5905513 SureCN7812571 SureCN8160827 SureCN140608 SureCN7836996 SureCN7838648 SureCN14785500 SureCN8122074 SureCN8139459 SureCN5173412 SureCN7837227 SureCN6582024 SureCN7835179 SureCN7071089 SureCN6168705 SureCN3862931 SureCN7838847 SureCN8477201 SureCN6383707 SureCN8127771 SureCN6581867 SureCN6325605 SureCN8122654 SureCN3862964 SureCN4436350 SureCN7837040 SureCN7075953 SureCN6587455 SureCN7828057 SureCN8133871 SureCN7840642 SureCN8127768 SureCN8184989 SureCN7826585 SureCN6586954 SureCN8474305 SureCN7075607 SureCN7075598 SureCN140609	SureCN9398351 SureCN5312626 SureCN2847926 SureCN9398356 SureCN6100418 SureCN9398351 SureCN1945816 SureCN114238 SureCN7247758 SureCN10643985 SureCN12537969 SureCN10643980 SureCN2583823 SureCN9398356 SureCN8301411 SureCN2847926 SureCN5312626 SureCN6100418	SureCN3074820 SureCN3076451 SureCN3070862 SureCN3067613 SureCN13799653 SureCN3067613 SureCN14010891 SureCN3070862 SureCN3074820 SureCN12538524 SureCN13139123 SureCN3076451		SureCN14510556 SureCN14510578 SureCN12538058 SureCN12538056 SureCN14510622 SureCN12538387 SureCN14510567 SureCN12538466 SureCN12538382 SureCN14510598 SureCN12538142 SureCN12538066 SureCN12538379 SureCN14510576 SureCN14510619 SureCN14510534 SureCN12538300 SureCN14510572 SureCN12538021 SureCN14510548 SureCN14510640 SureCN14510633 SureCN14510634 SureCN12538102 SureCN14510573 SureCN12538034 SureCN12538297 SureCN14510645 SureCN14510568			

Different salt forms of paroxetine

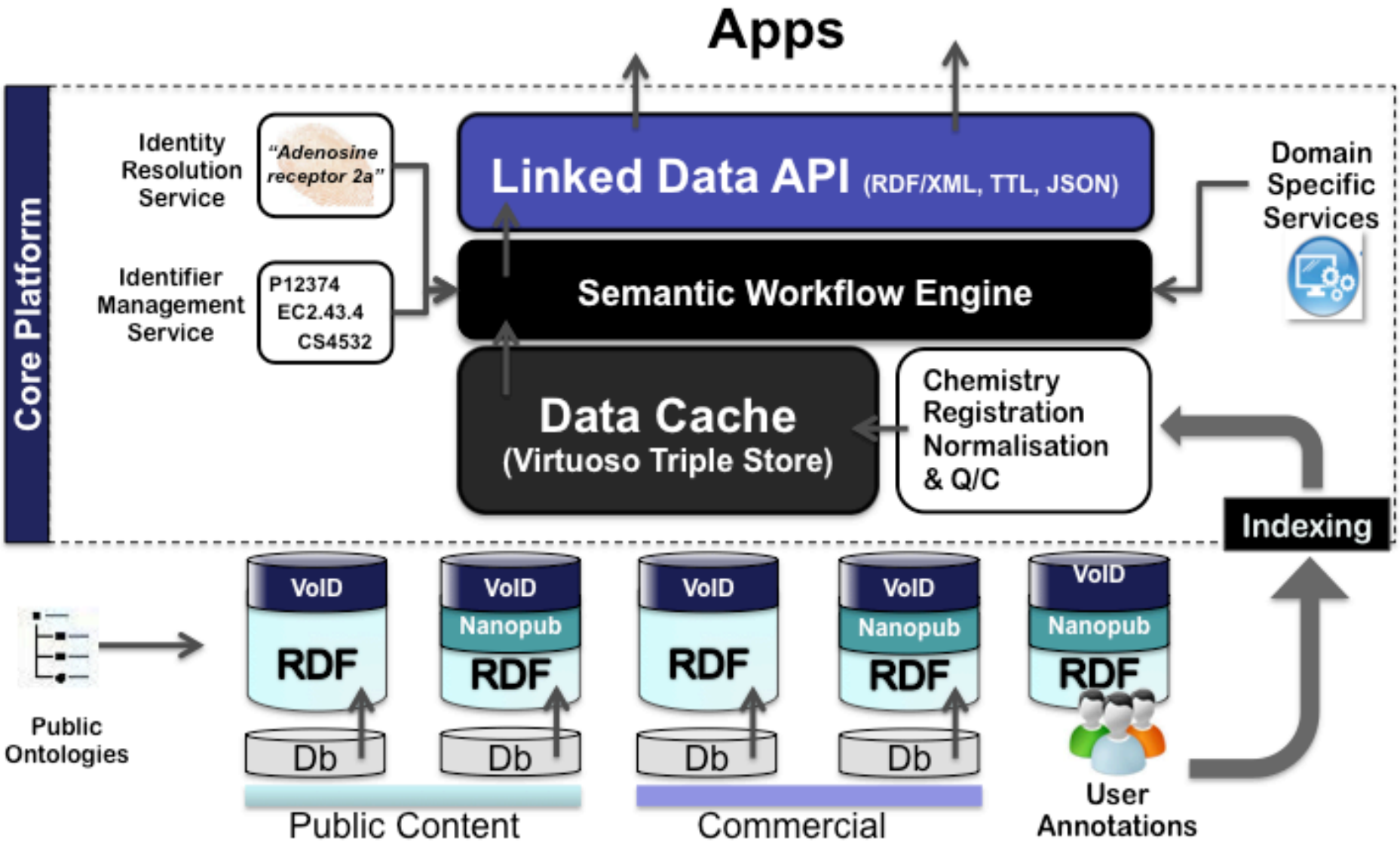
SureChEMBL Connectivity Overlap with Other Sources

Data Source	% in SureChEMBL (exact)	% in SureChEMBL (connectivity)	% in SureChEMBL (salts/mixtures)
ChEMBL	17.6	19.7	20.8
PubChem	19.2	21.7	21.7
Zinc	0.9	2	2
mcule	2.2	2.7	2.7
eMolecules	4.4	4.6	4.8
IBM patents	62.7	66.1	68.3
FDA SRS	51.1	56.5	70.7
DrugBank	50.3	63.4	70.6

Open PHACTS
Open Pharmacological Space

- Open PHACTS Discovery Platform is an open drug-discovery data integration platform
- Funded by Innovative Medicines Initiative (IMI) – public/ private partnership involving 29 partners
- Uses semantic web technology (e.g., Resource Description Framework)
- Integrates key data sources including ChEMBL, ChEBI, DrugBank, WikiPathways, Swiss-Prot, Gene Ontology
- Access provided by API and growing number of applications (e.g., PharmaTrek, ChemBioNavigator)

Open PHACTS Architecture



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

EMBL-EBI

Integration of SureChEMBL

- Extension of Open PHACTS project until Feb 2016
- Aim to enable additional use-cases
- Integration of patent data into platform is of interest
- EMBL-EBI partner on Open PHACTS extension to provide SureChEMBL data
- Will produce RDF export of data which will be available via Open PHACTS and EBI RDF-platform
- Challenge of integrating daily-updated resource
  - ~3500 new structures added per day to SureChEMBL
  - 12.4 million structures total (after filtering)

Entity Recognition – Beyond Chemistry

- Working with SciBite, also a partner on the Open PHACTS extension project
- Termite text-mining engine will be used to annotate other entity types within patent documents e.g.,
  - Human genes/proteins
  - Diseases
  - Cell lines
- These data will enable users to address more complex use-cases e.g., patents relating to a particular disease or target

Other Future Plans

- Annotation of database identifiers (e.g., ChEMBL, ChEBI, PubChem, UniProt)
- Extraction/searching of protein sequences
- Improved access via workflow tools (e.g., Knime)
- Extending coverage to earlier patents (e.g., image extraction pre-2007)
- Benchmarking, enhancements to recognition algorithms (entity recognition, name to structure, image to structure), open source solutions
- Application to literature (Europe PMC)
- Of interest but no immediate plans – extraction of bioactivity data, Markush structures!

Acknowledgements

- ChEMBL team
  - John Overington
  - Anne Hersey
  - Jon Chambers
  - George Papadatos
  - Mark Davies
- Digital Science
  - Nicko Goncharoff
  - James Siddle
  - Richard Koks
  - Tom Llewellyn
- SciBite
  - Lee Harland
- Open PHACTS consortium
  - http://www.openphacts.org/partners/consortium

Funding:

Innovative Medicines Initiative Joint Undertaking, grant agreement no. 115191 (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)

Software:

ChemAxon, DIGITAL science, ACD/Labs, keymodule, OpenEye Scientific Software, IFI CLAIMS® Patent Services, OPSIN, PerkinElmer, NextMove Software, Twin Dolphin Software, SciBite