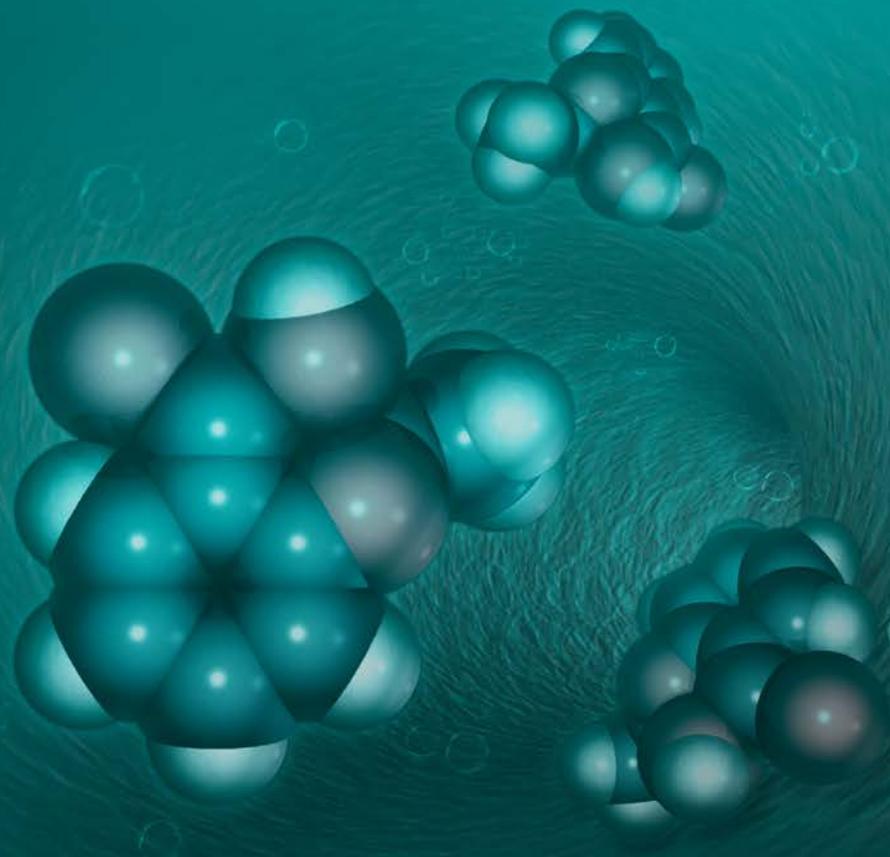


# ChEMBL – A Database of Bioactive Drug-like Small Molecules

Anne Hersey

ChEMBL Group, EMBL-EBI



# Outline

- ChEMBL
  - Bioactivity Database for Drug Discovery
- UniChem
  - Unified Chemical Structure Cross-referencing System

# What is ChEMBL

- Open access database for drug discovery
- Freely available (searchable and downloadable)
- Content:
  - Bioactivity data manually extracted from the primary medicinal chemistry literature from journals such as J. Med. Chem.
  - Subset of data from PubChem
  - Deposited data e.g. neglected disease screening, GSK kinase set
- Bioactivity data is associated with a biological target and a chemical structure
- Compounds are stored in a structure searchable format
- Protein targets are linked to protein sequences in UniProt
- Updated regularly with new data
- Secure searching (<https://www.ebi.ac.uk/chembl/db>)

# ChEMBL Content

ChEMBL16 (7th May 2013)
1,295,510 compounds
11,420,351 activities
712,836 assays
9,844 targets
50,095 documents

BioAssay Data Sources:	Assays:	Activities:
Scientific Literature	703,690	4,044,415
PubChem BioAssays	2,276	6,571,997
TP-search Transporter Database	3,592	6,765
Open TG-GATEs	1,376	179,525
Sanger Institute Genomics of Drug Sensitivity in Cancer	352	5,984
Guide to Receptors and Channels	344	801
DrugMatrix in vitro pharmacology assays	132	229,944
Millipore Kinase Screening	468	73,944
GSK Published Kinase Inhibitor Set	454	168,717
Drugs for Neglected Diseases Initiative (DNDi)	62	11,554
QSDD Malaria Screening	21	226
Screening	16	5,456
Malaria Screening	16	5,853
Box	14	3,788
Screening	6	81,198
Screening	6	27,888
Dis Screening	5	1,406
Screening	4	111
Elementary		
Data	2	779

Compound-Only Data Sources:	Compound Records:
USP Dictionary of USAN and International Drug Names	10,580
FDA Orange Book	2,012
Clinical Candidates	665
Manually Added Drugs	117

# Data Example

CC chemokine receptor-3 (CCR3) antagonists  
 J.R. Pruitt et al.  
 Bioorganic & Medicinal Chemistry Letters 17  
 (2007) 2992–2997

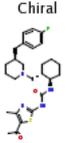
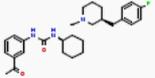
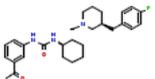
**Table 3.** In vitro binding and selectivity data for an initial set of heterocyclic DPC168 analogs

Compound	X <sup>3</sup>
3	CH <sub>2</sub>
4	O
5	CH <sub>2</sub>
6	CH <sub>2</sub>
7a	CH <sub>2</sub>
7b	CH <sub>2</sub>
8a	CH <sub>2</sub>
8b	CH <sub>2</sub>
9a	CH <sub>2</sub>
9b	CH <sub>2</sub>
9c <sup>d</sup>	CH <sub>2</sub>
9d <sup>d</sup>	CH <sub>2</sub>
10	CH <sub>2</sub>

**Table 5.** Comparison of potency, selectivity, pharmacokinetic, and in vivo efficacy results for DPC168 (1) and BMS-570520 (8i)<sup>a</sup>

Assay or PK parameter	Compound 1	Compound 8i	CYP2D6 IC <sub>50</sub> <sup>b</sup> (nM)
CCR3 IC <sub>50</sub> (nM)	2.0	1.9	
Chemotaxis IC <sub>50</sub> (nM)	0.034	0.068	
Ca <sup>2+</sup> mobilization IC <sub>50</sub> (nM)	8.0	2.6	
CYP2D6 IC <sub>50</sub> (nM)	30	1300	
5HT <sub>2A</sub> IC <sub>50</sub> (nM)	920	5300	
D <sub>2</sub> IC <sub>50</sub> (nM)	1000	640	
Serotonin transporter K <sub>i</sub> (nM)	2900	29,000	
Dopamine transporter K <sub>i</sub> (nM)	490	6500	
Norepinephrine transporter K <sub>i</sub> (nM)	950	7700	
hERG IC <sub>50</sub> (nM)	400	6000	
Mouse F (%)	20	25	
Mouse t <sub>1/2</sub> (h)	2.0	1.2	
Mouse CL (L/h/kg)	1.8	2.5	
Cyno F (%)	8	18	
Cyno t <sub>1/2</sub> (h)	4.0	5.5	
Cyno CL (L/h/kg)	2.1	0.87	
Chimp F (%)	22	7	
Chimp t <sub>1/2</sub> (h)	5.0	4.5	
Chimp CL (L/h/kg)	1.2	1.3	
Protein binding (human, %)	96.3	93.1	
Caco-2, Papp (cm/s)	11 × 10 <sup>-6</sup>	2.8 × 10 <sup>-6</sup>	
Intrinsic clearance (L/h/kg)	0.97	0.96	
Mouse CCR3 IC <sub>50</sub> (nM)	54	3.6	
Mouse chemotaxis IC <sub>50</sub> (nM)	41	7.0	
Mouse eotaxin challenge EC <sub>50</sub> (mg/kg)	20	1.5	
Mouse OVA challenge (% inhibition at 100 mg/kg bid)	86%	82%	
			240
			100
			490
			440
			570
			20,000
			160
			56,000
			800
			28,000
			1400
			>100,000
			2100

# View of data in ChEMBL

Ingredient	Molweight	Standard Type	Relation	Standard Value	Standard Units	Assay Type	Description	Assay Src Description	Assay Organism	Target Type	Target Name	Target Organism	Reference
Chiral  <a href="#">CHEMBL195433</a>	486.65	IC50	=	240	nM	A	<a href="#">Inhibition of human recombinant CYP2D6</a>	Scientific Literature	Homo sapiens	SINGLE PROTEIN	<a href="#">Cytochrome P450 2D6</a>	Homo sapiens	<a href="#">Bioorg. Med. Chem. Lett. (2007) 17:11:2992</a>
Chiral  <a href="#">CHEMBL250689</a>	465.6	IC50	=	1000	nM	B	<a href="#">Inhibition of dopamine D2 receptor</a>	Scientific Literature		SINGLE PROTEIN	<a href="#">Dopamine D2 receptor</a>	Homo sapiens	<a href="#">Bioorg. Med. Chem. Lett. (2007) 17:11:2992</a>
Chiral  <a href="#">CHEMBL250689</a>	465.6	CL	=	35	mL.min-1.kg-1	A	<a href="#">Clearance in cynomolgus monkey</a>	Scientific Literature	Macaca fascicularis	ORGANISM	<a href="#">Cynomolgus monkey</a>	Macaca fascicularis	<a href="#">Bioorg. Med. Chem. Lett. (2007) 17:11:2992</a>

Compound details

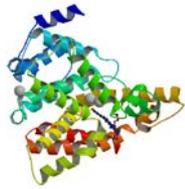
Activity details

Assay details

Target details Reference

# ChEMBL Targets:

## Protein



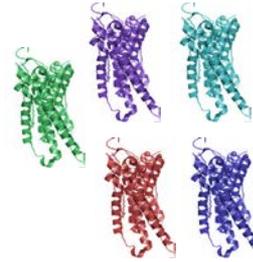
e.g., PDE5

## Protein complex



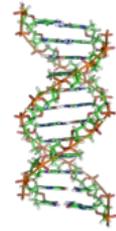
e.g., Nicotinic acetylcholine receptor

## Protein family



e.g., Muscarinic receptors

## Nucleic Acid



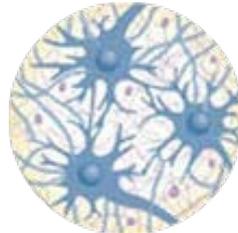
e.g., DNA

## Cell Line



e.g., HEK293 cells

## Tissue



e.g., Nervous

## Sub-cellular Fraction



e.g., Mitochondria

## Organism



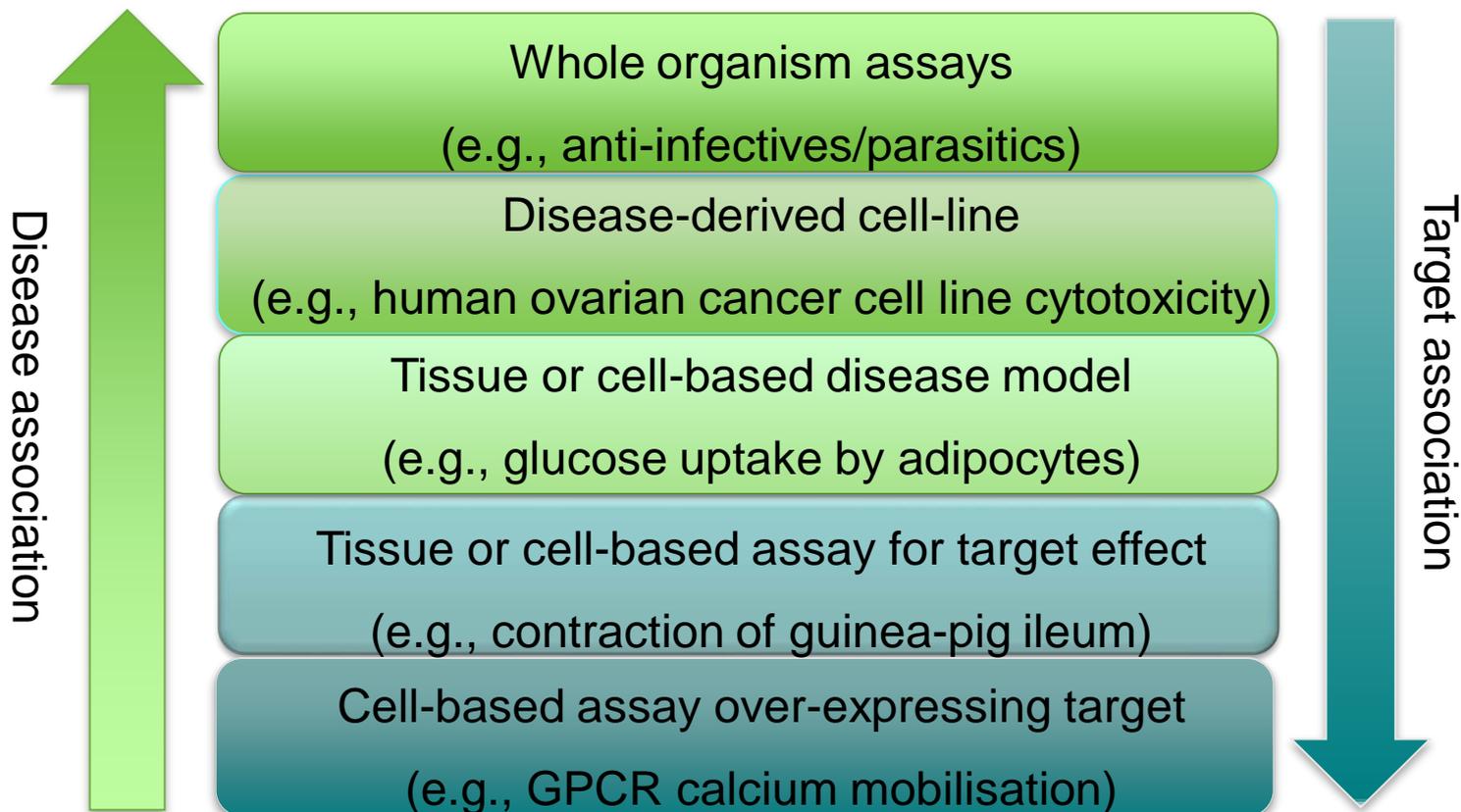
e.g., Drosophila

# ChEMBL Assays – Binding, Functional, ADMET

## Binding:

Assays which directly measure the binding of a compound to a particular target  
E.g., competition binding assays with a radioligand

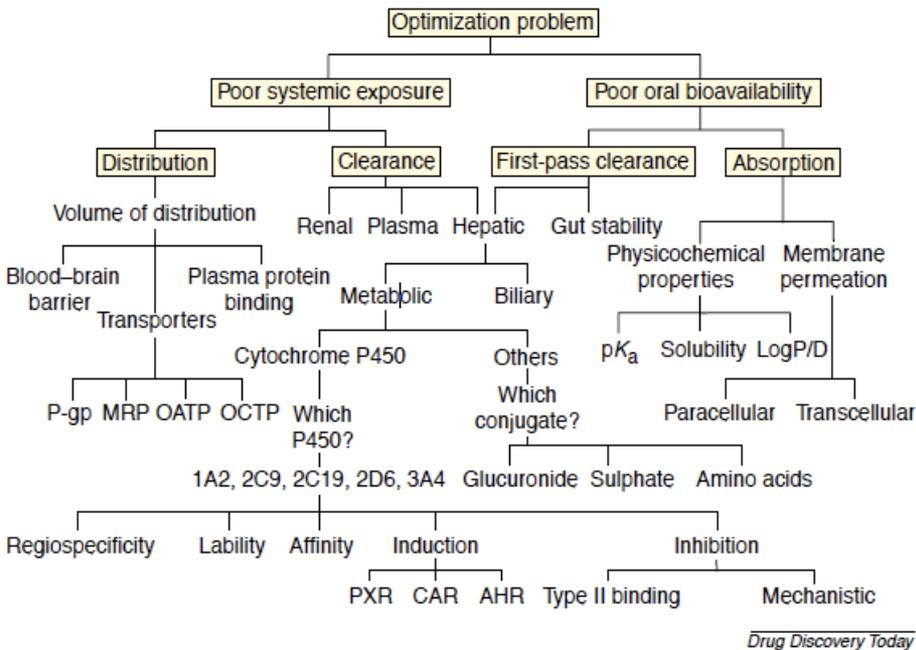
## Functional:



# ADMET:

## ADME

- Absorption, Distribution, Metabolism, Excretion



*Drug Discovery Today*

## Toxicity

- Hepatotoxicity, nephrotoxicity, cytotoxicity, cardiotoxicity (hERG, QT prolongation),

Cardiotoxicity in male Hartley guinea pig assessed as QT prolongation by electrocardiogram

Cardiotoxicity in guinea pig Langendorff heart model assessed as QT prolongation

Increase in QT prolongation in anesthetized dog up to 22 uM plasma level

Cardiotoxicity in dog assessed as QT prolongation at 100 mg/kg, po BID for 28 days

Increase in QT prolongation in anesthetized dog up to 0.60 uM plasma level

Hepatotoxicity using isolated ICR mouse hepatocytes by inclusion body formation assay (In Vitro)

Hepatotoxicity in Swiss albino mouse assessed as plasma ALT level at 3 mg/kg, po after 30 days

Hepatotoxicity in rat assessed as change in cell morphology after 30 hrs by HE staining

Cytotoxicity against human fibroblasts (MRC-5) cells

In vitro cytotoxicity against A-549 human lung cancer cells

# Accessing ChEMBL Data

The screenshot shows the ChEMBL website interface. On the left, there is a navigation menu with buttons for ChEMBLdb, Malaria Data, ChEMBL-NTD, Kinase SARfari, GPCR SARfari, DrugEBllity, ChEMBL Group, Downloads, Web Services, and FAQ. Below this is a 'ChEMBLdb Statistics' section with various metrics. A 'ChEMBL Blog' section is also visible. The main content area features a search bar and a 'ChEMBL Downloads' table. Two green arrows point from the 'Downloads' button in the navigation menu to the 'ChEMBL Downloads' table, and from the 'Web Services' button to the 'Getting Started' section below the table.

Name	Current Release	Last Update	Description
<a href="#">ChEMBLdb</a>	16	May 2013	ChEMBL Database downloads, which includes Oracle, MySQL and PostgreSQL versions of the database.
<a href="#">ChEMBLNTD</a>	NA	January 2013	Datasets made available via the ChEMBL-NTD website. Current dataset contributors include <a href="#">here</a>
<a href="#">ChEMBL-RDF</a>	16.0	May 2013	RDF Version of ChEMBL database. The file download format is turtle and the minor version is 0.
<a href="#">DrugEBllity</a>	2.0	September 2012	Flatfile downloads for DrugEBllity system. Main website link <a href="#">here</a>
<a href="#">GPCRSARfari</a>	3.00	June 2012	Flatfile downloads for GPCR SARfari system. Main website link <a href="#">here</a>
<a href="#">KinaseSARfari</a>	5.01	December 2011	Flatfile downloads for Kinase SARfari system. Main website link <a href="#">here</a>
<a href="#">MalariaData</a>	2.0	January 2013	Release notes for Malaria Data website, a searchable database which contains compounds and targets.
<a href="#">VEHICLe</a>	1.0	April 2010	Dataset described in 'Heteroaromatic Rings of the Future', by Pitt et al., more details <a href="#">here</a>

**Getting Started**

- Search [target data](#)
- Search [compound data](#)
- Search [associated data](#)

**Support and Feedback**

We positively encourage [read more](#).

**Staying in Touch**

To keep up to date with ChEMBL [read more](#).

**Training**

The group run a series of [webinars](#) [read more](#).

**Data Licensing**

Access to the web interface of [Attribution-Share Alike 3.0 Unported License](#)

**Acknowledgements**

Many [people](#) have contributed to ChEMBL, and also some [INTACT](#) teams, and also some [people](#)

**HTTP Response Code**

HTTP Response Code	Description
200	OK. The request to the web service completed successfully.
400	Bad request. The parameters passed to the API endpoint were deemed invalid. This response will be returned for invalid ChEMBLID's i.e. CHEMBLX1, invalid UniProt accession numbers, etc.
404	Not found. The resource corresponding to the supplied parameters does not exist. This response will be returned for requests for non-existent ChEMBL compound, target, or activity.
500	Service unavailable. An internal problem prevented us from fulfilling your request.

**How to use the ChEMBL REST API**

We have provided a [Java](#) client and also [Perl](#) and [Python](#) scripts to help get you started with using the ChEMBL RESTful Web Service API.

**Getting Started with Java**

The chemblRestClient java client contains all of the dependencies necessary for interacting with the ChEMBL REST Web Service API. Below is an example java application that demonstrates how to use the ChEMBL REST API client.

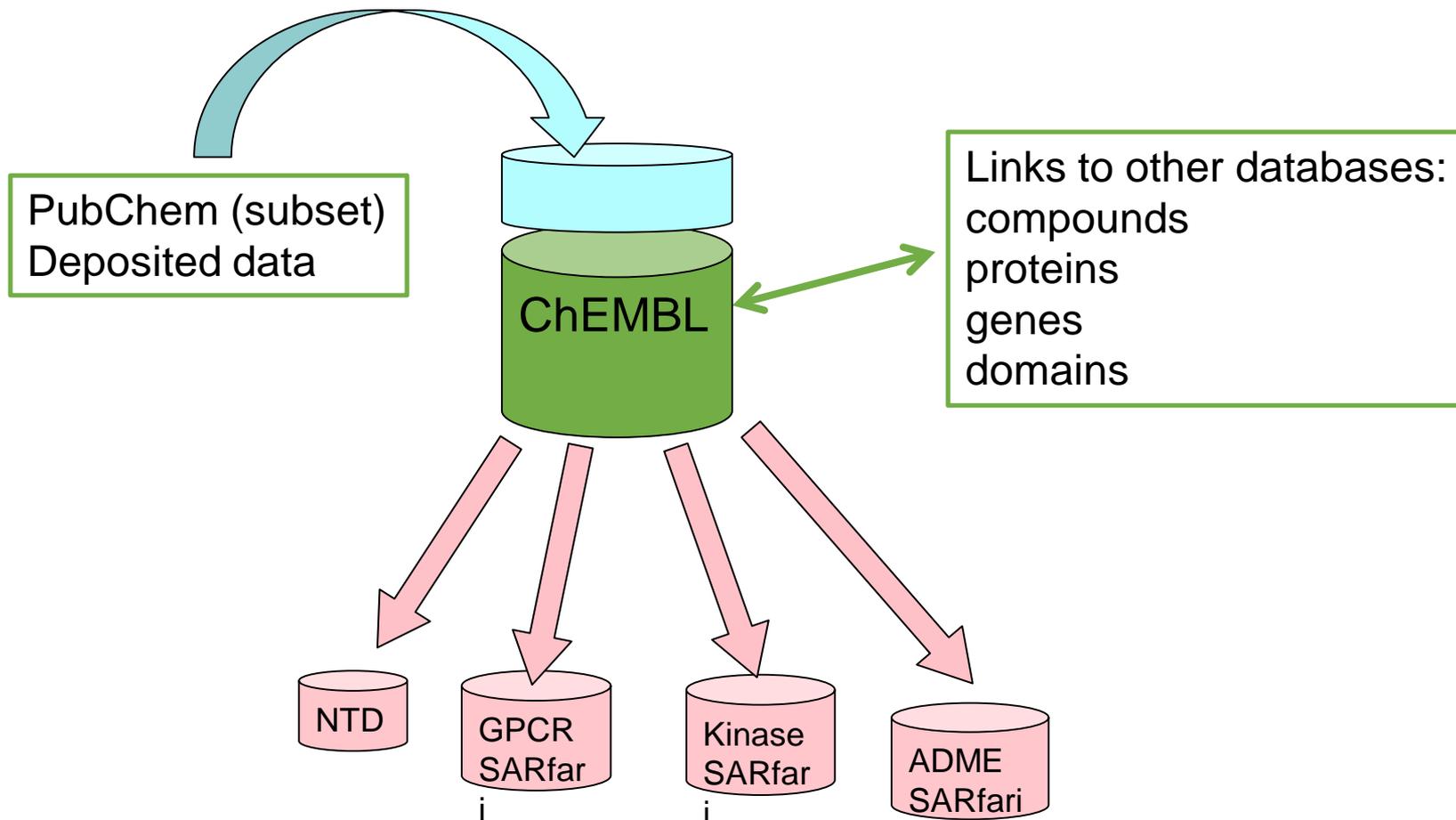
1. Download the [chemblRestClient](#) jar file
2. Copy the 'Example' class code below to 'Example.java'
3. Compile [Example.java](#)

```
javac -cp ./chemblRestClient-1.0.0.jar Example.java
```

4. Run the executable

```
java -cp ./chemblRestClient-1.0.0.jar Example
```

# Integration with other Public Resources



# Information on a Target – Target Report Card

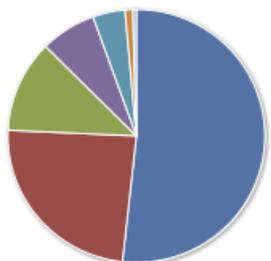
## Target Report Card

### Target Name and Classification

Target ID	CHEMBL182
Target Type	PROTEIN
Preferred Name	Phosphodiesterase 5
Synonyms	cGMP-specific phosphodiesterase type 5
Organism	Homo sapiens
Protein Target Classification	enzyme phosphodiesterase

### Target Associated Bioactivities

#### ChEMBL Activity Types for Target



Total: 2539

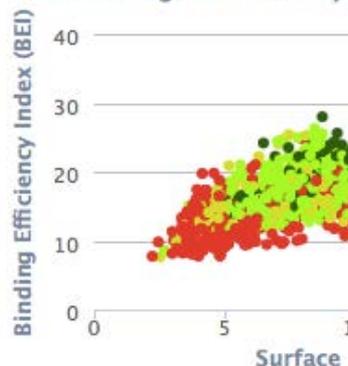
### Target Associated Assays

#### ChEMBL Assays for Target



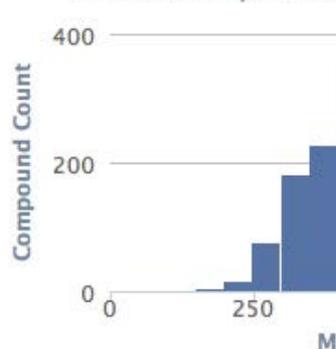
### Target Ligand Efficiencies

#### ChEMBL Ligand Efficiency



### Target Associated Compound Properties

#### ChEMBL Compounds



### Target Cross References - Gene

<a href="#">Array Express</a>	<a href="#">ENSG00000138735</a>
<a href="#">Ensembl</a>	<a href="#">ENSG00000138735</a>
<a href="#">GO Cellular Component</a>	<a href="#">GO:0005829</a> (cytosol)
<a href="#">GO Molecular Function</a>	<a href="#">GO:0008270</a> (zinc ion binding) <a href="#">GO:0030553</a> (cGMP binding) <a href="#">GO:0047555</a> (3',5'-cyclic-GMP phosphodiesterase activity)
<a href="#">GO Biological Process</a>	<a href="#">GO:0007165</a> (signal transduction) <a href="#">GO:0030168</a> (platelet activation)
<a href="#">Wikipedia</a>	<a href="#">cGMP-specific phosphodiesterase type 5</a>

### Target Cross References - Protein

<a href="#">Human Protein Atlas</a>	<a href="#">ENSG00000138735</a>
<a href="#">UniProt</a>	<a href="#">O76074</a> (A0AV69 A8K2C4 O75026 O75887 Q86UI0 Q86V66 Q9Y6Z6)
<a href="#">Reactome</a>	<a href="#">REACT_604</a> (Hemostasis.)

### Target Cross References - Domain

<a href="#">InterPro</a>	<a href="#">IPR002073</a> (PDEase_catalytic_dom.) <a href="#">IPR003018</a> (GAF.) <a href="#">IPR003607</a> (Metal-dep_PHydrolase_HD_dom.) <a href="#">IPR023088</a> (PDEase.) <a href="#">IPR023174</a> (PDEase_CS.)
<a href="#">Pfam</a>	<a href="#">PF00233</a> (PDEase_I) <a href="#">PF01590</a> (GAF)

### Target Cross References - Structure

<a href="#">PDBe</a>	<a href="#">1RKP</a> <a href="#">1T9R</a> <a href="#">1T9S</a> <a href="#">1TBF</a> <a href="#">1UDT</a> <a href="#">1UDU</a> <a href="#">1UHO</a> <a href="#">1XOZ</a> <a href="#">1XP0</a> <a href="#">2CHM</a> <a href="#">2H40</a> <a href="#">2H42</a> <a href="#">2H44</a> <a href="#">3B2R</a> <a href="#">3BJC</a> <a href="#">3HC8</a> <a href="#">3HDZ</a> <a href="#">3JWQ</a> <a href="#">3JWR</a> <a href="#">3LFV</a> <a href="#">3MFO</a>
----------------------	--

# Compound Search – Compound Report Card

### Compound Report Card

**Compound Name and Classification**

Compound ID	CHEMBL192
Compound Name	SILDENAFIL
Synonyms	UK-92480, Sildenafil, UK-9 SILDENAFIL, SILDENAFIL
Max Phase	4 (Approved)
Trade Names	Viagra, Revatio

**Compound Representations**

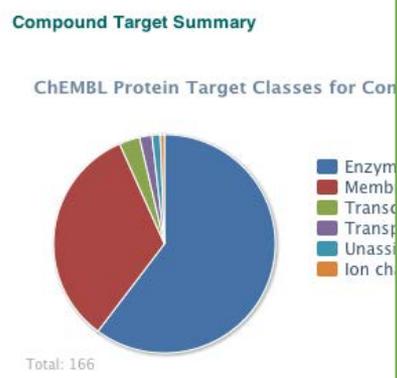
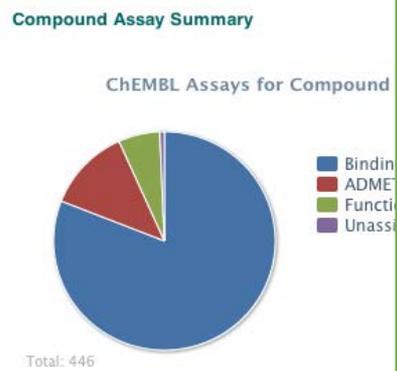
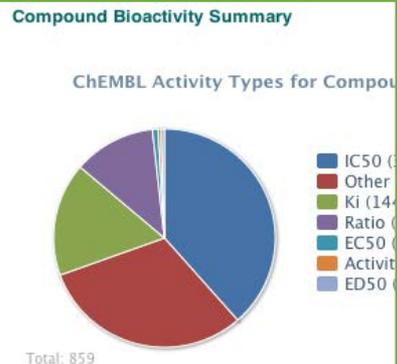
Molfile	<a href="#">Download MolFile</a>
Canonical SMILES	CCCc1nn(C)c2C(=O)N
Standard InChI	InChI=1S/C22H30N6O <a href="#">Download InChI</a>
Standard InChI Key	BNRNXUUZRGQAQC

**Molecule Features**

**Alternate Forms of Compound in ChEMBL**

CHEMBL192

CHEMBL173



**Clinical Trials for Compound**

Number of clinical trials registered at <a href="http://clinicaltrials.gov">clinicaltrials.gov</a>	<a href="#">240</a>
--	---------------------

**Calculated Compound Parent Properties**

Mol. Weight	ALogP	#Ro5 Violations	#Rotatable Bonds	Ro3	Med Chem Friendly	ACD Acidic pKa	ACD Basic pKa	ACD LogP	ACD LogD pH7.4	Molecular Species
474.6	2.25	0	7	No	Yes	10.05	6.03	2.47	2.45	NEUTRAL

**Compound Cross References**

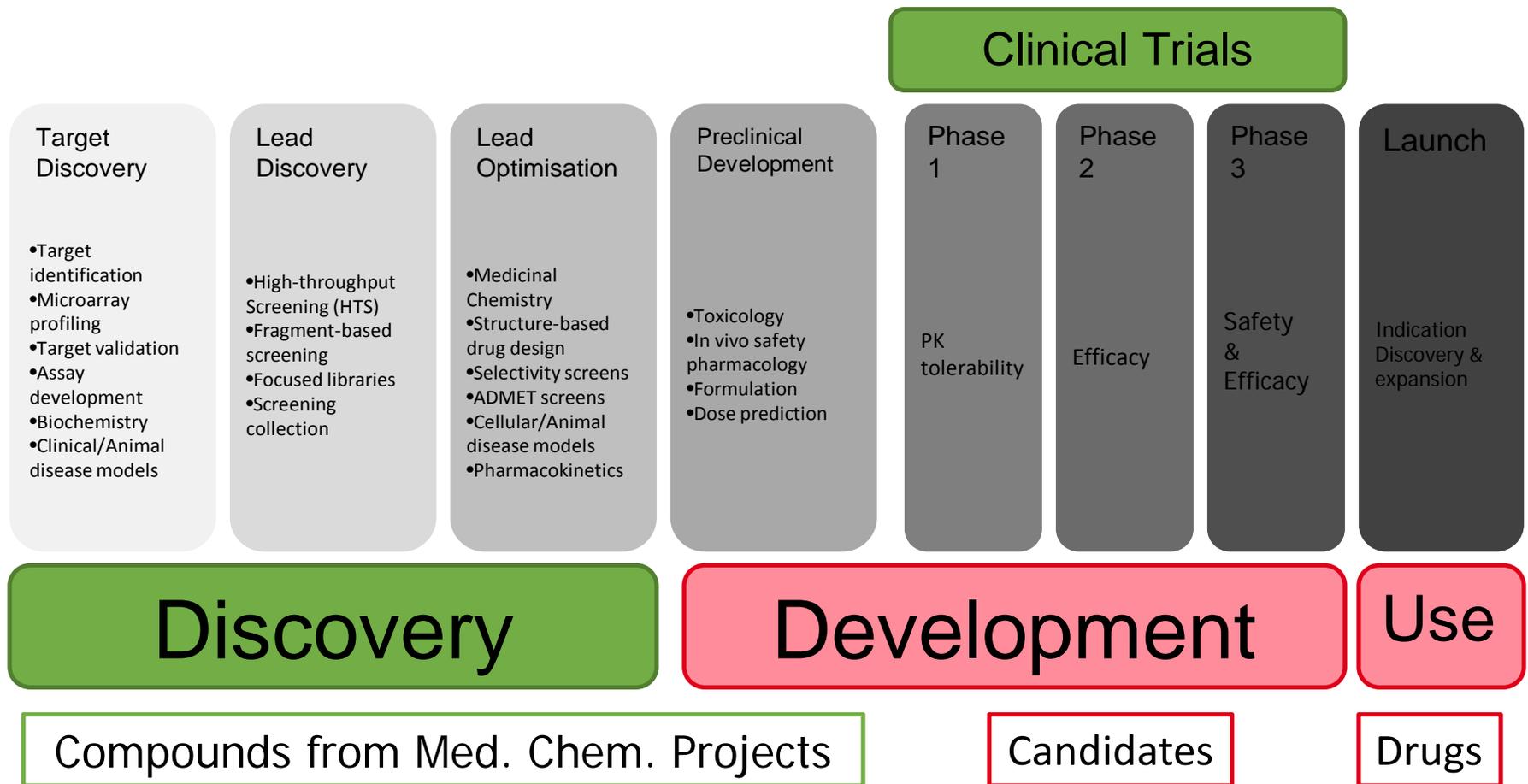
<a href="#">ChEBI</a>	<a href="#">ChEBI:9139</a>
<a href="#">ChemSpider</a>	<a href="#">ChemSpider:BNRNXXUZRGAQC-UHFFFAOYSA-N</a>
<a href="#">PubChem</a>	<a href="#">SID: 26748898</a> <a href="#">SID: 50085897</a>
<a href="#">Wikipedia</a>	<a href="#">Sildenafil</a>

**UniChem Cross References**

<a href="#">DrugBank</a>	<a href="#">DB00203</a>
<a href="#">PDBe</a>	<a href="#">VIA</a>
<a href="#">KEGG Ligand</a>	<a href="#">C07259</a>
<a href="#">ZINC</a>	<a href="#">ZINC19796168</a>
<a href="#">eMolecules</a>	<a href="#">902463</a>
<a href="#">IBM Patent System</a>	<a href="#">D814FFE26EDA163F0CDC1115AD9C7CC3</a>
<a href="#">IBM Strategic IP</a>	<a href="#">WO200007597A1</a> <a href="#">WO1999066933A1</a> <a href="#">EP1020190A3</a> <a href="#">US6066735</a> <a href="#">EP1027887A2</a> <a href="#">US6087362</a> <a href="#">EP0812845A1</a> <a href="#">WO2000054773A1</a> <a href="#">WO200000199A1</a> <a href="#">WO2000038655A1</a> <a href="#">WO2000054777A1</a> <a href="#">US5250534</a> <a href="#">WO2000054774A1</a> <a href="#">EP0995441A3</a> <a href="#">WO1999066924A1</a> <a href="#">WO2000004875A2</a> <a href="#">WO2000072827A2</a> <a href="#">WO1999059584A1</a> <a href="#">EP1027888A3</a> <a href="#">WO2000059475A1</a> <a href="#">WO2000012076A1</a> <a href="#">WO2000010542A2</a> <a href="#">WO2000003721A1</a> <a href="#">WO1999060985A2</a> <a href="#">WO1999064033A1</a> <a href="#">WO2000067735A2</a> <a href="#">WO1999067231A1</a> <a href="#">WO1999051252A1</a> <a href="#">WO1999030697A2</a> <a href="#">WO1999066870A1</a> <a href="#">EP1037616A2</a> <a href="#">WO2000044363A2</a> <a href="#">WO1999039763A1</a> <a href="#">WO2000007596A1</a> <a href="#">WO2000006121A1</a> <a href="#">WO2000078760A1</a> <a href="#">WO1999021558A2</a> <a href="#">WO2000000212A1</a> <a href="#">US6037346</a> <a href="#">WO1998055176A1</a> <a href="#">WO2000040226A2</a> <a href="#">EP1027054A1</a> <a href="#">EP1027887A3</a> <a href="#">WO2000045795A2</a> <a href="#">EP0967214A1</a> <a href="#">WO2000043012A1</a> <a href="#">WO2000051978A1</a> <a href="#">WO1999027905A1</a> <a href="#">WO2000057857A1</a> <a href="#">WO2000050007A1</a> <a href="#">EP0960621A2</a> <a href="#">WO2000074652A1</a> <a href="#">WO1999020251A1</a> <a href="#">US6075028</a> <a href="#">WO2000015233A1</a> <a href="#">WO1999030688A1</a> <a href="#">EP0812845B1</a> <a href="#">WO2000053148A2</a> <a href="#">EP0995441A2</a> <a href="#">US5955611</a> <a href="#">WO1999038507A1</a> <a href="#">EP0951908A2</a> <a href="#">WO2000066084A1</a> <a href="#">WO2000012110A2</a> <a href="#">EP1027888A2</a> <a href="#">WO2000042992A2</a> <a href="#">EP0916675A2</a> <a href="#">WO1999002161A1</a>

UniChem REST Service Call: [https://www.ebi.ac.uk/unicem/rest/verbose\\_inchikey/BNRNXXUZRGAQC-UHFFFAOYSA-N](https://www.ebi.ac.uk/unicem/rest/verbose_inchikey/BNRNXXUZRGAQC-UHFFFAOYSA-N)

# Discovery to Development to Market

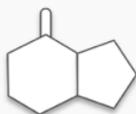
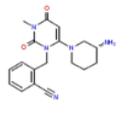
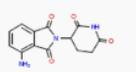


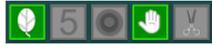
# Drugs and Clinical Candidates

Also added and annotated:

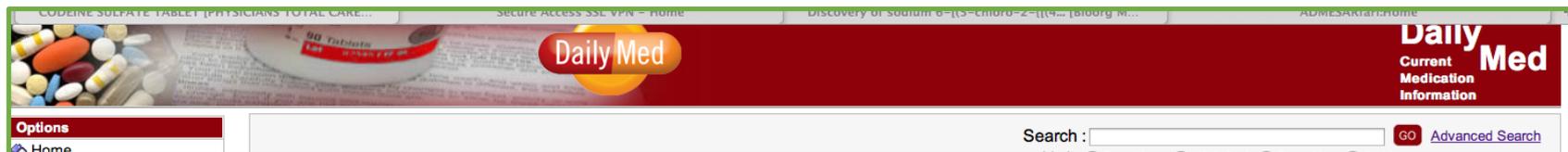
FDA approved drugs from orange book (~1800)

All compounds with USAN/INN names (~10,000)

Parent Molecule	Synonyms	Phase	Research Codes	Applicants	USAN Stem	USAN Year	First Approval	ATC Code	Icon
 CHEMBL2107849	Mipomersen Sodium (FDA, USAN)	4	ISIS 301012	Genzyme Corp	-rsen	2007	2013	C10AX11	
 CHEMBL376359	Alogliptin (INN) Alogliptin Benzoate (FDA, USAN)	4	SYR-322	Takeda Pharmaceuticals Usa Inc	-gliptin	2008	2013	A10BH04	
 CHEMBL43452	Pomalidomide (INN, USAN)	4	CC-4047 IMID 3		-domide	2006	2013		

 CHEMBL71752	Vinpocetine (JAN, USAN, INN)		AY-27,255		vin-	1981		N06BX18	
 CHEMBL2110748	Vinepidine (INN) Vinepidine Sulfate (USAN)			Lilly	vin-	1983			

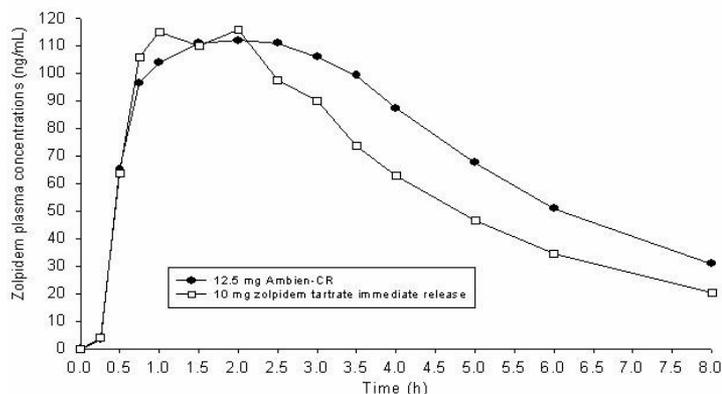
# DailyMed Data - Work in Progress



## 12.3 Pharmacokinetics

Ambien CR exhibits biphasic absorption characteristics, which results in rapid initial absorption from the gastrointestinal tract similar to zolpidem tartrate immediate-release, then provides extended plasma concentrations beyond three hours after administration. A study in 24 healthy male subjects was conducted to compare mean zolpidem plasma concentration-time profiles obtained after single oral administration of Ambien CR 12.5 mg and of an immediate-release formulation of zolpidem tartrate (10 mg). The terminal elimination half-life observed with Ambien CR (12.5 mg) was similar to that obtained with immediate-release zolpidem tartrate (10 mg). The mean plasma concentration-time profiles are shown in Figure 1.

Figure 1: Mean plasma concentration-time profiles for Ambien CR (12.5 mg) and immediate-release zolpidem tartrate (10 mg)



In adult and elderly patients treated with Ambien CR, there was no evidence of accumulation after repeated once-daily dosing for up to two weeks.

### Absorption:

Following administration of Ambien CR, administered as a single 12.5 mg dose in healthy male adult subjects, the mean peak concentration ( $C_{max}$ ) of zolpidem was 134 ng/mL (range: 68.9 to 197 ng/mL) occurring at a median time ( $T_{max}$ ) of 1.5 hours. The mean AUC of zolpidem was 740 ng·hr/mL (range: 295 to 1359 ng·hr/mL).

A food-effect study in 45 healthy subjects compared the pharmacokinetics of Ambien CR 12.5 mg when administered while fasting or within 30 minutes after a meal. Results demonstrated that with food, mean AUC and  $C_{max}$  were decreased by 23% and 30%, respectively, while median  $T_{max}$  was increased from 2 hours to 4 hours. The half-life was not changed. These results suggest that, for faster sleep onset, Ambien CR should not be administered with or immediately after a meal.

### Distribution:

Total protein binding was found to be  $92.5 \pm 0.1\%$  and remained constant, independent of concentration between 40 and 790 ng/mL.

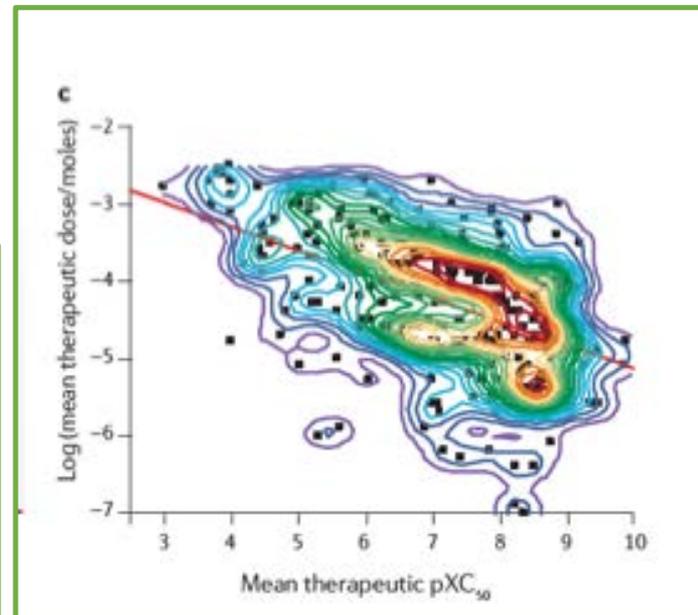
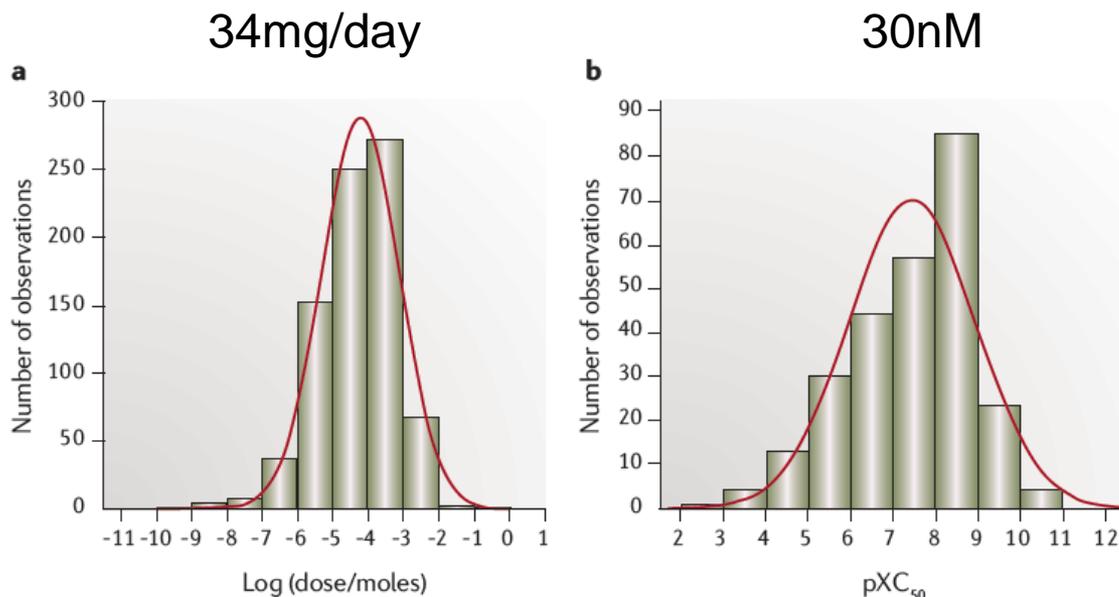
### Metabolism:

Zolpidem is converted to inactive metabolites that are eliminated primarily by renal excretion.

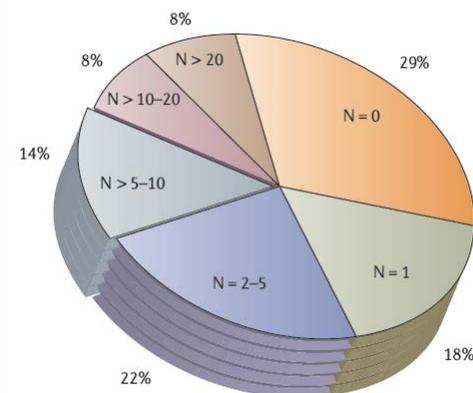
### Elimination:

When Ambien CR was administered as a single 12.5 mg dose in healthy male adult subjects, the mean zolpidem elimination half-life was 2.8 hours (range: 1.62 to 4.05 hr).

# Probing the Link between Potency, ADMET and Physicochemical Parameters



Number of targets with <1uM potency



Analysis of ChEMBL data shows:

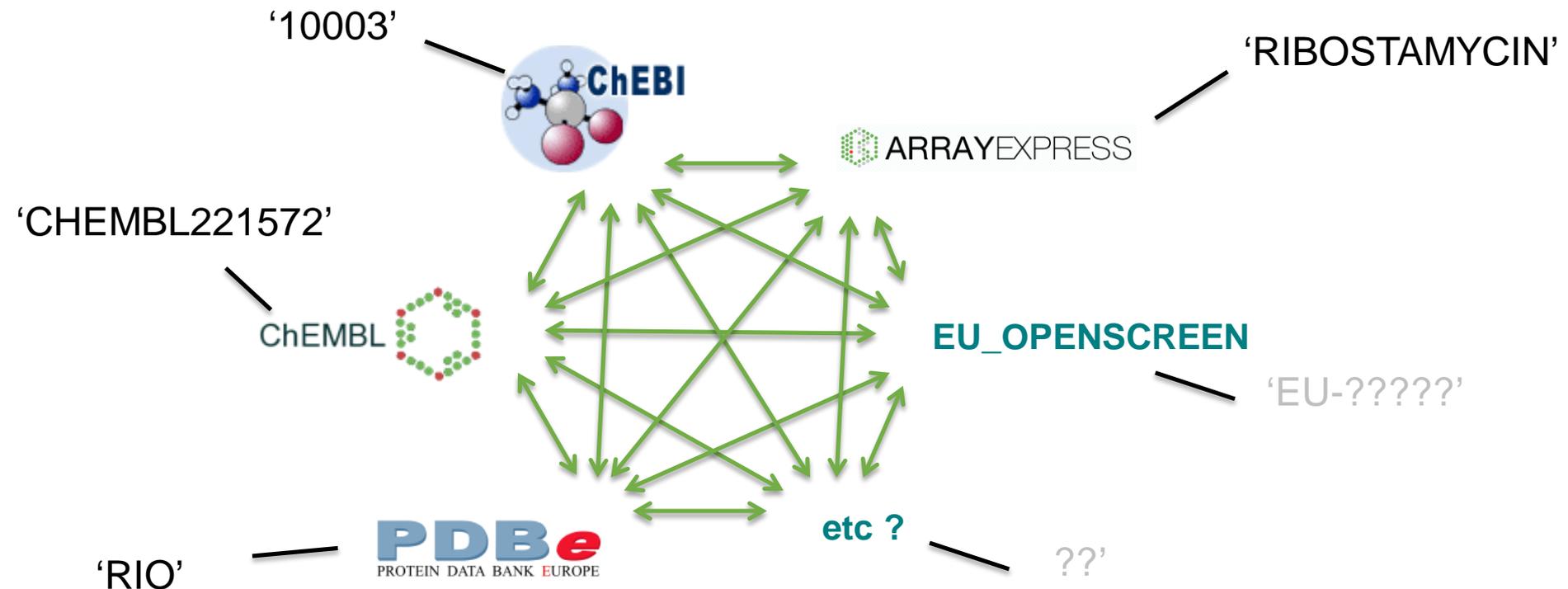
- Average dose for oral drugs is not <10mg/day
- Average potency for oral drugs is not 1nM
- Dose is only weakly correlated with potency
- ~50% of drugs bind to >1 target with <1uM potency

# UniChem

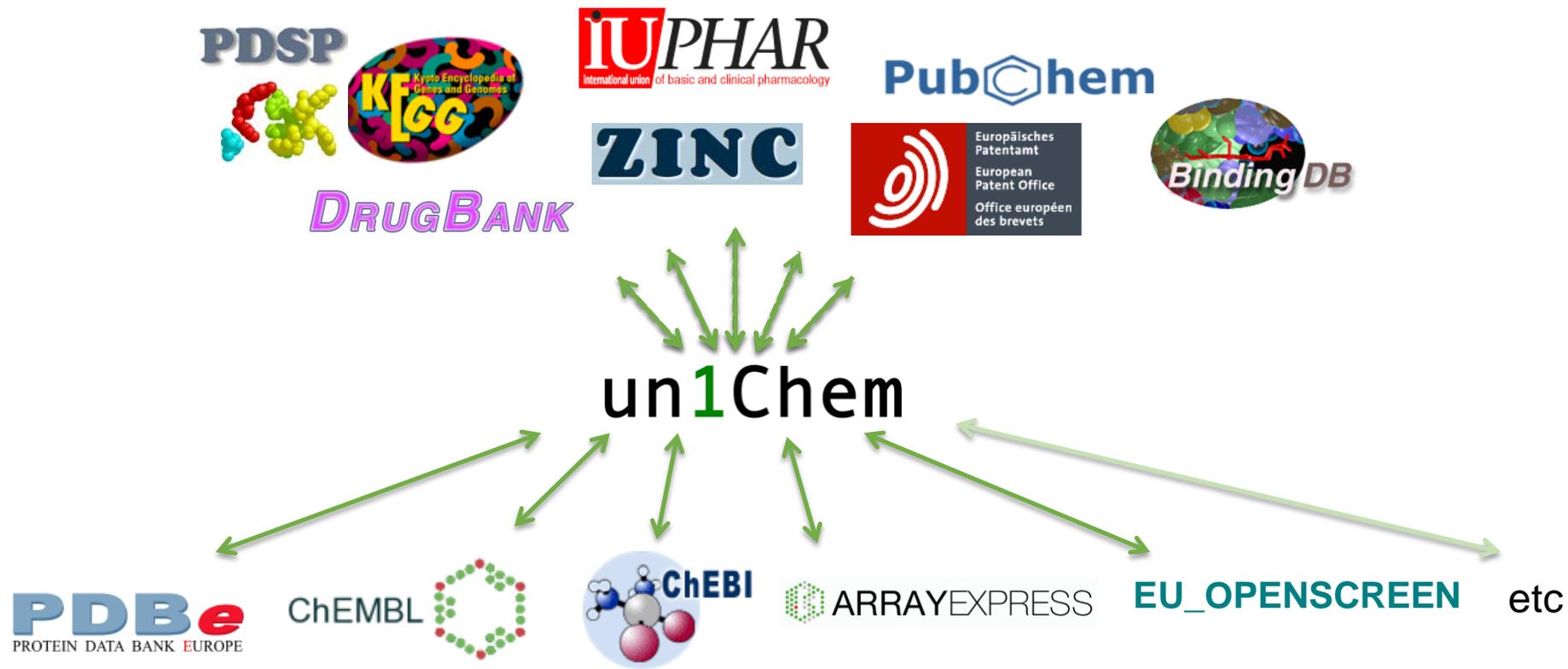
Unified Chemical Structure Cross-referencing System

# Multiple EBI Resources hold Compound Structure Data

- Maintaining links between DBs is a manual/time consuming for each source.
- Business rules for constructing identifiers not consistent – users confused.



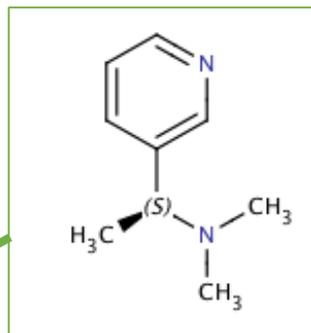
# Links to EBI and external sources are shared



- All EBI DBs share the maintenance overhead of creating links to each other
- All EBI DBs share the benefits of maintained links to external resources
- The UniChem mapping service is freely available to all external users

# UniChem

- Rapid cross-referencing of chemical structures and their identifiers between databases
- InChi based 'Unified Chemical Identifier' system



InChI

1S/C9H14N2/c1-8(11(2)3)9-5-4-6-10-7-9/h4-8H,1-3H3/t8-/m0/s1

Molecular  
Formula

Connectivity

Hydrogens

Stereochemistry

InChIKey

MSDVRBYHBSTAOM-QMMMGPBSA-N

DATABASE

Open Access

## UniChem: a unified chemical structure cross-referencing and identifier tracking system

Jon Chambers<sup>1\*</sup>, Mark Davies<sup>1</sup>, Anna Gaulton<sup>1</sup>, Anne Hersey<sup>1</sup>, Sameer Velankar<sup>2</sup>, Robert Petryszak<sup>3</sup>, Janna Hastings<sup>4</sup>, Louisa Bellis<sup>1</sup>, Shaun McGlinchey<sup>1</sup> and John P. Overington<sup>1</sup>

### Abstract

UniChem is a freely available compound identifier mapping service on the internet, designed to optimize the efficiency with which structure-based hyperlinks may be built and maintained between chemistry-based resources. In the past, the creation and maintenance of such links at EMBL-EBI, where several chemistry-based resources exist, has required independent efforts by each of the separate teams. These efforts were complicated by the different data models, release schedules, and differing business rules for compound normalization and identifier nomenclature that exist across the organization. UniChem, a large-scale, non-redundant database of Standard InChIs with pointers between these structures and chemical identifiers from all the separate chemistry resources, was developed as a means of efficiently sharing the maintenance overhead of creating these links. Thus, for each source represented in UniChem, all links to and from all other sources are automatically calculated and immediately available for all to use. Updated mappings are immediately available upon loading of new data releases from the sources. Web services in UniChem provide users with a single simple automatable mechanism for maintaining all links from their resource to all other sources represented in UniChem. In addition, functionality to track changes in identifier usage allows users to monitor which identifiers are current, and which are obsolete. Lastly, UniChem has been deliberately designed to allow additional resources to be included with minimal effort. Indeed, the recent inclusion of data sources external to EMBL-EBI has provided a simple means of providing users with an even wider selection of resources with which to link to, all at no extra cost, while at the same time providing a simple mechanism for external resources to link to all EMBL-EBI chemistry resources.

**Keywords:** UniChem, InChI, InChIKey, Chemical databases, Data integration

### Background

There is much data available in the public domain on the structures, effects and interactions of small molecules with biological systems. Many research projects benefit from scientists having easy access to data from these diverse

An alternative to such full-scale integration is to simply provide the user with links or bridges between the separate resources. This alternative suffers from the shortfall of not providing the user with a single point from which all integrated resources can be searched, and requires the

# UniChem - Sources

UniChem				
Table of Sources...				
UniChem currently contains data from the sources listed below. Follow the links on the short names for more detailed information on each source...				
Show	25	entries	Apply filter:	...to whole table
src_id	Short name	Full name	Description	Process of Data Acquisition
1	<a href="#">chembl</a>	ChEMBL	A database of bioactive drug-like small molecules and associated bioactivities abstracted from the scientific literature	Standard InChIs and Keys provided on ftp site for each release
2	<a href="#">drugbank</a>	DrugBank	A database of drugs (i.e. sequences)	
3	<a href="#">pdb</a>	PDBe (Protein Data Bank Europe)	The European macromolecular	
4	<a href="#">iuphar</a>	International Union of Basic and Clinical Pharmacology	A resource for small molecule	
5	<a href="#">pubchem_dof</a>	PubChem ('Drugs of the Future' subset)	A subset of the	
6	<a href="#">kegg_ligand</a>	KEGG (Kyoto Encyclopedia of Genes and Genomes) Ligand	KEGG LIGAND and ENZYME	
7	<a href="#">chebi</a>	ChEBI (Chemical Entities of Biological Interest).	ChEBI is a free	
8	<a href="#">nih_ncc</a>	NIH Clinical Collection	Collections of Assembled	
9	<a href="#">zinc</a>	ZINC	A free database Laboratory in (UCSF). [In	
10	<a href="#">emolecules</a>	eMolecules	A free chemical availabilities	
11	<a href="#">ibm</a>	IBM strategic IP insight platform and the National Institutes of Health	A massive, scientific and scientific	
12	<a href="#">atlas</a>	Gene Expression Atlas	The Gene Expression Atlas is a semantically enriched database of meta-analysis based summary statistics over a curated subset of ArrayExpress Archive, servicing queries for condition-specific gene expression patterns as well as broader exploratory searches for biologically interesting genes/samples.	Currently extracted from compound names.
13	<a href="#">patents</a>	IBM strategic IP insight platform and the National	Data, provided by IBM-NIH, was originally extracted from patents from three publishing bodies (US, EPO and WIPO) with publication dates through (including) 2000-12-31. For UniChem, these data were parsed to include only whole molecules present in either the title or claims fields. Further filters included removal of: 1. All molecules mapping to > 10,000 patents. 2. Non-organic molecules. 3. Small molecules (mw < 90	SMILES download available. Converted to InChi in house. Patent Ids used for Ids instead of cpd_ids. Data set filtered to remove compounds not appearing in the

## UniChem

### Stats Summary of current UniChem Content.

A number of global parameters are measured after each data load...

Show 25 entries Apply filter: ...to whole table

Parameter No.	Parameter	Value
1	Last updated	17-MAY-2013
2	Total number of Structures	31992347
3	Total number Assignments*	37554538
4	Number Current Assignments	37527241
5	Number Obsolete Assignments	27297
6	Number of Sources	16



# UniChem – InChi or Identifier Searching

un1Chem

- Home
- Sources
- Stats
- Whole source mapping
- Web Services
- General Information
  - Rules for Loading
  - Getting in touch
  - FAQ

EBI > Databases > Small Molecules > UniChem

## UniChem

UniChem is a 'Unified Chemical Identifier' system, designed to assist in the rapid cross-referencing of chemical structures, and their identifiers.

Use the query form below to search UniChem with structures (InChIs or InChIKeys) or with [src\\_compound\\_id](#)'s from various [sources](#). Some example queries are given at the foot of the page to help you get started. For formatting tips go [here](#).

Queries return a list of `src_compound_id`-to-structure 'assignments' ( [What's an 'assignment?'](#) ) and related information.

For larger queries, users are strongly advised to use the [web services](#) instead.

Fields that are **highlighted** are required.

Query:

BNRNXUZRQQAQC-UHFFFAOYSA-N

Type of data:

`src_compound_id`  InChI  InChIKey

\* Also, for '`src_compound_id`' queries....

Source:

-select-

Include obsolete `src_compound_id` assignments:  Yes

(Recommended default: leave unchecked. See footnote\*)

Search

\* Footnote: Querying with `src_compound_id`'s requires additional information to be specified (ignored for InChI and InChIKey queries).

1. Since `src_compound_id`'s may be ambiguous between different sources, it is necessary to specify the source of `src_compound_id`'s when a

# UniChem – InChi or Identifier Searching

## InChi/InChiKey Searching

### Identifier Searching

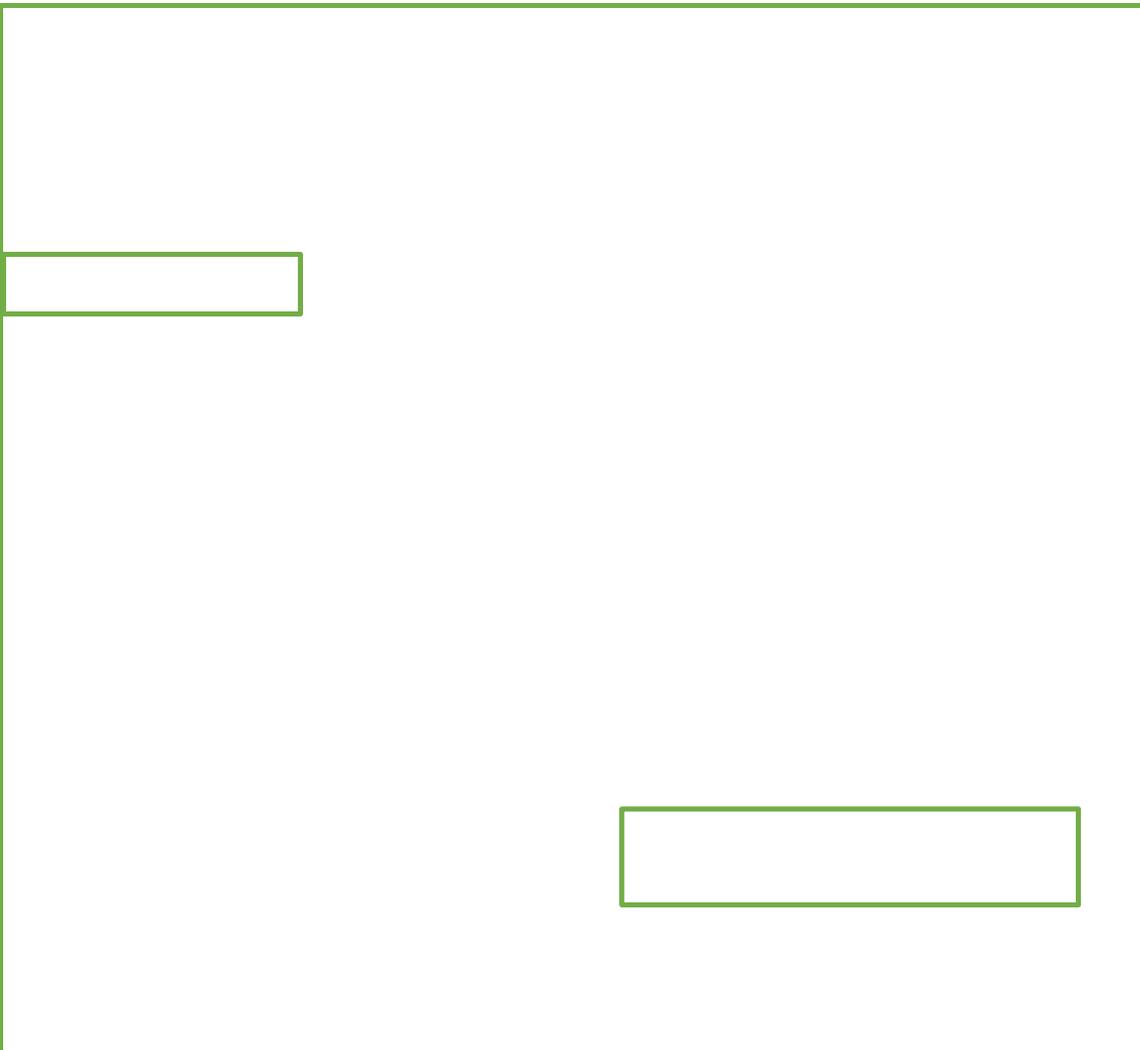
src_id	Source Name	src_compound_id	Currently Assigned	LR *	UCI **	Standard InChiKey
1	chembl	<a href="#">CHEMBL741</a>	Yes		619287	PYZRQGJRPPTADH-UHFFFAOYSA-N
2	drugbank	<a href="#">DB00555</a>	Yes		619287	PYZRQGJRPPTADH-UHFFFAOYSA-N
4	iuphar	<a href="#">2622</a>	Yes		619287	PYZRQGJRPPTADH-UHFFFAOYSA-N
5	pubchem_dotf	<a href="#">12013463</a>	Yes		619287	PYZRQGJRPPTADH-UHFFFAOYSA-N
7	ebi	<a href="#">6267</a>	Yes		619287	PYZRQGJRPPTADH-UHFFFAOYSA-N
8			Yes		619287	PYZRQGJRPPTADH-UHFFFAOYSA-N
9			Yes		619287	PYZRQGJRPPTADH-UHFFFAOYSA-N
10	emolecules	<a href="#">578746</a>	Yes		619287	PYZRQGJRPPTADH-UHFFFAOYSA-N
11	ibm	<a href="#">8678217FE9CF4D9BDABB88EC196F5EC7</a>	Yes		619287	PYZRQGJRPPTADH-UHFFFAOYSA-N

Show 15 entries

Apply filter:

src_id	Source Name	src_compound_id	Currently Assigned	LR *	UCI **	Standard InChiKey
1	chembl	<a href="#">CHEMBL192</a>	Yes		584480	BNRNXUUZRGQAQC-UHFFFAOYSA-N
2	drugbank	<a href="#">DB00203</a>	Yes		584480	BNRNXUUZRGQAQC-UHFFFAOYSA-N
3	pdb	<a href="#">VIA</a>	Yes		584480	BNRNXUUZRGQAQC-UHFFFAOYSA-N
6	kegg_ligand	<a href="#">C07259</a>	Yes		584480	BNRNXUUZRGQAQC-UHFFFAOYSA-N
9	zinc	<a href="#">ZINC19796168</a>	Yes		584480	BNRNXUUZRGQAQC-UHFFFAOYSA-N
10	emolecules	<a href="#">902463</a>	Yes		584480	BNRNXUUZRGQAQC-UHFFFAOYSA-N
11	ibm	<a href="#">D814FFE26EDA163F0CDC1115AD9C7CC3</a>	Yes		584480	BNRNXUUZRGQAQC-UHFFFAOYSA-N
13	patents	<a href="#">WO2000007596A1</a>	Yes		584480	BNRNXUUZRGQAQC-UHFFFAOYSA-N
13	patents	<a href="#">WO2000010542A2</a>	Yes		584480	BNRNXUUZRGQAQC-UHFFFAOYSA-N
13	patents	<a href="#">US6066735</a>	Yes		584480	BNRNXUUZRGQAQC-UHFFFAOYSA-N
13	patents	<a href="#">WO2000044363A2</a>	Yes		584480	BNRNXUUZRGQAQC-UHFFFAOYSA-N

# UniChem – Source Mapping



## UniChem

### Mapping Results...

The results of your query are shown below. '3926' records were returned. ...

Record No	'From' source ids	'To' source ids
1	DB04129	HWD
2	DB08302	NRO
3	DB07249	7X1
4	DB07236	797
5	DB07755	FBL
6	DB07097	4BB
7	DB01802	7A8
8	DB07220	740
9	DB00163	VIV
10	DB07959	IDZ
11	DB08033	K02
12	DB07112	4HG
13	DB07346	AEH
14	DB02919	DTM
15	DB04541	GIO
16	DB02580	PG6
17	DB07912	HPO
18	DB06957	2D9
19	DB08594	T2M
20	DB06924	23N

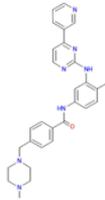
# UniChem and ChEMBL

- Compound cross references dynamically called from report card page

KTUFNOKKBVMGRW-UHFFFAOYSA-N

un1Chem

**Compound Name and Classification**

Compound ID	CHEMBL941	 CHEMBL941
Compound Name	IMATINIB	
Synonyms	IMATINIB, IMATINIB, Gleevec, STI-571, Gleevec, IMATINIB MESYLATE	
Max Phase	4 (Approved)	
Trade Names	Gleevec	

**Compound Representations**

Molfile	<a href="#">Download MolFile</a>
Canonical SMILES	<chem>CN1CCN(Cc2ccc(cc2)C(=O)Nc3ccc(C)c(Nc4nccc(n4)c5ocnc5)c3)CC1</chem>
Standard InChI	InChI=1S/C29H31N7O/c1-21-5-10-25(18-27(21)34-29-31-13-11-26( ... <a href="#">Download InChI</a>
Standard InChI Key	KTUFNOKKBVMGRW-UHFFFAOYSA-N

**Molecule Features**



**UniChem Cross References**

<a href="#">DrugBank</a>	<a href="#">DB00619</a>
<a href="#">PDBe</a>	<a href="#">STI</a>
<a href="#">ChEBI</a>	<a href="#">45783</a>
<a href="#">ZINC</a>	<a href="#">ZINC19632618</a>
<a href="#">eMolecules</a>	<a href="#">876446</a>
<a href="#">IBM Patent System</a>	<a href="#">2C349E68BE42FC7B2B0FDD5080E27BB3</a>
<a href="#">Atlas</a>	<a href="#">imatinib</a>
<a href="#">FDA SRS</a>	<a href="#">BKJ8M8G5HI</a>
<a href="#">SureChem</a>	<a href="#">SureCN3827</a>
<a href="#">PharmGKB</a>	<a href="#">PA10804</a>

[https://www.ebi.ac.uk/unichem/rest/verbose\\_inchikey/KTUFNOKKBVMGRW-UHFFFAOYSA-N](https://www.ebi.ac.uk/unichem/rest/verbose_inchikey/KTUFNOKKBVMGRW-UHFFFAOYSA-N)

# Useful Information

ChEMBL Website

<https://www.ebi.ac.uk/chembl/db>

## ChEMBL Paper (NAR 2011)

Nucleic Acids Research Advance Access published September 23, 2011

*Nucleic Acids Research*, 2011, 1-8  
doi:10.1093/nar/gkr777

### ChEMBL: a large-scale bioactivity database for drug discovery

Anna Gaulton<sup>1</sup>, Louisa J. Bellis<sup>1</sup>, A. Patricia Bento<sup>1</sup>, Jon Chambers<sup>1</sup>, Mark Davies<sup>1</sup>, Anne Hersey<sup>1</sup>, Yvonne Light<sup>1</sup>, Shaun McGlinchey<sup>1</sup>, David Michalovich<sup>2</sup>, Bissan Al-Lazikani<sup>3</sup> and John P. Overington<sup>1,\*</sup>

<sup>1</sup>EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, <sup>2</sup>David Michalovich Scientific Consulting, London and <sup>3</sup>Cancer Research UK Cancer Therapeutics Unit, Institute of Cancer Research, 15 Cotswold Road, Belmont, Surrey, SM2 5NG, UK

Received August 15, 2011; Accepted September 5, 2011

#### ABSTRACT

ChEMBL is an Open Data database containing binding, functional and ADMET information for a large number of drug-like bioactive compounds. These data are manually abstracted from the primary published literature on a regular basis, then further curated and standardized to maximize their quality and utility across a wide range of chemical biology and drug-discovery research problems. Currently, the database contains 5.4 million bioactivity measurements for more than 1 million compounds and 5200 protein targets. Access is available through a web-based interface, data downloads and web services at: <https://www.ebi.ac.uk/chembl/db>.

However, bioactivity data published in journal articles are usually found in a relatively unstructured format and are labour-intensive to search and extract. For example, compound structures are frequently depicted only as images and are not therefore searchable, protein targets may be referred to by a variety of synonyms or abbreviations with no reference to any database identifiers, and details of assays may be included only in Supplementary Data or by reference to previous publications. In addition, there is not currently any requirement by most journals for authors to deposit small-molecule assay results in public databases (as is the case for sequence, protein structure and gene expression data). Historically, therefore, the majority of the published small-molecule bioactivity data have only been readily available via commercial products.

In recent years, in response to the growing demand for open access to this kind of information, a variety of

ChEMBL Blog:

<http://chembl.blogspot.com>

## The ChEMBL-og - Open Data For Drug Discovery

The news, progress, whereabouts, and ephemera from the Computational Chemical Biology group at the EMBL-EBI.

Resources: [ChEMBL database](#) [ChEMBL-NTD](#) [ChEMBL-Malaria](#) [GPCR SARfari](#) [Kinase SARfari](#) [UniChem](#) [DrugEPI](#)

19

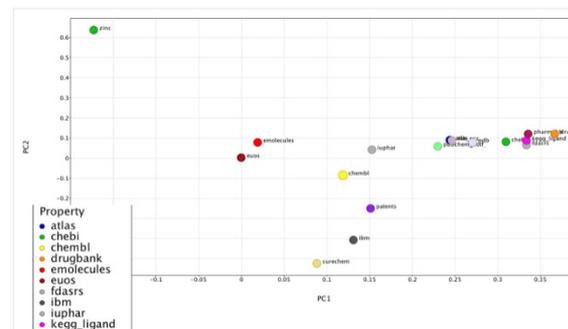
Saturday, 15 June 2013

### Clustering of a few chemical databases

#### About the ChEMBL-og

The ChEMBL-og covers the activities of the Computational Chemical Biology Group at the EMBL-EBI in Hinxton. Our activities include **Drug Discovery, Bioinformatics, Chemoinformatics, Structural Biology, Genetics, Pharmacogenomics, Toxicology, Open Data, The Semantic Web, and Data Integration**. Resources we produce include...

- ChEMBL - a drug discovery bioactivity database



For ChEMBL news and data releases subscribe to:

<http://listserver.ebi.ac.uk/mailman/listinfo/chembl-announce>

# Acknowledgements

- John Overington
- Anna Gaulton
- Mark Davies
- Patricia Bento
- Jon Chambers
- Francis Atkinson
- Louisa Bellis
- Yvonne Light
- George Papadatos
- Shaun McGlinchey
- Nathan Dedman
- Michal Nowotka
- Ruth Akhtar

**welcome**trust

