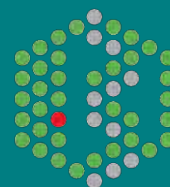


# ChEMBL Database:

## Meeting Chemical and Biological Information needs of Scientists of the Future

Anne Hersey

EMBL-EBI

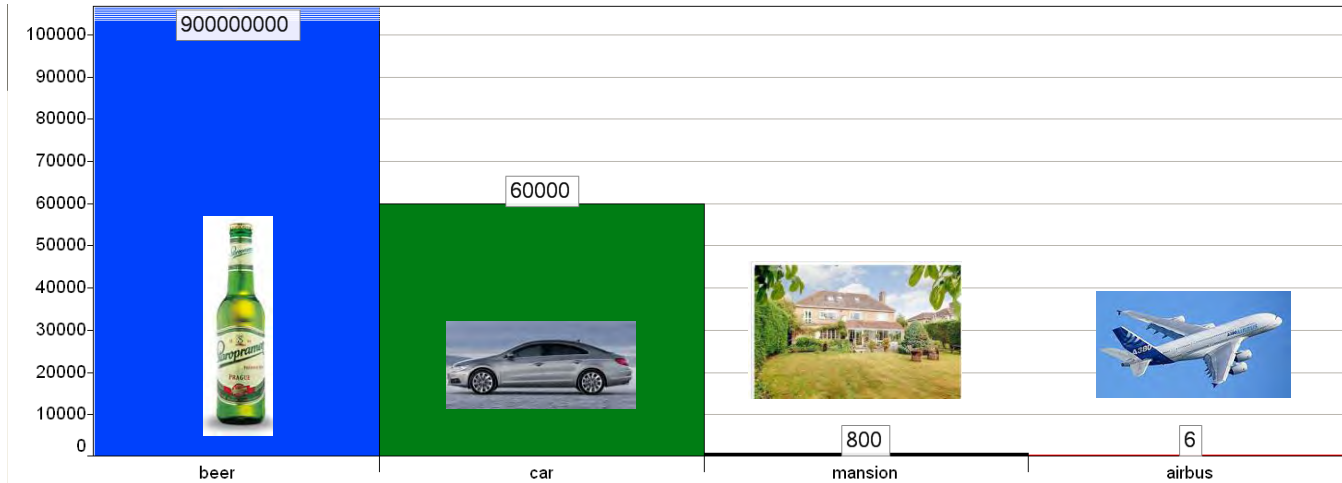


# Outline

- Background
- ChEMBL
  - What is it
  - Applications
  - Data Integration
- Challenges for the future

# Getting Drugs to Market is difficult and costly

- Cost of bringing a new drug to market is \$1.8billion



- Clinical development (Phase 1-3) accounts for 63% of cost
- Only 8% of new molecules make it from candidate selection to a marketed drug
- Takes 13.5 years to discover and develop a drug
- Need ~10 molecules a year entering clinical development to achieve 1 launched drug per year



# Changing Landscape

- Patent Cliff
- Pharma Site Closures

- More academic, SME research
- Precompetitive initiatives

## GlaxoSmithKline to shed 380 research jobs in Harlow

A pharmaceutical giant is to close one of its research units in Essex with the loss of about 380 jobs.

One third of the workforce is to be made redundant after projects for pain relief, anxiety and depression drugs



BBC Essex  
Sport, travel, what to do, features and more

SEE ALSO

Pistoia Alliance | Lowering Barriers to R&D Innovation

Search

Newsletter Sign-Up | Contact Us | Join | Member Login

ABOUT WORKING GROUPS INDUSTRY CHALLENGES NEWS & EVENTS OUTPUTS BLOG

Precompetitive Collaboration.

Infosys, AstraZeneca, the Royal Society of Chemistry, and EMBL-EBI are doing it.

Are you ready to collaborate?

## AstraZeneca announces plans to close Loughborough site

One of Leicestershire's major employers is to shut its operation with the loss of almost 1,200 jobs.

Managers at AstraZeneca's research facility on the outskirts of Loughborough are to be made redundant.

The firm has said it will be looking for other sites to move its research and development work to.



WHERE I LIVE



BBC Leicester  
Sport, travel, weather, things to do, features and much more

amid profit rise

1 February 2011 Last updated at 18:02

3.3K Share

## Pfizer to close UK research site

Drug maker Pfizer is to close its research and development (R&D) facility in Kent, which employs 2,400 people.

The move has raised concerns that the UK is losing highly-skilled jobs and about the private sector's ability to absorb cuts in the public sector.

The Unite union said the roles were "exactly the sort of jobs we need to keep in this country".



Pfizer described the facilities in Sandwich as "excellent"

Business Secretary Vince Cable said the firm's

## GSK and Online Communities Create Unique Alliance to Stimulate Open Source Drug Discovery for Malaria

Like Tweet Share +1

-- GSK becomes first company to freely share chemical structures on 13,500 molecules from its compound library

-- Alliances formed with leading scientific research communities from private industry and public-domain data providers

# ChEMBL Database

## Press Release

EMBL-EBI



### Open access to large-scale drug discovery data

Data on drugs and small molecules is placed in the public domain, helping the discovery and development of new medicines

**Hinxton, 23 July 2008** – The Wellcome Trust has awarded £4.7 million (€5.8 million) to EMBL's European Bioinformatics Institute (EMBL-EBI) to support the transfer of a large collection of information on the properties and activities of drugs and a large set of drug-like small molecules from publicly listed company Galapagos NV to the public domain. It will be incorporated into the EMBL-EBI's collection of open-access data resources for biomedical research and will be maintained by a newly established team of scientists at the EMBL-EBI. These data lie at the heart of translating information from the human genome into successful new drugs in the clinic.

The human genome sequence provided a molecular 'parts list' for a human being, comprising all the genes and proteins that are encoded by our genetic blueprint. But to develop new medicines, it is important to catalogue how each of these 'parts' interacts with drugs and drug-like molecules. This interface of the genome with chemistry is a core part of the new scientific area of chemogenomics. For the past eight years, researchers at BioFocus DPI, the service division of Galapagos, have been integrating the existing collections of information in these two areas to develop a set of well-structured chemogenomic databases that can be used to help determine whether a particular molecule has the right properties to make an effective drug. BioFocus DPI licensed this information to pharmaceutical and biotech companies worldwide. As part of the Wellcome Trust grant announced today, the EBI will obtain the rights to the databases from BioFocus DPI. The award will make it possible to provide free access to this information for all researchers. "The scientific community worldwide will greatly benefit from unrestricted access to these data. It will aid their efforts in predictive drug discovery," says Galapagos CEO Onno

van de Stolpe. "Galapagos has successfully accelerated its research programs with these, and BioFocus DPI used the data to deliver on its contracts with customers. After this transfer, which we hope will contribute to the advancement of drug discovery research by improving access to the data that we have collected, we will continue to use these resources."

The transfer will empower academia to participate in the first stages of drug discovery for all therapeutic areas, including major diseases of the developing world. In future it could also result in improved prediction of drug side-effects. "We are excited to be able to provide information that defines the effects of a large number of small molecules on the body, and link this to the proteins that these molecules interact with, as part of our mission to provide wide access to bioinformatics tools to promote scientific progress and disseminate cutting-edge technologies to industry," says EMBL-EBI Director Janet Thornton. "With this transfer, we aim to facilitate faster and better drug discovery. It speaks to the importance of this information for translational research that the Wellcome Trust has chosen to support this particular transfer with sufficient long-term funding."

This unprecedented transfer of pharmaceutical data resources from the private sector to the public domain will have the greatest impact on researchers in academia and in small companies on limited budgets. "The Wellcome Trust has a strong commitment to making vital research tools freely available to the academic research community," says Dr Alan Schafer, Head of Molecular and Physiological Sciences at the Wellcome Trust. "Enabling these previously proprietary data to enter the public domain will allow researchers worldwide to make free use of knowledge essential for drug discovery." ●

- Open access bioactivity database (since 2009)
- Manually extracted data from Med Chem literature
- Gives users access to curated SAR data
- Open repository for deposited datasets
- Integration of subset of PubChem data



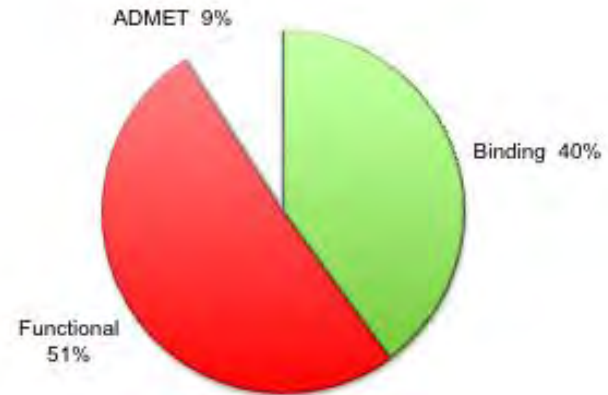


# ChEMBL Database

- Freely available (searchable and downloadable) resource for drug discovery
- Updated regularly with new data
- Secure searching <https://www.ebi.ac.uk/chembl/db>

Assays are classified as:

- Binding measurements
- Functional assays
- ADME/toxicity data



## ChEMBL11

Targets: 8,603

Compounds: 1,060,258

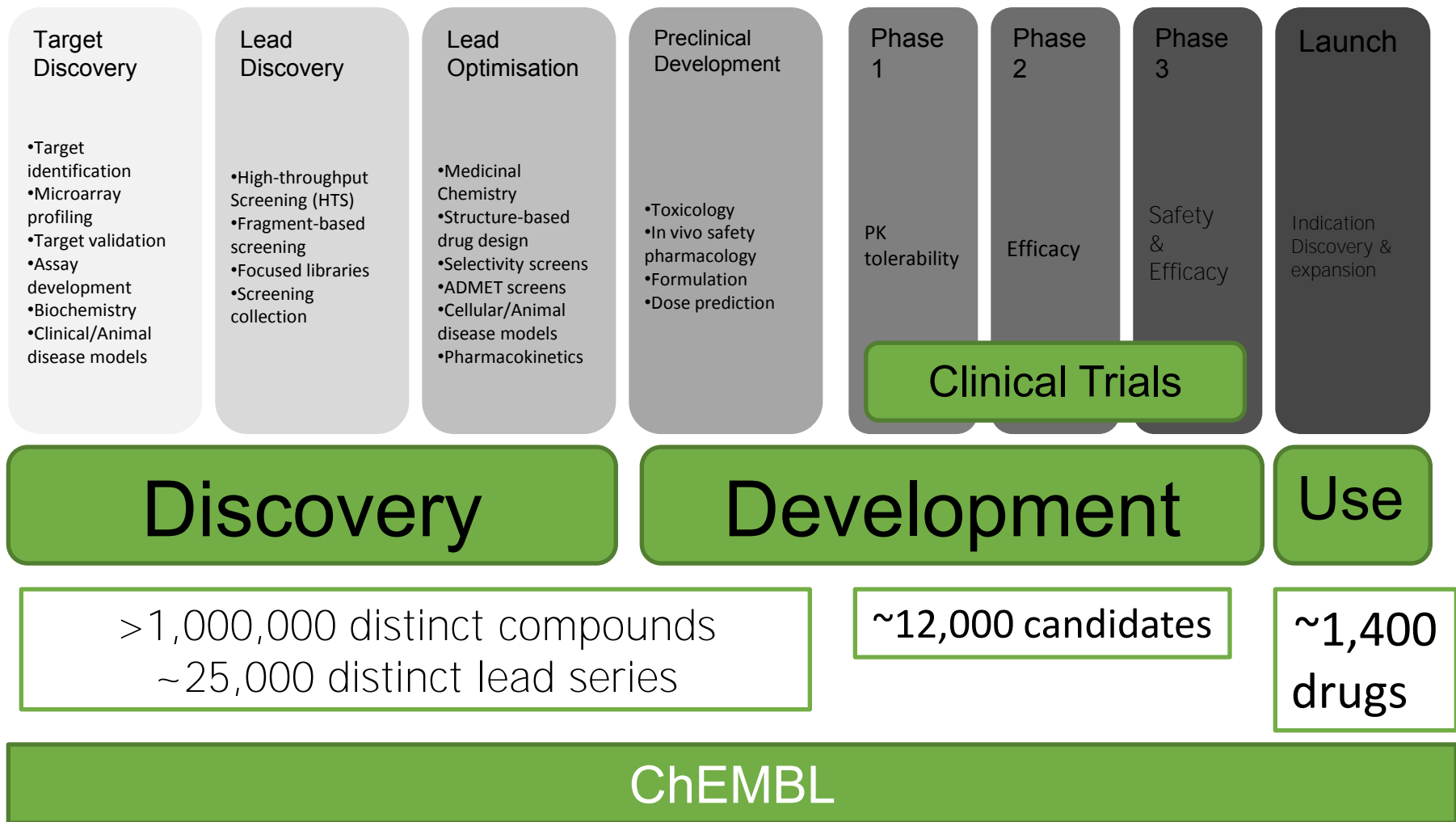
(~350K from PubChem)

Activities: 5,479,146

Publications: 42,516

5420 proteins  
1481 organisms  
1431 cell lines

# Drug Discovery Data in ChEMBL



# Accessing ChEMBL Data

The image shows a composite of three screenshots illustrating how to access ChEMBL data. The top-left screenshot shows the ChEMBL website's navigation menu with 'Web Services' highlighted. The top-right screenshot shows an FTP directory listing for 'ftp://ftp.ebi.ac.uk/pub/databases/chembl/'. The bottom screenshot shows the 'ChEMBL Web Services' page with a list of REST API methods and a code example for using the Java client.

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

Name	Size	Last Modified
BioTorrents		22/02/2011 11:32:00
ChEMBLNTD		19/05/2010 00:00:00
ChEMBLdb		07/02/2011 13:51:00
GPCRSARfari		
KinaseSARfari		
VEHICLE		

**ChEMBL Web Services**

**ChEMBL REST API Methods**

Using the ChEMBL web service API users can retrieve data from the ChEMBL database in a programmatic fashion. The following list defines the methods available:

1. [Get compound by ChEMBLID](#)
2. [Get compound by Standard InChiKey](#)
3. [Get individual compound bioactivities](#)
4. [Get target by ChEMBLID](#)
5. [Get target by UniProt Accession Identifier](#)
6. [Get target by RefSeq Accession Identifier](#)
7. [Get individual target bioactivities](#)
8. [Get all targets](#)
9. [Get assay by ChEMBLID](#)
10. [Get individual assay bioactivities](#)

**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 are the following steps:

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

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

4. Run the executable

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

```
import java.util.List;
import org.springframework.context.ApplicationContext;
import org.springframework.context.support.ClassPathXmlApplicationContext;
import uk.ac.ebi.chemblservice.restclient.ChemblRestClient;
import uk.ac.ebi.chemblservice.model.Assay;
import uk.ac.ebi.chemblservice.model.Compound;
import uk.ac.ebi.chemblservice.model.Target;
import uk.ac.ebi.chemblservice.model.Bioactivity;

public class Example
{
    public static void main( String[] args )
    {
        ApplicationContext applicationContext = new ClassPathXmlApplicationContext( "applicationContext.xml" );
        ChemblRestClient chemblClient = applicationContext.getBean( "chemblRestClient", ChemblRestClient.class );

        /***** Uncomment sections as required *****/
    }
}
```

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

# Use Case 1 - Searching by Target

- What is known about chemical structures that bind to a specific protein (ZAP70)?
- What is known about their potency/selectivity/ADMET Properties
- Is there any protein structure data?

# Use Case 1 Searching by target in ChEMBL

EMBL-EBI   [Help](#) [Feedback](#)

[Databases](#) [Tools](#) [Research](#) [Training](#) [Industry](#) [About Us](#) [Help](#) [Site Index](#)

ChEMBL

ChEMBLdb  
ChEMBL-NTD  
Kinase SARfari  
GPCR SARfari  
DrugEBllity  
ChEMBL Group  
Downloads  
Web Services  
FAQ

ChEMBLdb Statistics

- DB: ChEMBL\_11
- Targets: 8,603
- Compound records: 1,195,368
- Distinct compounds: 1,060,258
- Activities: 5,479,146
- Publications: 42,516

ChEMBL Blog

- [Lipinski Seminar - Cancelled](#)
- [Opinion: The Inventive and Creative Industries - Patents and Copyright](#)

EBI > Databases > Small Molecules > ChEMBL Database > Target Search > Target Classification Hierarchy

[Activity Source Filter](#)

Browse  Protein Target Tree  Taxonomy Tree

Click arrows to navigate tree

- Enzyme (3219)
  - Kinase (595)
    - Protein Kinase (593)
      - Ser Thr (408)
      - Tyr (133)
      - Ser Thr Tyr (32)
      - His (10)
      - Guanylate cyclase (4)
      - Ser (3)
      - Endoribonuclease (2)
      - Ser Thr (1)
    - Protease (402)
    - Phosphatase (70)
    - Phosphodiesterase (57)
    - Cytochrome P450 (55)
    - Aminoacyltransferase (1)
    - Reductase (1)
  - Membrane receptor (555)
  - Ion channel (338)
  - Transporter (136)
  - Transcription Factor (102)
  - Cytosolic other (102)
  - Secreted (57)
  - Structural (29)
  - Surface antigen (26)
  - Membrane other (16)
  - Adhesion (14)
  - Nuclear other (13)

Choose Sources to include in search

**Selected Bioactivity Sources**

Selected	Source	Counts
<input checked="" type="checkbox"/>	Scientific Literature	3079587 (65.97%)
<input checked="" type="checkbox"/>	PubChem BioAssays	1473189 (31.56%)
<input checked="" type="checkbox"/>	GSK Malaria Screening	81198 (1.74%)
<input checked="" type="checkbox"/>	Novartis Malaria Screening	22788 (0.49%)
<input checked="" type="checkbox"/>	Sanger Institute Genomics of Drug Sensitivity in Cancer	5984 (0.13%)
<input checked="" type="checkbox"/>	St Jude Malaria Screening	5456 (0.12%)



# Retrieving Bioactivity Data

**Target Report Card**

**Target Details**

Target ID	CHEMBL2803
Target Type	PROTEIN
Preferred Name	Tyrosine-protein kinase ZAP-70
Synonyms	Tyrosine-protein kinase ZAP-70; 70 kDa zeta-associated protein; Syk-related tyrosine kinase
Organism	Homo sapiens
Description	Tyrosine-protein kinase ZAP-70
Protein Target Classification	enzyme kinase protein kinase tyr tk syk

**Database Links**

UniProt	<a href="#">P43403</a>
PDBe	<a href="#">1M61</a> <a href="#">1U59</a> <a href="#">2OQ1</a> <a href="#">2OZO</a>

**Bioactivity Summary**

ChEMBL Activity Types for Target CHEMBL2803

IC50 (252)
Inhibition (135)
Activity (62)
Other (61)
KI (21)
EC50 (21)

Total: 532

**Assay Summary**

ChEMBL Assays for Target CHEMBL2803

Binding (107)
Functional (3)
Unassigned (1)

Total: 111

Summary of bioactivity/assay/compound data available for target

Sequence data

3D Structures

Display all bioactivity data for target

Click pie chart to retrieve particular end-points



Select the activities and conditions you desire:

Activity (#Endpoints)	Condition	Value	
Ki (nM) (21)	<	10000	Add

Activity	Condition	Value	
IC50	lessthan	10000	Delete
Ki	lessthan	10000	Delete

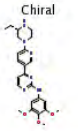
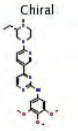
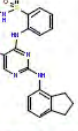
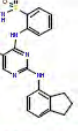
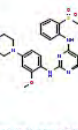
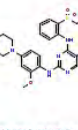
Step 2:

Filter on IC50 <10uM

168 ZAP70 IC50 or Ki values <10uM

ChEMBL Bioactivity Search Results: 168

1 2 3 4 5 6 [Next] [End] Please select...

Parent	Ingredient	Bioactivity	Activity Comment	Operator	Value	Units	Assay ChEMBL ID	Assay Source	Assay Type	Description	ChEMBL Target ID	Target Name	Organism	Target Mapping	Curated By	Reference	Name in Reference
 CHEMBL112346	 CHEMBL112346	IC50		=	8	nM	<a href="#">CHEMBL824029</a>	Scientific Literature	B	Inhibition of Zeta-chain (TCR) associated protein kinase 70 kDa phosphorylation of polyGly-Tyr.	<a href="#">CHEMBL2803</a>	Tyrosine-protein kinase ZAP-70	Homo sapiens	Homologous protein	Expert	<a href="#">Bioorg. Med. Chem. Lett. (1999) 9:23:3351</a>	20
 CHEMBL575048	 CHEMBL575048	IC50		=	10	nM	<a href="#">CHEMBL1037893</a>	Scientific Literature	B	Inhibition of human GST-fused ZAP-70 expressed in Sf9 cells	<a href="#">CHEMBL2803</a>	Tyrosine-protein kinase ZAP-70	Homo sapiens	Protein	Intermediate	<a href="#">Eur. J. Med. Chem. (2009) 44:12:4793</a>	26
 CHEMBL573483	 CHEMBL573483	IC50		=	10	nM	<a href="#">CHEMBL1037893</a>	Scientific Literature	B	Inhibition of human GST-fused ZAP-70 expressed in Sf9 cells	<a href="#">CHEMBL2803</a>	Tyrosine-protein kinase ZAP-70	Homo sapiens	Protein	Intermediate	<a href="#">Eur. J. Med. Chem. (2009) 44:12:4793</a>	1

Compound structures

Activity values

Assay details

Target details

References



# Use Case 2 – Searching by Structure

- What compounds contain a particular substructure?
- What is known about their bioactivities?
- What other data exists
  - Clinical Trials, 3D structures, drug data etc

# Use Case 2 - Structure Searching

The screenshot shows the ChEMBL Compound Search page. At the top, there is a navigation bar with 'Databases', 'Tools', 'Research', 'Training', 'Industry', 'About Us', and 'Help'. Below this is a search bar with the text 'Search ChEMBLdb...' and buttons for 'Compounds', 'Targets', and 'Assays'. A green box labeled 'name' points to the search bar. Below the search bar are several tabs: 'ChEMBLdb', 'Compound Search', 'Protein Target Search', 'Browse Targets', 'Browse Drugs', and 'Drug Approvals'. The 'Compound Search' tab is active. In the center, there is a chemical structure viewer showing a benzimidazole derivative. Below the viewer is a 'Compound Sketcher' dropdown menu with options: 'Please select...', 'JME', 'Marvin', and 'JDraw'. A green box labeled 'Different sketchers' points to this menu. To the right of the structure viewer is a 'List Search' section with radio buttons for 'SMILES Search', 'ChEMBL ID Search', and 'Keyword Search'. A text input field contains the text 'Please enter a list of Compound IDs, keywords, or SMILES separated by ...'. A green box labeled 'Lists of Identifiers' points to this input field. Below the input field is a 'Fetch Compounds' button. On the left side of the page, there is a sidebar with various links and statistics.

Lists of Identifiers

Types of synonyms:

- Research codes
- Trade names
- INN, USAN

Different sketchers

# Compound Report Card

## Summary

Compound ID	CHEMBL192
Compound Name	sildenafil
Synonyms	UK-92480, Sildenafil, UK-92480-10, Sildenafil Citrate, Sildenafil citrate
Approved Drug	Yes
Trade Names	Viagra, Revatio



CHEMBL192

names

## Approved Drug Features



Clinical Trials

## Clinical Trials

Number of clinical trials registered at <a href="http://clinicaltrials.gov">clinicaltrials.gov</a>	324
--	-----

## Parent Properties

Mol. Weight	ALogP	Num Ro5 Violations	Num Rotatable Bonds	Passes Rule-Of-Three	Med Chem Friendly
474.6	2.25	0	7	No	Yes

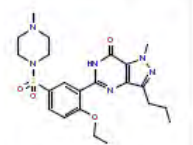
## Parent ACD Properties

Acidic pKa	Basic pKa	LogP	LogD	Species
10.052	6.027	2.468	2.449	NEUTRAL

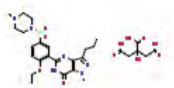
## Representations

Molfile	<a href="#">Download Molfile</a>
Molecular Formula	C22H30N6O4S
Canonical SMILES	<chem>CCCC1nn(C)c2C(=O)NC(=Nc12)c3ccc(ccc3OCC)S(=O)(=O)N4CCN(C)CC4</chem> <a href="#">Download SMILES</a>
Standard InChI	InChI=1S/C22H30N6O4S/c1-5-7-17-19-20(27(4)25-17)22(29)24-21( ... <a href="#">Download InChI</a>
Standard InChI Key	BNRNXUZRGAQC-UHFFFAOYSA-N <a href="#">Download InChI Key</a>

## Molecule Forms



CHEMBL192



CHEMBL1737

Structure

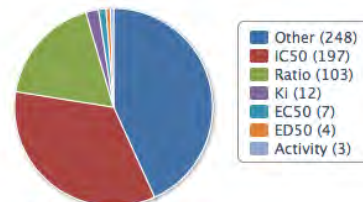
## Database Links

ChEBI	<a href="#">ChEBI:9139</a>
ChemSpider	<a href="#">ChemSpider:BNRN XUZRGAQC-UHFFFAOYSA-N</a>
DrugBank	<a href="#">DB00203</a>
PDBe	<a href="#">VIA - VIA (PDBe Entries)</a>
PubChem	<a href="#">SID: 26748898</a> <a href="#">SID: 50085897</a>
Wikipedia	<a href="#">Sildenafil</a>

Links

## Bioactivity Summary

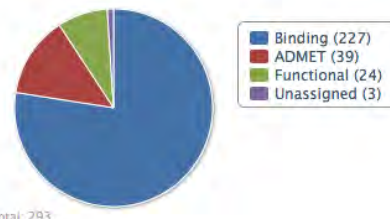
ChEMBL Activity Types for Compound CHEMBL192



Bioactivities

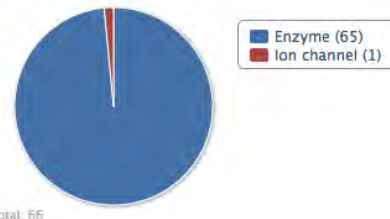
## Assay Summary

ChEMBL Assays for Compound CHEMBL192



## Protein Target Summary

ChEMBL Protein Target Classes for Compound CHEMBL192



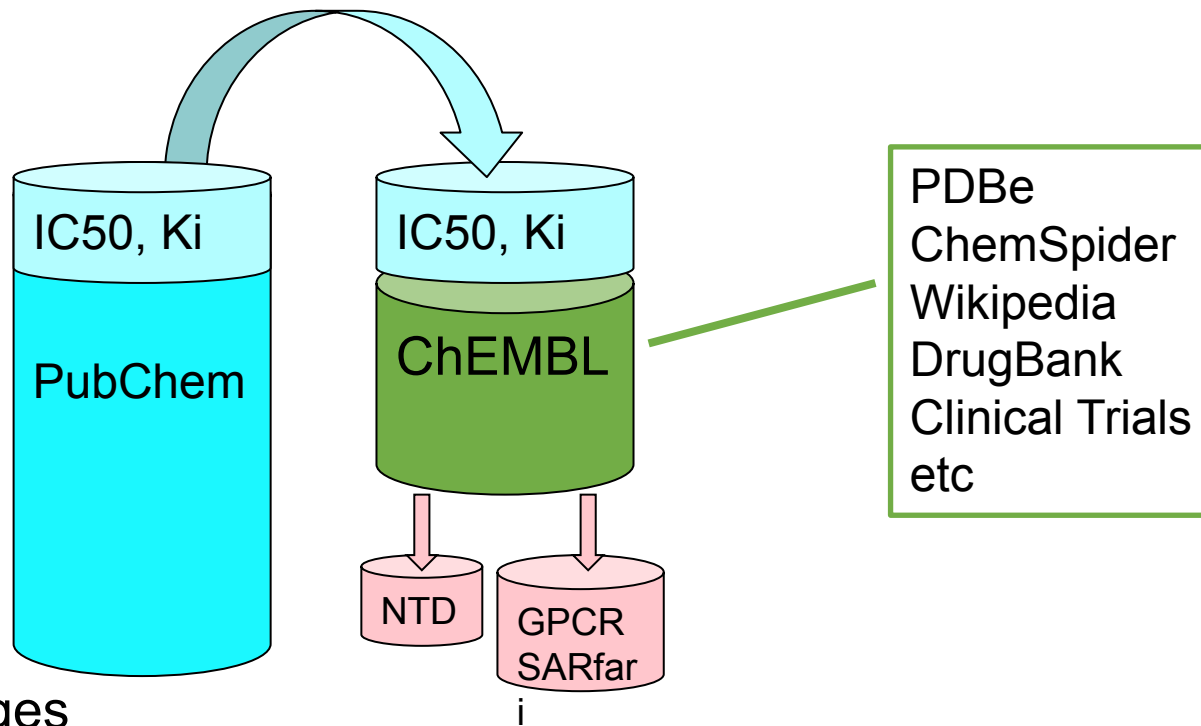
# Chemistry Challenges

- Standardise representations of chemical structures
- Needs to be largely automated (1 chemistry curator)
- Will be mistakes - some from abstraction but some in papers
- Software packages interpret structures differently particularly when converting between structure formats

## Standardisation Protocol

- Based on FDA Substance Registration System User's Guide
  - <http://www.fda.gov/ForIndustry/DataStandards/SubstanceRegistrationSystem-UniqueIngredientIdentifierUNII/default.htm>
- Standard InChI is used to identify unique structures
- Parent compounds and salts are registered with different ChEMBL\_IDs and the relationship is recorded in database
- Stereochemistry is retained and recorded where known
- Tautomers of the same compound are registered with a single ChEMBL ID (merged on standard InChI)

# Integration with other Resources



## Challenges

- Data from different sources have different business rules
- Maintaining and updating database with new data
- Links to other resources vs integration
- More data sources becoming available
- Scalability

# UniChem

FRIDAY, 25 NOVEMBER 2011

UniChem - An EBI compound structure cross-referencing resource

The logo for UniChem, featuring the text 'un1Chem' in a sans-serif font. The '1' is green, while 'un' and 'Chem' are black.

We have faced for some time some issues with compound integration with ChEMBL - specifically the loading of compound sets into ChEMBL for cross referencing, between for example, ChEBI, PDBe compounds, *etc.* The ChEMBL update cycle is relatively slow with respect to some other resources, and there is inevitable thrash with compounds not being present, especially for exciting new data. Without doing something different for compound integration, we were starting to face a scenario where we had a compound table with many millions of compounds without any bioactivity data, and following this the inevitable slowdown in searching, *etc.*

We also had some issues facing us about curation of other people's primary data, changing compound structures, or their rendering, *etc.*

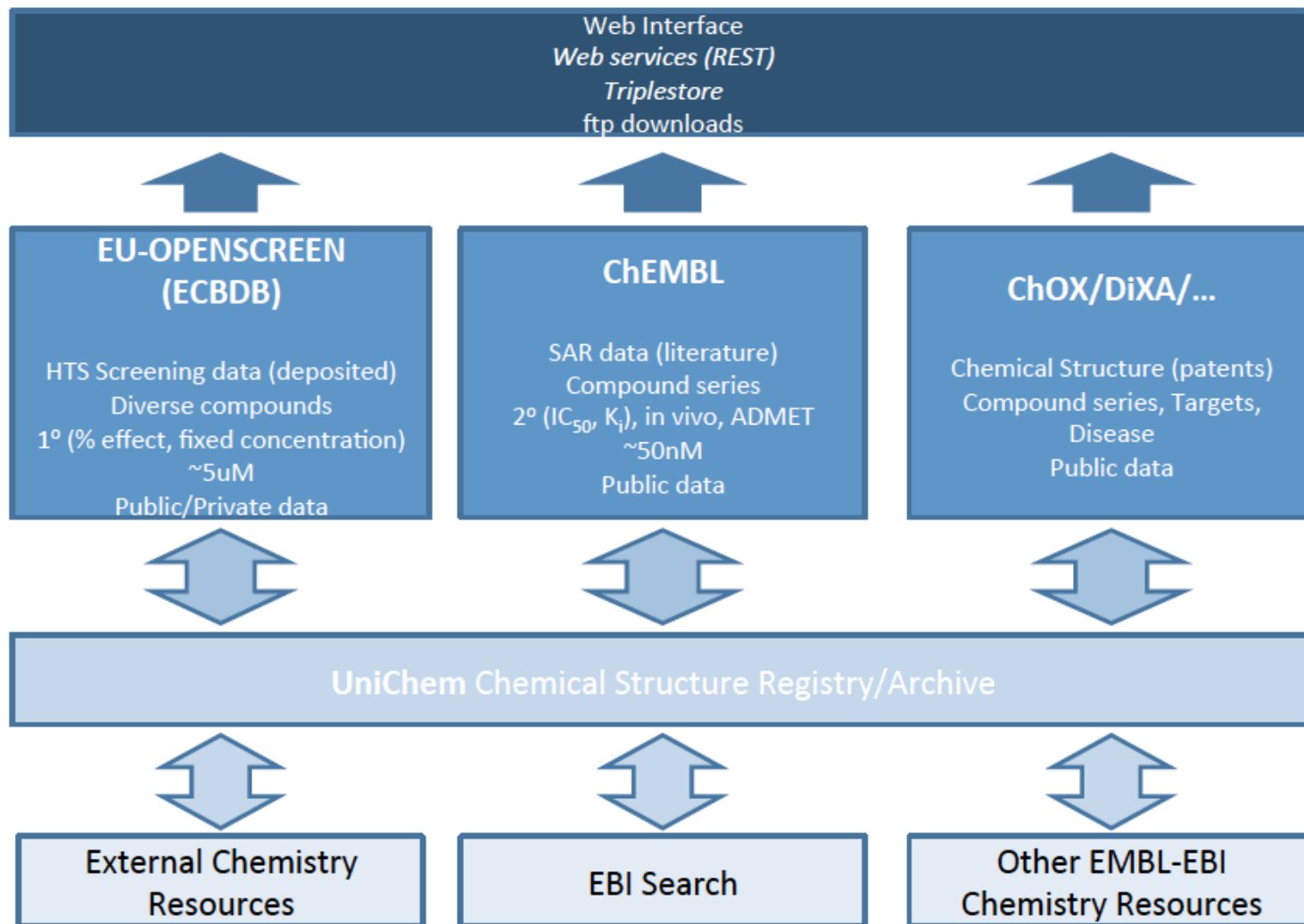
So, we decided to set up an external system to resolve cross-references between various databases. This is a very simple Standard InChI lookup, containing compounds from resources such as ChEMBL, ChEBI, PDBe, DrugBank, KEGG, BindingDB, PubChem, and so forth. UniChem can also handle versioning of the contained resources. We will be migrating various components of the current ChEMBL interface across to use web services on UniChem, this way, the cross links will always be fresh and correct, and we can focus on curation and optimisation of ChEMBL content. There are some other resources, like ZINC, STITCH, and ChemSpider, for example, that would be great to integrate, if we can get hold of the required data.

## What is UniChem ?

- A 'Unified Chemical Identifier' system
- A system for cross-referencing chemical structures and their identifiers between databases
- Enables tracking of 'id-to-structure' assignments over time
- Uses standard InChI to compare chemical structures across databases



# UniChem





# Summary

- ChEMBL is publicly available database of drugs, drug-like small molecules and their bioactivity data
- Useful resource for anyone involved in drug discovery particularly those in academia or SMEs
- As well as enabling users to access ChEMBL data they can link to other resources where information about common chemical structures (or targets) exists
- Standardising chemical structures and comparing across databases is challenging
- Standard InChi is a simple method to reference compounds across different databases

# Acknowledgements

## **ChEMBL Group**

John Overington

Anne Hersey

Anna Gaulton

Mark Davies

Jon Chambers

Louisa Bellis

Kazuyoshi Ikeda

Patricia Bento

Shaun McGlinchey

Yvonne Light

Felix Krueger

Ben Stauch

Ruth Akhtar

Francis Atkinson

Rita Santos

## **EMBL-EBI**

Samuel Kerrien, Sandra Orchard,  
Bruno Aranda, Rafael Jimenez,  
Reactome, UniProt and ChEBI teams

## **Collaborators**

Imperial Cancer Research, University  
of Dundee, University of Cambridge,  
Sanger Centre, University of  
Maryland, NCBI, TDR, IUPHAR,  
Bayer-Schering, Pfizer, GSK,  
Schering-Plough, MMV, Novartis, St  
Jude Children's Research Hospital

## **Former Inpharmatica colleagues**

**welcome**trust

EMBL

