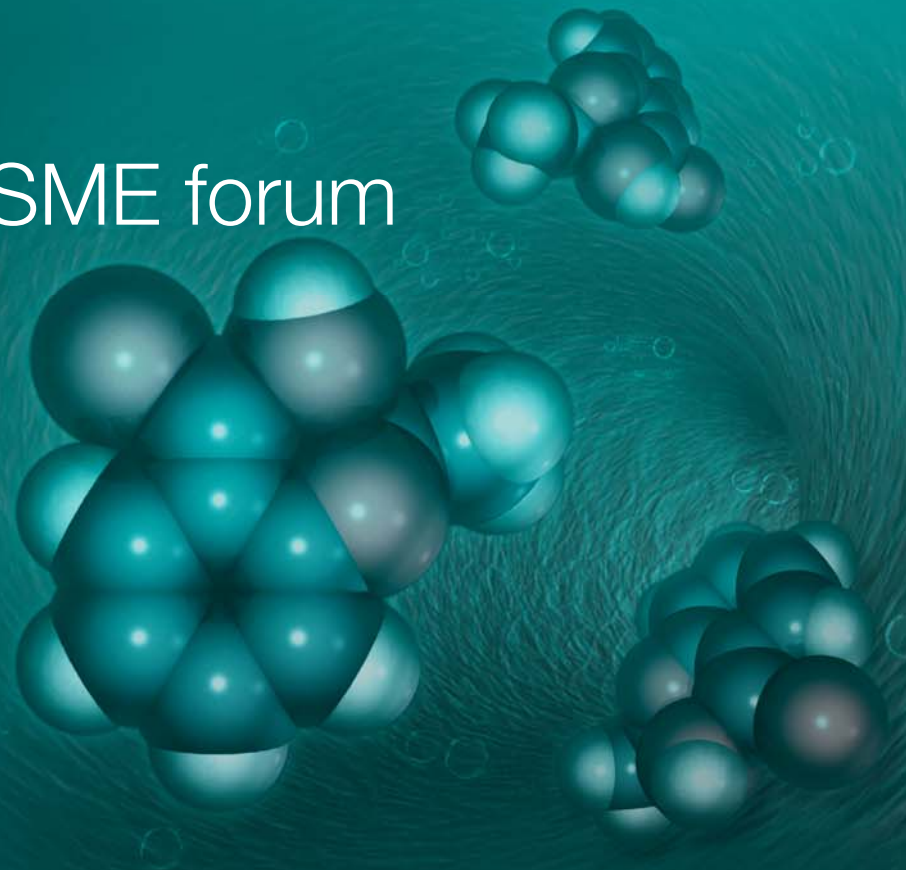
ChEMBL – Tech Track Session

ELIXIR Innovation and SME forum

Mark Davies

Technical Lead

ChEMBL Group



EMBL-EBI

Outline

- Background
- ChEMBL Data – *in more detail*
- ChEMBL Database Schema
- ChEMBL Web Services
- ChEMBL Web Portals
- UniChem
- ChEMBL RDF
- myChEMBL
- Exercises

What is EMBL-EBI?

- Part of the European Molecular Biology Laboratory
- International, non-profit research institute
- Europe's hub for biological data services and research
- 500 members of staff from 53 nations.

EMBL-EBI resources & groups

Genes, genomes & variation

European Nucleotide Archive
1000 Genomes
Ensembl
Ensembl Genomes
European Genome-phenome Archive
Metagenomics portal

Gene, protein & metabolite expression

ArrayExpress
Expression Atlas
Metabolights
PRIDE

Literature & ontologies

Europe PubMed Central
Gene Ontology
Experimental Factor Ontology

Protein sequences, families & motifs

InterPro
Pfam
UniProt

Molecular structures

Protein Data Bank in Europe
Electron Microscopy Data Bank

Chemical biology

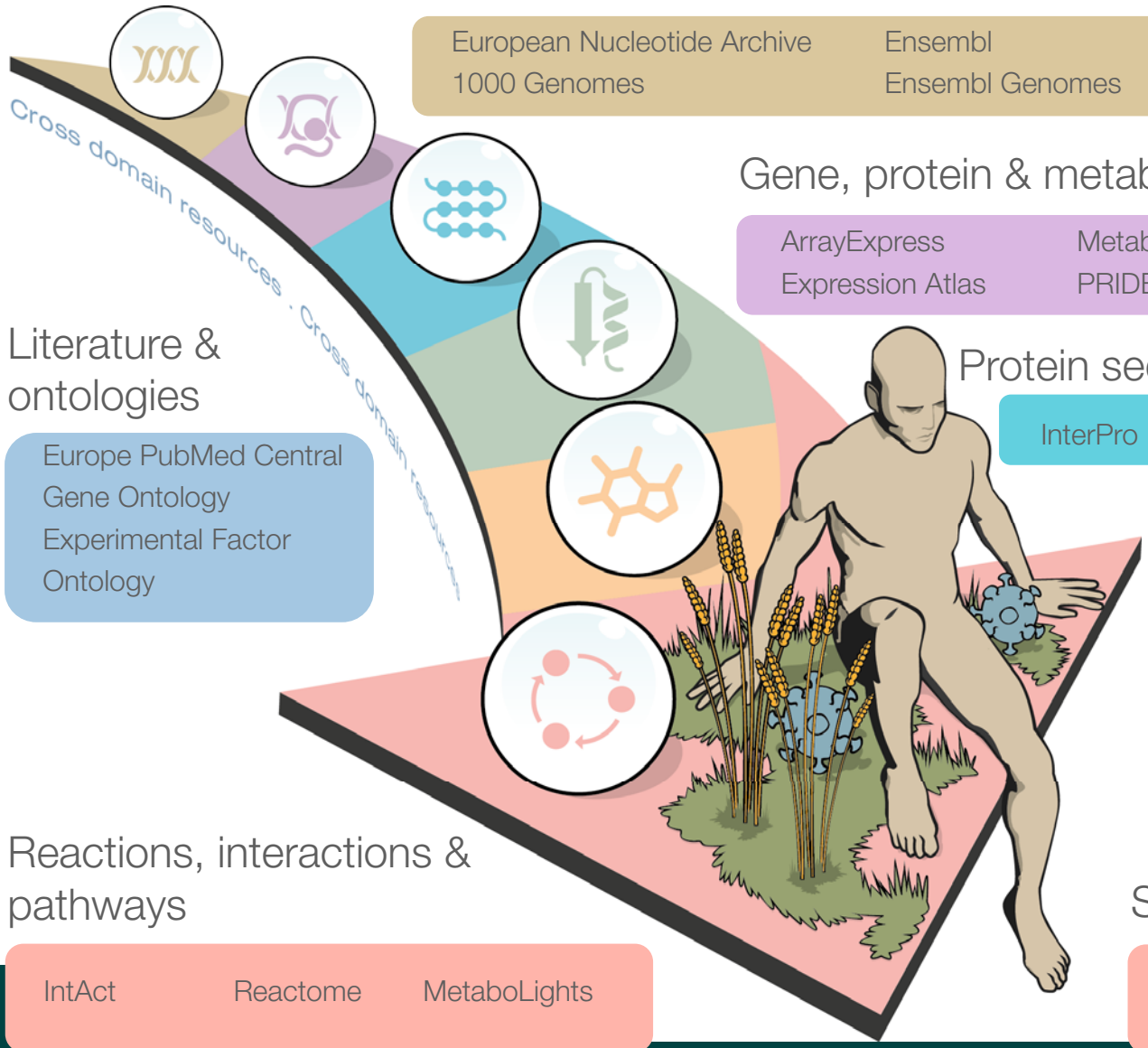
ChEMBL
ChEBI

Reactions, interactions & pathways

IntAct
Reactome
MetaboLights

Systems

BioModels
Enzyme Portal
BioSamples

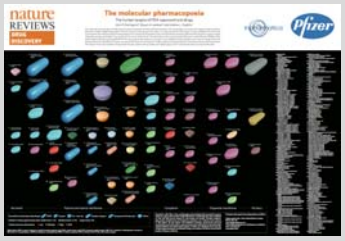


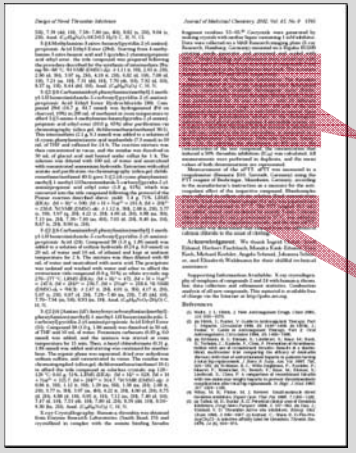
ChEMBL: Data for drug discovery

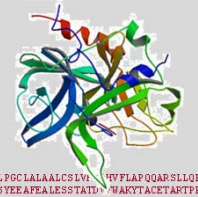
1. Scientific facts

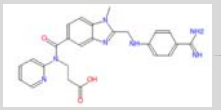

Design of Selective Thrombin Inhibitors Based on the (R)-Phe-Pro-Arg Sequence

3. Insight, tools and resources for translational drug discovery









>Thrombin
MAHVRGLQLPGCLALAALCSLVHSQHVFLAPQQARSLLQRVRRANTFLEEVRKGNLE
RECVEETCSYEEAFEALESSTATDVFWAKYTACETARTPRDKLAACLEGNCAEGLGT
NYRGHVNITRSGIECQLWRSRYPHKPEINSTTHPGADLQENFCRNPDSSTTGP
TDPTVRRQECSIPVCGQDQVTVAMTPRSEGSSVNLSPPLEQCVPDRGQQYQGR
THGLPCLAWASAQAKALSKHQDFNSAVQLVENFCRNPDGDEEGVWCYVAGKPG
CDLNYCEEAVEEETGDGLDEDSDRAIEGRTATSEYQTFFNPRTFGSGEADCGL
EKKSLEDKTERELLESYIDGRIVEGSDAEIGMSPWQVMLFRKSPQELLCGASL
WVLTAAHCLLYPPWDKNFTENDLLVRIGKHSRTRYERNIEKISMLEKIYIHPR
ENLDRDIALMKLKKPVAFSDYIHPVCLPDRETAASLLQAGYKGRVTGWGNLKE
NVGKGQPSVLQVVNLPIVERPVCKDSTRIRITDNMFCAGYKPDEGKRGDACEG
PFVMKSPFNNRWYQMGIVSWGEGCDRDGKYGFYTHVFRLKKWIQKVIDQFGE

Compound

K_i = 4.5nM

Bioactivity data

APTT = 11 min.

Assay/Target

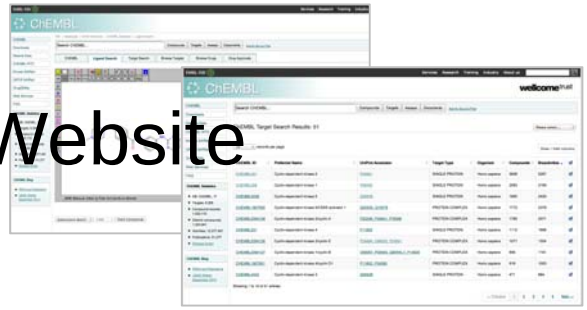
2. Organization, integration, curation and standardization of pharmacology data

Who works with ChEMBL?

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

How to access ChEMBL data

Web Services

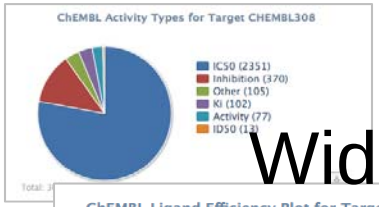


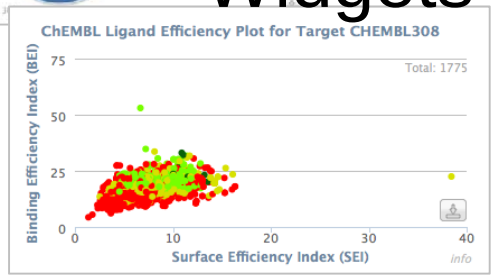
Website

Virtual Machine

ChEMBL 19 contains:

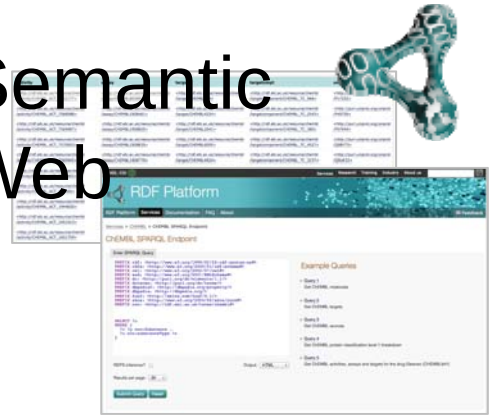
- 1,404,752 compounds
- 12,843,338 activities
- 1,106,285 assays
- 10,579 targets
- 57,156 documents
- 19 bioactivity sources





Widgets

Downloads



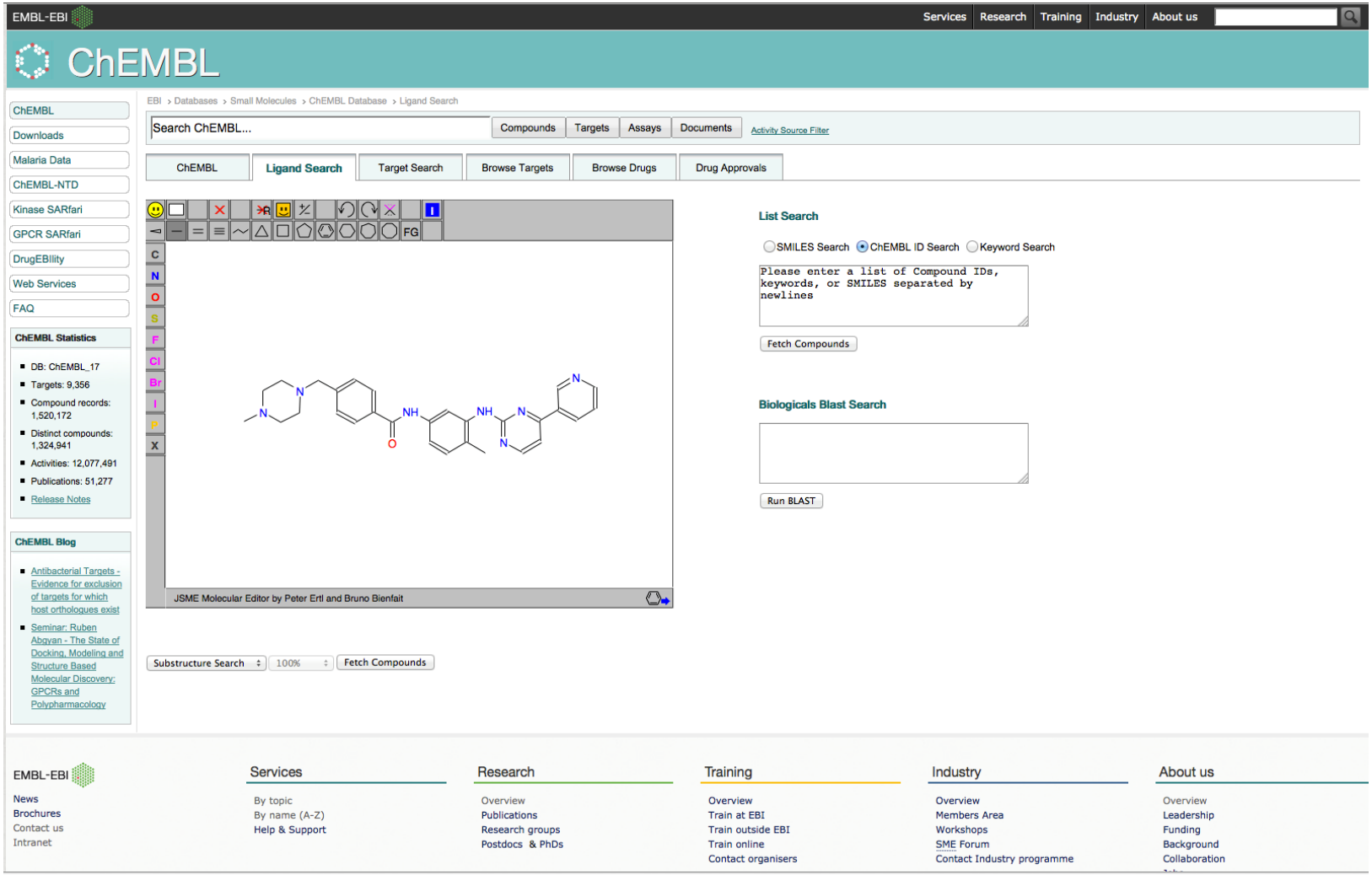
Semantic Web

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

EMBL-EBI

ChEMBL Data – *in more detail*

Where to start?



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

EMBL-EBI

Data sets in ChEMBL19

Bioactivity Source	Assays	Activities
Scientific Literature	826,105	4,828,175
PubChem BioAssays	2,308	7,036,920
Drugs for Neglected Diseases Initiative (DNDi)	225	13,752
MMV Malaria Box	30	7,667
Open Source Malaria Screening	22	344
WHO-TDR Malaria Screening	16	5,853
St Jude Malaria Screening	16	5,456
Novartis Malaria Screening	6	27,888
GSK Malaria Screening	6	81,198
Harvard Malaria Screening	4	111
GSK Tuberculosis Screening	5	1,406
TP-search Transporter Database	3,592	6,765
Open TG-GATEs	158,199	158,199
DrugMatrix	113,678	350,929
Guide to Receptors and Channels	344	801
GSK Published Kinase Inhibitor Set	456	169,451
Millipore Kinase Screening	468	73,944
Sanger Institute Genomics of Drug Sensitivity in Cancer	714	73,169
Deposited Supplementary Bioactivity Data	5	1,310

Compound only Sources	Compound Records
Orange Book (FDA Approved Drugs)	1975
USP Dictionary of USAN & International Drug Names	10822

Targets: 10,579
Compounds: 1,411,786
Activities: 12,843,338
Publications: 57,156

Literature source data example

EP1 Antagonists for Inflammatory Pain
A. Hall et al.
Bioorg. Med. Chem. Lett. 19 (2009) 2599–2603

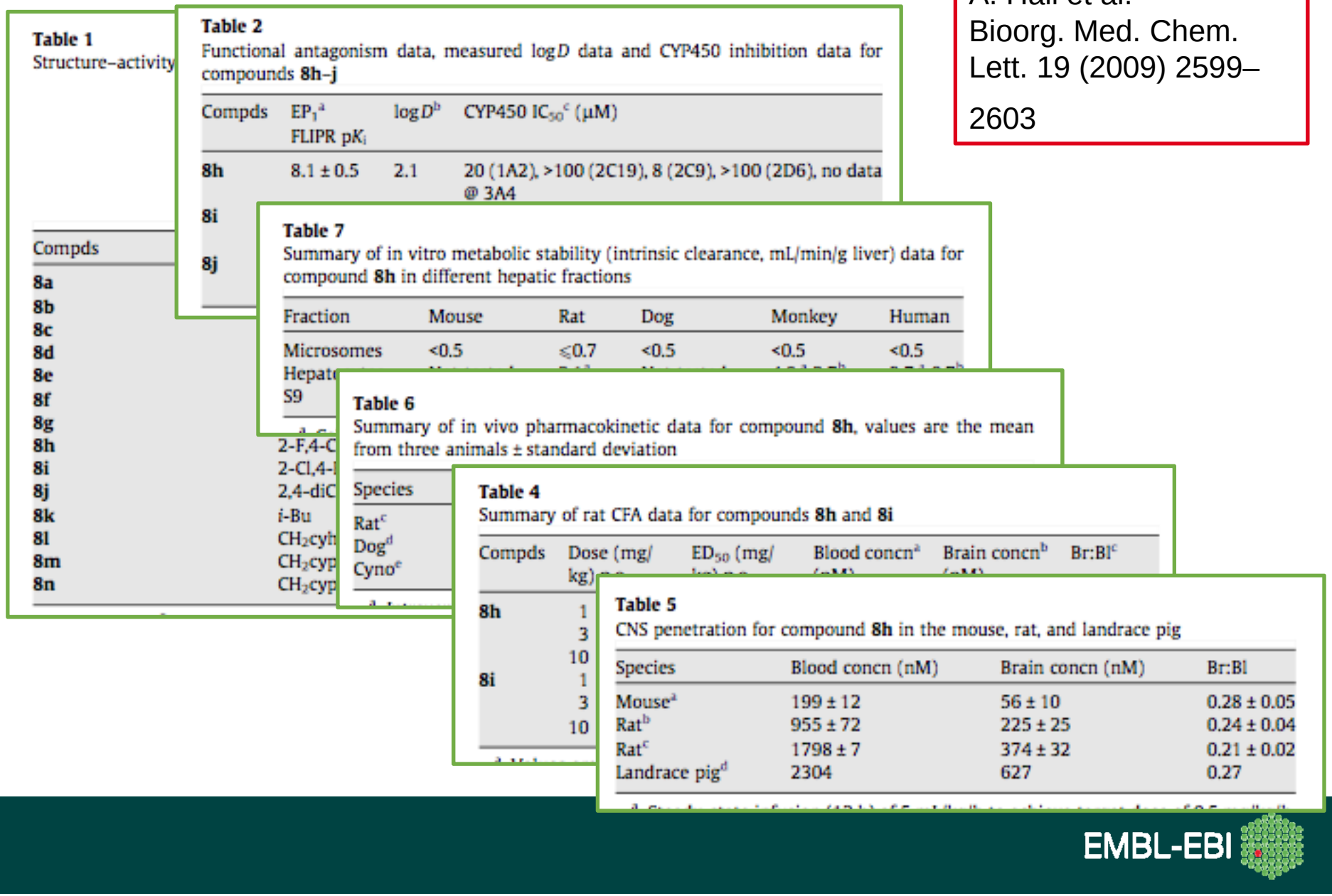


Table 1
Structure–activity

Compds
8a
8b
8c
8d
8e
8f
8g
8h
8i
8j
8k
8l
8m
8n

Table 2
Functional antagonism data, measured log*D* data and CYP450 inhibition data for compounds **8h–j**

Compds	EP_1[a] FLIPR pK_i	log*D*[b]	CYP450 IC_{50}[c] (μM)
8h	8.1 ± 0.5	2.1	20 (1A2), >100 (2C19), 8 (2C9), >100 (2D6), no data @ 3A4
8i			
8j			

Table 7
Summary of in vitro metabolic stability (intrinsic clearance, mL/min/g liver) data for compound **8h** in different hepatic fractions

Fraction	Mouse	Rat	Dog	Monkey	Human
Microsomes	<0.5	≤0.7	<0.5	<0.5	<0.5
Hepat					
S9					

Table 6
Summary of in vivo pharmacokinetic data for compound **8h**, values are the mean from three animals ± standard deviation

Species
Rat[c]
Dog[d]
Cyno[e]

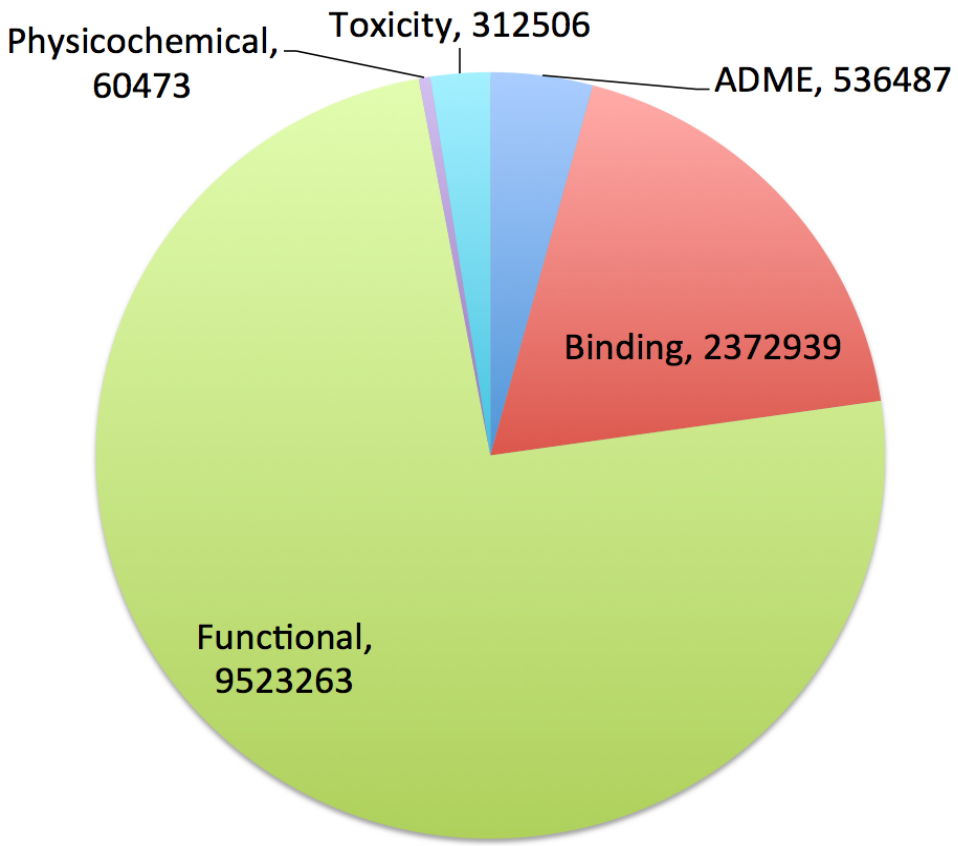
Table 4
Summary of rat CFA data for compounds **8h** and **8i**

Compds	Dose (mg/kg)	ED_{50} (mg/	Blood concn[a]	Brain concn[b]	Br:Bl[c]
8h	1				
	3				
	10				
8i	1				
	3				
	10				

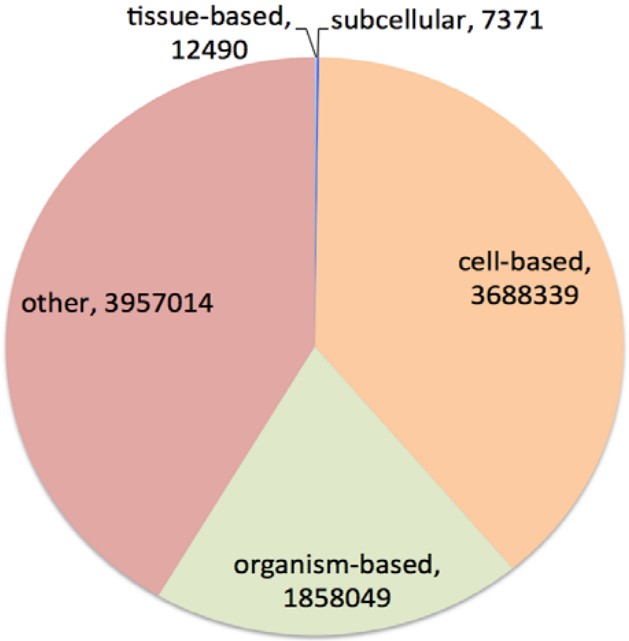
Table 5
CNS penetration for compound **8h** in the mouse, rat, and landrace pig

Species	Blood concn (nM)	Brain concn (nM)	Br:Bl
Mouse[a]	199 ± 12	56 ± 10	0.28 ± 0.05
Rat[b]	955 ± 72	225 ± 25	0.24 ± 0.04
Rat[c]	1798 ± 7	374 ± 32	0.21 ± 0.02
Landrace pig[d]	2304	627	0.27

Types of assay data in ChEMBL



Functional Data is comprised of:



Activity data curation

Standardise activity types
Including:
Antilog pKi, -logIC50 data etc

STANDARD_TYPE	PUBLISHED_ACTIVITY_TYPE
IC50	IC50_Mean
IC50	IC50 app
IC50	IC50 max

MOLREGNO	PREF_NAME	STANDARD_TYPE	STANDARD_VALUE	STANDARD_UNITS	PCHEMBL_VALUE	JOURNAL	YEAR	POTENTIAL_DUPLICATE
34997	Protein kinase C delta	Ki	66.2	nM	7.18	J. Med. Chem.	1999	
34997	Protein kinase C delta	Ki	66.2	nM	7.18	J. Med. Chem.	2001	1

IC50	0.000000038	nM

Outside typical range

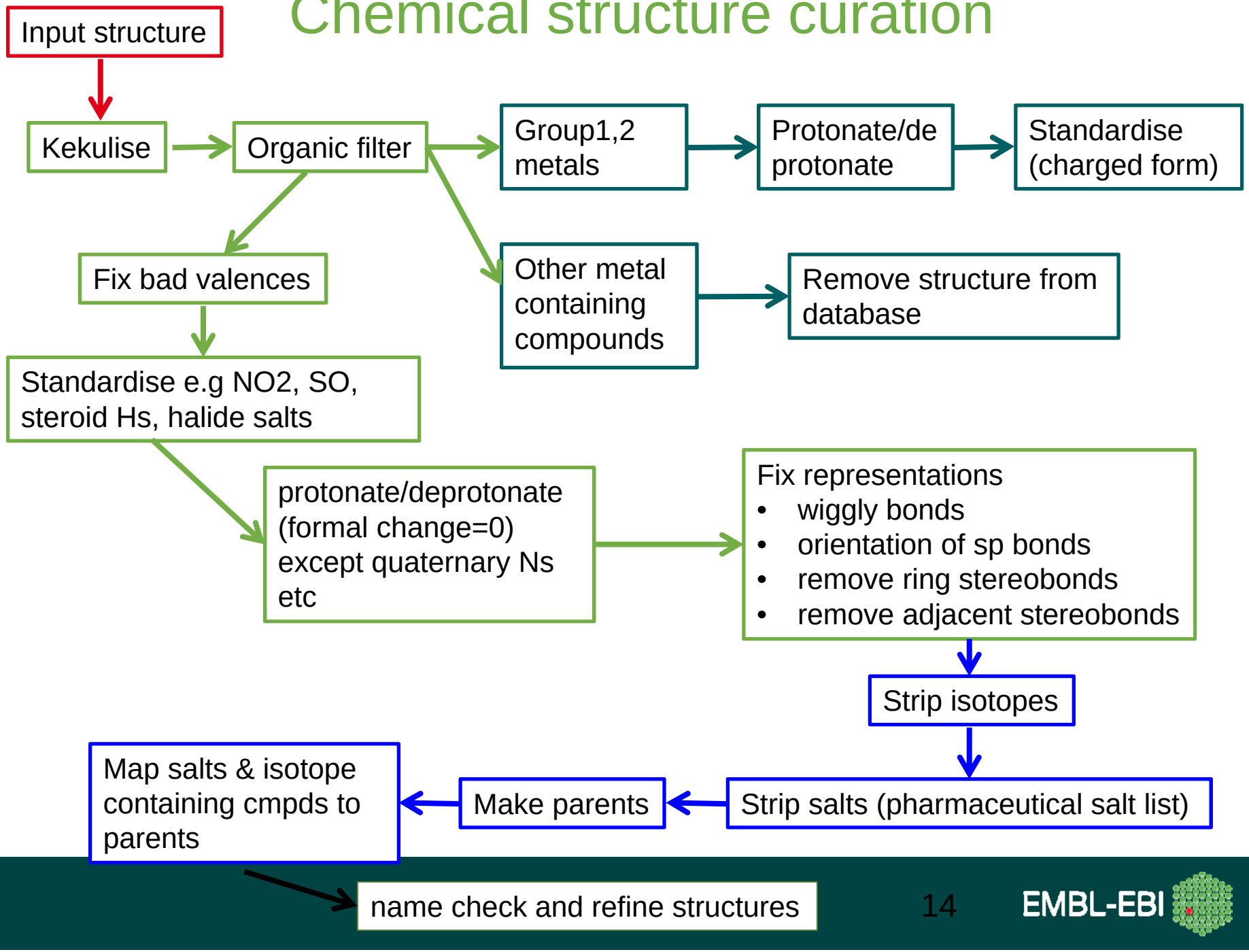
Potent

nM	nmol/l	10'-11M	um
nM	NM	10^-2nmol	nM/ml
nM	M,mol dme-3	10^-3nmol	10^-3M
nM	10'5nM	10^-5/M	10'-5M

CL	Clpl
CL	ClT
CL	Cl
CL	Total clearance

mL.min-1.kg-1	mL min-1 kg-1
mL.min-1.kg-1	ml/min.Kg

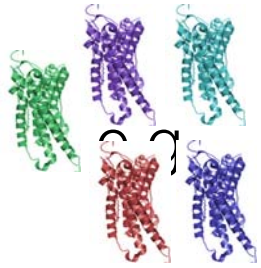
Chemical structure curation



Rationale for ChEMBL target model

- Many drug-discovery relevant targets are protein complexes e.g.,
  - Nicotinic acetylcholine receptors
  - GABA-A receptors
  - Cyclin-dependent kinases
  - Integrins

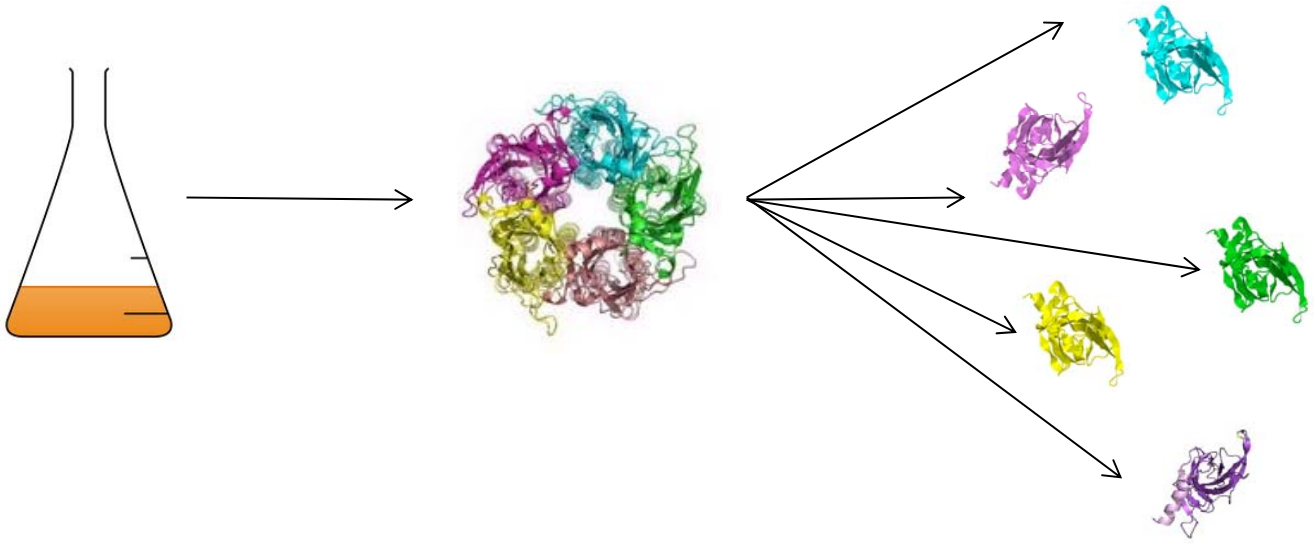
- Also some drugs e.g., monoclonal antibodies



- Many biological assays are non-specific
  ‘muscarinic receptors in guinea-pig ileum’

ChEMBL target data model

- Distinction between ‘target’ and ‘protein’
- Multi-protein targets created with different types
- Assays mapped to a single target



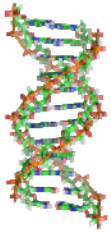
Assay Target Components

ChEMBL targets

Molecular

Non-molecular

Nucleic acid



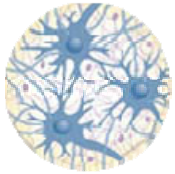
DNA

Protein

Cell-line

HEK293 cells

Tissue



Nervous

Subcellular-fraction

Mitochondria

Organism

Drosophila

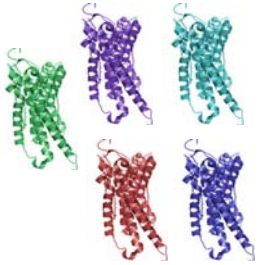
Single Protein

PDE5

Protein Complex

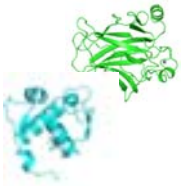
Nicotinic acetylcholine receptor

Protein Family



Muscarinic receptors

Protein-Protein Interaction



p53-MDM2

Overview of ChEMBL targets (ChEMBL_19)

Target Type	Count
Single Protein	5856
Organism	2122
Cell-Line	1633
Protein Complex	248
Tissue	240
Protein Family	213
Selectivity Group	97
Protein Complex Group	43
Nucleic-Acid	29
Small Molecule	20
Unknown	18
Protein-Protein Interaction	17
Other	40

Binding site model

- Binding sites relate to a target and can be defined at various levels:
  - The protein (e.g., subunit) to which the compound binds
  - The domain within the protein to which the compound binds
  - The specific residues that make up the binding site
  - Or may be represented just by a label (e.g., allosteric site)
- A binding site can consist of multiple components:
  - The interface between two subunits
  - The interface between two domains
- Used to annotate binding subunit for approved drugs and predicted pfam domain for binding of small molecules

Compound data

Compound Name and Classification

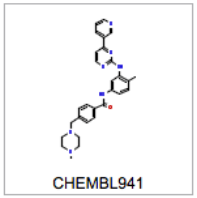
Compound ID	CHEMBL941
Compound Name	IMATINIB
Synonyms	IMATINIB, IMATINIB, Gleevec, STI-571, Gleevec, IMATINIB MESYLATE
Max Phase	4 (Approved)
Trade Names	Gleevec

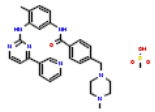
Compound Representations

Molfile	Download MolFile
Canonical SMILES	CN1CCN(Cc2ccc(cc2)C(=O)Nc3ccc(C
Standard InChI	InChI=1S/C29H31N7O/c1-21-5-10-25(Download InChI
Standard InChI Key	KTUFNOKKBVMGRW-UHFFFAOYSA

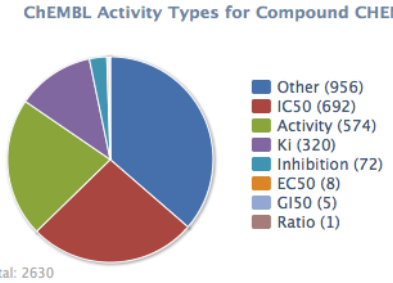
Molecule Features

Alternate Forms of Compound in ChEMBL

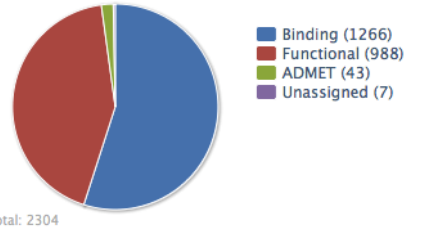

CHEMBL941


CHEMBL1642

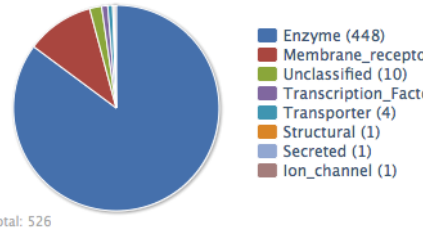
Compound Bioactivity Summary



Compound Assay Summary



Compound Target Summary



Clinical Trials for Compound

Number of clinical trials registered at clinicaltrials.gov	505

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
493.6	4.22	0	7	No	Yes	13.28	7.55	2.89	2.49	NEUTRAL

Compound Cross References

ChEBI	ChEBI:45783
ChemSpider	ChemSpider:KTUFNOKKBVMGRW-UHFFFAOYSA-N
PubChem	SID: 103905596 SID: 124892207 SID: 124892208 SID: 26755316 SID: 29215405 SID: 50100104
Wikipedia	Imatinib

UniChem Cross References

DrugBank	DB00619
PDBe	STI
ChEBI	45783
ZINC	ZINC19632618
eMolecules	876446
IBM Patent System	2C349E68BE42FC7B2B0FDD5080E27BB3
Atlas	imatinib

EMBL-EBI

Target data

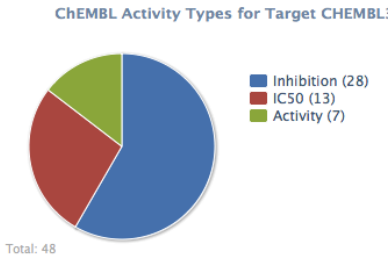
Target Name and Classification

Target ID	CHEMBL3099
Target Type	SINGLE PROTEIN
Preferred Name	Tyrosine-protein kinase ABL
Synonyms	
Organism	Mus musculus
Protein Target Classification	enzyme kinase protein kinase tyr tk ab

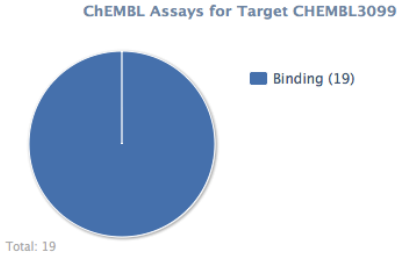
Target Components

Component Description	Rela
Tyrosine-protein kinase ABL1	SINGLE

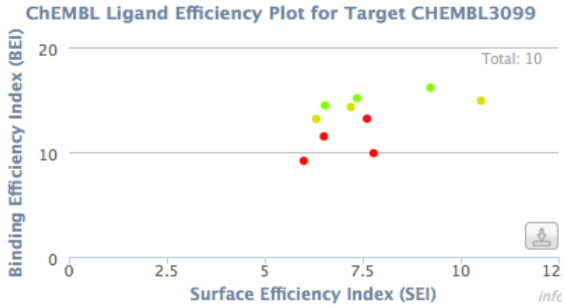
Target Associated Bioactivities



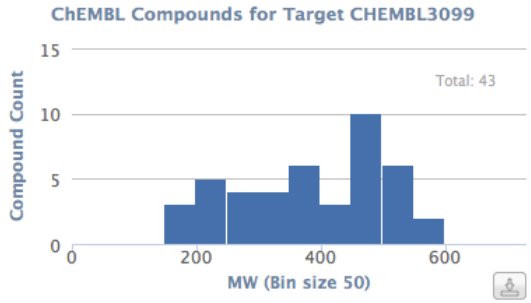
Target Associated Assays

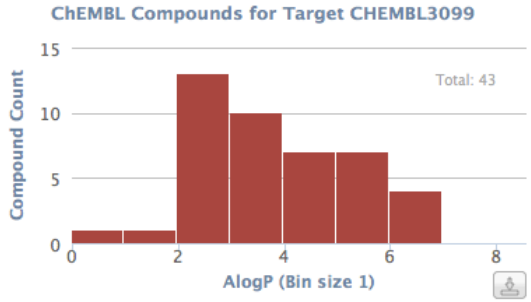


Target Ligand Efficiencies



Target Associated Compound Properties





Target Cross References - Gene

Array Express	ENSMUSG00000026842
Ensembl	ENSMUSG00000026842
GO Cellular Component	GO:0005634 (nucleus) GO:0005829 (cytosol) GO:0005856 (cytoskeleton)
GO Molecular Function	GO:0000287 (magnesium ion binding) GO:0004715 (non-membrane spanning protein tyrosine kinase activity) GO:0005524 (ATP binding) GO:0019904 (protein domain specific binding) GO:0030145 (manganese ion binding)
GO Biological Process	GO:0007155 (cell adhesion) GO:0018108 (peptidyl-tyrosine phosphorylation) GO:0030036 (actin cytoskeleton organization) GO:0051726 (regulation of cell cycle)

Target Cross References - Protein

canSAR	P00520
IntAct	P00520
UniProt	P00520 P97896 Q61252 Q61253 Q61254 Q61255 Q61256 Q61257 Q61258 Q61259 Q61260 Q61261 Q6PCM5 Q8C1X4

Target Cross References - Domain

InterPro	IPR000719 (Prot_kinase_cat_dom.) IPR000980 (SH2.) IPR001245 (Ser-Thr/Tyr-Pkinase.) IPR001452 (SH3_domain.) IPR008266 (Tyr_prot_kinase_AS.) IPR011009 (Kinase-like_dom.) IPR015015 (F-actin_binding.) IPR017441 (Protein_kinase_ATP_BS.) IPR020635 (Tyr_Pkinase_cat_dom.)
Pfam	PF00017 (SH2) PF00018 (SH3_1) PF07714 (Pkinase_Tyr) PF08919 (F_actin_bind)

Target Cross References - Structure

PDBe	1ABO 1ABQ 1FPU 1IEP 1M52 1OPJ 1OPK 2HZN 2QOH 2Z60 3DK3 3DK6 3DK7 3IK3 3K5V 3KF4 3KFA 3MS9 3MSS 3OXZ 3OY3

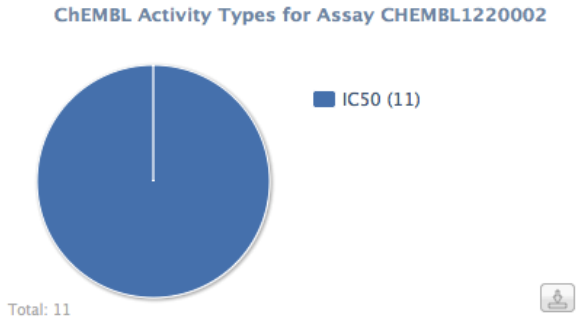
Assay data

Assay ID	CHEMBL1220002
Assay Type	Binding
Journal	Bioorg. Med. Chem., (2010); 18:15:5738
Assay Organism	
Assay Strain	PUBLICATION
Assay Description	Inhibition of mouse recombinant N-terminal His6-tagged Abl by radiometric kinase assay
No of Bioactivities	11

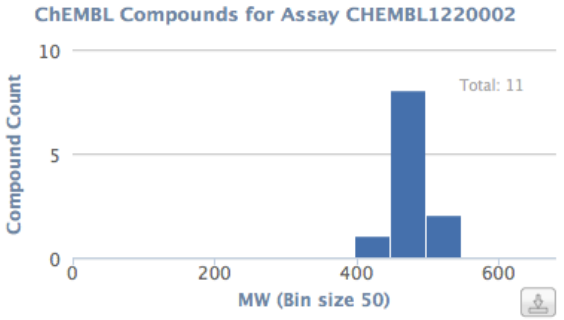
Curation Summary

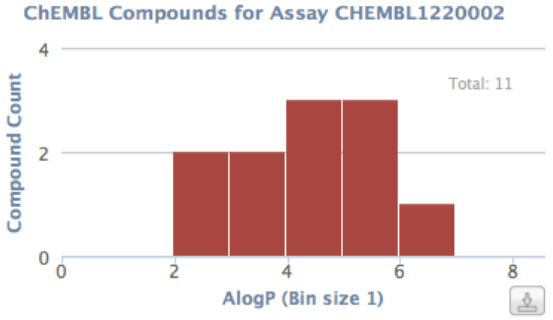
Target	Target Type	Target Description
Tyrosine-protein kinase ABL	SINGLE PROTEIN	

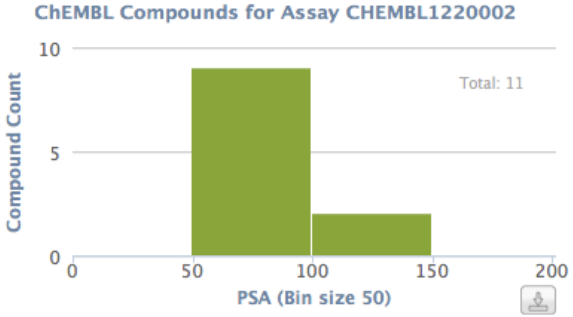
Bioactivity Summary



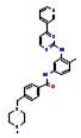
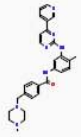
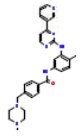
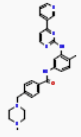
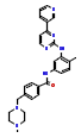
Compound Summaries





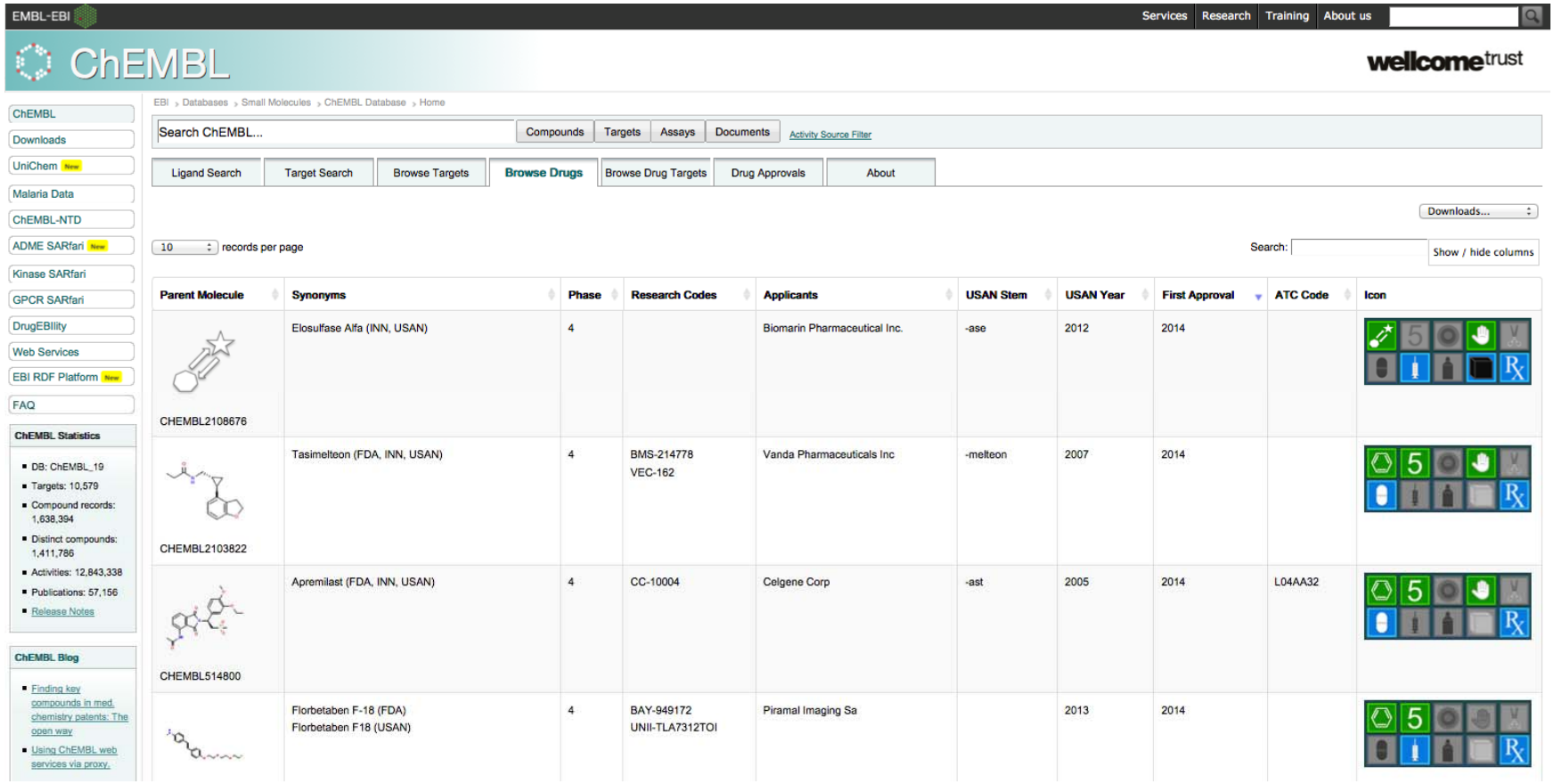


Activity data

Compound	Bioactivity					Assay			Target			Ref
Ingredient	Molweight	Standard Type	Relation	Standard Value	Standard Units	Assay Type	Description	Assay Src Description	Assay Organism	Target Type	Target Name	Target Organism	Reference
 CHEMBL941	493.6	IC50	>	30000	nM	B	Inhibition of NPM/ALK L256T mutant autophosphorylation activity in BaF3 cells by antiphosphotyrosine immunoblotting assay	Scientific Literature	Homo sapiens	PROTEIN COMPLEX	NPM/ALK (Nucleophosmin/ALK tyrosine kinase receptor)	Homo sapiens	J. Med. Chem., (2006) 49:19:5759
 CHEMBL941	493.6	IC50	=	500	nM	F	TP_TRANSPORTER: inhibition of ATPase activity in BCRP-expressing Sf9 cells	TP-search Transporter Database		SINGLE PROTEIN	ATP-binding cassette sub-family G member 2	Homo sapiens	Mol. Pharmacol., (2004) 65:1:1485
 CHEMBL941	493.6	IC50	=	25000	nM	F	Cytotoxicity against human HCT15 cells after 72 hrs by alamar-blue cell viability assay	Scientific Literature	Homo sapiens	CELL-LINE	HCT-15	Homo sapiens	J. Med. Chem., (2009) 52:8:2265
 CHEMBL941	493.6	IC50	=	10000	nM	F	Cytotoxicity against human HCT116 cells after 72 hrs by alamar-blue cell viability assay	Scientific Literature	Homo sapiens	CELL-LINE	HCT-116	Homo sapiens	J. Med. Chem., (2009) 52:8:2265
	493.6	IC50	=	5000	nM	B	Inhibition of recombinant Syk	Scientific Literature		SINGLE PROTEIN	Tyrosine-protein kinase SYK	Homo sapiens	Bioorg. Med. Chem. Lett., (2009) 19:7:1944

Drug data in ChEMBL

- ChEMBL has data on FDA marketed drugs, compounds with USAN/INN names



https://www.ebi.ac.uk/chembl/drugstore

EMBL-EBI

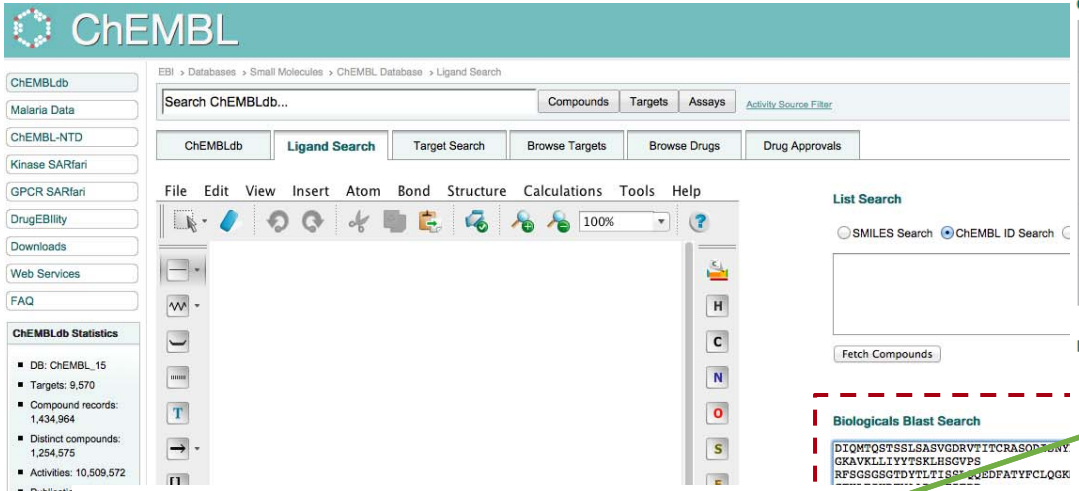
Biotherapeutic drugs

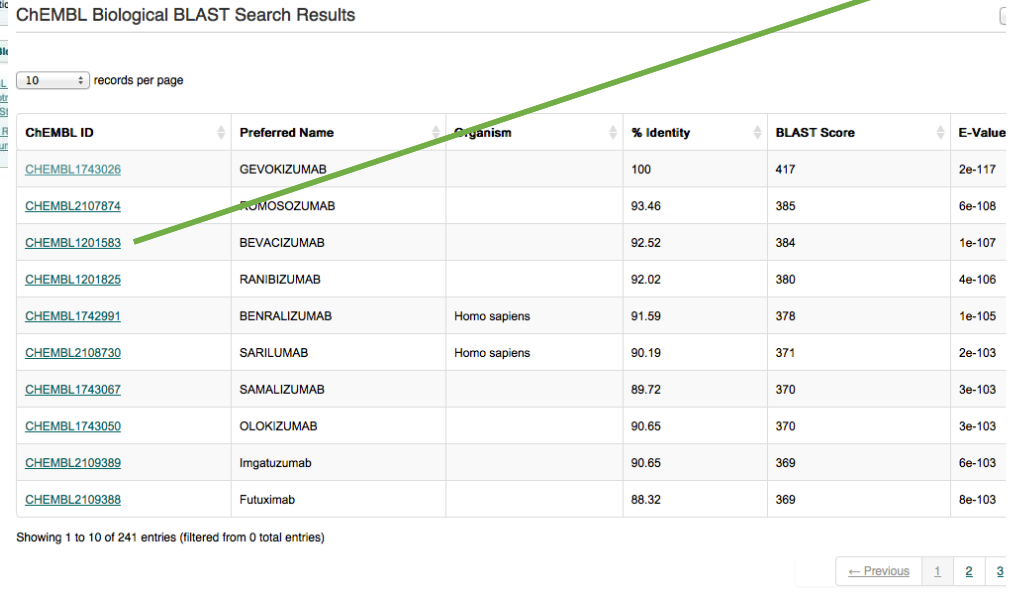
- Similar data model adopted to store information for biotherapeutic drugs (which can be protein complexes).

- Compound identifier (CHEMBL_ID) assigned to the whole drug. Structural information stored as protein sequence rather than molfile.

- Links to multiple protein components where there are multiple chains (e.g., antibody heavy/light chains).

Biotherapeutics drug searching





Compound Name and Classification

Compound ID	CHEMBL1201583
Compound Name	BEVACIZUMAB
Synonyms	R-435, Bevacizumab, VEGF, L01XC07, 12-IGG1, RG-435, RHUMAB-
Max Phase	4 (Approved)
Trade Names	Avastin

CHEMBL1201583

Molecule Features

Biological Sequence

Description	Sequence
Bevacizumab heavy chain	EVQLVESGGGLVQPGGSLRLSCAASGYTFTNYGMNWVRQAPGKGLEWVGWINTYTGEPTY AADFKRRFTFSLDTSKSTAYLQMNSLRAEDTAVYYCAKYPHYYGSSHWYFDVWGQGTLVT VSSASTKGPSVFPLAPSSKSTSGGTAALGCLVKDYFPEPVTVSWNSGALTSGVHTFPAVL QSSGLYSLSSVVTVPSSSLGTQTYICNVNHKPSNTKVDKKVEPKSCDKTHT
Bevacizumab light chain	DIQMTQSPSSLSASVGDRVTITCSASQDISNYLNWYQQKPGKAPKVLIYFTSSLHSGVPS RFSGSGSGTDFTLTISSLQPEDFATYYCQQYSTVPWTFGQGTKVEIKRTVAAPSVFIFPP SDEQLKSGTASVVCLLNNFYPREAKVQWKVDNALQSGNSQESVTEQDSKDSTYSLSSTLT LSKADYEKHKVYACEVTHQGLSSPVTKSFNRGEC

Clinical Trials for Compound

Number of clinical trials registered at clinicaltrials.gov	1810

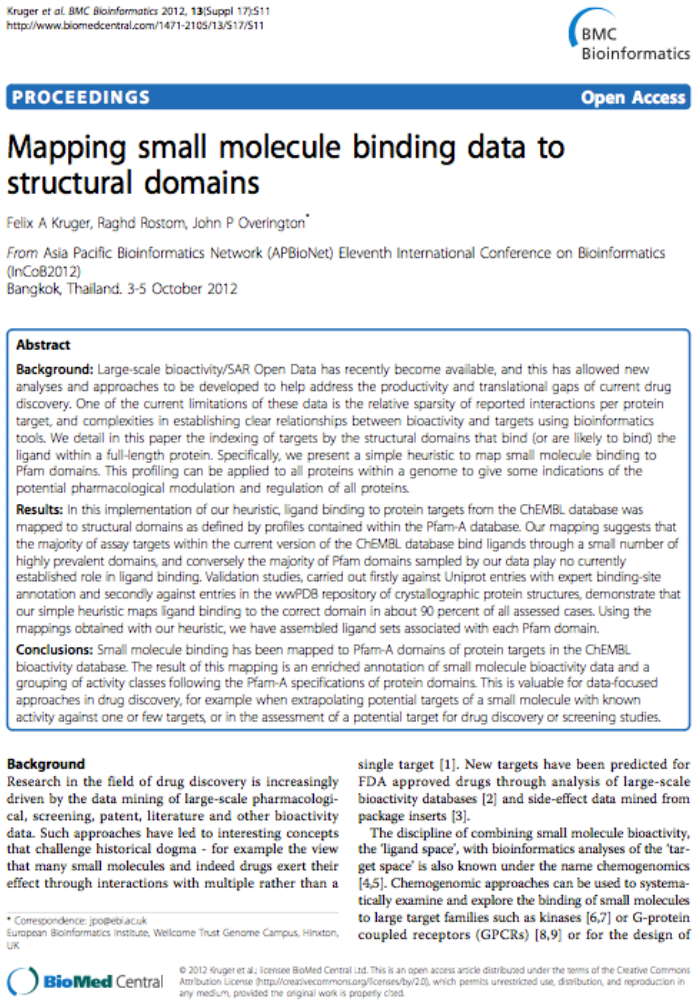
Compound Cross References

ATC	L ANTINEOPLASTIC AND IMMUNOMODULATING AGENTS L01 ANTINEOPLASTIC AGENTS L01X OTHER ANTINEOPLASTIC AGENTS L01XC Monoclonal antibodies L01XC07 bevacizumab
ChemSpider	ChemSpider Identifier not yet assigned
Wikipedia	Bevacizumab

EMBL-EBI

Predicted binding domains

- ChEMBL research project to identify likely binding domain on each target through analysis of protein domain structure
- For subset of activities in ChEMBL predicted compound-binding Pfam domain has been annotated.
- New data model allows this data to be integrated into ChEMBL and displayed



Kruger et al. BMC Bioinformatics 2012, 13(Suppl 17):S11
http://www.biomedcentral.com/1471-2105/13/S17/S11

BMC Bioinformatics

PROCEEDINGS Open Access

Mapping small molecule binding data to structural domains

Felix A Kruger, Raghd Rostom, John P Overington*

From Asia Pacific Bioinformatics Network (APBioNet) Eleventh International Conference on Bioinformatics (InCoB2012)
Bangkok, Thailand. 3-5 October 2012

Abstract

Background: Large-scale bioactivity/SAR Open Data has recently become available, and this has allowed new analyses and approaches to be developed to help address the productivity and translational gaps of current drug discovery. One of the current limitations of these data is the relative sparsity of reported interactions per protein target, and complexities in establishing clear relationships between bioactivity and targets using bioinformatics tools. We detail in this paper the indexing of targets by the structural domains that bind (or are likely to bind) the ligand within a full-length protein. Specifically, we present a simple heuristic to map small molecule binding to Pfam domains. This profiling can be applied to all proteins within a genome to give some indications of the potential pharmacological modulation and regulation of all proteins.

Results: In this implementation of our heuristic, ligand binding to protein targets from the ChEMBL database was mapped to structural domains as defined by profiles contained within the Pfam-A database. Our mapping suggests that the majority of assay targets within the current version of the ChEMBL database bind ligands through a small number of highly prevalent domains, and conversely the majority of Pfam domains sampled by our data play no currently established role in ligand binding. Validation studies, carried out firstly against Uniprot entries with expert binding-site annotation and secondly against entries in the wwPDB repository of crystallographic protein structures, demonstrate that our simple heuristic maps ligand binding to the correct domain in about 90 percent of all assessed cases. Using the mappings obtained with our heuristic, we have assembled ligand sets associated with each Pfam domain.

Conclusions: Small molecule binding has been mapped to Pfam-A domains of protein targets in the ChEMBL bioactivity database. The result of this mapping is an enriched annotation of small molecule bioactivity data and a grouping of activity classes following the Pfam-A specifications of protein domains. This is valuable for data-focused approaches in drug discovery, for example when extrapolating potential targets of a small molecule with known activity against one or few targets, or in the assessment of a potential target for drug discovery or screening studies.

Background

Research in the field of drug discovery is increasingly driven by the data mining of large-scale pharmacological, screening, patent, literature and other bioactivity data. Such approaches have led to interesting concepts that challenge historical dogma - for example the view that many small molecules and indeed drugs exert their effect through interactions with multiple rather than a single target [1]. New targets have been predicted for FDA approved drugs through analysis of large-scale bioactivity databases [2] and side-effect data mined from package inserts [3].

The discipline of combining small molecule bioactivity, the 'ligand space', with bioinformatics analyses of the 'target space' is also known under the name chemogenomics [4,5]. Chemogenomic approaches can be used to systematically examine and explore the binding of small molecules to large target families such as kinases [6,7] or G-protein coupled receptors (GPCRs) [8,9] or for the design of

* Correspondence: jpo@ebi.ac.uk
European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, UK

BMC Bioinformatics (2012), 13(S17), S11.

Predicted binding domains

Ingredient	Standard Type	Relation	Standard Value	Standard Units	Assay Type	Description	Assay Src Description	Target Type	Protein Accession	Target Name	Target Organism	Predicted Domain	Domain Confidence	Reference
Chiral  CHEMBL162	IC50	=	510	nM	B	Inhibition of human Fyn expressed in Sf9 cells after 20 mins by ELISA in presence of 100 umol/L ATP	Scientific Literature	SINGLE PROTEIN	P06241	Tyrosine-protein kinase FYN	Homo sapiens	Pkinase_Tyr	medium	J. Med. Chem., (2007) 50:6:1090
Chiral  CHEMBL162	IC50	=	6	nM	B	Inhibition of JAK3 expressed in Sf21 cells	Scientific Literature	SINGLE PROTEIN	P52333	Tyrosine-protein kinase JAK3	Homo sapiens	Pkinase_Tyr	medium	Bioorg. Med. Chem. Lett., (2007) 17:2:326
Chiral  CHEMBL162	IC50	=	1920	nM	B	Inhibition of SRC kinase at 1.0 mM ATP	Scientific Literature	SINGLE PROTEIN	P12931	Tyrosine-protein kinase SRC	Homo sapiens	Pkinase_Tyr	medium	Bioorg. Med. Chem. Lett., (2006) 16:7:2000
Chiral  CHEMBL162	IC50	=	3000	nM	B	Inhibition of human p60 c-src	Scientific Literature	SINGLE PROTEIN	P12931	Tyrosine-protein kinase SRC	Homo sapiens	Pkinase_Tyr	medium	J. Med. Chem., (1993) 36:1:21
Chiral  CHEMBL162	IC50	=	83	nM	B	Inhibition of human Fyn expressed in Sf9 cells after 20 mins by ELISA in presence of 10 umol/L ATP	Scientific Literature	SINGLE PROTEIN	P06241	Tyrosine-protein kinase FYN	Homo sapiens	Pkinase_Tyr	medium	J. Med. Chem., (2007) 50:6:1090

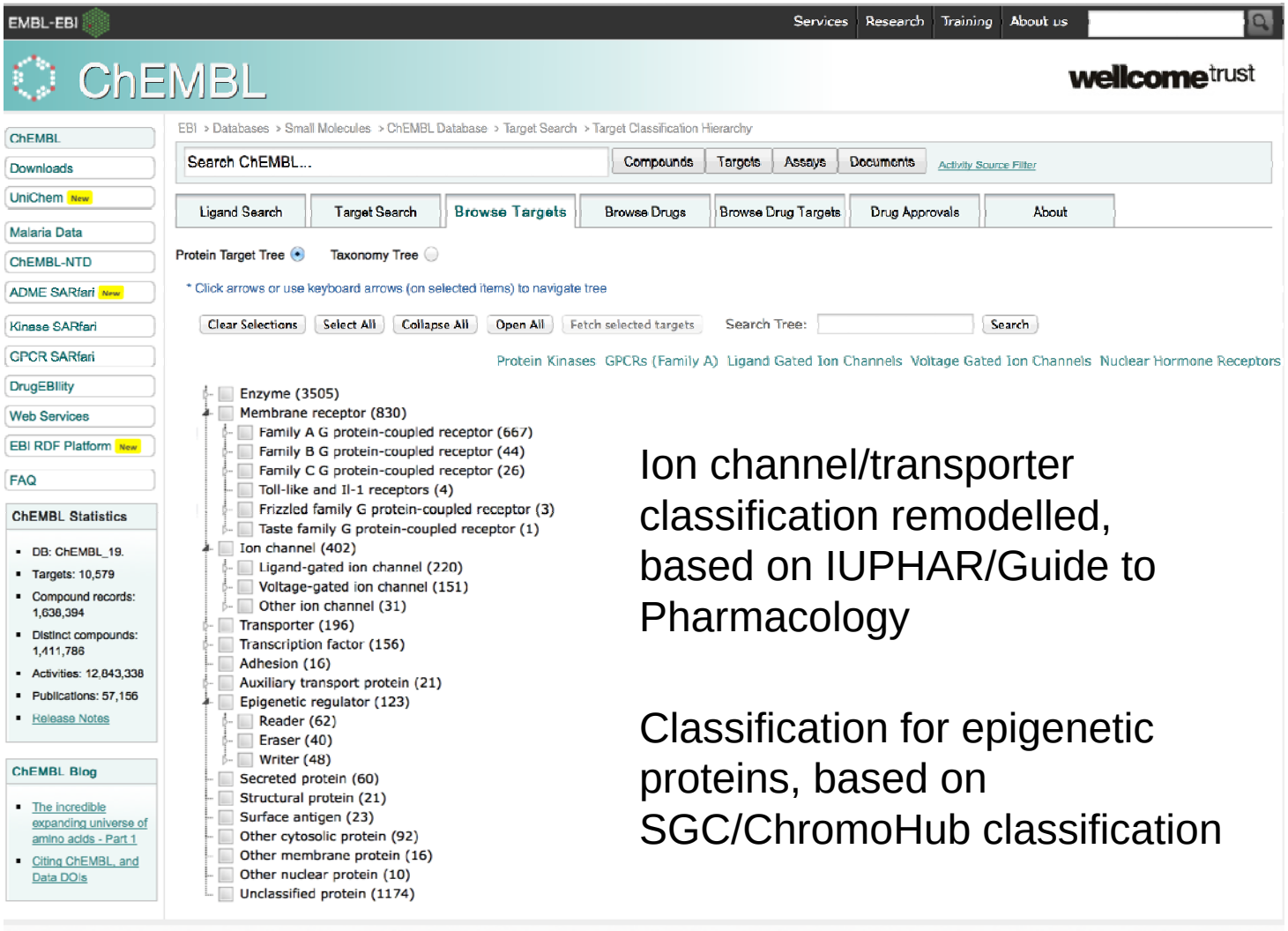
Pfam

EMBL-EBI

Target family classification

- ChEMBL provides protein family classification of for major drug-target families
- Aim to follow widely-used community ‘standards’ for individual families (e.g., IUPHAR, MEROPS, ‘kinome’), providing broad coverage without reinventing the wheel
- Feedback from Open PHACTS and pharma companies that it is a useful classification for drug-discovery
- Manually curated and maintained

Target family classification



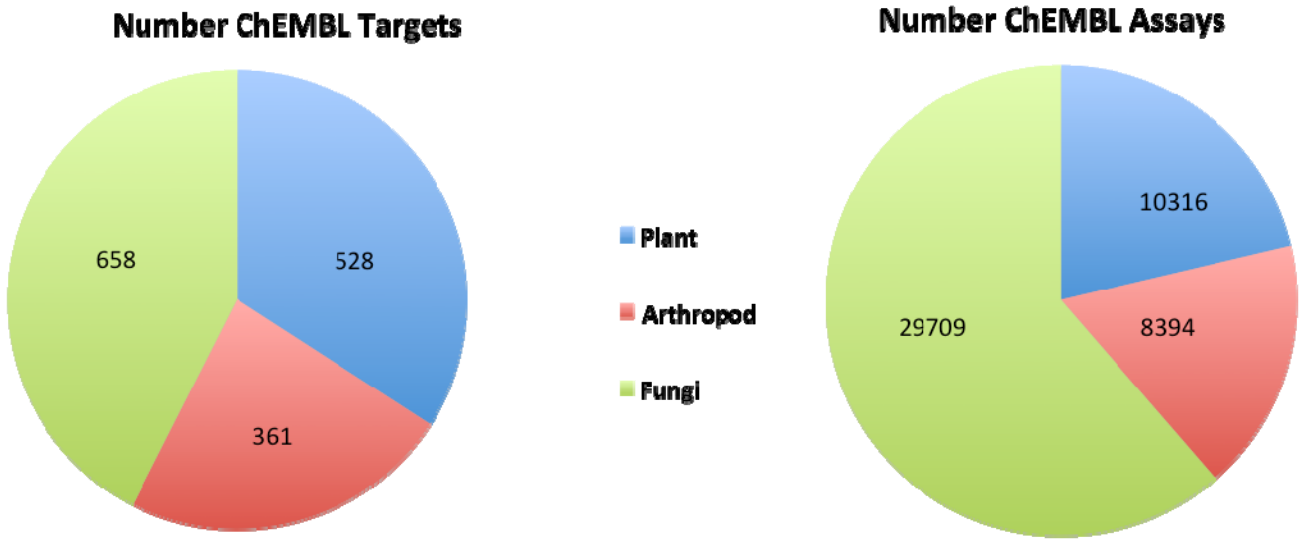
Ion channel/transporter classification remodelled, based on IUPHAR/Guide to Pharmacology

Classification for epigenetic proteins, based on SGC/ChromoHub classification

https://www.ebi.ac.uk/chembl/target/browser

More new data in CHEMBL_19

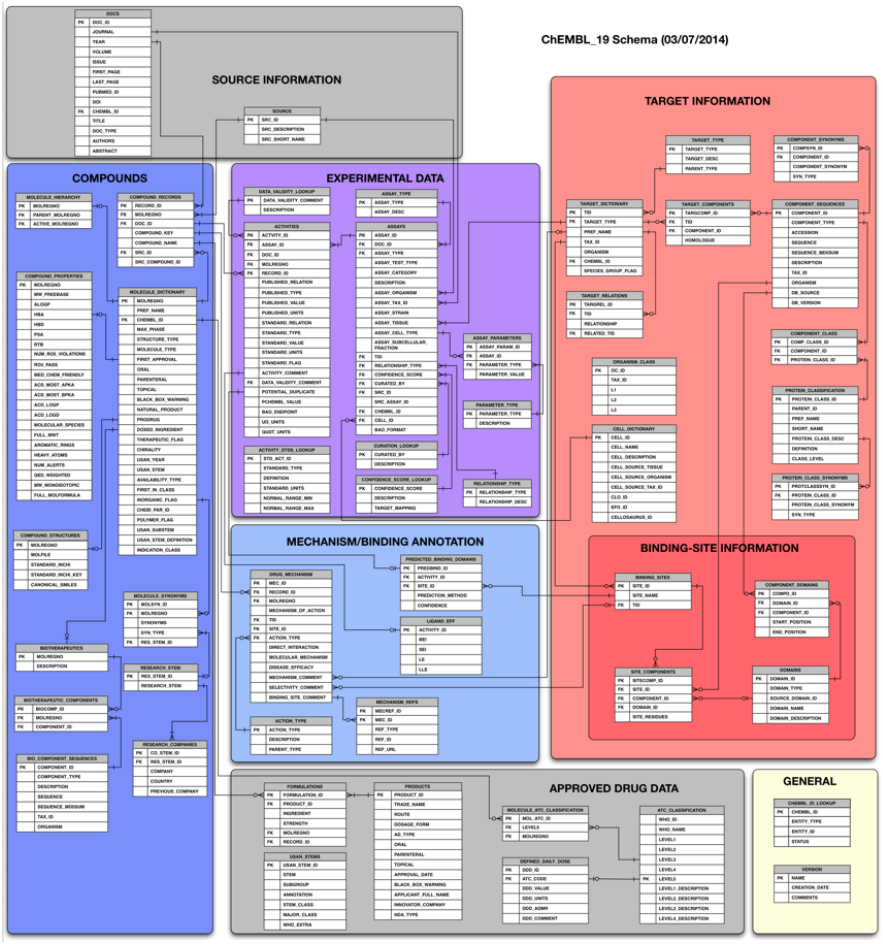
- Agrochemical data funded by Syngenta



ChEMBL Assay ID	Assay Source	Assay Type	Assay Organism	Description	Activity Count	Reference	
CHEMBL3065390	Scientific Literature	F	Helicoverpa assulta	Insecticidal activity against sixth-instar larvae of Helicoverpa assulta SD1 at 1.8 to 183360 mg/l after 48 hr by chi-squared analysis	2	Crop Protection, (2009) 28:2:162	☑
CHEMBL3060077	Scientific Literature	F	Spodoptera exigua	Insecticidal activity against third-instar Spodoptera exigua larvae assessed as mortality at 4 g/l measured after 3 days	1	J Pesticide Sci, (2010) 35:4:483	☑
CHEMBL3060078	Scientific Literature	F	Spodoptera exigua	Insecticidal activity against third-instar Spodoptera exigua larvae assessed as mortality at 2 g/l measured after 3 days	1	J Pesticide Sci, (2010) 35:4:483	☑
CHEMBL3065595	Scientific Literature	F	Lepidoptera	Insecticidal activity against Lepidoptera infected rice plant assessed as increase in grain yield at 1 kg ai/ha applied as granules on 30 to 50 day...	3	Crop Protection, (2006) 25:5:409	☑
CHEMBL3053939	Scientific Literature	F	Rhopalosiphum padi	Insecticidal activity against adult Rhopalosiphum padi infested wheat plant assessed as reduction in xylem ingestion at 0.4 mg/l applied to the roo...	1	Pest Manag Sci, (2010) 66:7:779	☑
CHEMBL3060362	Scientific Literature	F	Apolygus lucorum	Insecticidal activity against third-instar Apolygus lucorum reared in asparagus bean pod assessed as mortality administered by dipping pods in comp...	8	J Pesticide Sci, (2012) 37:2:135	☑

ChEMBL Database Schema

The ChEMBL schema



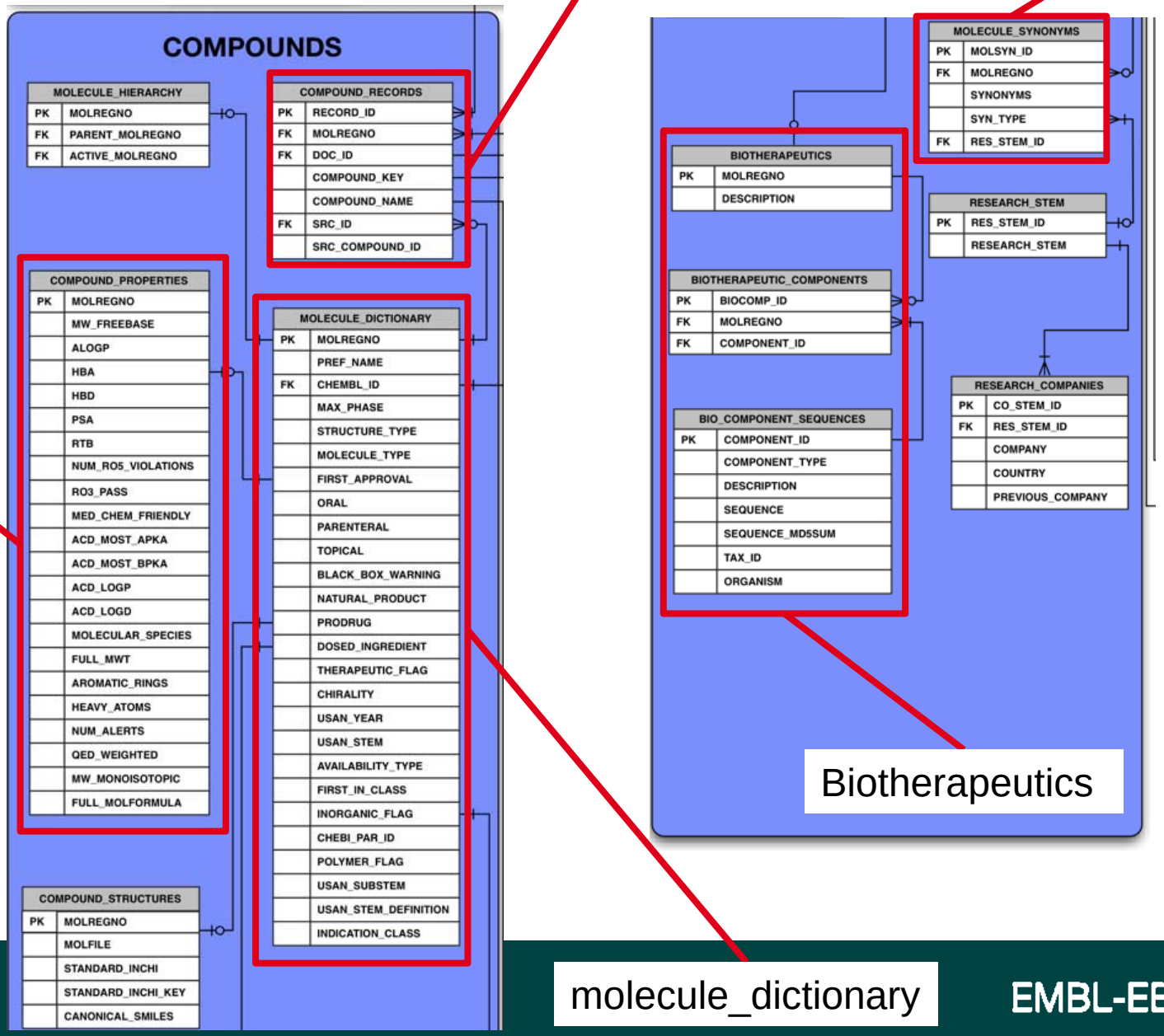
ftp://ftp.ebi.ac.uk/pub/databases/chembl/ChEMBLdb/latest/chembl_19_schema.png

Compound data

compound_records

compound_synonyms

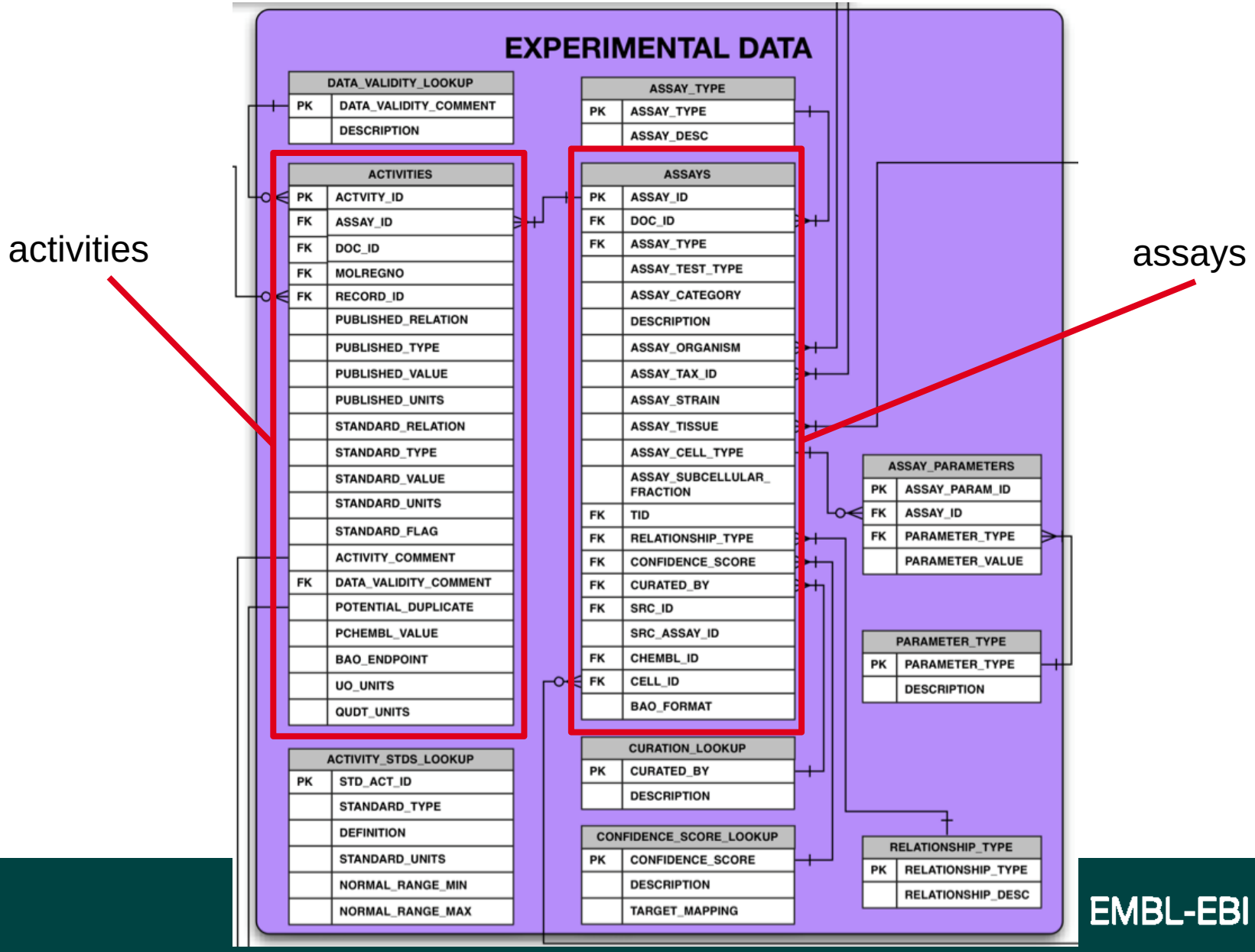
compound_
properties



molecule_dictionary

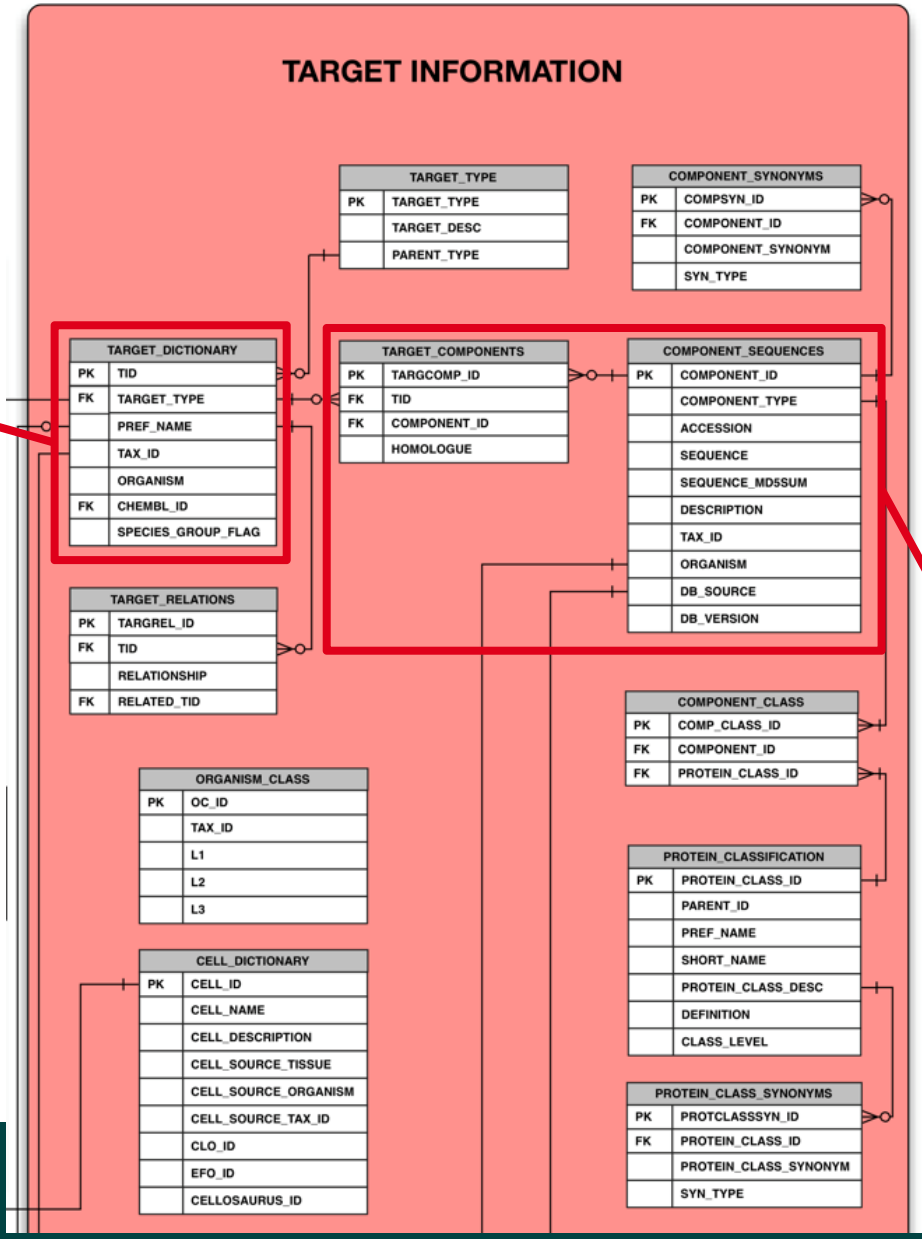
EMBL-EBI

Bioactivity data



Target data

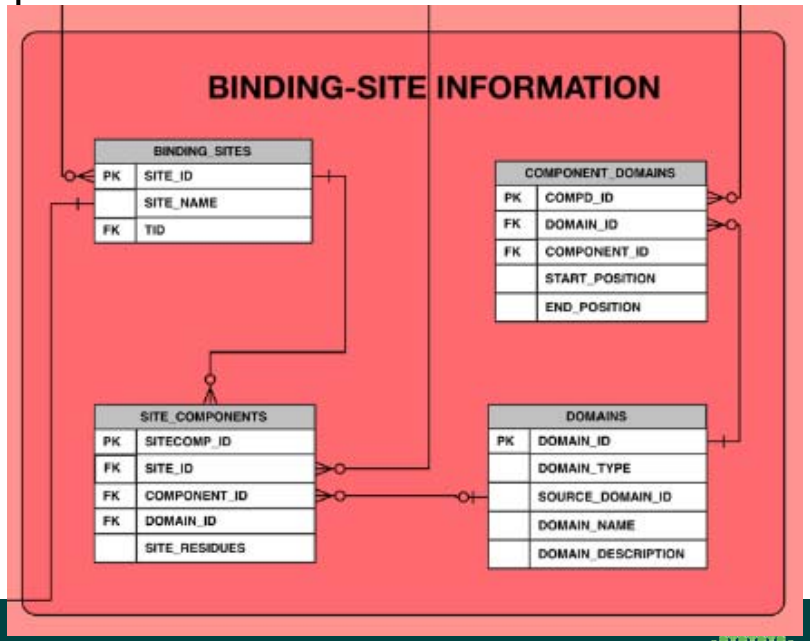
target_dictionary



Target Components

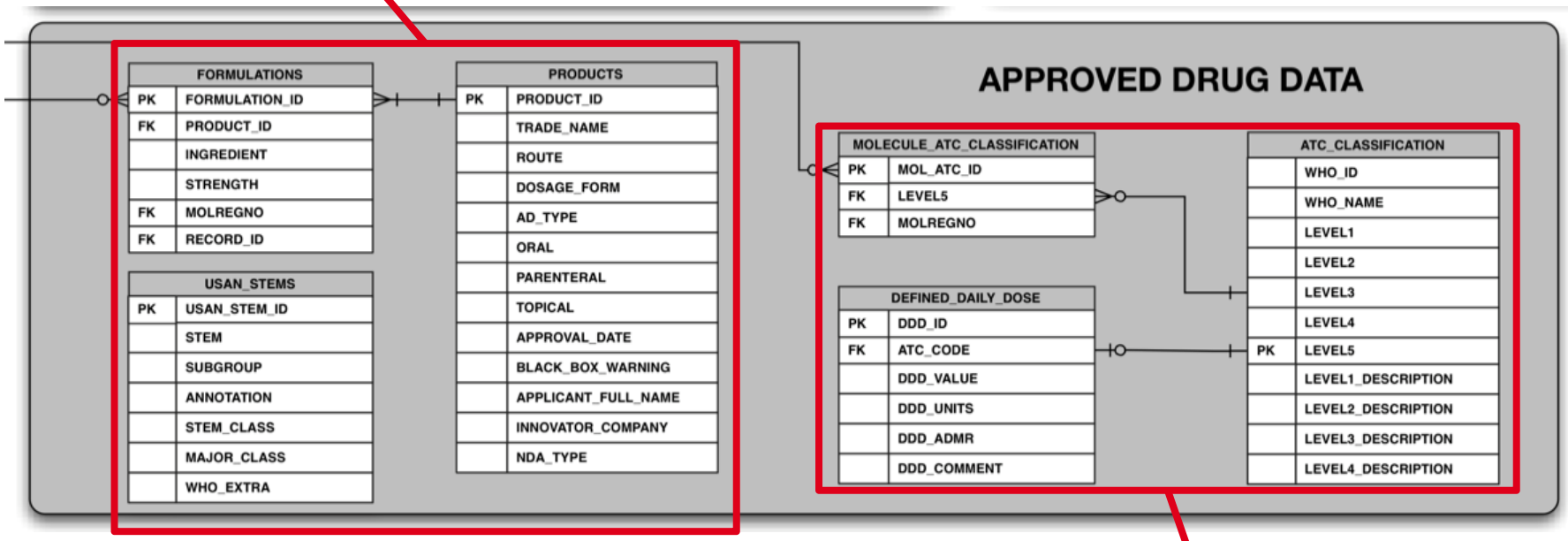
Binding site information

- While representing the biological form of the target is important, also desirable to identify the subunit/domain/site at which the compound is binding
- E.g., If assay identifies compounds active against CDK4/cyclin D1, need to know whether these compounds are binding to CDK4 or cyclin D1
- Target data model allows modelling of both the biological target and the location of the binding site



Approved drug data

ChEMBL to FDA Orange Book



ChEMBL to ATC Classification

ChEMBL Web Services

ChEMBL web services

- Provide access to ChEMBL data programmatically
- Language-agnostic access *via* RESTful HTTP interface
  - Python requests, Perl LWP *etc.*
- Interactive online documentation
- Two types of web service available:

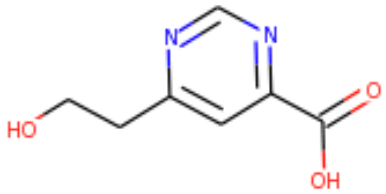
ChEMBL Data

CHEMBL941

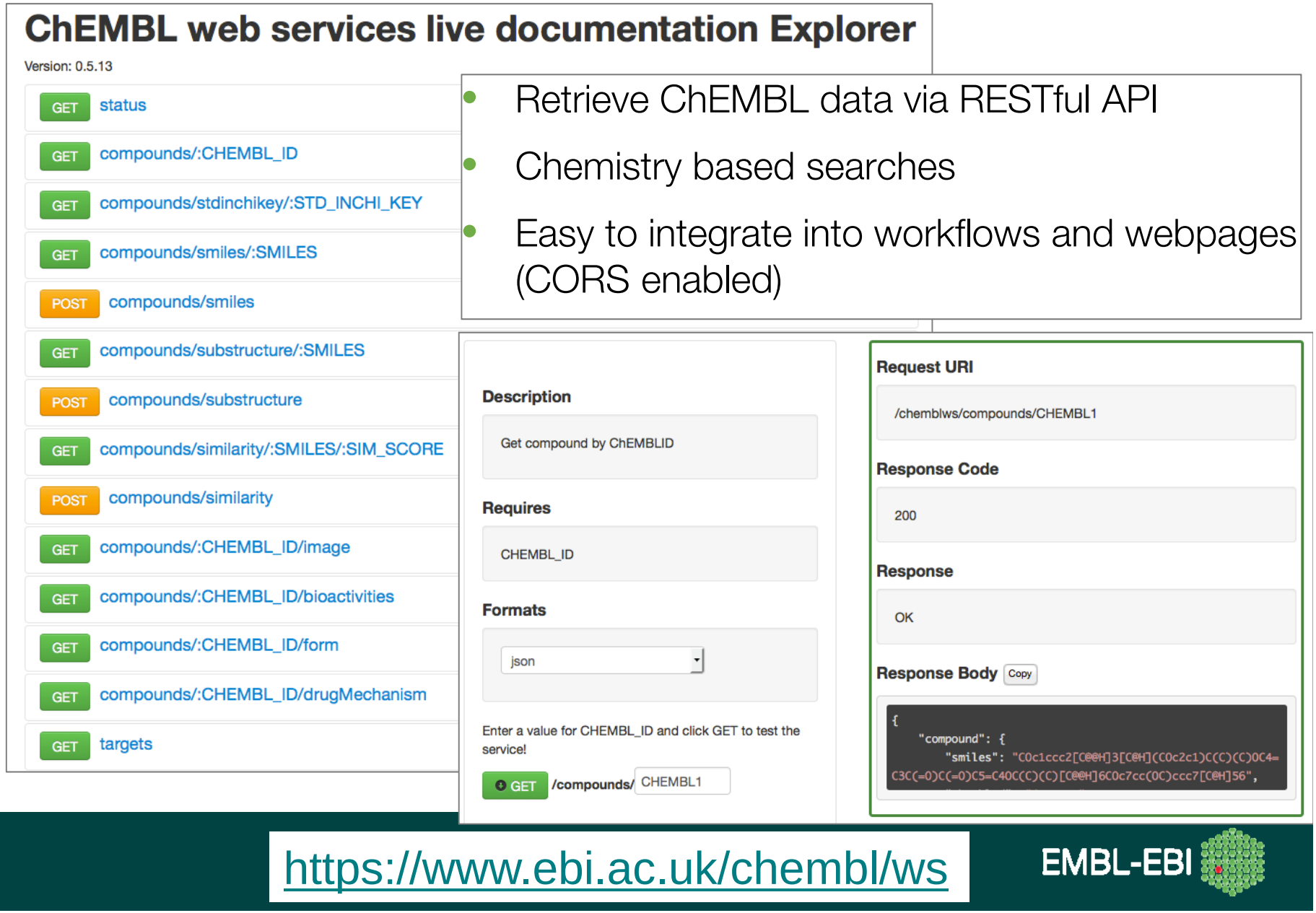

```
{
  - compounds: {
        chemblId: "CHEMBL941",
        numRo5Violations: 0,
        molecularWeight: 493.60273,
        preferredCompoundName: "IMATINIB",
        alogp: 3.583,
        knownDrug: "Yes",
        medChemFriendly: "Yes",
        rotatableBonds: 7,
        passesRuleOfThree: "No",
        molecularFormula: "C29 H31 N7 O",
        smiles: "CN1CCN(Cc2ccc(cc2)C(=O)Nc3ccc(C)c(Nc4nccc(n4)c5
        species: "NEUTRAL",
        acdAcidicPka: 13.279,
        acdBasicPka: 7.553,
        acdLogp: 2.887,
        acdLogd: 2.486
    }
}
```


Cheminfomatic Utilities

c1c(CCO)ncnc1C(=O)O



Web services – Data

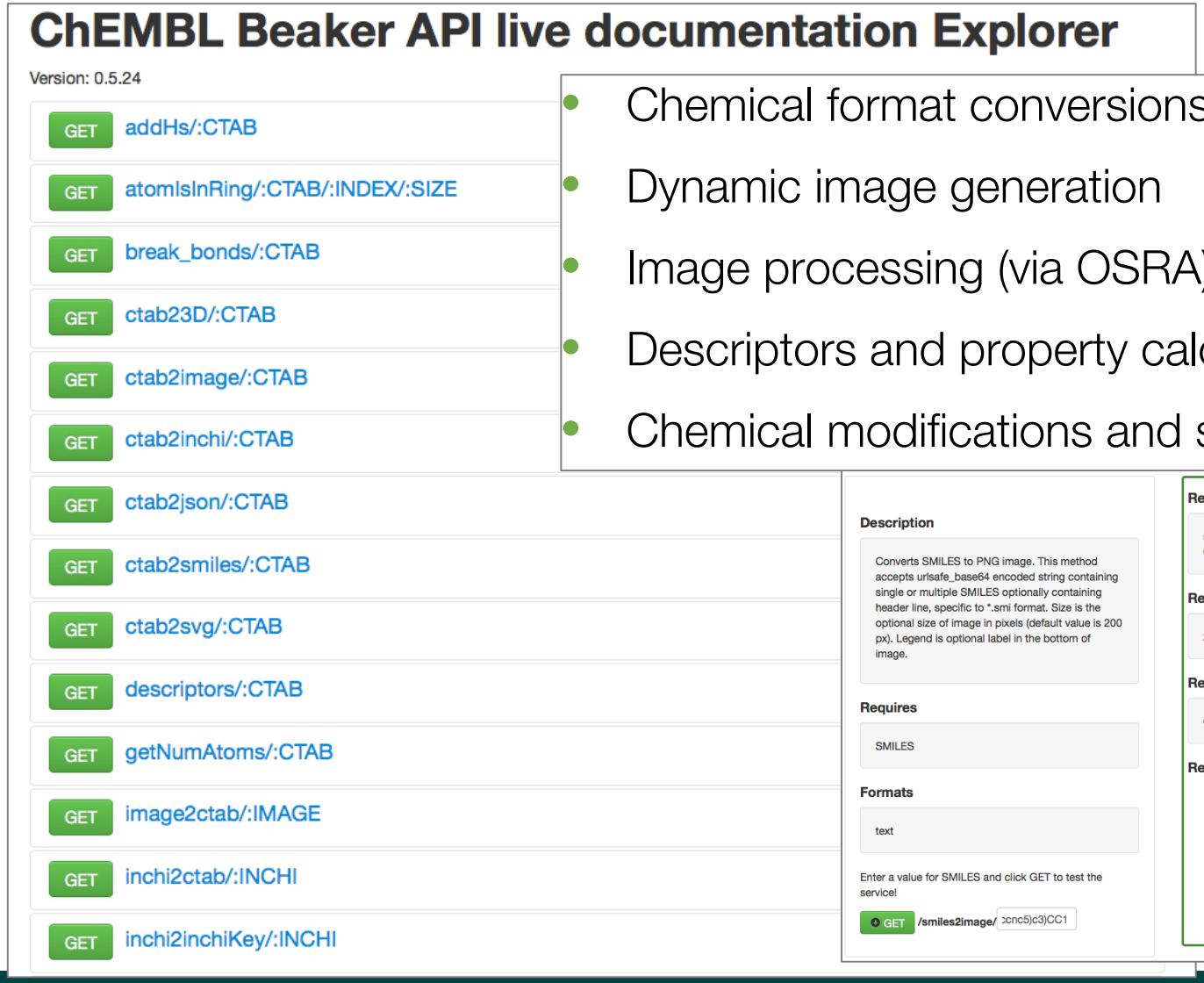


- Retrieve ChEMBL data via RESTful API
- Chemistry based searches
- Easy to integrate into workflows and webpages (CORS enabled)

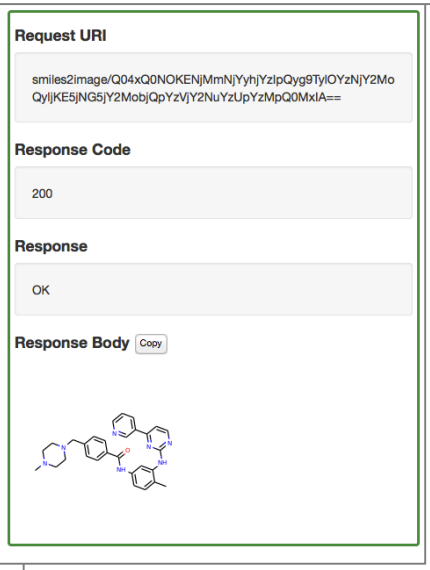
https://www.ebi.ac.uk/chembl/ws

EMBL-EBI

Web services – Utilities (aka ‘Beaker’)



- Chemical format conversions
- Dynamic image generation
- Image processing (via OSRA)
- Descriptors and property calculations
- Chemical modifications and standardization



https://www.ebi.ac.uk/chembl/api/utils/docs

ChEMBL web service developments

- Start of 2014 migrated existing technology stack from Java to Python
  - Developed ChEMBL data model and API using Python django framework
- Currently undertaking comprehensive review of existing web service data endpoints
  - Improved ‘Data’ endpoints available Q1 2015
- Web service python client available
  - https://github.com/chembl/chembl_webresource_client
  - Documentation, ChEMBL-og and myChEMBL

ChEMBL Web Portals

- ADME SARfari
- Neglected Disease Portals

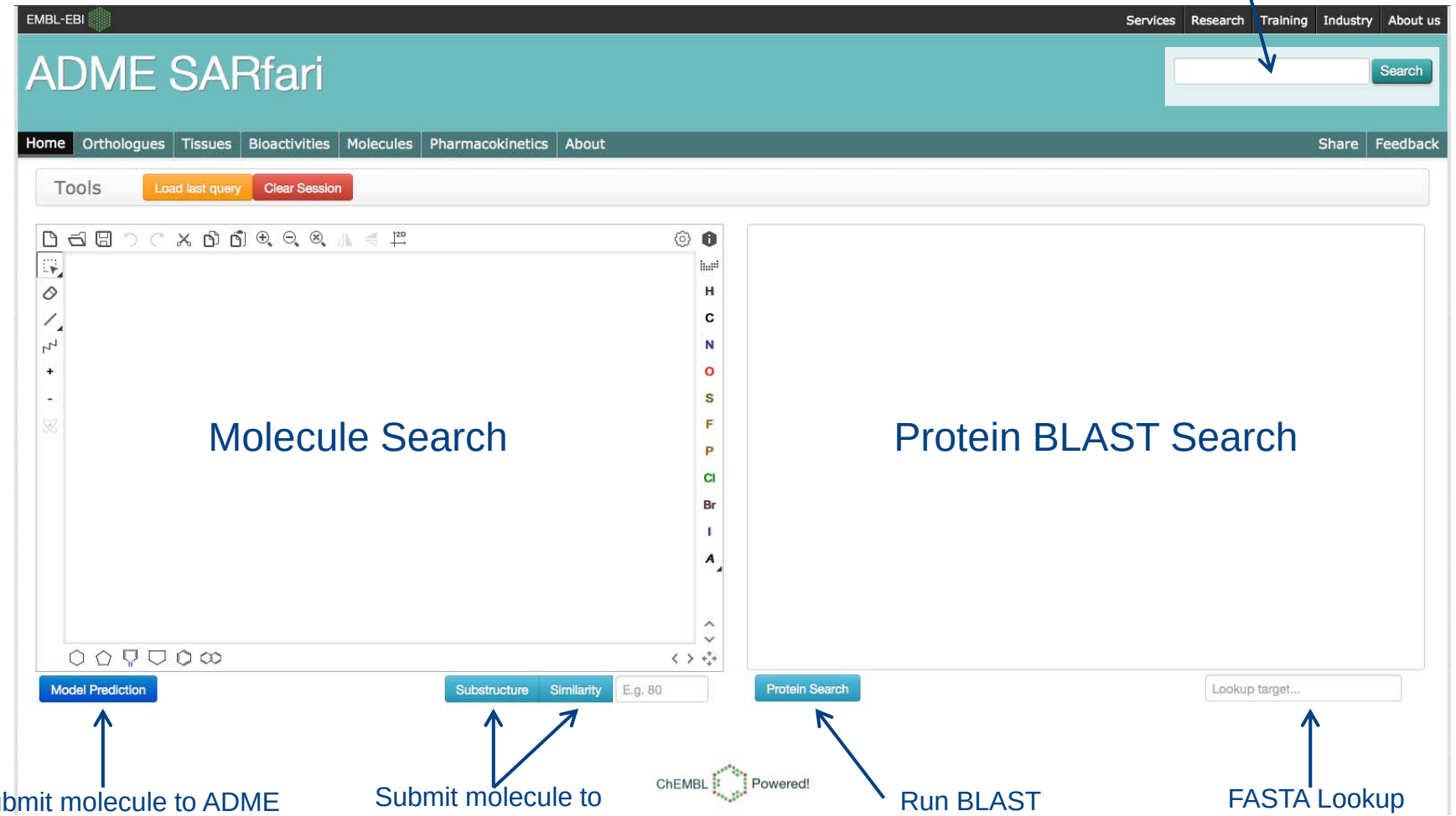
ADME SARfari

- A central resource that integrates ADME related data across multiple EMBL-EBI resources and external resources
- Highlight specific differences relevant for a particular chemotype in relation to known or predicted metabolic route
- Protein and chemical focused searches
- Model based prediction of small molecule protein interactions
- Identify and highlight cross species differences
- Resource developed in collaboration with GSK

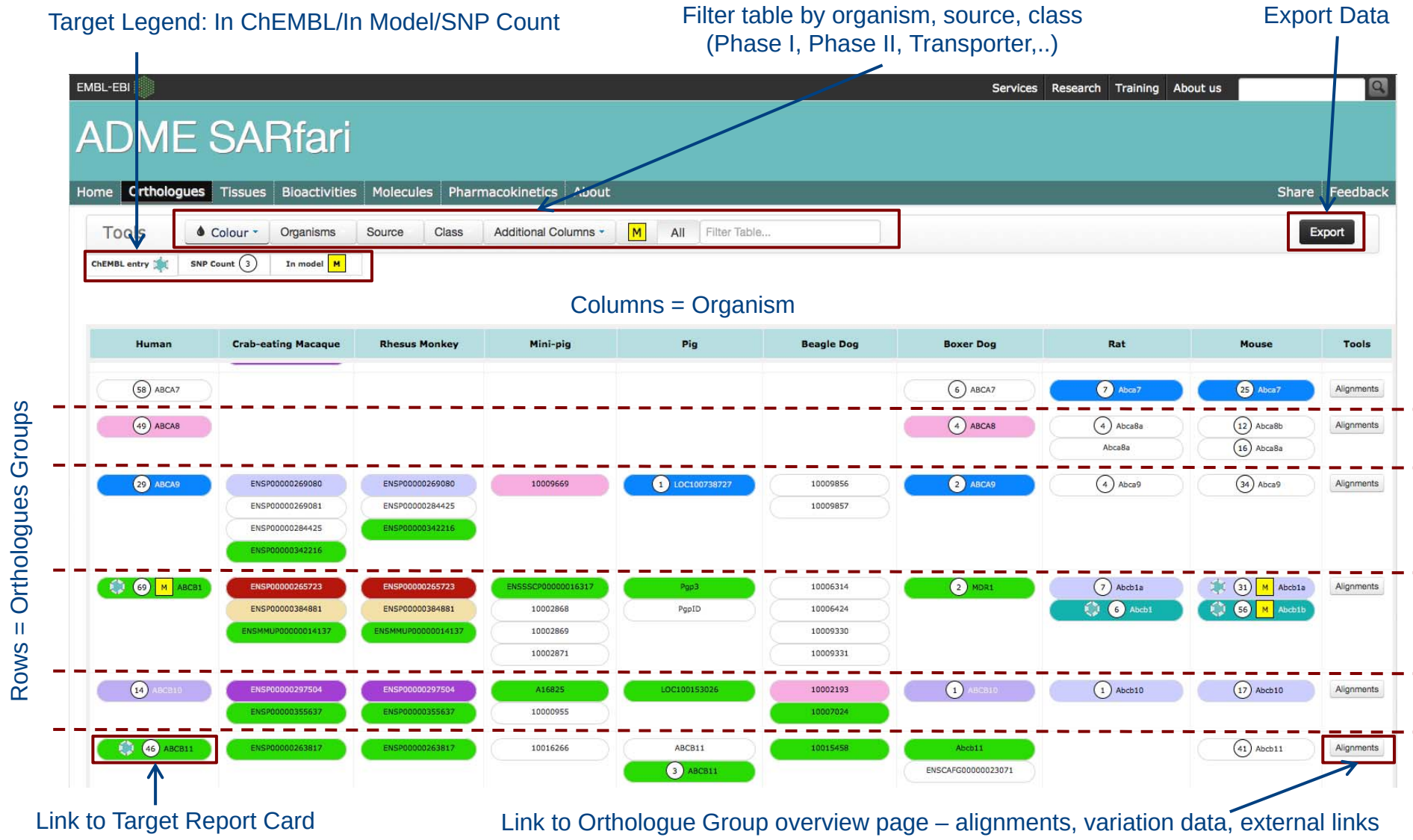
ADME SARfari components

- **Orthologues** – ADME protein targets from human and model organisms
- **Tissues** – Tissue distribution data of human ADME proteins
- **Bioactivities** – ADME Related in vitro protein interaction data
- **Molecules** – ADME related molecules
- **Pharmacokinetics** – PK summary view for a set of molecules (Clearance, Bioavailability, Cmax, Tmax, Vd)
- **Target Report Card** – Protein summary page
- **Orthologue Report Card** – Orthologue set report card page, contains alignment and SNPs details

Homepage


Keyword Search
EMBL-EBI
Services
Research
Training
Industry
About us
ADME SARfari
Search
Home
Orthologues
Tissues
Bioactivities
Molecules
Pharmacokinetics
About
Share
Feedback
Tools
Load last query
Clear Session
Molecule Search
Protein BLAST Search
Model Prediction
Substructure
Similarity
E.g. 80
Protein Search
Lookup target...
ChEMBL
Powered!
Submit molecule to ADME Target Model Prediction
Submit molecule to substructure/similarity search
Run BLAST Search
FASTA Lookup

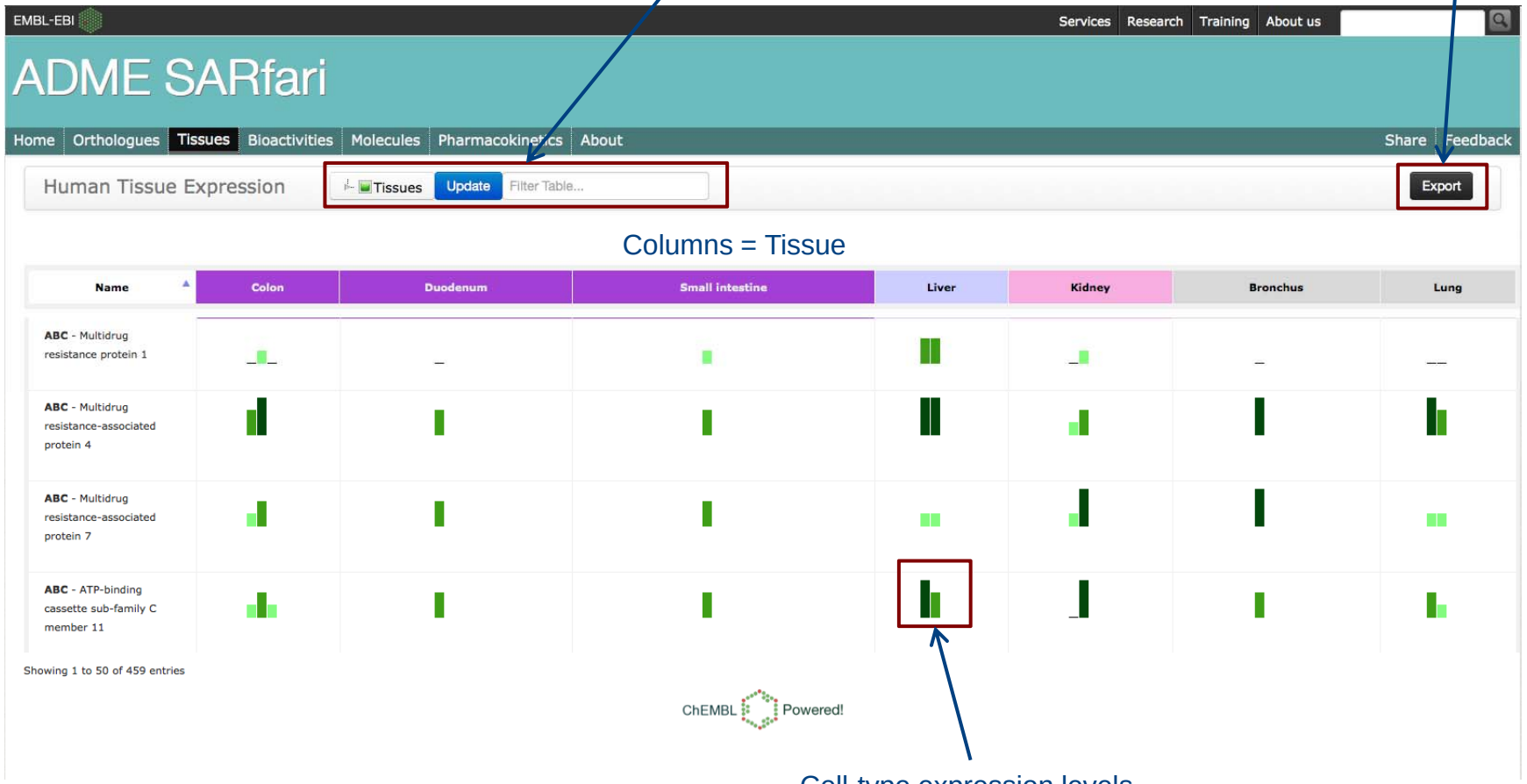
Orthologues – All Data



Tissues – All Data

Select Tissues and keyword filter search

Export Data



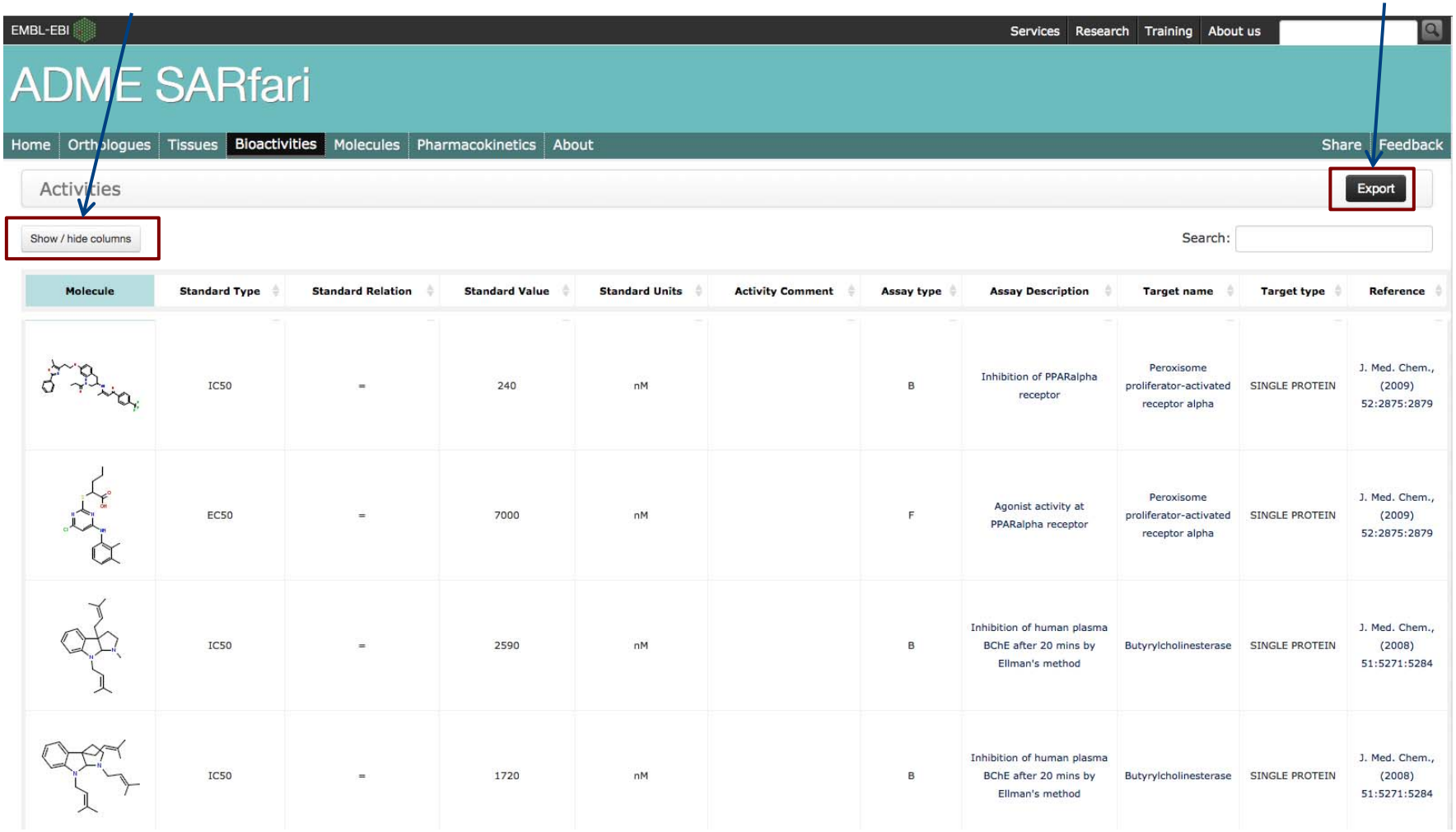
Rows = Human Targets

Cell-type expression levels, hover over to get more details

Bioactivities – All Data

Hide/Show Extra Columns

Export Data



Molecule	Standard Type	Standard Relation	Standard Value	Standard Units	Activity Comment	Assay type	Assay Description	Target name	Target type	Reference
	IC50	=	240	nM		B	Inhibition of PPARalpha receptor	Peroxisome proliferator-activated receptor alpha	SINGLE PROTEIN	J. Med. Chem., (2009) 52:2875:2879
	EC50	=	7000	nM		F	Agonist activity at PPARalpha receptor	Peroxisome proliferator-activated receptor alpha	SINGLE PROTEIN	J. Med. Chem., (2009) 52:2875:2879
	IC50	=	2590	nM		B	Inhibition of human plasma BChE after 20 mins by Ellman's method	Butyrylcholinesterase	SINGLE PROTEIN	J. Med. Chem., (2008) 51:5271:5284
	IC50	=	1720	nM		B	Inhibition of human plasma BChE after 20 mins by Ellman's method	Butyrylcholinesterase	SINGLE PROTEIN	J. Med. Chem., (2008) 51:5271:5284

All ChEMBL ADME Related Bioactivity Data

Molecules – All Data

Display compound properties (MW, LogP,..) or raw pharmacokinetics data (Cl, Cmax, Vd,..)

Export Data (can also export SDF)

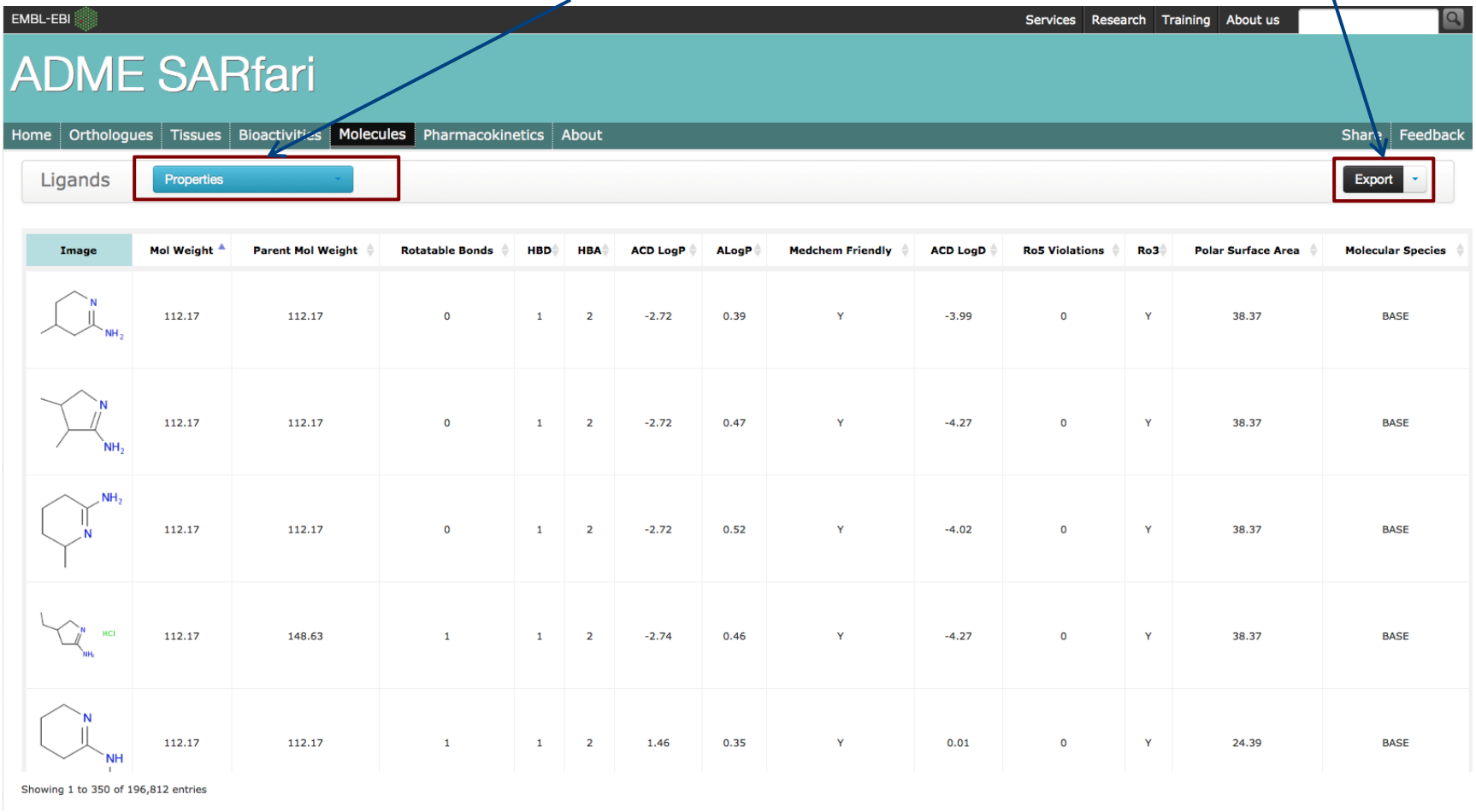
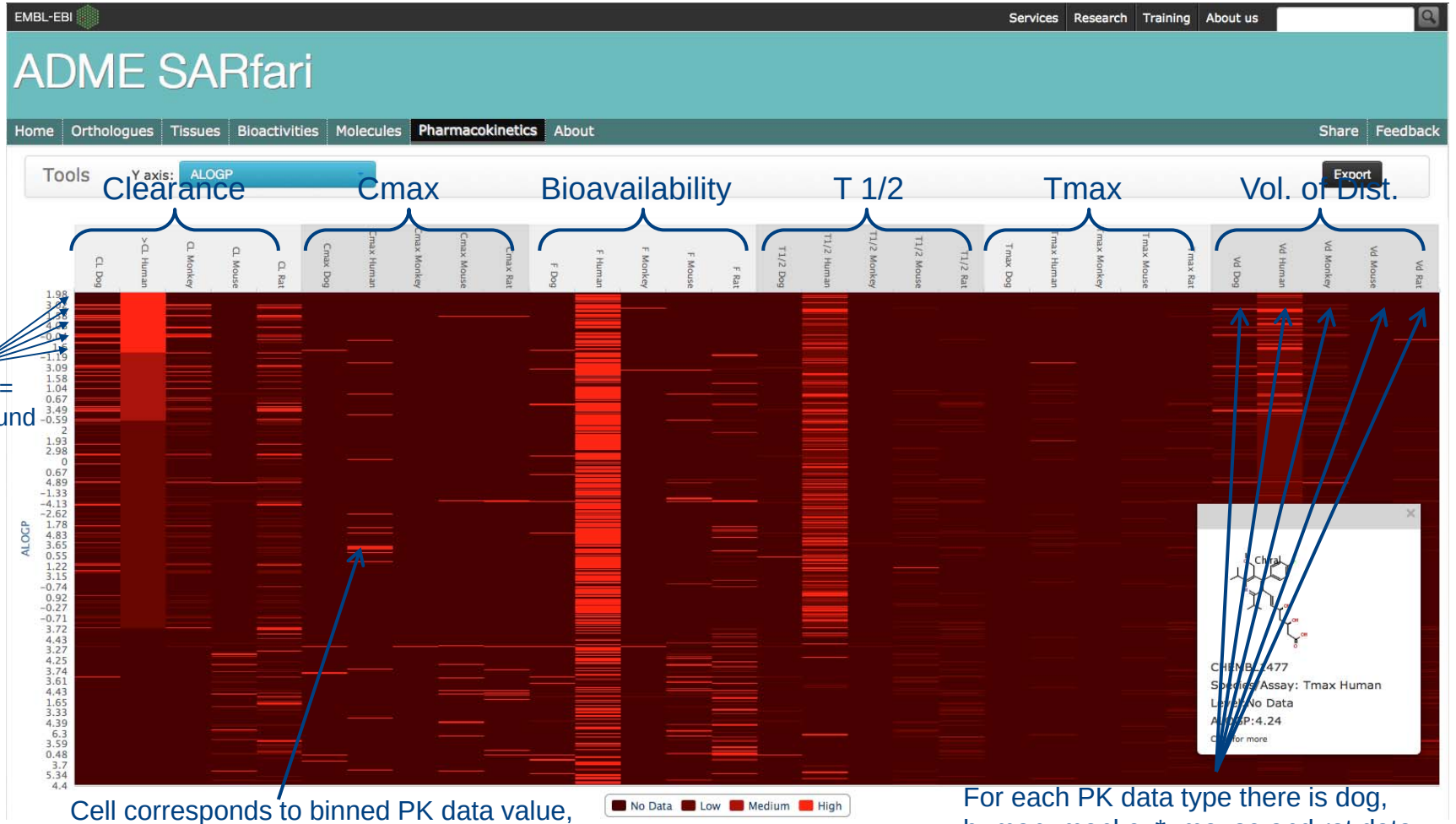


Image	Mol Weight	Parent Mol Weight	Rotatable Bonds	HBD	HBA	ACD LogP	ALogP	Medchem Friendly	ACD LogD	Ro5 Violations	Ro3	Polar Surface Area	Molecular Species
	112.17	112.17	0	1	2	-2.72	0.39	Y	-3.99	0	Y	38.37	BASE
	112.17	112.17	0	1	2	-2.72	0.47	Y	-4.27	0	Y	38.37	BASE
	112.17	112.17	0	1	2	-2.72	0.52	Y	-4.02	0	Y	38.37	BASE
	112.17	148.63	1	1	2	-2.74	0.46	Y	-4.27	0	Y	38.37	BASE
	112.17	112.17	1	1	2	1.46	0.35	Y	0.01	0	Y	24.39	BASE

Showing 1 to 350 of 196,812 entries

All ChEMBL ADME Related Compound Data

Pharmacokinetics



1 Row = Compound

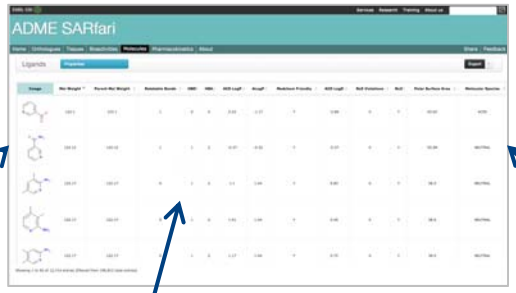
Cell corresponds to binned PK data value, which can be High, Medium, Low or No Data

For each PK data type there is dog, human, monkey*, mouse and rat data

**monkey = all monkey-like species in ChEMBL*

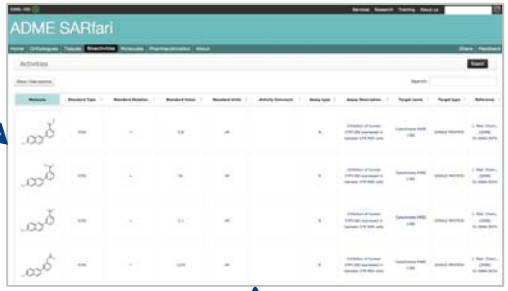
Compound Search Workflow

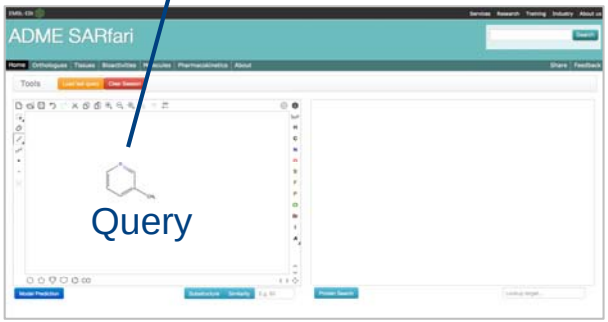
Default Results Page: Molecules



Pharmacokinetics Overview

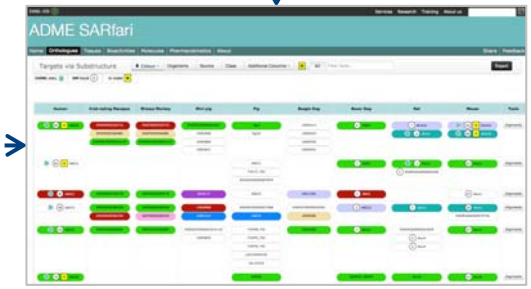
Related Bioactivities





Query

Protein Tissue Expression

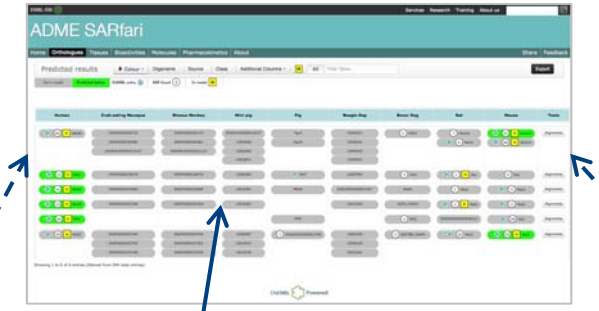


Related Target Orthologues

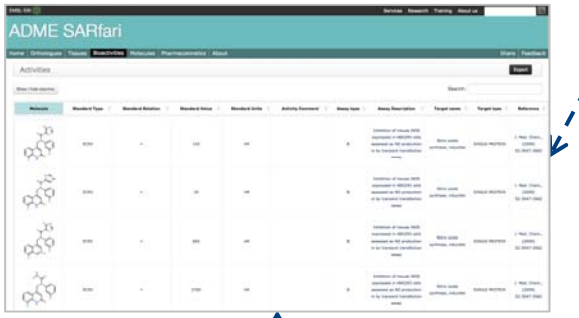
Model Search Workflow

**Orthologue page default colouring changes when running a model search. Targets are coloured green predicted to bind user submitted compound structure.*

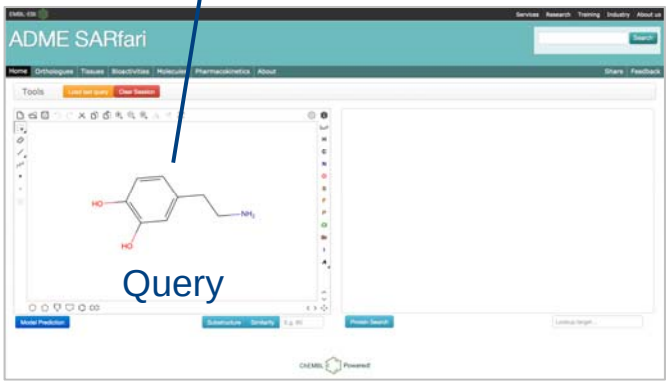
Default Results Page: Orthologues*



Related Target Bioactivities



Protein Tissue Expression



Query

Related Compounds

Pharmacokinetics Overview

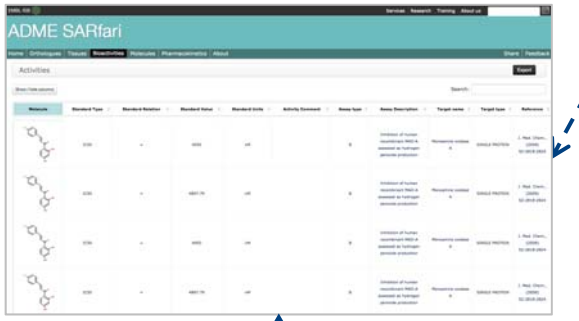
EMBL-EBI

Target Search (BLAST) Workflow

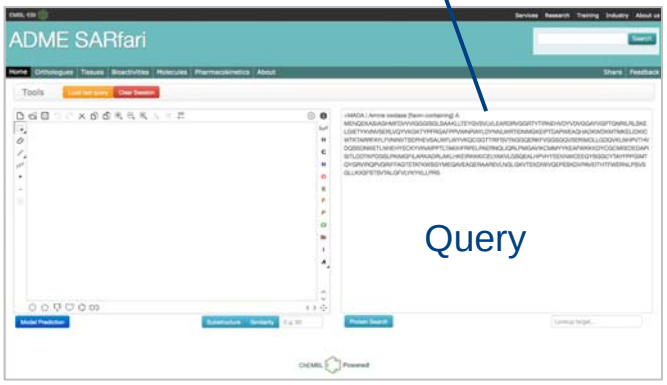
**Orthologue page default colouring changes when running a BLAST search. Targets are coloured different shades of green, dependent on sequence identity to user submitted sequence.*

Default Results Page: Orthologues*

Related Target Bioactivities



Protein Tissue Expression



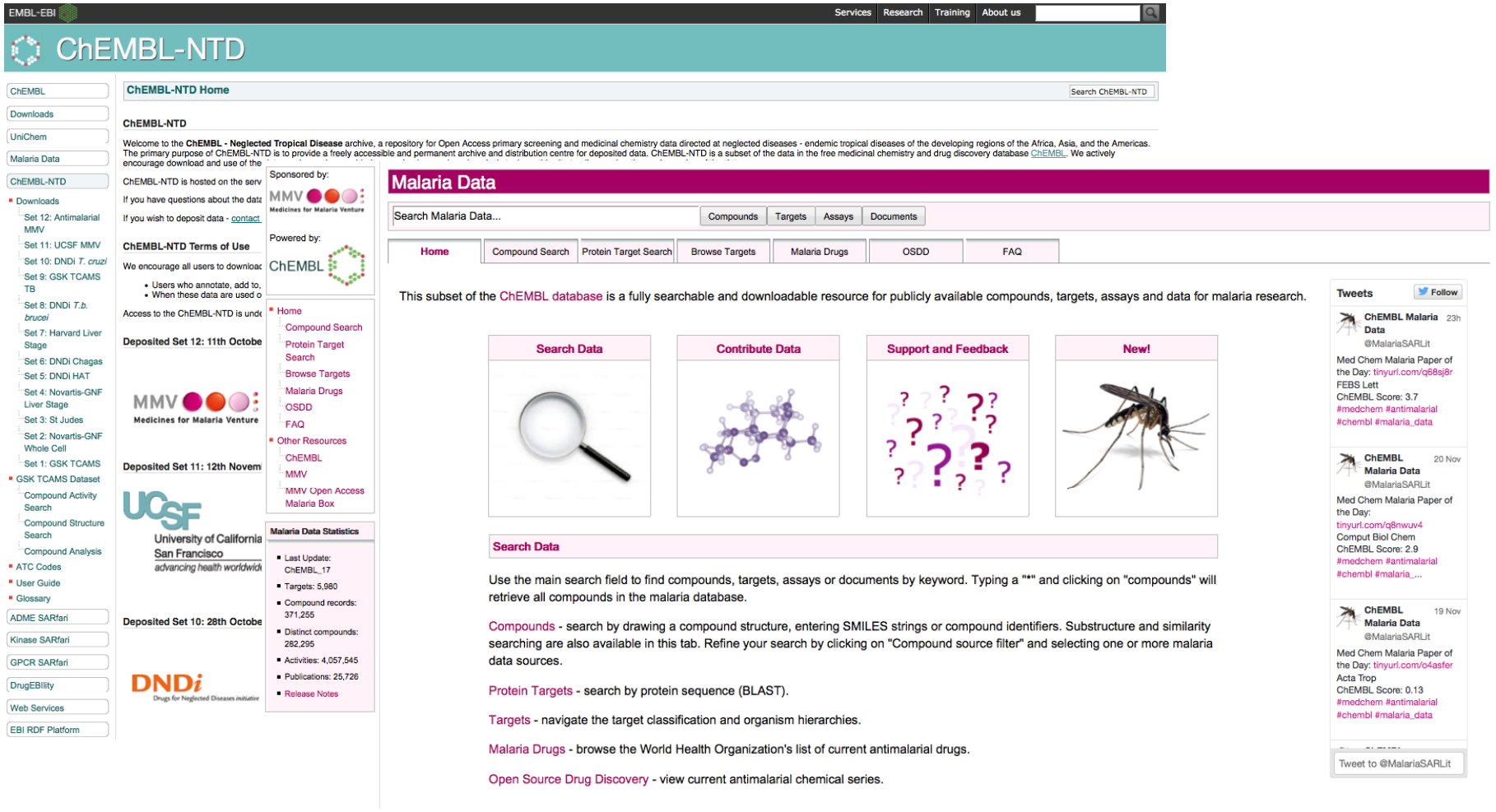
Query

Related Compounds

Pharmacokinetics Overview

EMBL-EBI

Neglected disease portals



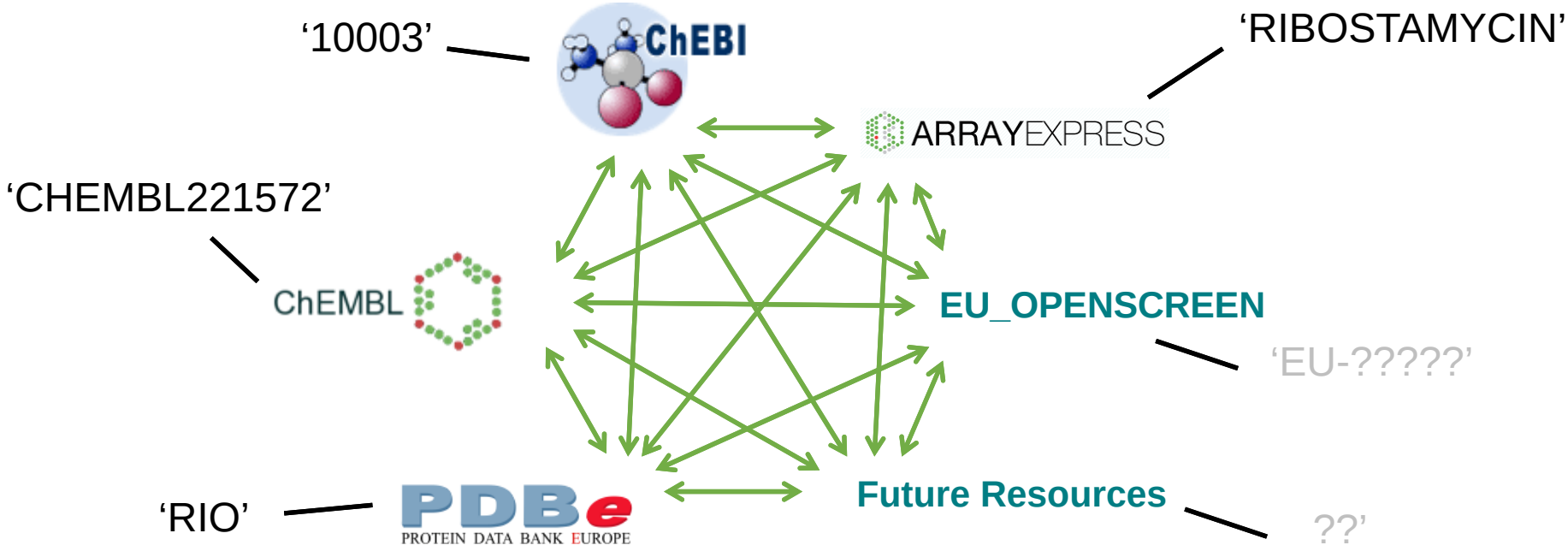
https://www.ebi.ac.uk/chemblntd/

https://www.ebi.ac.uk/chembl/malaria/

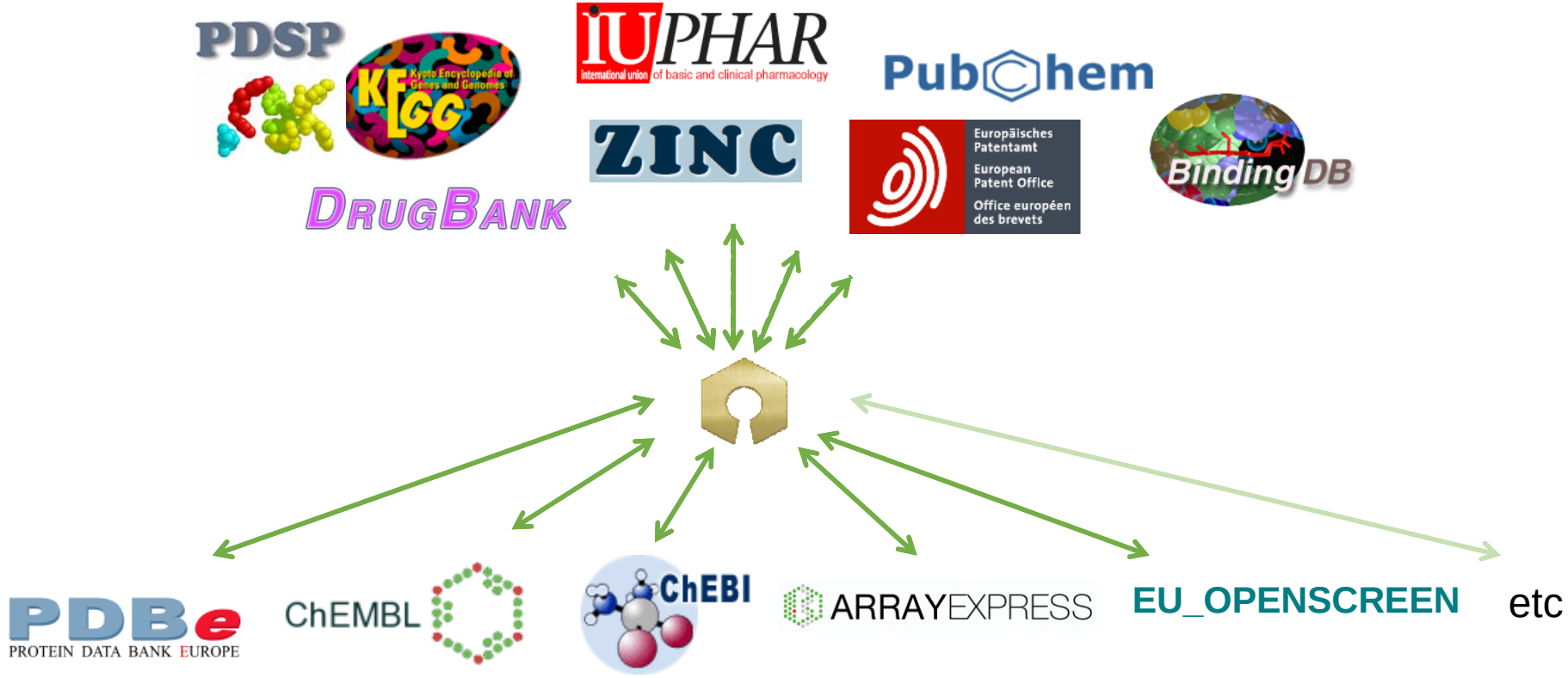
UniChem

Multiple EBI resources hold compound data

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



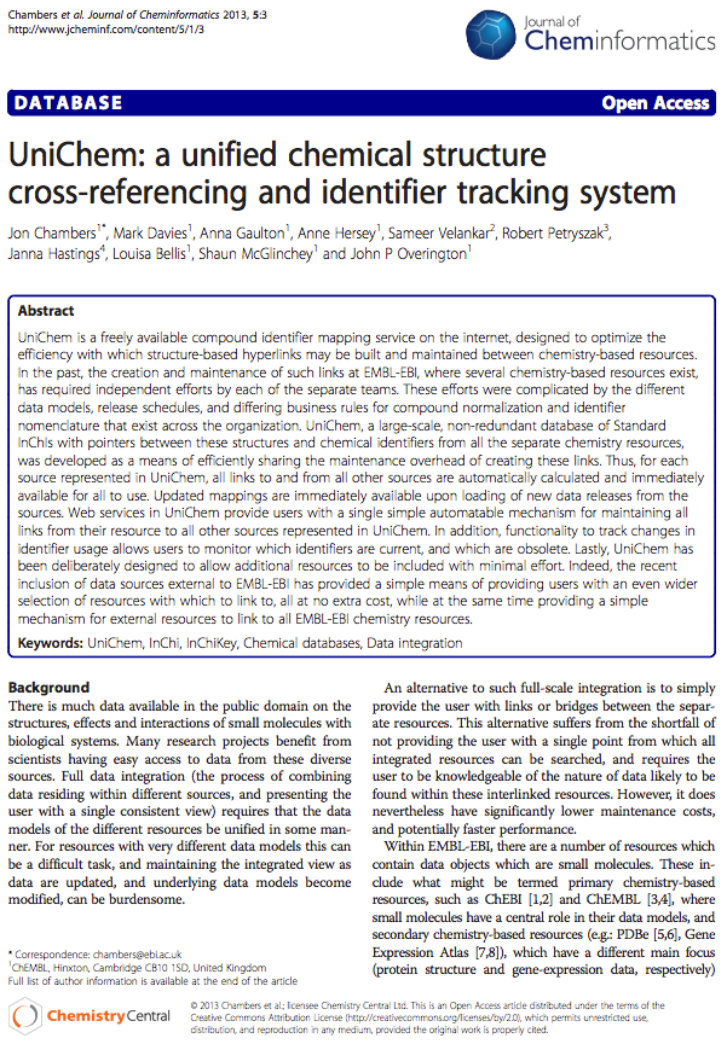
This is why we built UniChem



- 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

What is UniChem?

- InChi based ‘Unified Chemical Identifier' system
- Rapid cross-referencing of chemical structures and their identifiers between databases, tracking changes to ‘id-to-structure’ assignments over time
- Chemical equivalent to ‘UniParc’, but using InChI instead of protein sequence



Chambers et al. Journal of Cheminformatics 2013, 5:3
http://www.jcheminf.com/content/5/1/3

Journal of Cheminformatics

DATABASE Open Access

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

Jon Chambers[1*], Mark Davies[1], Anna Gaulton[1], Anne Hersey[1], Sameer Velankar[2], Robert Petryszak[3], Janna Hastings[4], Louisa Bellis[1], Shaun McGlinchey[1] and John P Overington[1]

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 sources. Full data integration (the process of combining data residing within different sources, and presenting the user with a single consistent view) requires that the data models of the different resources be unified in some manner. For resources with very different data models this can be a difficult task, and maintaining the integrated view as data are updated, and underlying data models become modified, can be burdensome.

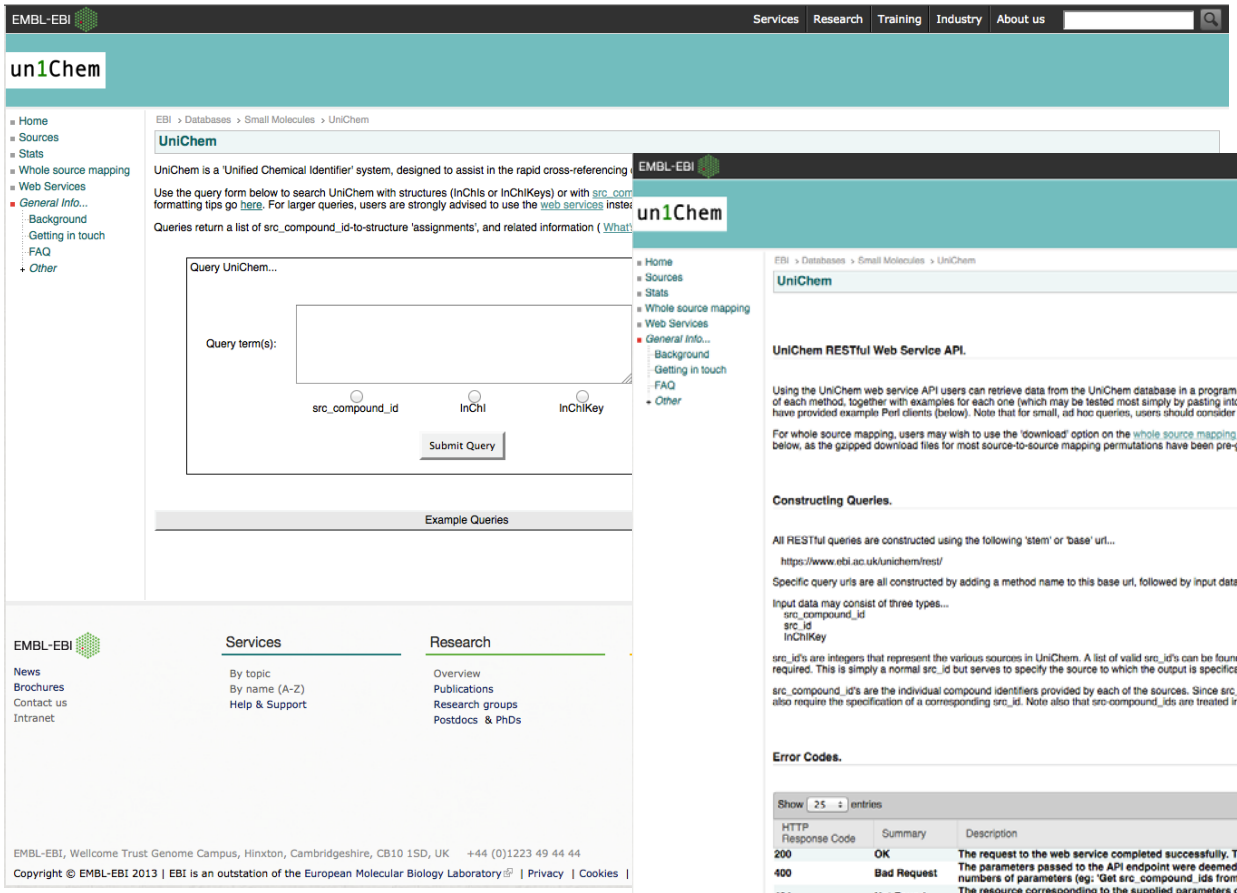
An alternative to such full-scale integration is to simply provide the user with links or bridges between the separate resources. This alternative suffers from the shortfall of not providing the user with a single point from which all integrated resources can be searched, and requires the user to be knowledgeable of the nature of data likely to be found within these interlinked resources. However, it does nevertheless have significantly lower maintenance costs, and potentially faster performance.

Within EMBL-EBI, there are a number of resources which contain data objects which are small molecules. These include what might be termed primary chemistry-based resources, such as ChEBI [1,2] and ChEMBL [3,4], where small molecules have a central role in their data models, and secondary chemistry-based resources (e.g.: PDBe [5,6], Gene Expression Atlas [7,8]), which have a different main focus (protein structure and gene-expression data, respectively)

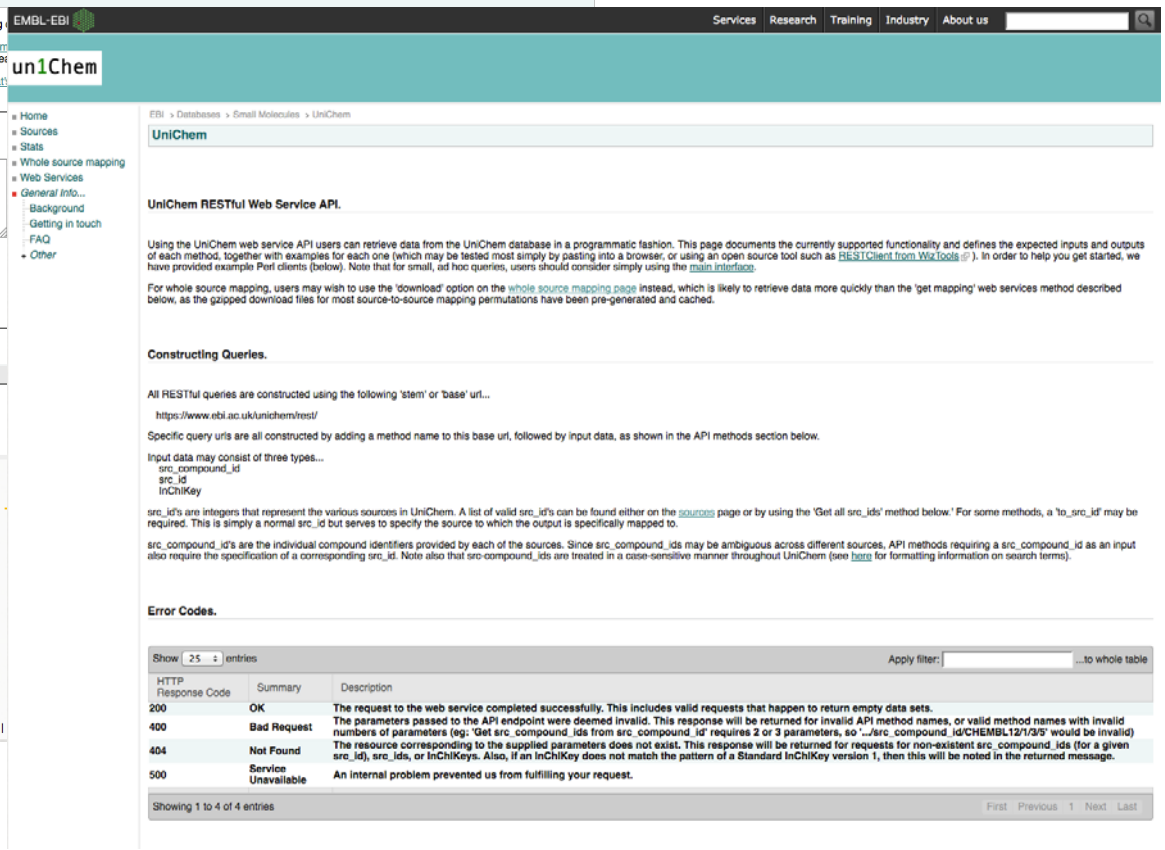
* Correspondence: chambers@ebi.ac.uk
[1]ChEMBL, Hinxton, Cambridge CB10 1SD, United Kingdom
Full list of author information is available at the end of the article

JChemInf (2013), 5(3)

UniChem in action



Web Interface



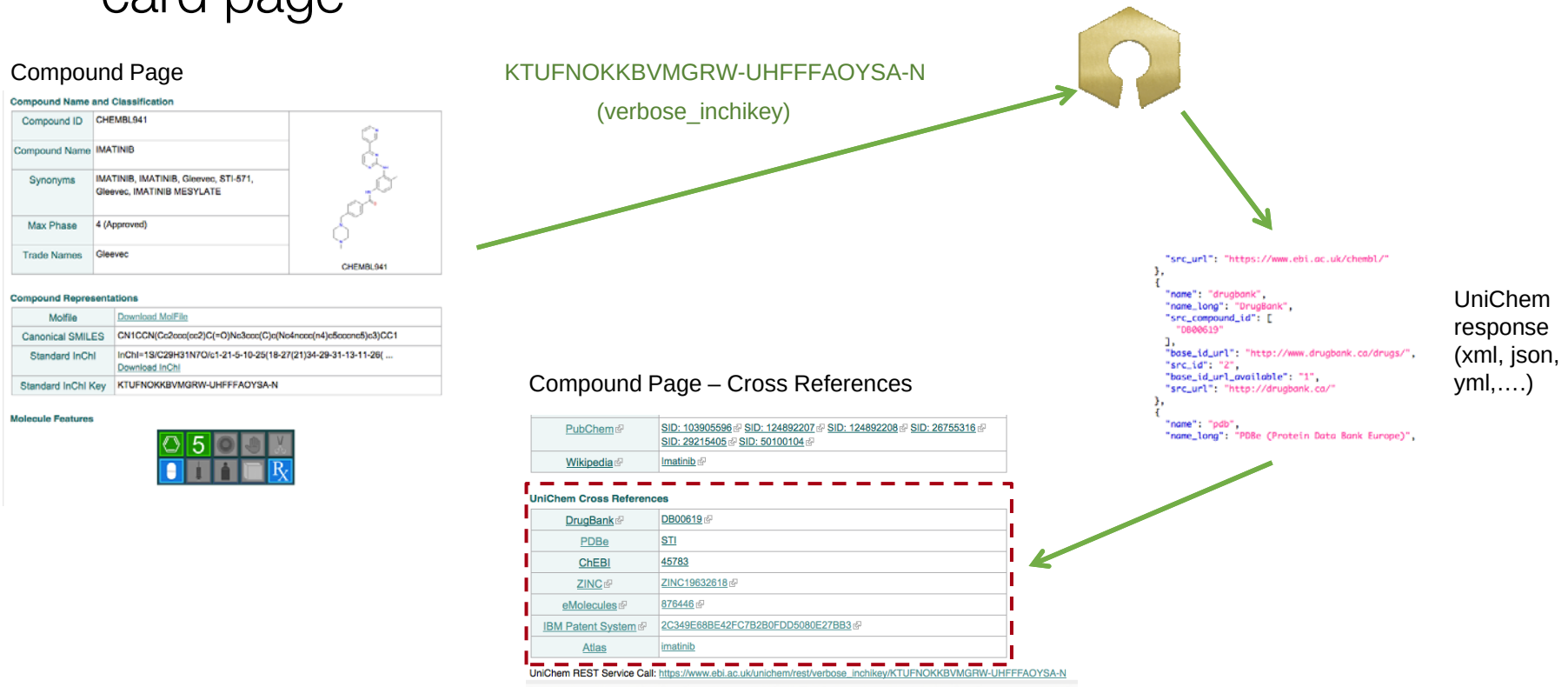
Web Services

https://www.ebi.ac.uk/unichem/

EMBL-EBI

UniChem and ChEMBL

- Compound cross references dynamically called from report card page



https://www.ebi.ac.uk/unichem/rest/verbose_inchikey/KTUFNOKKBVMGRW-UHFFFAOYSA-N

EMBL-EBI chemistry resources

RDF and REST API interfaces

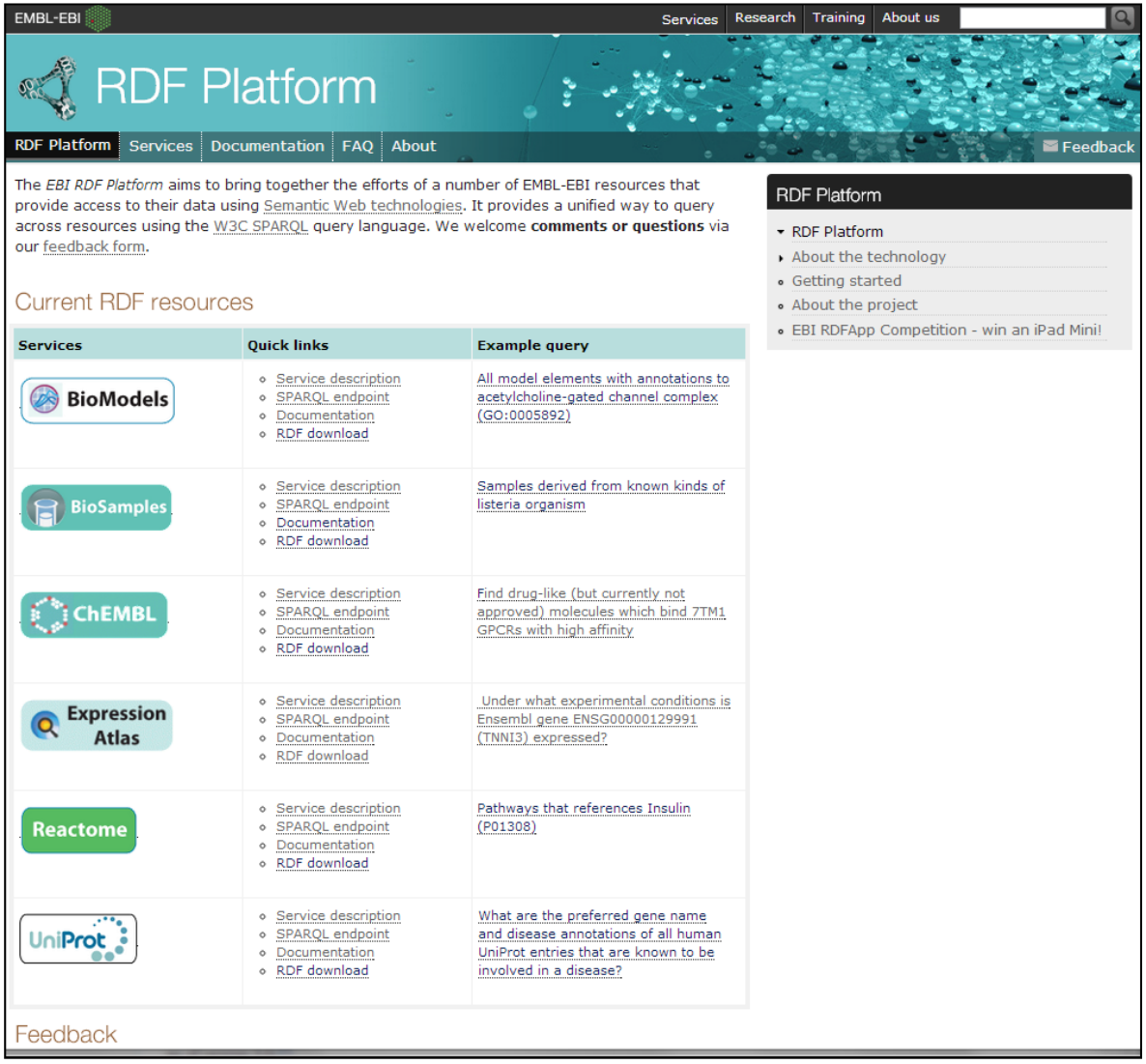
Atlas	PDBe	ChEBI	ChEMBL	SureChEMBL	3rd Party Data
Ligand induced transcript response	Ligand structures from structurally defined protein complexes	Nomenclature of primary and secondary metabolites. Chemical Ontology	Bioactivity data from literature and depositions	Chemical structures from patent literature	ZINC, PubChem, ThomsonPharma DOTF, IUPHAR, DrugBank, KEGG, NIH NCC, eMolecules, FDA SRS, PharmGKB, Selleck, ….
750	15K	24K	1.5M	**~16M**	~55M

UniChem – InChI-based chemical resolver (full + relaxed 'lenses') >70M

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

ChEMBL RDF

EBI-RDF Platform + ChEMBL-RDF



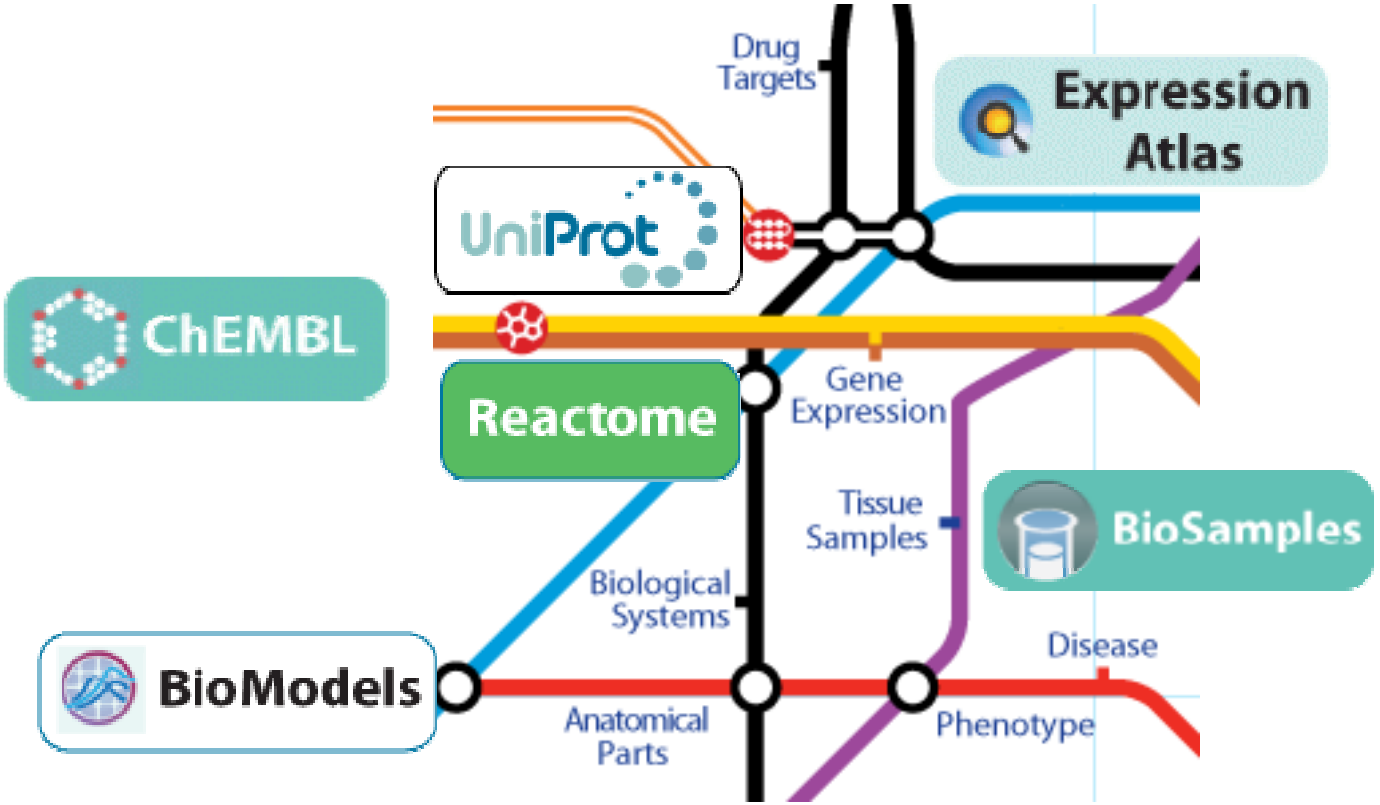
https://www.ebi.ac.uk/rdf/

EMBL-EBI

Why RDF at EBI?

- Interest for a number of years from some of our user base particularly some industry partners
- Betas and pilots starting up within individual EBI projects without coordination
- Overall feeling that technology is maturing
- And community is growing
- EBI RDF has been available elsewhere but users had concerns of stability and faithfulness to source
- Great timing for ChEMBL group as coincided with Open PHACTS – allowed us to ‘publish’ ChEMBL-RDF

Which resources are involved?

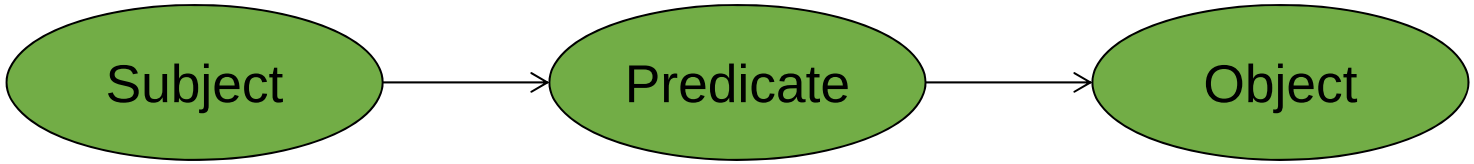


Ensembl coming soon

The Triple

ChEMBL941 → hasName → Imatinib

ChEMBL941 → hasMW → 493.6



ChEMBL941 → a → Molecule

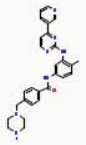
ChEMBL-RDF

```
chembl_molecule:CHEMBL941 a cco:SmallMolecule ;
    rdfs:label "IMATINIB" ;
    skos:prefLabel "IMATINIB" ;
    cco:highestDevelopmentPhase "4"^^xsd:int ;
    cco:hasDocument chembl_document:CHEMBL1138583 .
```

```
chembl_activity:CHEMBL_ACT_7569852 a cco:Activity ;
    rdfs:label "CHEMBL_ACT_7569852" ;
    cco:publishedType "Kd" ;
    cco:publishedRelation "=" ;
    cco:publishedValue "1.1"^^xsd:double ;
    cco:publishedUnits "nM" ;
    cco:standardType "Kd" ;
    cco:standardRelation "=" ;
    cco:standardValue "1.1"^^xsd:double ;
    cco:standardUnits "nM" ;
    cco:hasUnitOnto uo:UO_0000065 ;
    cco:hasQUDT ops:Nanomolar ;
    cco:hasMolecule chembl_molecule:CHEMBL941 ;
    cco:pChembl "8.96"^^xsd:double ;
    bao:BAO_00000208 bao:BAO_0000034 .
```

```
chembl_document:CHEMBL1908390 rdfs:label "CHEMBL1908390" ;
    cco:documentType "PUBLICATION" ;
    dcterms:title "Comprehensive analysis of kinase inhibitor sel
    bibo:issue "11" ;
    bibo:volume "29" ;
    bibo:pageStart "1046" ;
    bibo:pageEnd "1051" ;
    dcterms:date "2011"^^xsd:int ;
    bibo:pmid pubmed:22037378 .
```

Compound | Bioactivity | Assay | Target | Ref

Ingredient	Molweight	Standard Type	Relation	Standard Value	Standard Units	Assay Type	Description	Assay Src Description	Assay Organism	Target Type	Target Name	Target Organism	Reference
 CHEMBL941	493.6	Kd	=	1.1	nM	B	Binding constant for ABL1-nonphosphorylated kinase domain	Scientific Literature		SINGLE PROTEIN	Tyrosine-protein kinase ABL	Homo sapiens	Nat. Biotechnol., (2011) 29:11:1046

```
chembl_assay:CHEMBL1908688 a cco:Assay ;
    rdfs:label "CHEMBL1908688" ;
    dcterms:description "Binding constant for ABL1-nonphosphorylated
    cco:assayType "Binding" ;
    cco:assayTestType "In vitro" ;
    cco:targetConfDesc "Direct single protein target assigned" ;
    cco:targetConfScore "9"^^xsd:int ;
    cco:targetRelType "D" ;
    cco:targetRelDesc "Direct protein target assigned" ;
    cco:hasTarget chembl_target:CHEMBL1862 .
```

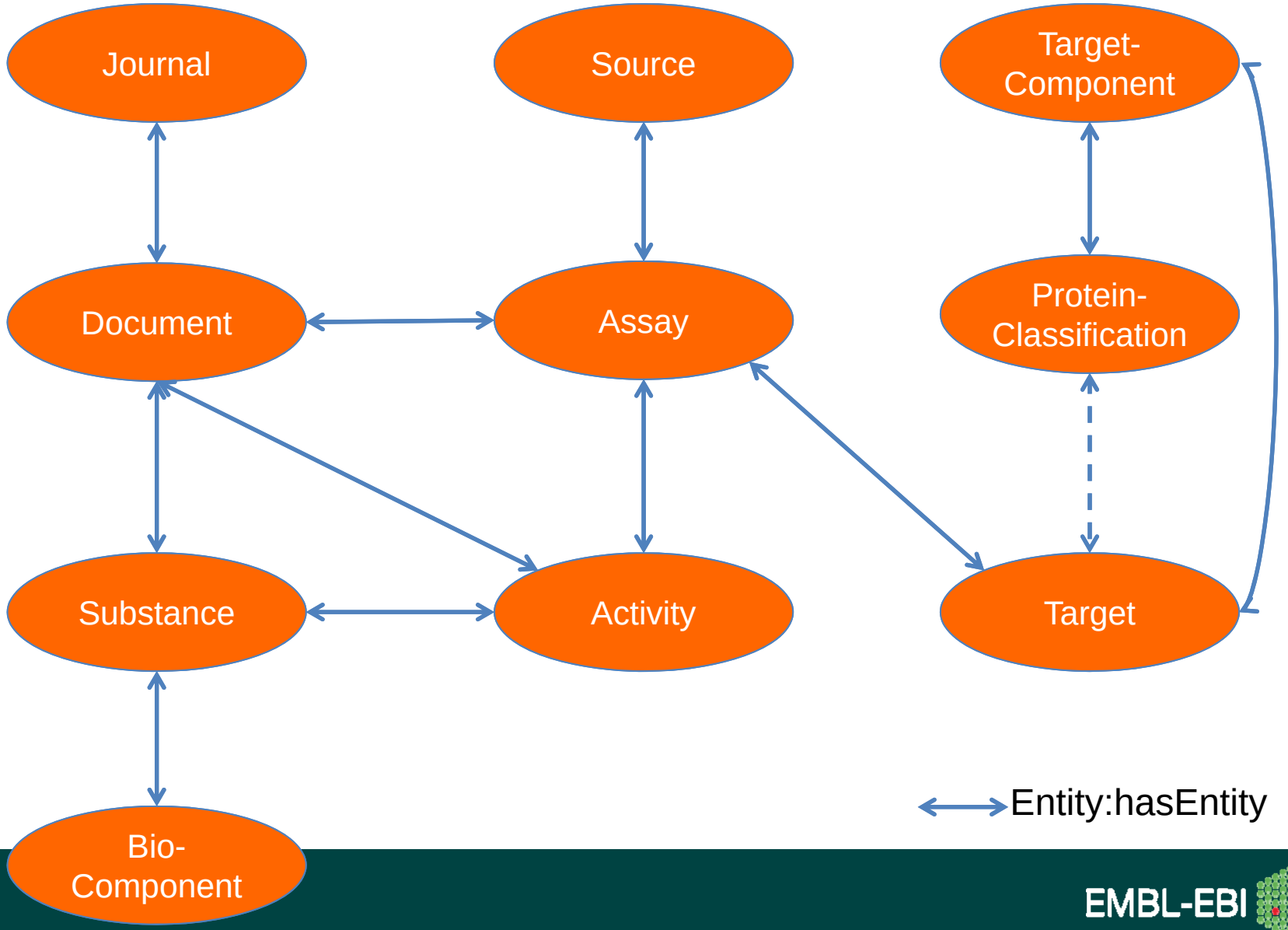
```
chembl_target:CHEMBL1862 a cco:SingleProtein ;
        rdfs:label "Tyrosine-protein kinase ABL" ;
        dcterms:title "Tyrosine-protein kinase ABL" ;
        cco:targetType "SINGLE PROTEIN" ;
        cco:taxonomy ncbitax:9606 , iotax:9606 ;
        cco:organismName "Homo sapiens" ;
        cco:hasTargetComponent chembl_target_cmpt:CHEMBL_TC_173 .
```

ftp://ftp.ebi.ac.uk/pub/databases/chembl/ChEMBL-RDF/

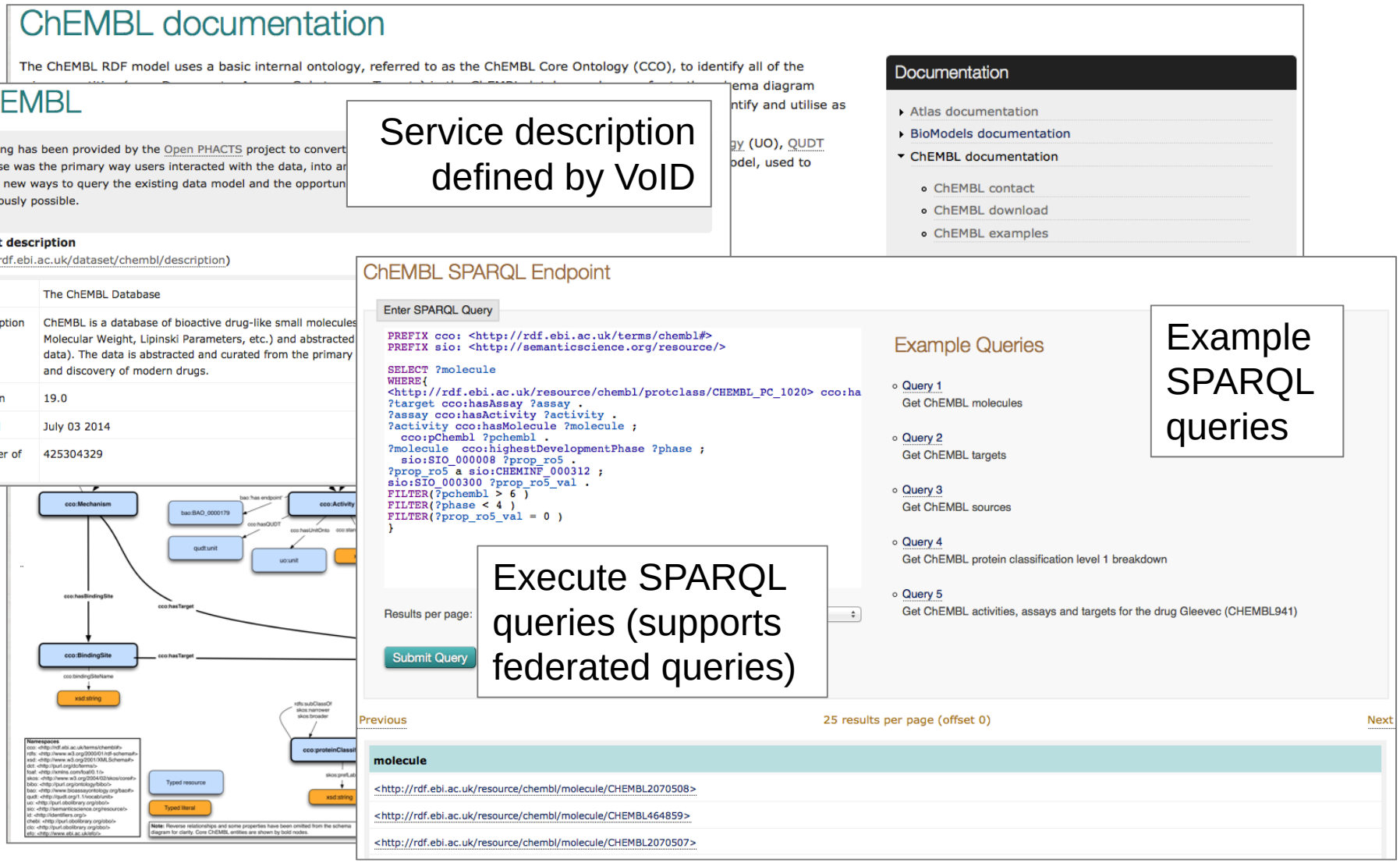
ChEMBL Core Ontology (CCO)

- The skeleton schema used to store ChEMBL classes, object properties and datatype properties
  - The file is also RDF, so can be queried independent of an instances
  - ftp://ftp.ebi.ac.uk/pub/databases/chembl/ChEMBL-RDF/18.0/cco.ttl.gz
  - Namespace: http://rdf.ebi.ac.uk/terms/chembl#
  - Initial focus on Substance (Molecule) and Target Classification
  - In future an additional mapping file may be provided, which maps/aligns ChEMBL classes and properties to external resources

ChEMBL entities relationships



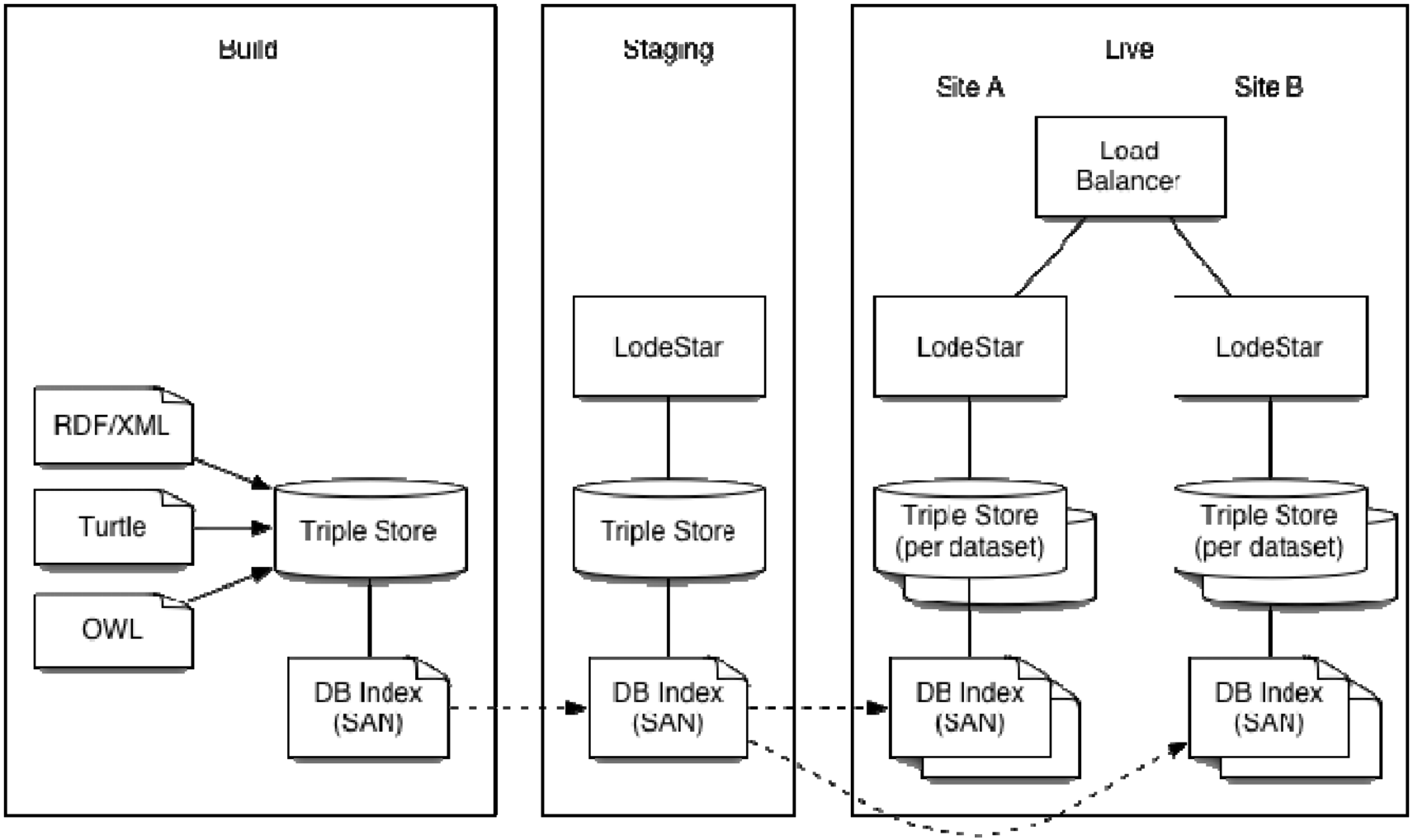
ChEMBL @ EBI-RDF Platform



https://www.ebi.ac.uk/rdf/services/chembl/

MBL-EBI

EBI-RDF Platform infrastructure overview



myChEMBL

What is myChEMBL?

- A Virtual Machine, preloaded with…
- A complete version of the ChEMBL database
  - Chemical structure searching
  - GUI & web services for accessing the database
- A suite of cheminformatics tools
- Tutorials on a range of topics
  - Using ChEMBL data
  - Cheminformatics, machine learning *etc.*
- Completely free and open

What would I use it for?

- On a server
  - Run searches securely, behind a firewall
- On a laptop
  - Offline access: never be without ChEMBL again!
- Easy access to cheminformatics and bioinformatics tools
  - No need to install & maintain them locally
- Education
  - Learn about ChEMBL, cheminformatics and bioinformatics

Key technologies

- Oracle VirtualBox
  - Free to download
  - All major platforms supported
  - Runs the myChEMBL VM
- Ubuntu Linux
  - Operating system used inside myChEMBL VM
  - You never *need* to interact with this directly
- PostgreSQL
  - Open-source database engine

ubuntu

PostgreSQL

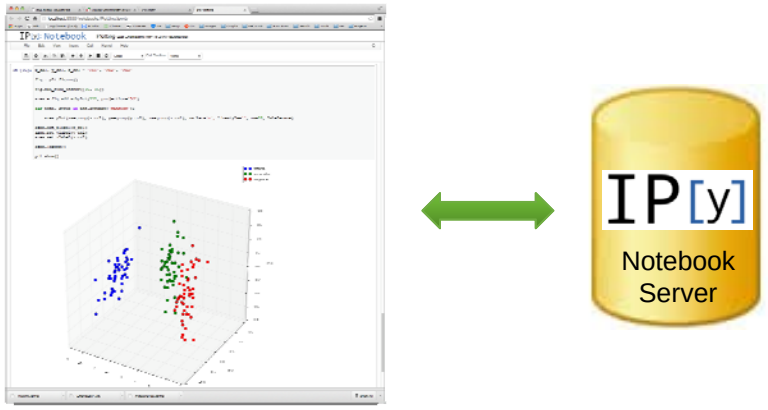
Key technologies

- RDKit
  - Open-source cheminformatics toolkit
  - Enables structure-based searching of database
  - and much more
- The IPython Notebook
  - Used to provide interactive tutorials
- and also…

IPython Notebook

IP[y]: IPython
Interactive Computing

- http://ipython.org/notebook.html
- A browser-based interactive tool combining, in a single document…
  - code execution
  - marked-up text
  - plots
  - images
  - mathematics

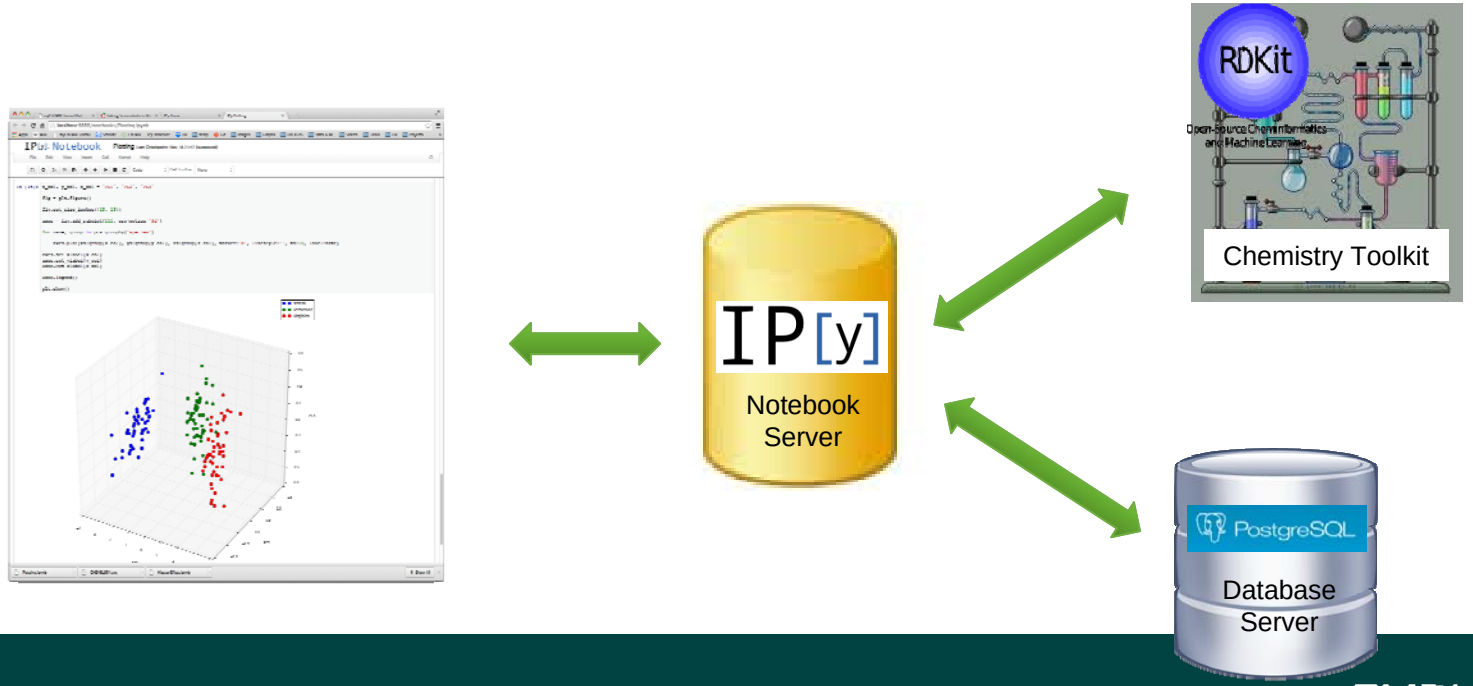


- Notebooks are just text files (JSON) and can easily be shared or viewed online

Notebook Server

IP[y]: IPython
Interactive Computing

- Browser communicates with a Notebook server
  - on local machine, server or inside myChEMBL
- This server communicates with the databases, chemistry toolkit *etc.*



RDKit

- rdkit.org
- Open source cheminformatics toolkit
  - Used extensively inside Novartis
- Fully featured
  - molfile handling, SMARTS, reactions, fingerprints, images, conformers, forcefield, database *etc.*
- Written in C++
  - Accessible from Python ☺
- Tight integration with IPython & Pandas
- Excellent online documentation & tutorials

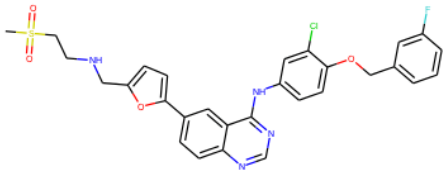
RDKit

Open-Source Cheminformatics and Machine Learning

```
In [1]: from rdkit import Chem
        from rdkit.Chem import AllChem
```

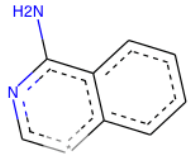
```
In [2]: smiles = 'CS(=O)(=O)CCNCc1oc(cc1)c2ccc3ncnc(Nc4ccc(OC
        mol = Chem.MolFromSmiles(smiles)
        mol
```

Out[2]:



```
In [3]: smarts = 'Nc1ncac2ccccc12'
        query = Chem.MolFromSmarts(smarts)
        query
```

Out[3]:

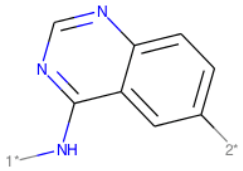



```
In [4]: mol.HasSubstructMatch(query)
```

Out[4]: True

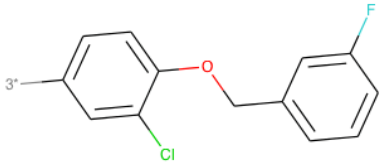
```
In [5]: rxn = AllChem.ReactionFromSmarts('[*:3][N:1]c1ncnc2cc
        mols = rxn.RunReactants((mol, ))[0]
        mols[0]
```

Out[5]:



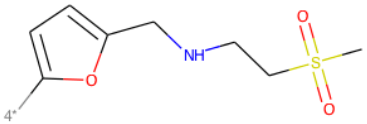

```
In [6]: mols[1]
```

Out[6]:



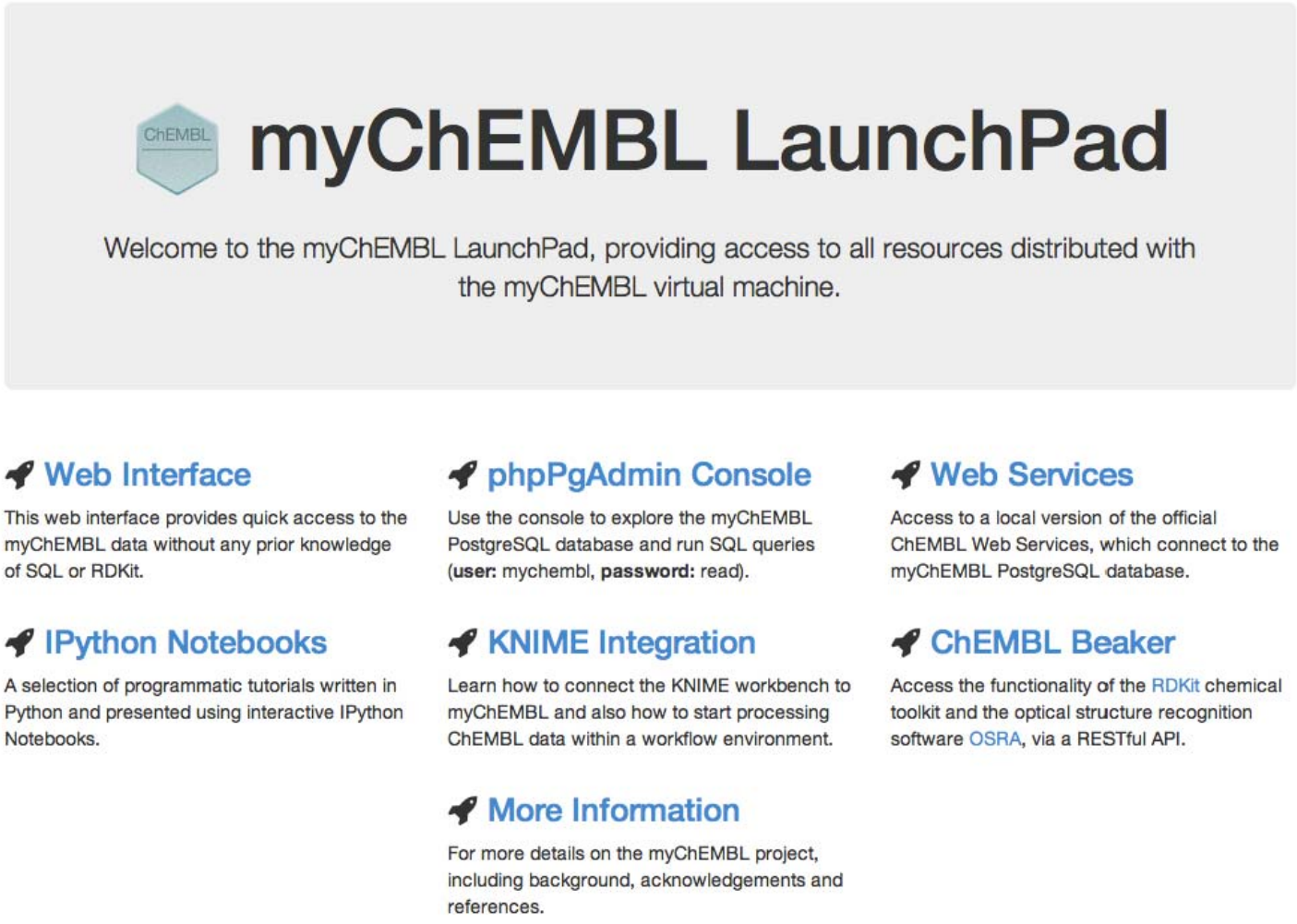

```
In [7]: mols[2]
```

Out[7]:



EMBL-EBI

myChEMBL LaunchPad



myChEMBL installation

- Download files from our ftp site and follow the latest INSTALL instructions
  - ftp://ftp.ebi.ac.uk/pub/databases/chembl/VM/myChEMBL
- Vagrant installation also available, more details on our blog
  - http://chembl.blogspot.co.uk/2014/06/how-to-install-mychembl-using-two.html
- The project is also available on github
  - https://github.com/chembl/mychembl/

myChEMBL Demo

Publications

2014 Chemoinformatics Tool of the Year Contest

OPEN ACCESS
challenges
ISSN 2078-1547
www.mdpi.com/journal/challenges

Communication

myChEMBL: A virtual platform for distributing cheminformatics tools and open data

Mark Davies[1], Michał Nowotka[1], George Papadatos[1], Francis Atkinson[1], Gerard JP van Westen[1], Nathan Dedman[1], Rodrigo Ochoa[2] and John P. Overington[1,*]

1 European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, CB10 1SD, UK; E-Mails: mdavies@ebi.ac.uk; mnowotka@ebi.ac.uk; georgep@ebi.ac.uk; francis@ebi.ac.uk; gerardvw@ebi.ac.uk; ndedman@ebi.ac.uk
2 Programa de Estudio y Control de Enfermedades Tropicales (PECET), Universidad de Antioquia, Medellín, Colombia; E-Mail: rodrigo.ochoa@udea.edu.co

Communication

ChEMBL Beaker: A lightweight web framework providing robust and extensible cheminformatics services.

Michał Nowotka[1], Mark Davies[1], George Papadatos[1] and John P. Overington[1,*]

1 European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, CB10 1SD, UK; E-Mails: mnowotka@ebi.ac.uk; mdavies@ebi.ac.uk; georgep@ebi.ac.uk

Future Outlook

- Common web interface with ChEMBL
  - Web services are already the same
  - New interface being built on top of web services
- Compound registration facility
  - Upload proprietary data into myChEMBL schema
  - Standardise chemical structures as in ChEMBL
- Bioactivity & structure curation interface
  - Will make curators tasks much easier

Exercises

Exercises

1. Find the ChEMBL_ID, molecule weight and InChI Key for the compound **GSK2606414** in the ChEMBL database
2. Does the compound **GSK2606414 exist in any other online chemical resources?**
3. Using the ChEMBL database can you predict what target **GSK2606414 inhibits?**
4. **What protein family does the target of GSK2606414 belong to? And how many other members of this family exist in the ChEMBL database?**
5. How many compounds are similar to **GSK2606414 in the ChEMBL database? Also, export an SD File containing these compounds.** ***(Similar: >=80% Tanimoto)***

ChEMBL support

chembl-help@ebi.ac.uk

mychembl@ebi.ac.uk

ChEMBL-og

Monday, 10 November 2014

Finding key compounds in med. chemistry patents: The open way

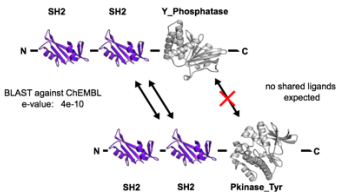
Tuesday, 4 November 2014

An overview and invitation to contribute to ChEMBL curation with PPDMs

PPDMs has been in the making for more than a year and is a follow-up on a conference paper we published in 2012. As in 2012, our objective is to map small molecule binding sites to protein domains, the structural units that form recurring building blocks in the evolution of proteins. An application note describing PPDMs is just out in *Bioinformatics*.

Mapping small molecule binding to protein domains

The mapping facilitates the functional interpretation of small molecule-protein interactions - if you understand which domain in a protein is targeted, you are in a better position to anticipate the downstream effect. Mapping small molecule binding to protein domains also provides a technical advantage to machine-learning approaches that incorporate protein sequence information as a descriptor to predict small molecule bioactivity. Reducing the sequence descriptor to the part that mediates small molecule binding increases the informative content of the descriptor. This is best exemplified by the domain-poisoning problem, illustrated below.



Result of a hypothetical query using as input the rat Tyrosine-protein phosphatase Syp (P35235) - and one of the hits, retrieved from a BLAST query against the ChEMBL target dictionary - the rat Tyrosine-protein kinase SYK (Q64725). The significant e-value for this query results from high scoring alignments of the SH2 domains. At the same time, the overlap between small molecules binding both proteins is expected to be low.

A simple heuristic

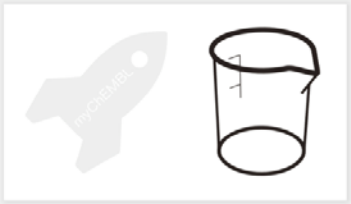
For individual experiments, it is often quite trivial to decide which domain was targeted. For example, medicinal chemists know whether their compound is a kinase inhibitor or one of a handful of SH2 inhibitors. This knowledge, while easily gleaned by the expert, is implicit and cannot be accessed programmatically. Hence we

Monday, 27 October 2014

Django model describing ChEMBL database.

Wednesday, 22 October 2014

myChEMBL 19 Released



We are very pleased to announce that the latest myChEMBL release, based on the ChEMBL 19 database, is now available to download. In addition to the extra data, you will also find a number a great new features. So what's new then?

More core chemoinformatics tools

We have included OSRA (Optical Structure Recognition), which is useful for extracting compound structures from images. OSRA can be accessed from the command line or by very convenient web interface, provided by Beaker (described below). We've also added OpenBabel - another great open source cheminformatics toolkit. This means you can now experiment with both RDKit and OpenBabel and use whichever you prefer.

ChEMBL Beaker

myChEMBL now ships with a local instance the ChEMBL Beaker service. For those not familiar with Beaker, the service provides users with an array of chemoinformatics utilities via a RESTful API. Under the hood, Beaker is using RDKit and OSRA to carry out its methods. With the addition of Beaker in myChEMBL, users can now carry out the following tasks in secure local environment:

- Convert chemical structure bewteen multiple formats
- Extract compound information from images and pdfs
- Generate compound images in raster (png) and vector (svg) forms
- Generate HTML5 ready representation of compound structure
- Generate compound fingerprints
- Generate compound descriptors
- Identify Maximum Common Substructure

http://chembl.blogspot.co.uk/

EMBL-EBI

ChEMBL references

Published online 7 November 2013 Nucleic Acids Research, 2014, Vol. 42, Database issue D1083–D1090
doi:10.1093/nar/gkt1031

The ChEMBL bioactivity database: an update

A. Patrícia Bento, Anna Gaulton, Anne Hersey, Louisa J. Bellis, Jon Chambers, Mark Davies, Felix A. Krüger, Yvonne Light, Lora Mak, Shaun McGlinchey, Michal Nowotka, George Papadatos, Rita Santos and John P. Overington*

European Molecular Biology Laboratory European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK

Received September 30, 2013; Accepted October 7, 2013

ABSTRACT

ChEMBL is an open large-scale bioactivity database (https://www.ebi.ac.uk/chembl), previously described in the 2012 Nucleic Acids Research Database Issue. Since then, a variety of new data sources and improvements in functionality have contributed to the growth and utility of the resource. In particular, more comprehensive tracking of compounds from research stages through clinical development to market is provided through the inclusion of data from United States Adopted Name applications; a new richer data model for representing drug targets has been developed; and a number of methods have been put in place to allow users to more easily identify reliable data. Finally, access to ChEMBL is now available via a new Resource Description Framework format, in addition to the web-based interface, data downloads and web services.

INTRODUCTION

ChEMBL is an open large-scale bioactivity database containing information largely manually extracted from the medicinal chemistry literature. Information regarding the compounds tested (including their structures), the biological or physicochemical assays performed on these and the targets of these assays are recorded in a structured form, allowing users to address a broad range of drug discovery questions. Applications of the data include the identification of suitable chemical tools for a target; investigation of the selectivity and off-targets effects of drugs; large-scale data mining, such as the construction of predictive models for targets and identification of bioisostere replacements or activity cliffs (1–4); and as a key component of integrated drug discovery platforms (5–7). In addition to literature-extracted information, ChEMBL also integrates deposited screening results and bioactivity data from other key public databases [e.g. PubChem BioAssay (8)], and information about approved drugs from resources such as the U.S. Food and Drug Administration (FDA) Orange Book (9) and DailyMed (http://dailymed.nlm.nih.gov/dailymed). Details of the data extraction process, curation and data model have been published previously (10); therefore, the current article focuses on recent enhancements to ChEMBL.

DATA CONTENT

Release 17 of the ChEMBL database contains information extracted from >51 000 publications, together with bioactivity data sets from 18 other sources (depositors and databases). In total, there are now >1.3 million distinct compound structures and 12 million bioactivity data points. The data are mapped to >9000 targets, of which 2827 are human protein targets. Data sets added over the past 2 years include the following: neglected disease screening results from projects funded by Medicines for Malaria Venture (11), Drugs for Neglected Diseases initiative (http://www.dndi.org), World Health Organization TDR programme (WHO-TDR) (12), Open Source Malaria (http://opensourcemalaria.org), Harvard University (13) and GlaxoSmithKline (14); kinase screening results from Millipore (15), and several groups using the Protein Kinase Inhibitor Set compound collection (16); supplementary bioactivity data associated with publications from GlaxoSmithKline (17–19); and information from several other databases including DrugMatrix (https://ntp.niehs.nih.gov/drugmatrix/index.html), TP-search (20) and Open TG-GATEs (21).

NEW DEVELOPMENTS

Tracking compound progression

Although the extraction of structure–activity relationship data from medicinal chemistry literature provides a good overview of drug discovery research, a fuller picture of drugs in development and marketed products is obtained only by combining literature data with other information

*To whom correspondence should be addressed. Tel: +44 1223 492666; Fax: +44 1223 494468; Email: jpo@ebi.ac.uk

Bento *et al.* 2014

D1100–D1107 Nucleic Acids Research, 2012, Vol. 40, Database issue Published online 23 September 2011
doi:10.1093/nar/gkr777

ChEMBL: a large-scale bioactivity database for drug discovery

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

[1]EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, [2]David Michalovich Scientific Consulting, London and [3]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/chembldb.

INTRODUCTION

A wealth of information on the activity of small molecules and biotherapeutics exists in the literature, and access to this information can enable many types of drug discovery analysis and decision making. For example: selection of tool compounds for probing targets or pathways of interest; identification of potential off-target activities of compounds which may pose safety concerns, explain existing side effects or suggest new applications for old compounds; analysis of structure–activity relationships (SAR) for a compound series of interest; assessment of *in vivo* absorption, distribution, metabolism, excretion and toxicity (ADMET) properties; or construction of predictive models for use in selection of compounds potentially active against a new target (1–5). Access to this information is especially important due to the continuing shift in fundamental research on disease mechanisms from the private to public sectors.

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 public-domain bioactivity resources have been developed. PubChem BioAssay (6) and ChemBank (7) are large archival databases providing access to millions of deposited screening results, typically from high-throughput screening (HTS) experiments. A number of other primary resources extract bioactivity data from literature, but tend to focus on particular thematic areas, and primarily on binding affinity information. For example, BindingDB contains quantitative binding constants manually extracted from publications, focusing chiefly on proteins that are considered to be potential drug targets (8). PDBBind (9), Binding MOAD (10) and AffinDB (11) contain binding affinity information for protein–ligand complexes found in the Protein Data Bank (PDB, 12). PDSP Ki database stores screening data from the National Institute of Mental Health's Psychoactive Drug Screening Program (13). BRENDA provides binding constants for enzymes (14), IUPHAR contains ligand information for receptors and ion channels (15), while GLIDA

*To whom correspondence should be addressed. Tel: + 44 (0) 1223 492 666; Fax: + 44 (0) 1223 494 468; Email: jpo@ebi.ac.uk

Gaulton *et al.* 2012

EMBL-EBI

Acknowledgements

ChEMBL team:

- John Overington
- Anna Gaulton
- Mark Davies
- Patricia Bento
- Jon Chambers
- Francis Atkinson
- Louisa Bellis
- George Papadatos
- Nathan Dedman
- Michal Nowotka
- Ines Smit
- Gerard van Westen
- Grace Mugumbate
- Joey Bach Hardie
- Yvonne Light
- Shaun McGlinchey
- Ruth Akhtar
- Rita Santos
- Felix Krueger

Answers

1. Go to the ChEMBL homepage and carry out compound keyword search for **GSK2606414 -> Go to compound report card page for result:**
   - ChEMBL_ID: CHEMBL2171124
   - MW: 451.4
   - InChI Key: SIXVRXARNAVBTC-UHFFFAOYSA-N

2. Go to the 'UniChem Cross Reference' section on the **GSK2606414** report card. Cross references (23/11/14):
   - PDBe - 0WH
   - SureChEMBL - SCHEMBL868254
   - Thomson Pharma (PubChem) - 126495735
   - PubChem - 53469448

Answers

3. Go to the ‘Bioactivity Summary’ section on the **GSK2606414** report card and click on the IC50 pie chart segment -> Sort the resulting bioactivity table by Standard Value in ascending order -> Target with lowest IC50 value (0.4 nM):
   - PERK (CHEMBL6030)

4. Go to the ‘Protein Target Classification’ section on the PERK report card and copy a classification level -> Go to target browser page (https://www.ebi.ac.uk/chembl/target/browser) and search for chosen classification:
   - Protein Kinase and 613 members of this family in ChEMBL

Answers

3. Go to ChEMBL homepage and draw* **GSK2606414** structure -> Set search type to ‘Similarity’, set Tanimoto cut-off to ‘>= 80%’ and click ‘Fetch Compounds’ button

   - 34 compounds are returned (includes query, so 33). To export SDF (or XLS) use dropdown menu top-right of results table

You can get copy structure form **GSK2606414/CHEMBL2171124 report card page*