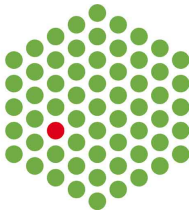
The ChEMBL Database

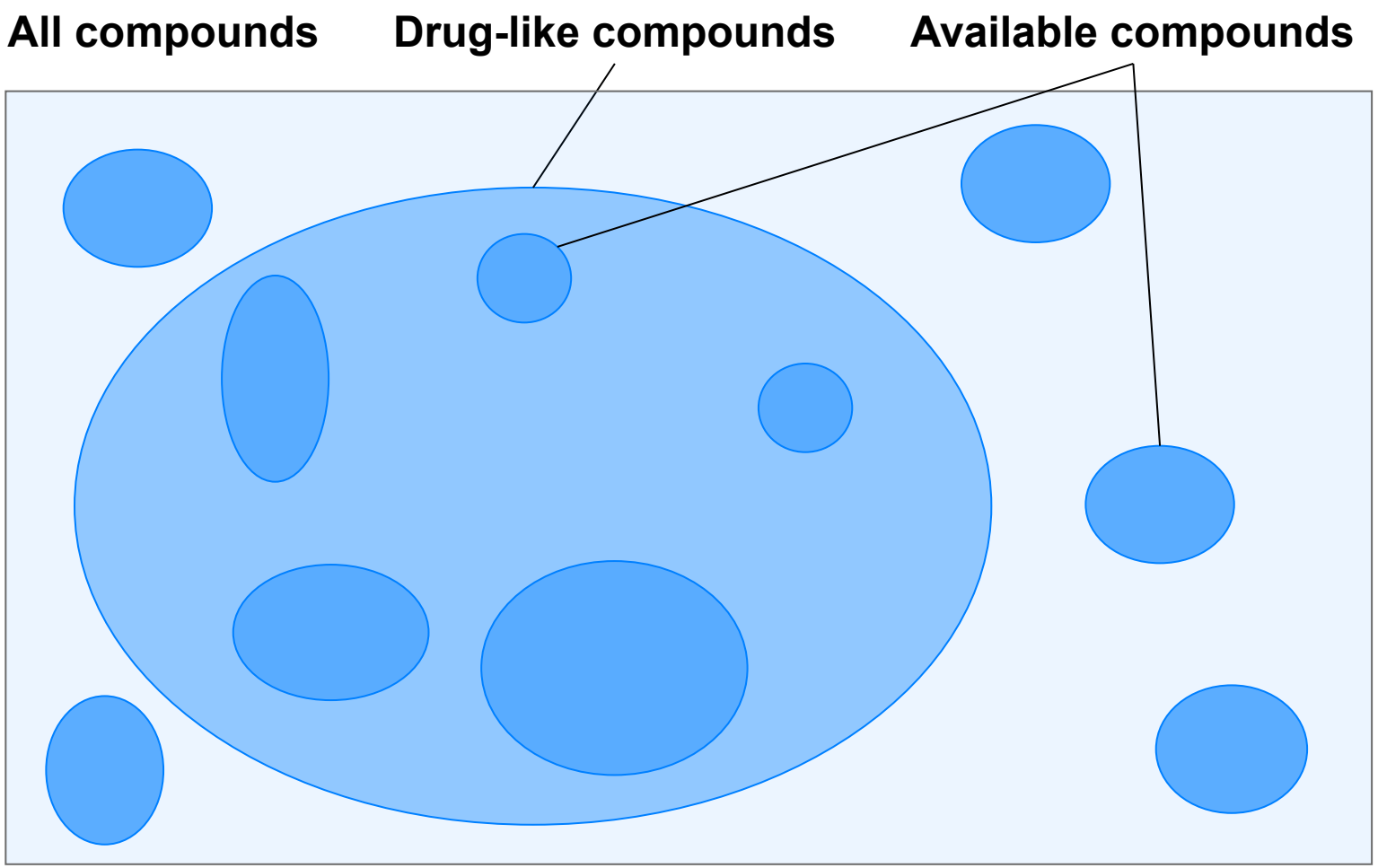
ICIC 2012
Berlin, Germany
October 2012

John P. Overington
EMBL-EBI

jpo@ebi.ac.uk

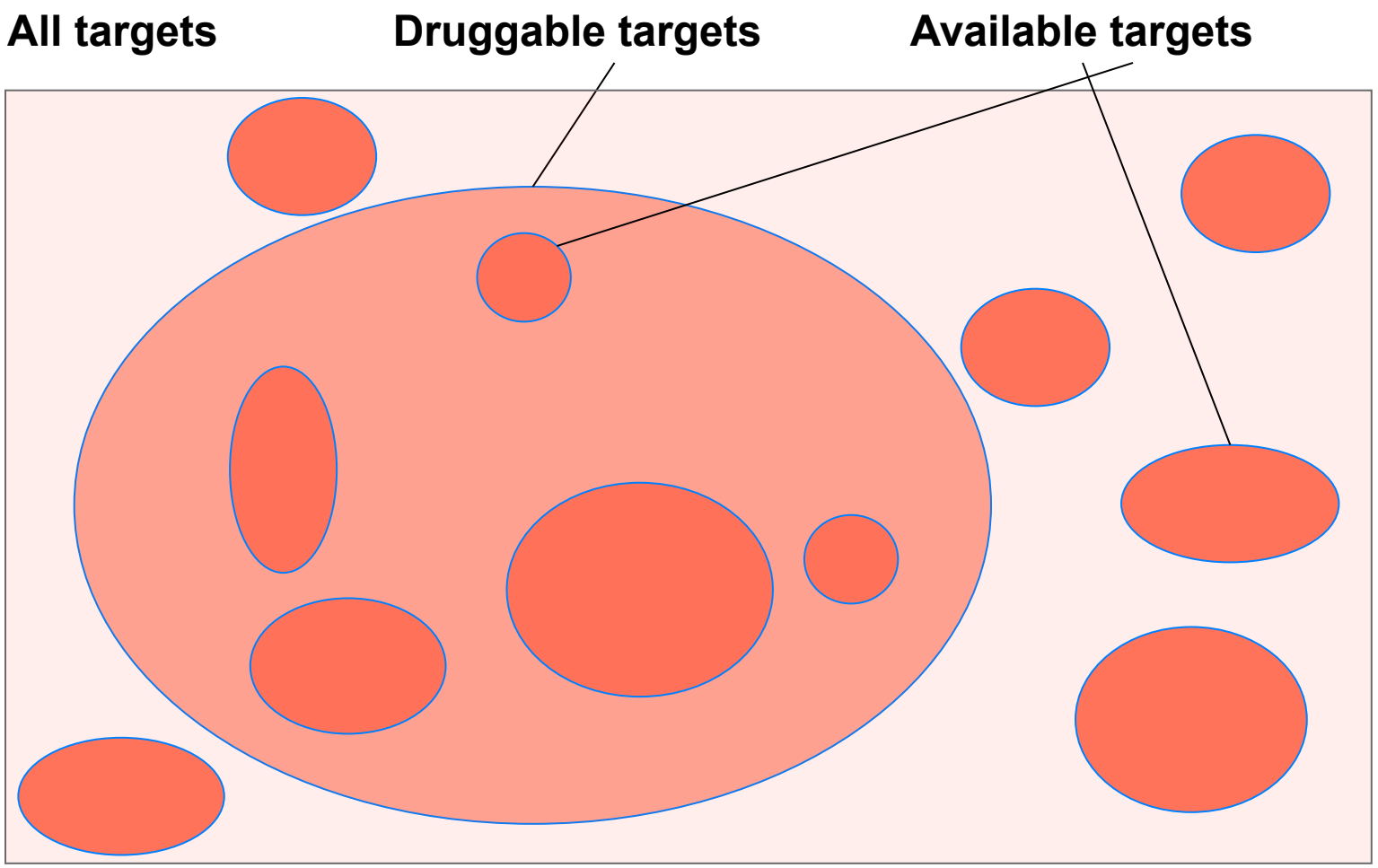
EMBL 

Chemical Space



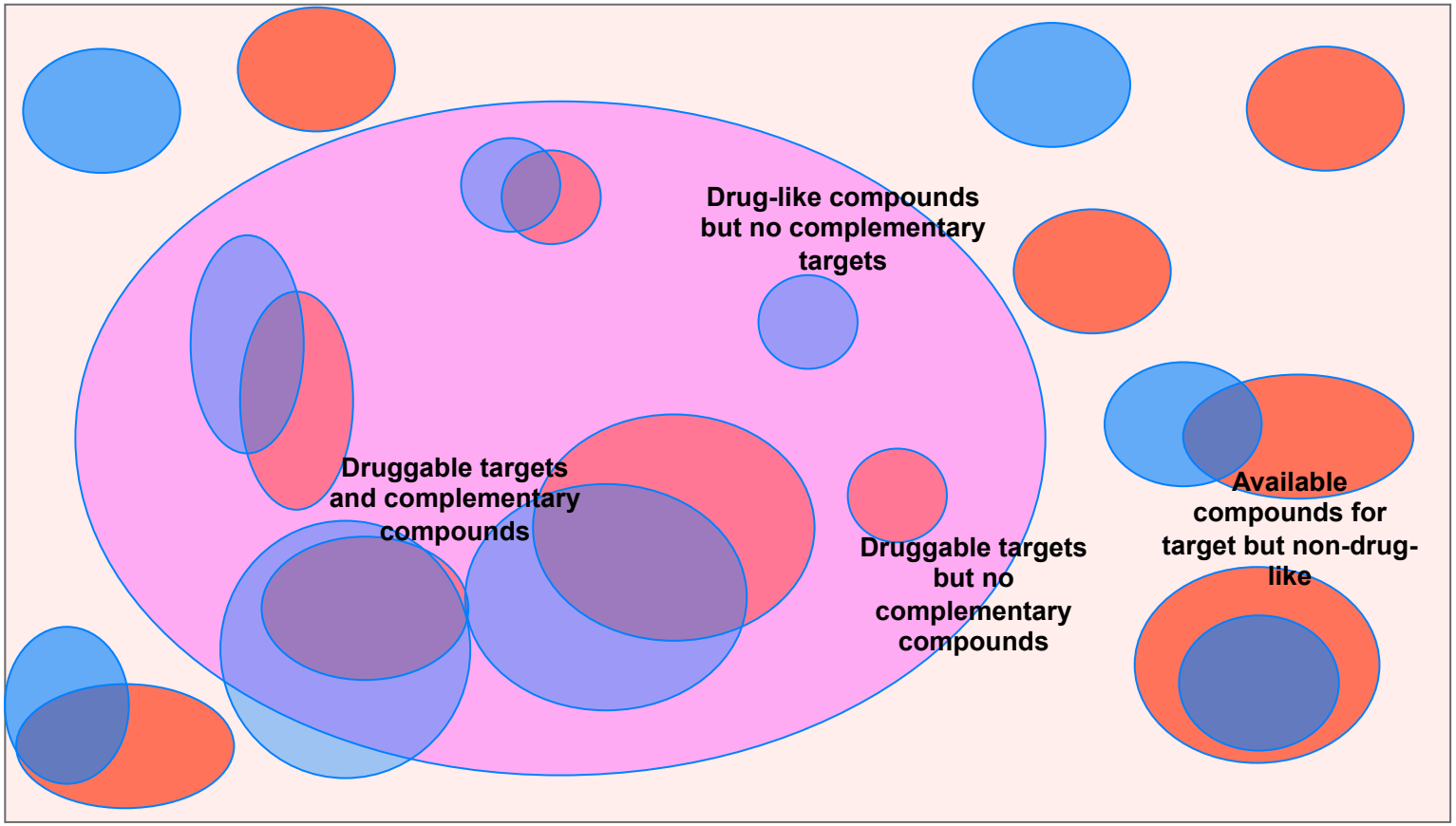
Only certain molecules have features consistent with good pharmacological properties

Target Space



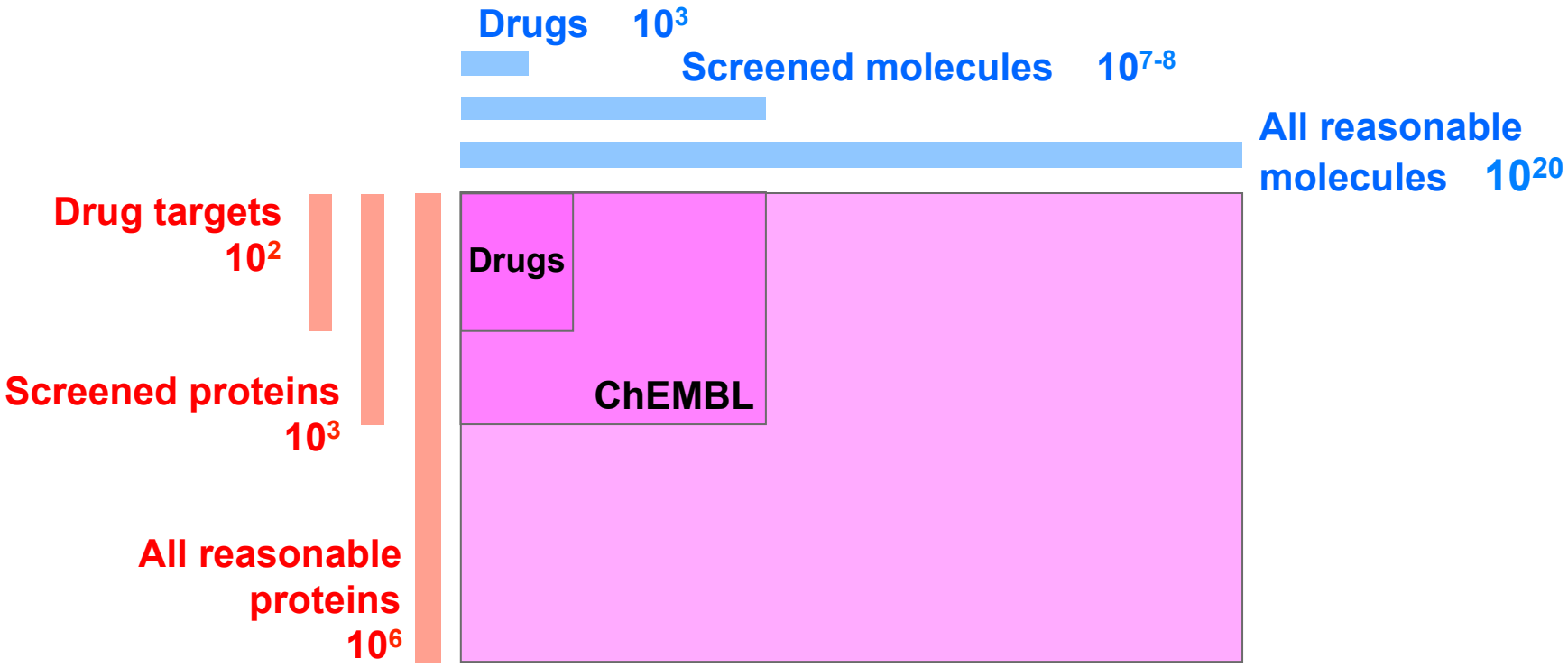
Only certain targets have binding sites capable of ligand efficient binding of drug-like ligands

Accessible Pharmacological Space



Chemogenomics

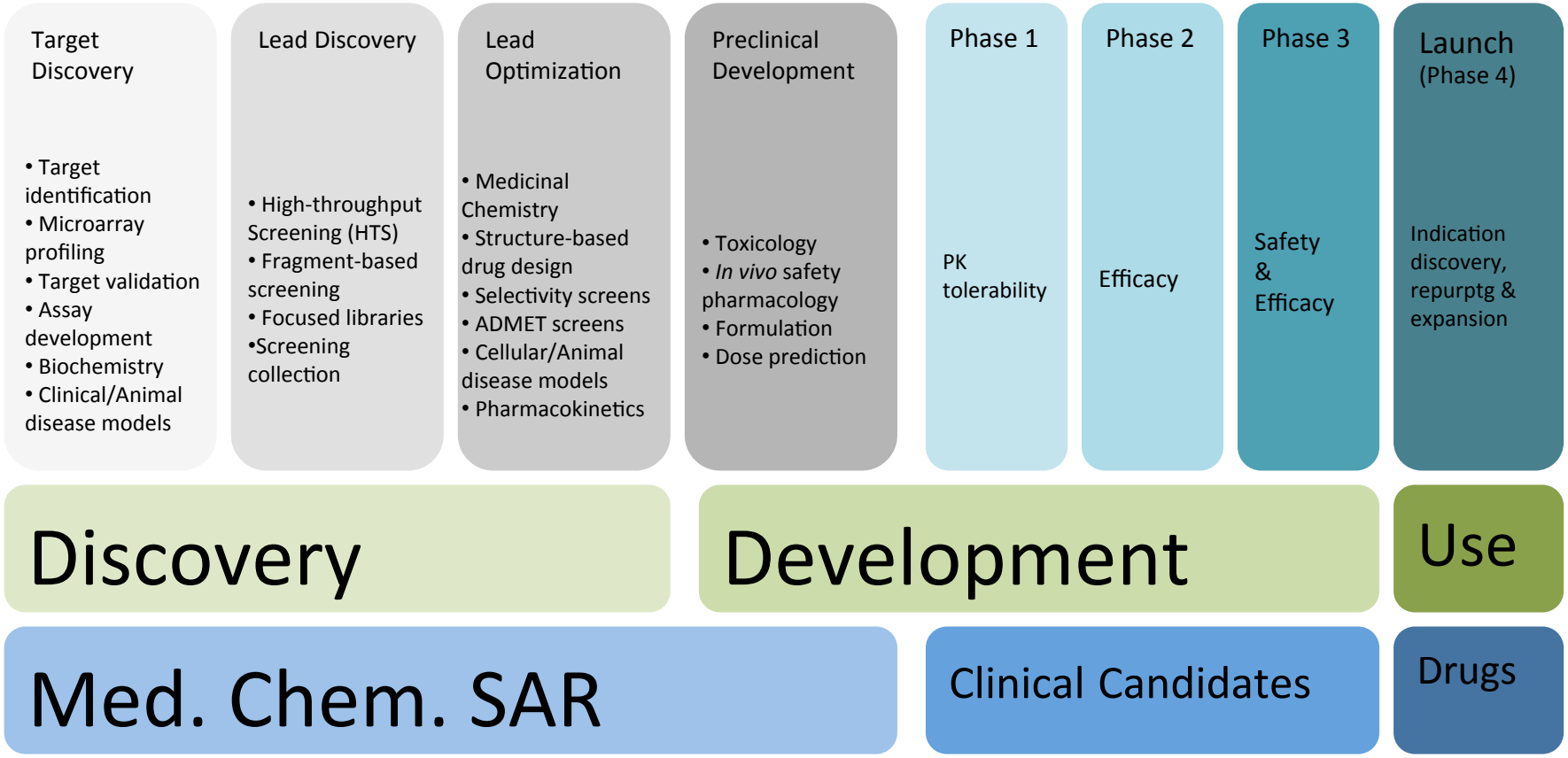
Exploration of bioactivity space at genomic scale
Structure Activity Relationship (SAR)



ChEMBL Database

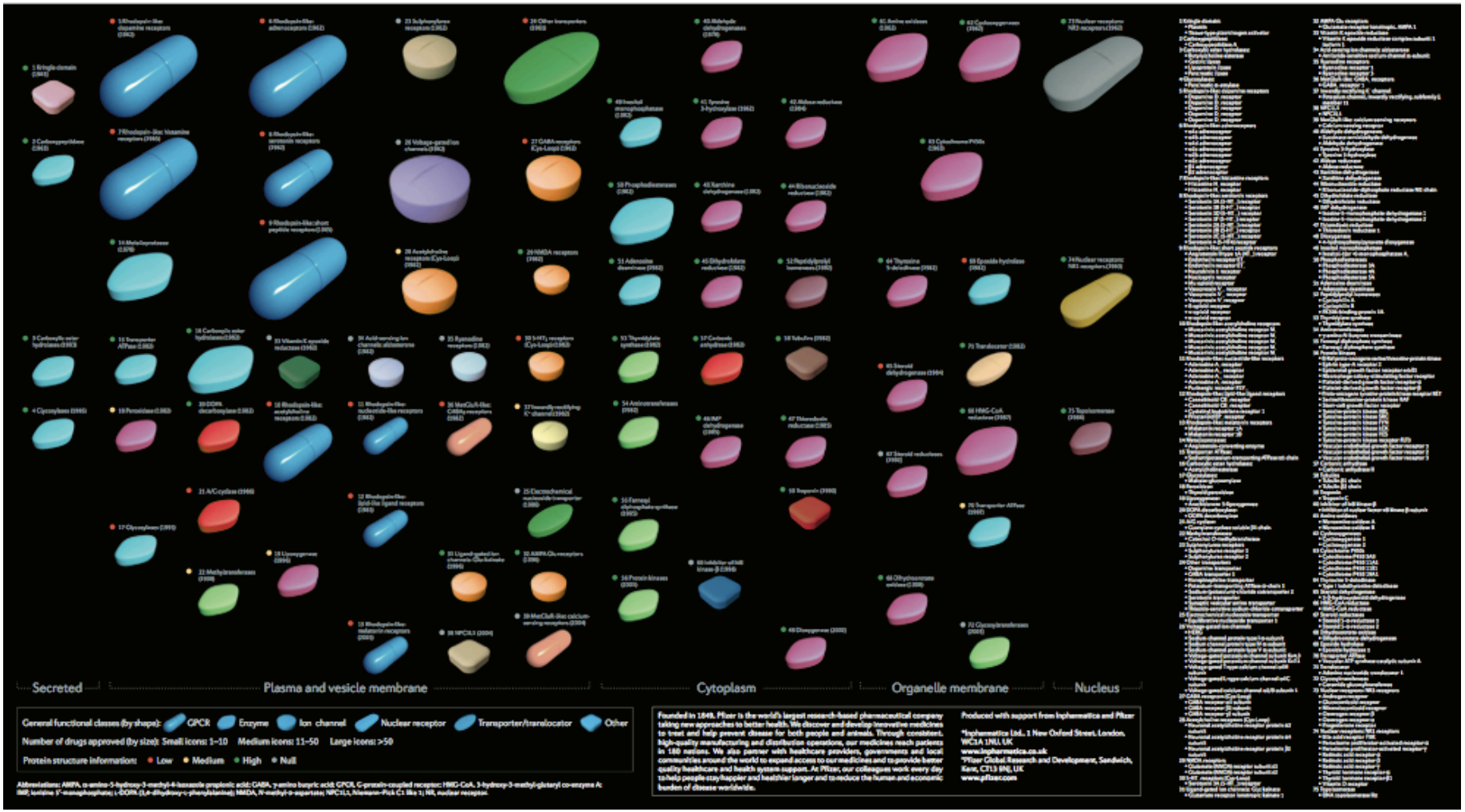
- http://www.ebi.ac.uk/chembl
- Funded by a Strategic Award from the Wellcome Trust
- World's largest primary source of Open pharmacology/drug discovery data
  - Contains synthetic small molecules, natural products and biologicals
  - Strong integration and annotation of chemical *and* biological data
  - OSINT approach to data gathering
  - Tight integration with other EBI resources
    - Ensembl, 1000 Genomes, UniProt, PDBe, ArrayExpress, Atlas….
  - Data sharing agreements in place with key public resources, *e.g.* PubChem
- Open Data – CC-BY-SA licence
- Free downloads, secure private searching,…
- REST web service API

Drug Discovery

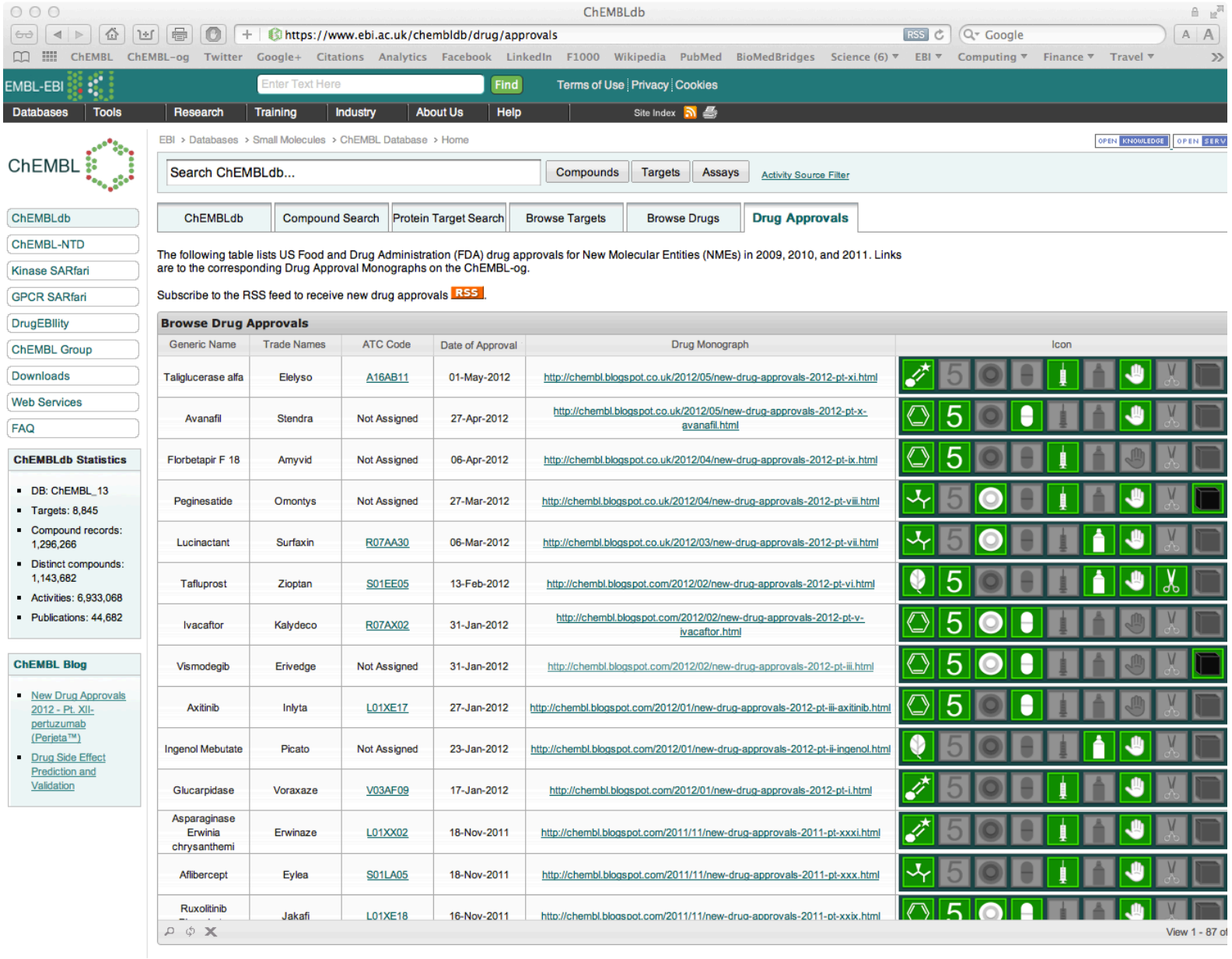


ChEMBL content			
~1,400,000 compound records	>10,000,000 bioactivities	~46,000 abstracted papers	~9,000 targets
~12,000 clinical candidates			
~1,600 drugs			

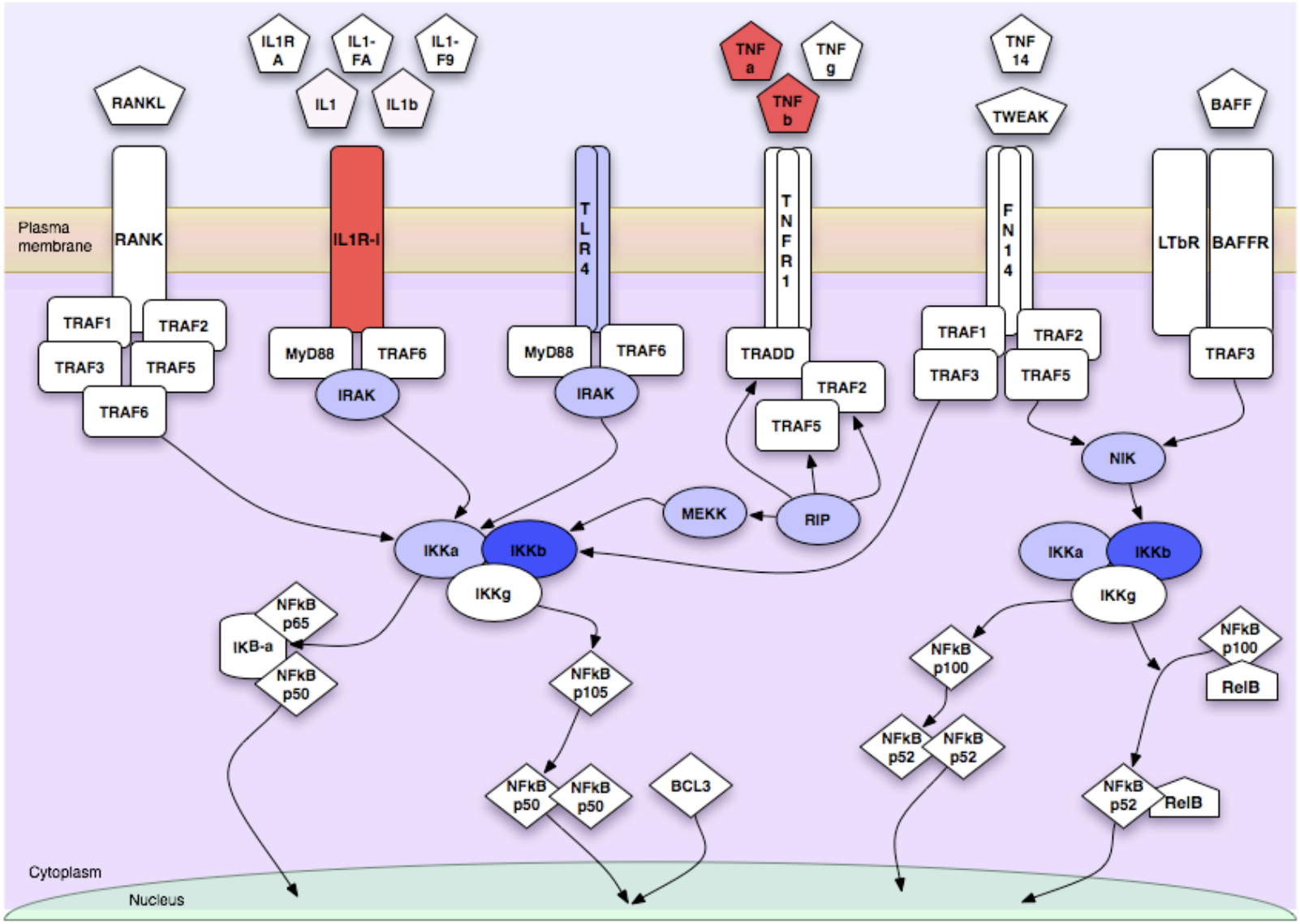
Only ~1% of Genome is a Drug Target



Drug Approvals



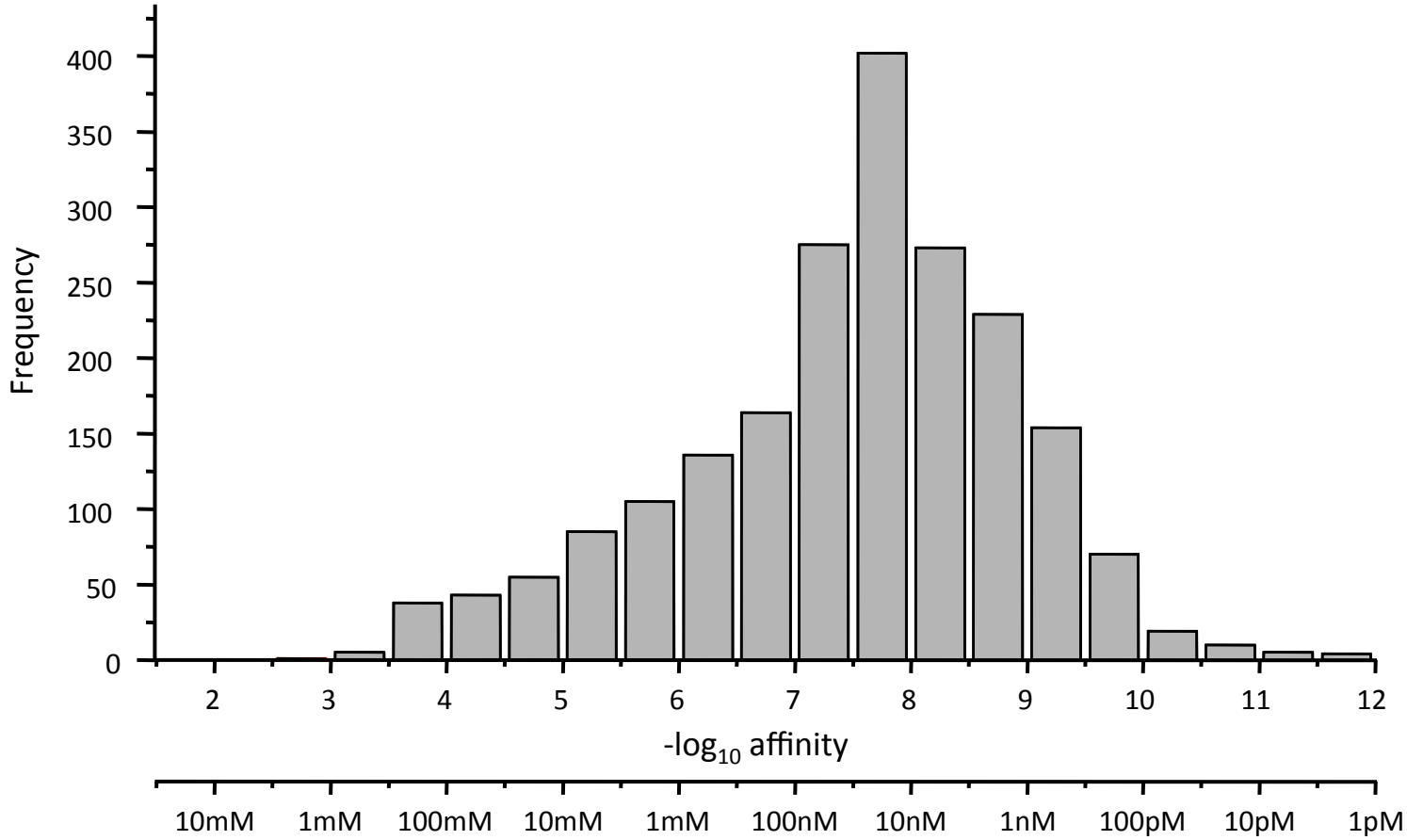
FDA Approved Drugs



NFκB Pathway – key control mechanism for inflammation

Affinity of Drugs for their 'Targets'

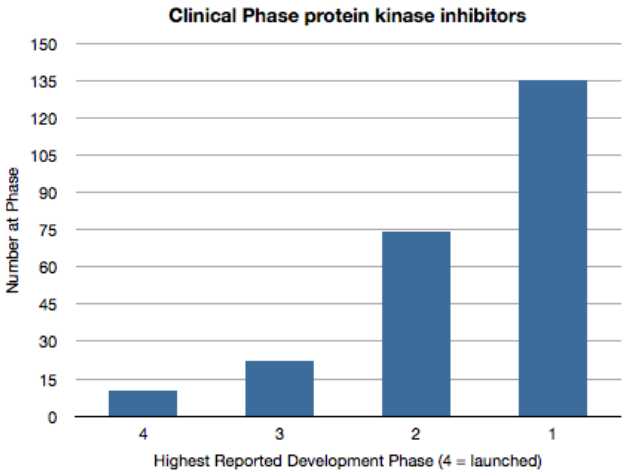
K_i, K_d, IC_{50}, EC_{50}, & pA_2 endpoints for drugs against their 'efficacy targets'

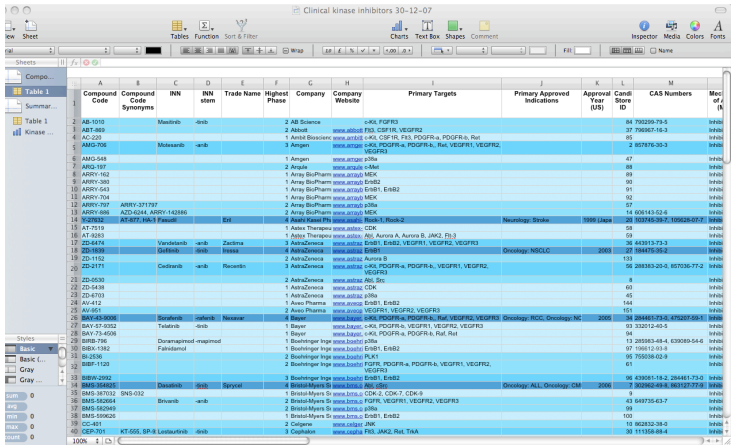


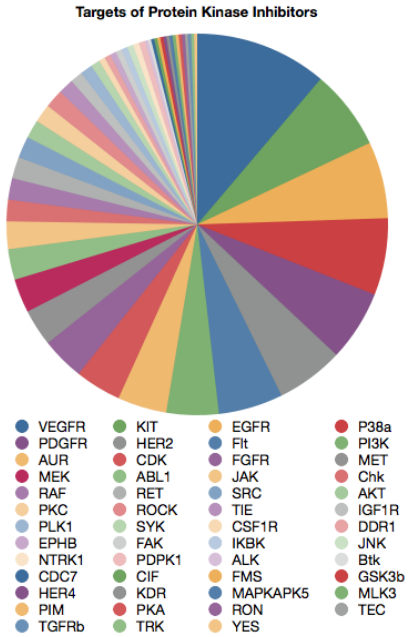
Overington, *et al, Nature Rev. Drug Discov*. **5** pp. 993-996 (2006)
Gleeson *et al*, *Nature Rev. Drug Discov*. **10** pp. 197-208 (2011)

Clinical Candidates

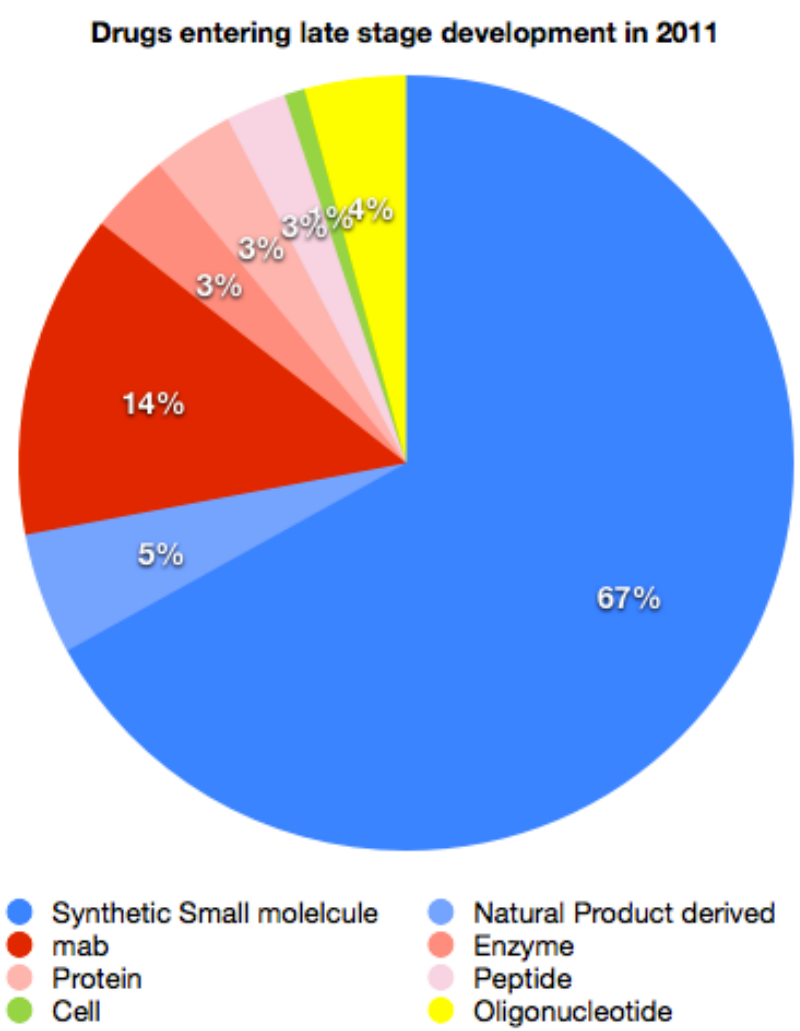
- Collection of clinical development candidates
  - Contains ~12,000 2-D structures/sequences
    - Estimated size ~35-45,000 compounds
  - Work in progress
    - *e.g.* Protein kinases, 393 distinct clinical candidates

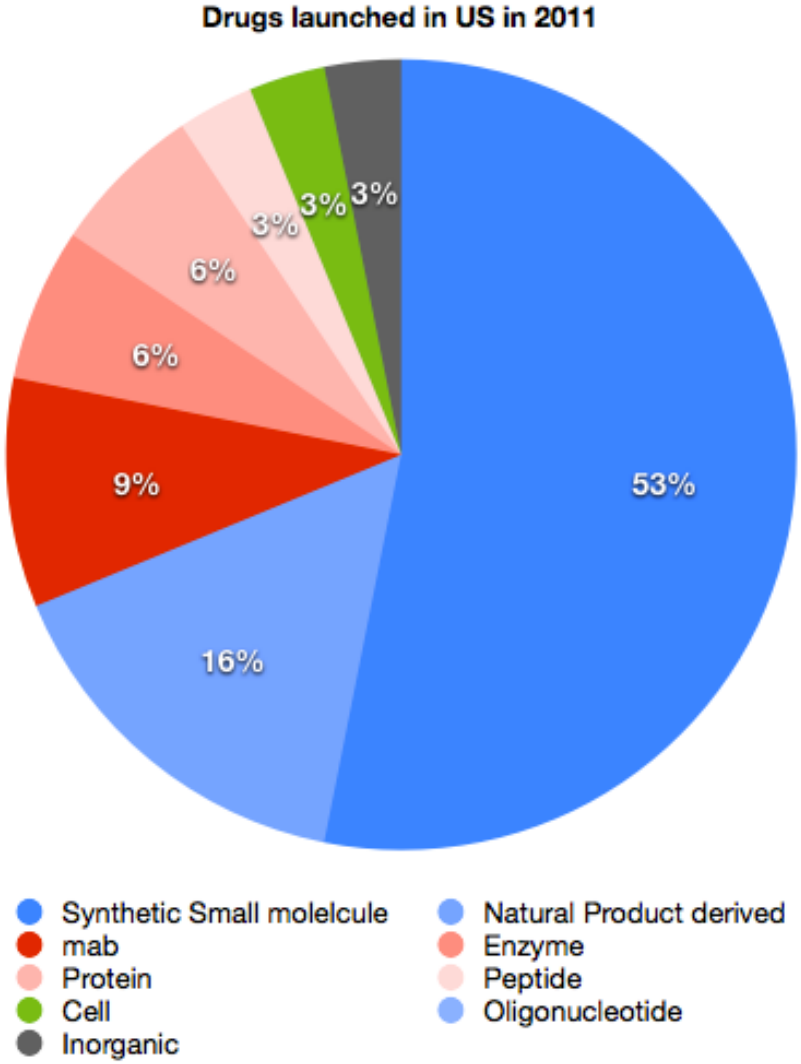




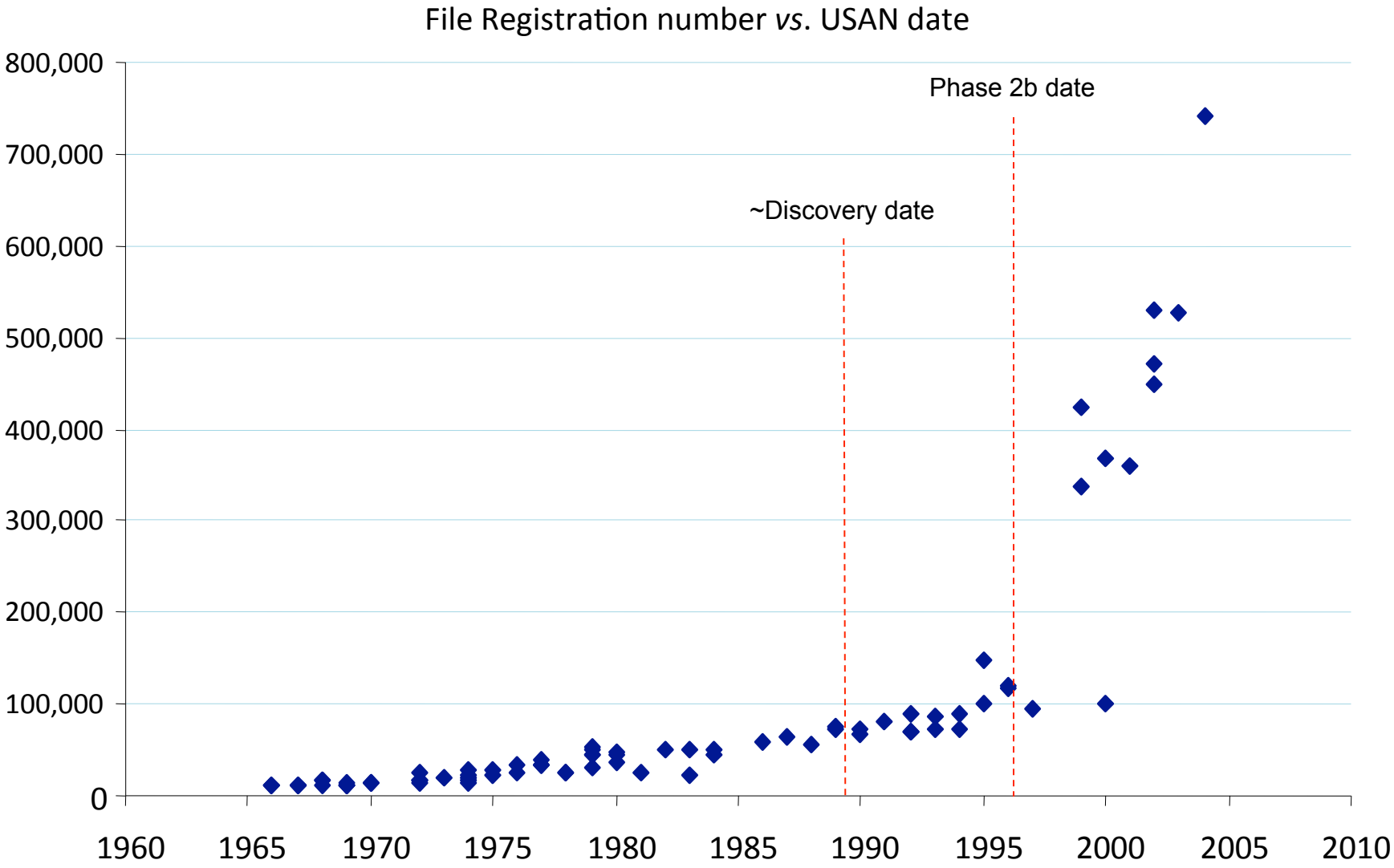


Different Types of Drugs





Pharma Industry Productivity

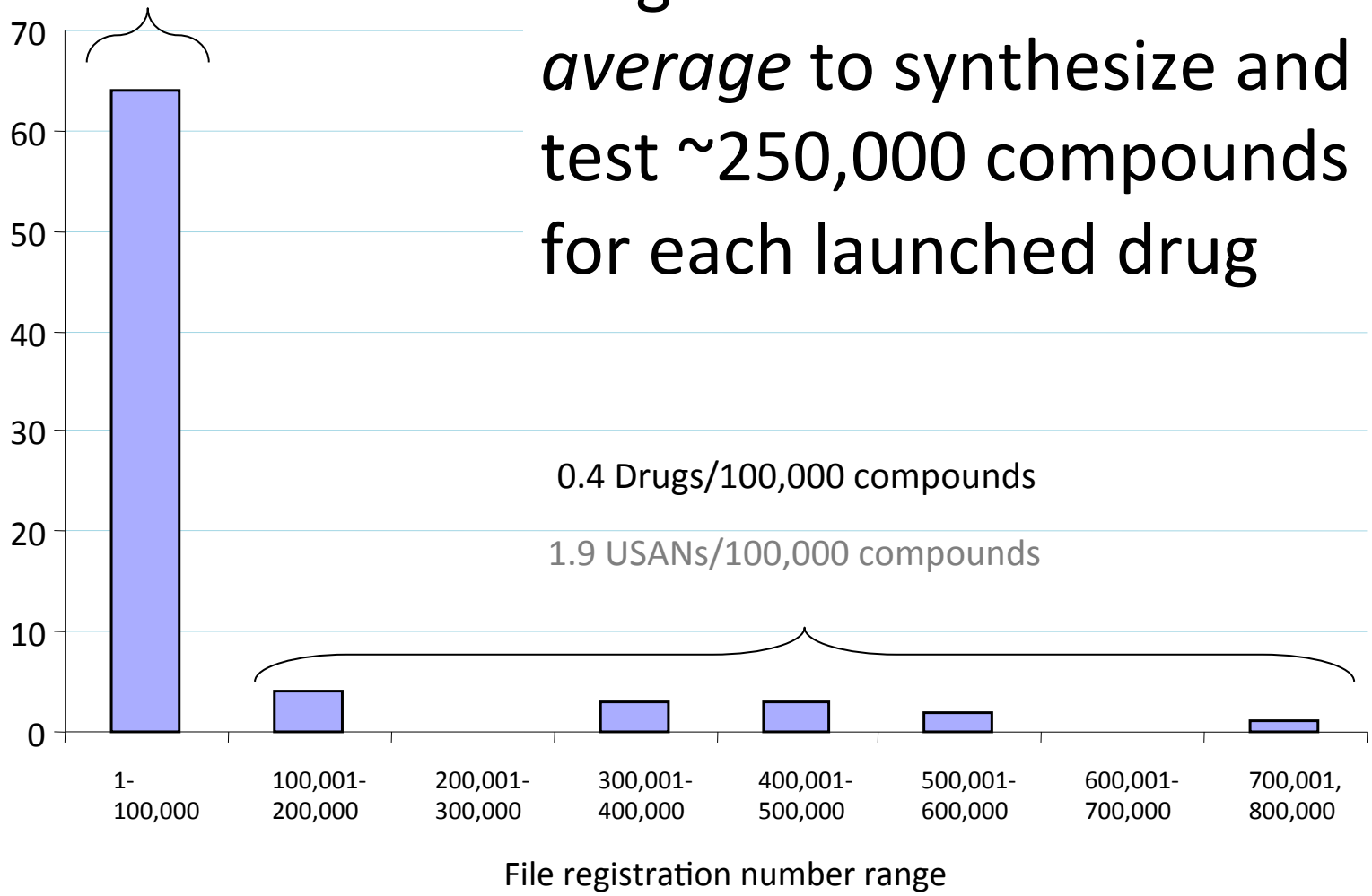


Overington, *unpublished*

Pharma Industry Productivity

16 Drugs/100,000 compounds

64 USANs/100,000 compounds



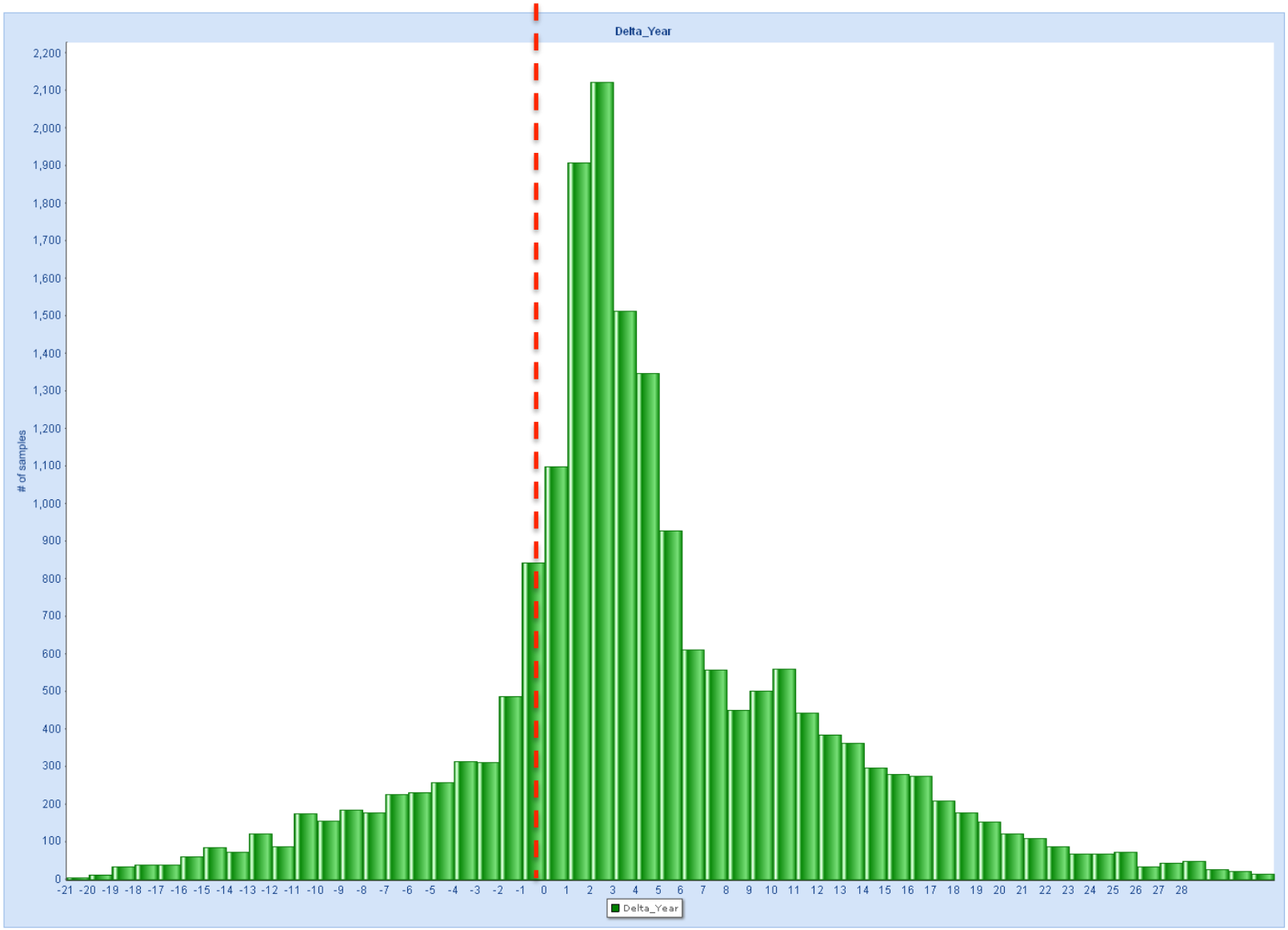
Large Pharma needs *on average* to synthesize and test ~250,000 compounds for each launched drug

0.4 Drugs/100,000 compounds

1.9 USANs/100,000 compounds

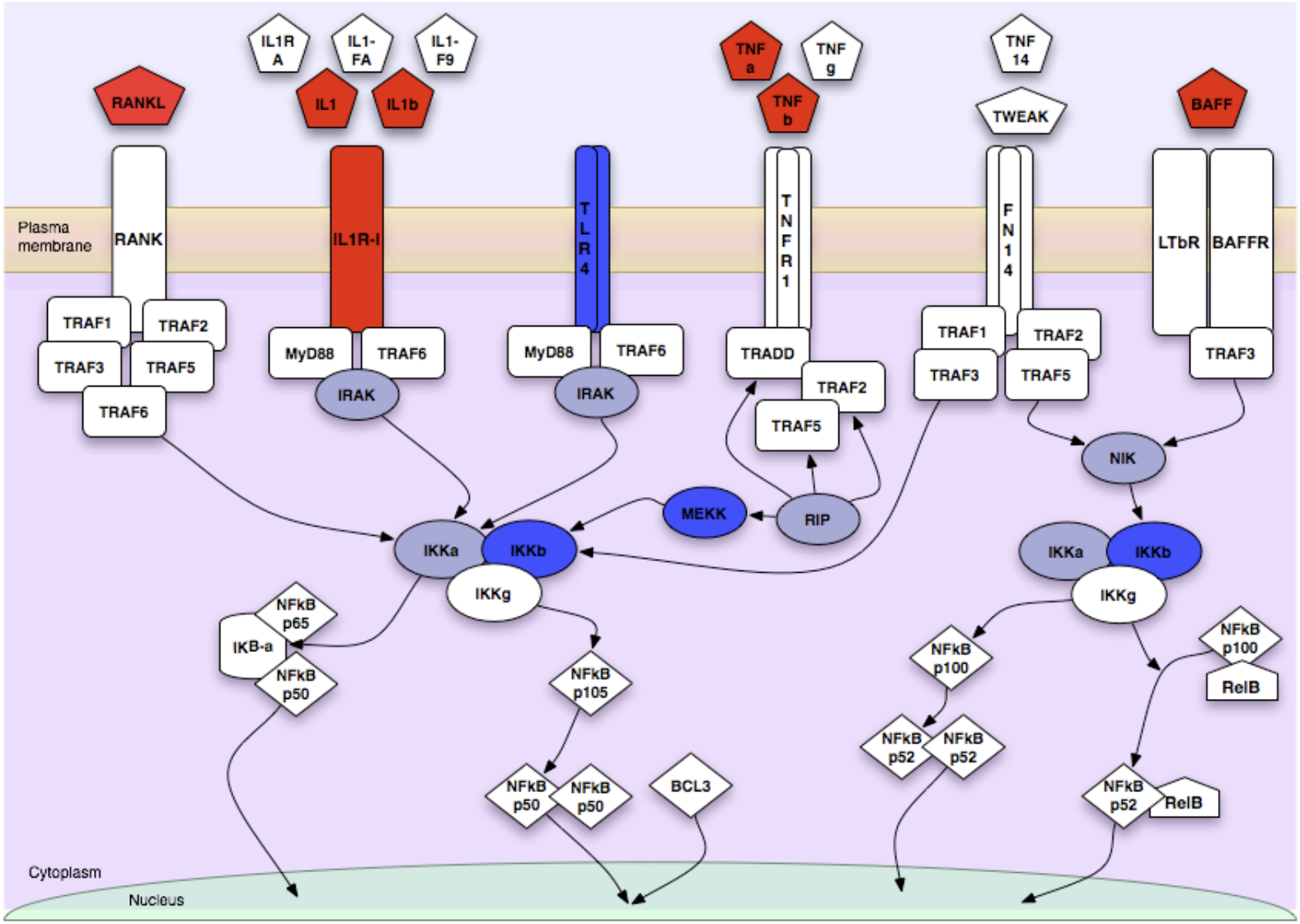
Overington, *unpublished*

Patent and Publication Lag



IBM Patent data and ChEMBL

Clinical Candidates



What Is the ChEMBL Data?

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

John C. Danilewicz,[†,‡] Stuart M. Abel,[§] Alan D. Brown,[‡] Paul V. Fish,*[,‡] Edward Hawkeswood,[‖] Stephen J. Holland,[‖] Keith James,[†] Andrew B. McElroy,[†] John Overington,[⊥] Michael J. Powling,[‖] and David J. Rance[§]

Departments of Discovery Chemistry, Drug Metabolism, Discovery Biology, and Molecular Informatics Structure and Design, Pfizer Global Research and Development, Sandwich, Kent CT13 9NJ, United Kingdom

Received December 21, 2001

Potent and selective inhibitors of thrombin were sought based on the (*R*)-Phe-Pro-Arg sequence. The objective was to generate similar binding interactions to those achieved by potent competitive inhibitors of the argatroban type, so eliminating the need for covalent interaction with the catalytic serine function, as utilized by aldehyde and boronic acid type inhibitors. Improving the S_1 subsite interaction by substitution of arginine with a 4-alkoxybenzamidine residue provided potent lead 2 (K_i = 0.37 nM). Though an amide bond, which H-bonds to the active site, is lost, modeling indicated that a new H-bond is generated between the alkoxy oxygen atom and the catalytic Ser-195 hydroxyl group. Substitution of the benzamidine system by 1-amidinopiperidine then gave compound 4, which provided a further gain in selectivity over trypsin. However, previous work had shown that these compounds were likely to be too lipophilic (Log *D* +0.4 and +0.2, respectively) and to suffer rapid hepatic extraction, presumably via biliary elimination. Accordingly, both proved short-acting when administered intravenously to rats and showed poor activity when given intraduodenally. The aim was then to reduce lipophilicity below a log *D* of −1.2, which in a previously reported series had been effective in preventing rapid clearance. It was anticipated that compounds of this type would rely on the cation selective paracellular route of absorption from the gastrointestinal tract. Potent polar analogues with selectivity >1000 over trypsin were obtained. The best in vivo activity was shown by compound 12. However, in the final analysis, its oral bioavilability proved poor, relative to analogues with similar physicochemical properties derived from argatroban, consistent with the hypothesis that molecular shape is an additional important determinant of paracellular absorption.

Introduction

The search for potent selective and orally active thrombin inhibitors has gathered momentum in recent years.[1] Thrombin is the last in a cascade of trypsin-like plasma serine proteases, which by catalyzing the conversion of fibrinogen to fibrin, activation of FXIII and inducing platelet aggregation is a key enzyme in haemostasis and thrombus formation. The inhibition of a single enzyme in the cascade, and in particular thrombin, has been an attractive goal in that it could also provide superior antithrombotic therapy by increasing efficacy and safety as compared to heparin and the coumarins. Additionally, by keeping molecular size small, the opportunity exists for obtaining oral activity.[1,2]

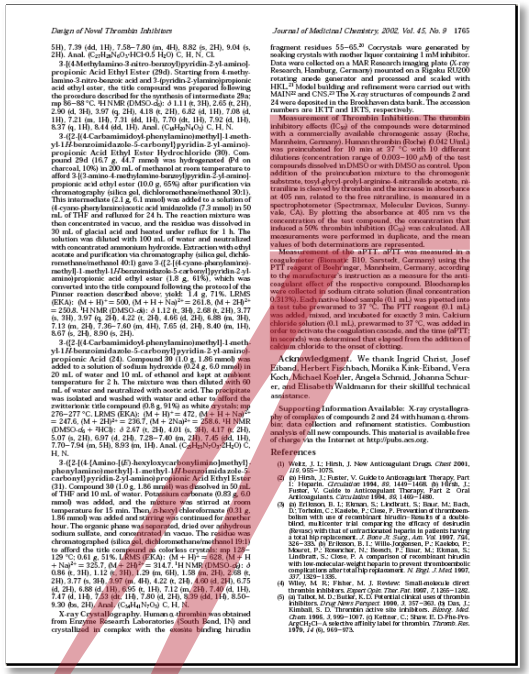
Two small molecular weight inhibitor types are emerging as structure−activity relationships are explored. The first is of the argatroban[3] and NAPAP[4] type (Chart 1), where lipophilic groups on either side of the basic P_1 side chain pack together to interact with the hydrophobic S_2 site.[5−7] Napsagatran (Ro 46-6240), developed by Hilpert et al.,[8] though having a more complex P_1 residue, can nevertheless be viewed as belonging to this group. The only interaction with the catalytic serine residue is via a hydrogen bond to the carboxylate function in both argatroban and napsagatran. Unfortunately, none of these compounds is orally active due to either poor absorption from the gastrointestinal tract and/or rapid clearance via the bile.[4,9]

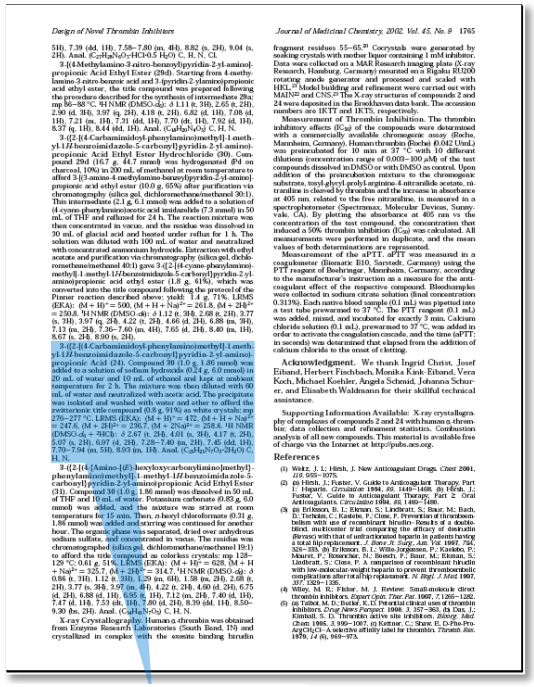
A second inhibitor type is based on the substrate-derived irreversible chloromethyl ketone inhibitor PPACK and includes compounds such as DuP-714[10,11] and efegatran (GYKI-14 766).[12] These compounds interact covalently with the hydroxyl group of the catalytic serine residue. The neighboring proline ring and (*R*)-Phe side chain cooperate to fill the S_2 site in a similar fashion as the two distal lipophilic groups of the first series.[5] Though oral activity has been claimed for these compounds, we were concerned that high enzyme selectivity might not be obtainable when substantial affinity is derived by interacting covalently with the ubiquitous active site serine function. In the case of aldehyde type inhibitors, there is also the potential problem of achieving adequate optical and chemical stability.

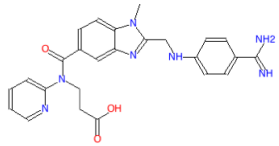
* To whom correspondence should be addressed. Tel.: 44 1304 644589. Fax: 44 1304 656433. E.mail: paul_fish@sandwich.pfizer.com.
† Senior author.
‡ Department of Discovery Chemistry.
§ Department of Drug Metabolism.
‖ Department of Discovery Biology.
⊥ Molecular Informatics Structure and Design.

10.1021/jm011133d CCC: $22.00
Published on Web 05/11/2002

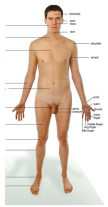
What Is the ChEMBL Data?







>Thrombin
MAHVRGLQLPGCLALAALCSLVHSQHVFLAPQQARSLLQRVRRANTFLEEVRKGNLERECVEETCS
YEEAFEALESSTATDVFWAKYTACETARTPRDKLAACLEGNCAEGLGTNYRGHVNITRSGIECQLW
RSRYPHKPEINSTTHPGADLQENFCRNPDSSTTGPWCYTTDPTVRRQECSIPVCGQDQVTVAMTPR
SEGSSVNLSPPLEQCVPDRGQQYQGRLAVTTHGLPCLAWASAQAKALSKHQDFNSAVQLVENFCRN
PDGDEEGVWCYVAGKPGDFGYCDLNYCEEAVEEETGDGLDEDSDRAIEGRTATSEYQTFFNPRTFGS
SGEADCGLRPLFEKKSLEDKTERELLESYIDGRIVEGSDAEIGMSPWQVMLFRKSPQELLCGASLI
SDRWVLTAAHCLLYPPWDKNFTENDLLVRIGKHSRTRYERNIEKISMLEKIYIHPRYNWRENLDRD
IALMKLKKPVAFSDYIHPVCLPDRETAASLLQAGYKGRVTGWGNLKETWTANVGKGQPSVLQVVNL
PIVERPVCKDSTRIRITDNMFCAGYKPDEGKRGDACEGDSGGPFVMKSPFNNRWYQMGIVSWGEGC
DRDGKYGFYTHVFRLKKWIQKVIDQFGE



Inhibition of human Thrombin

PTT (partial thromboplastin time)

Compound

Assay

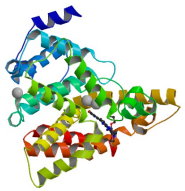
SAR Data

K_i=4.5 nM

ED_2=230 nM

ChEMBL Target Types

Protein

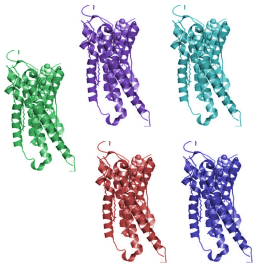


e.g. PDE5

Protein complex

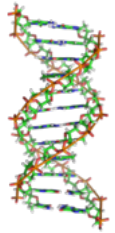
e.g. Nicotinic acetylcholine receptor

Protein family



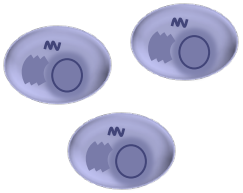
e.g. Muscarinic receptors

Nucleic Acid



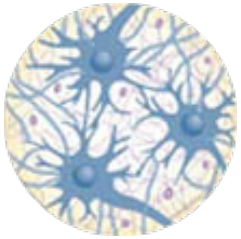
e.g. DNA

Cell line



e.g. HEK293 cells

Tissue



e.g. Trachea

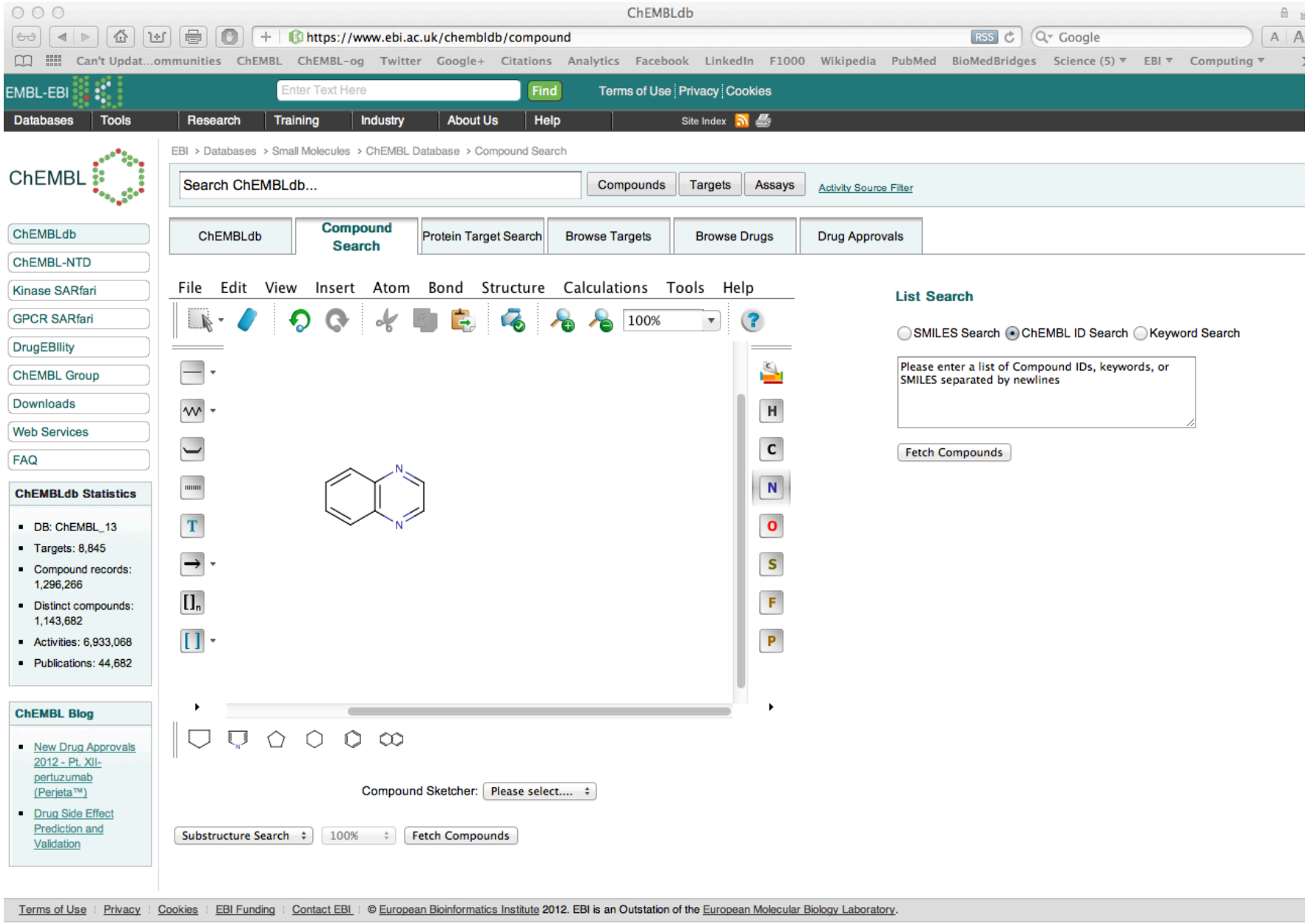
Sub-cellular fraction

e.g. Mitochondria

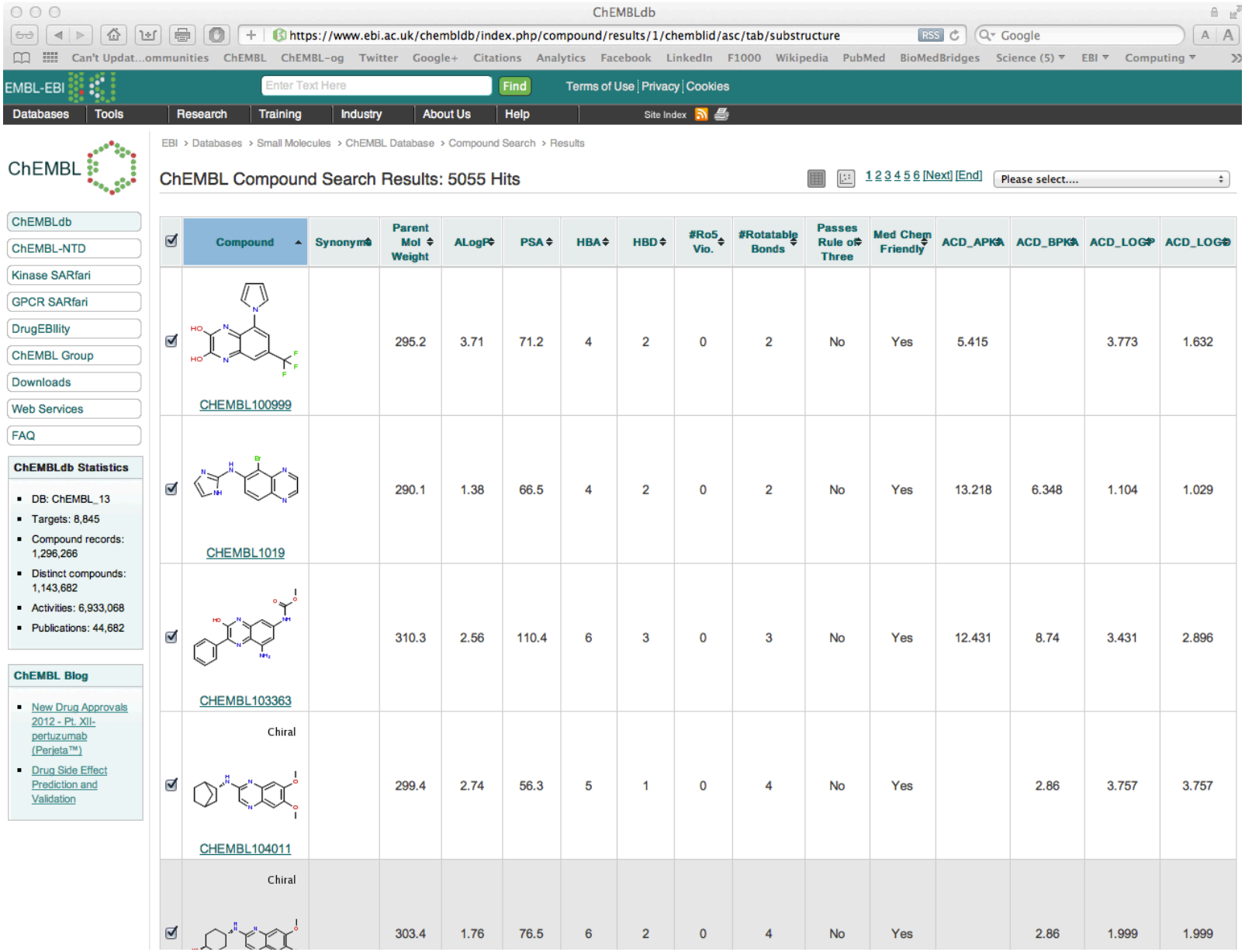
Organism

e.g. Drosophila

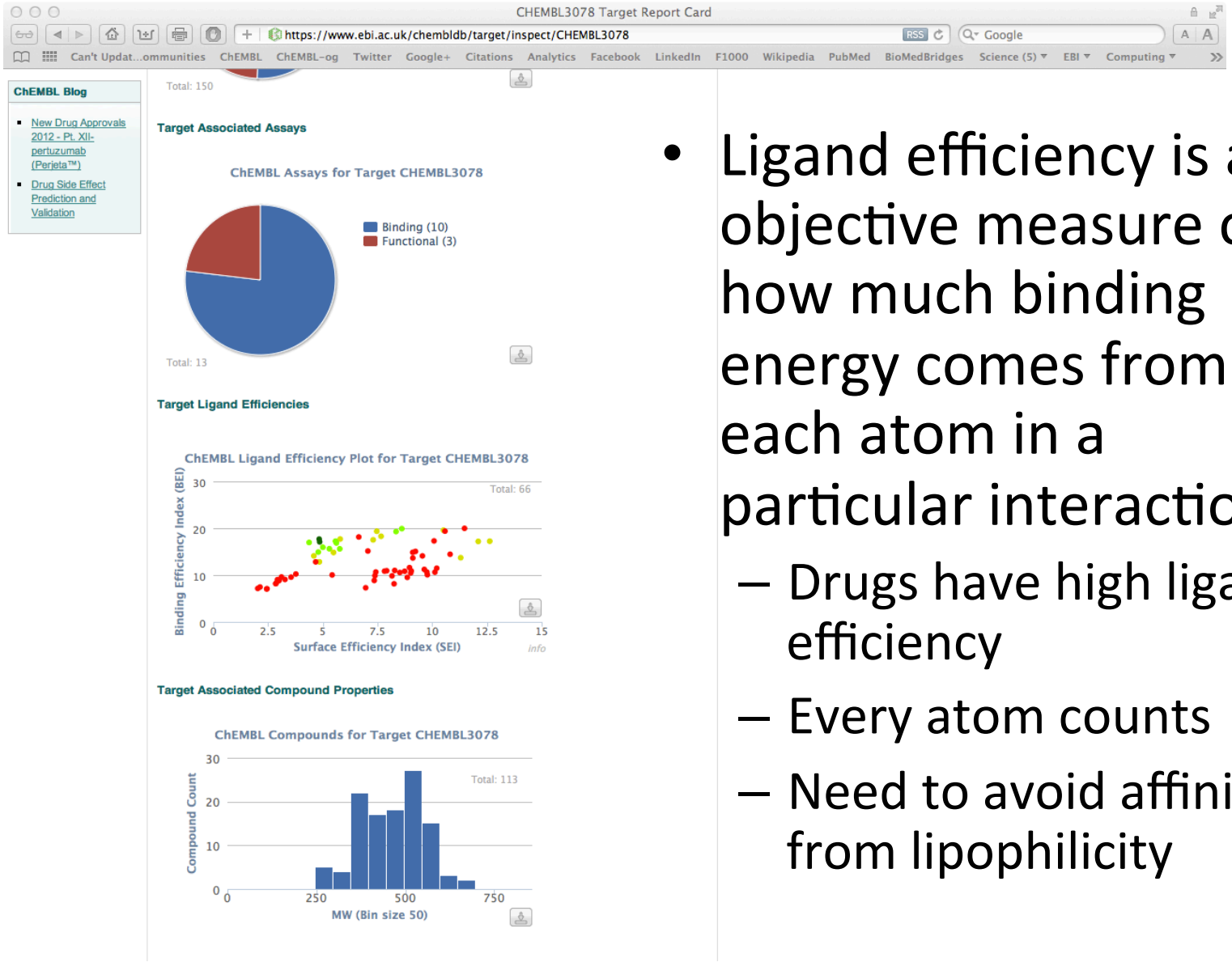
Compound Searching



Spreadsheet Views

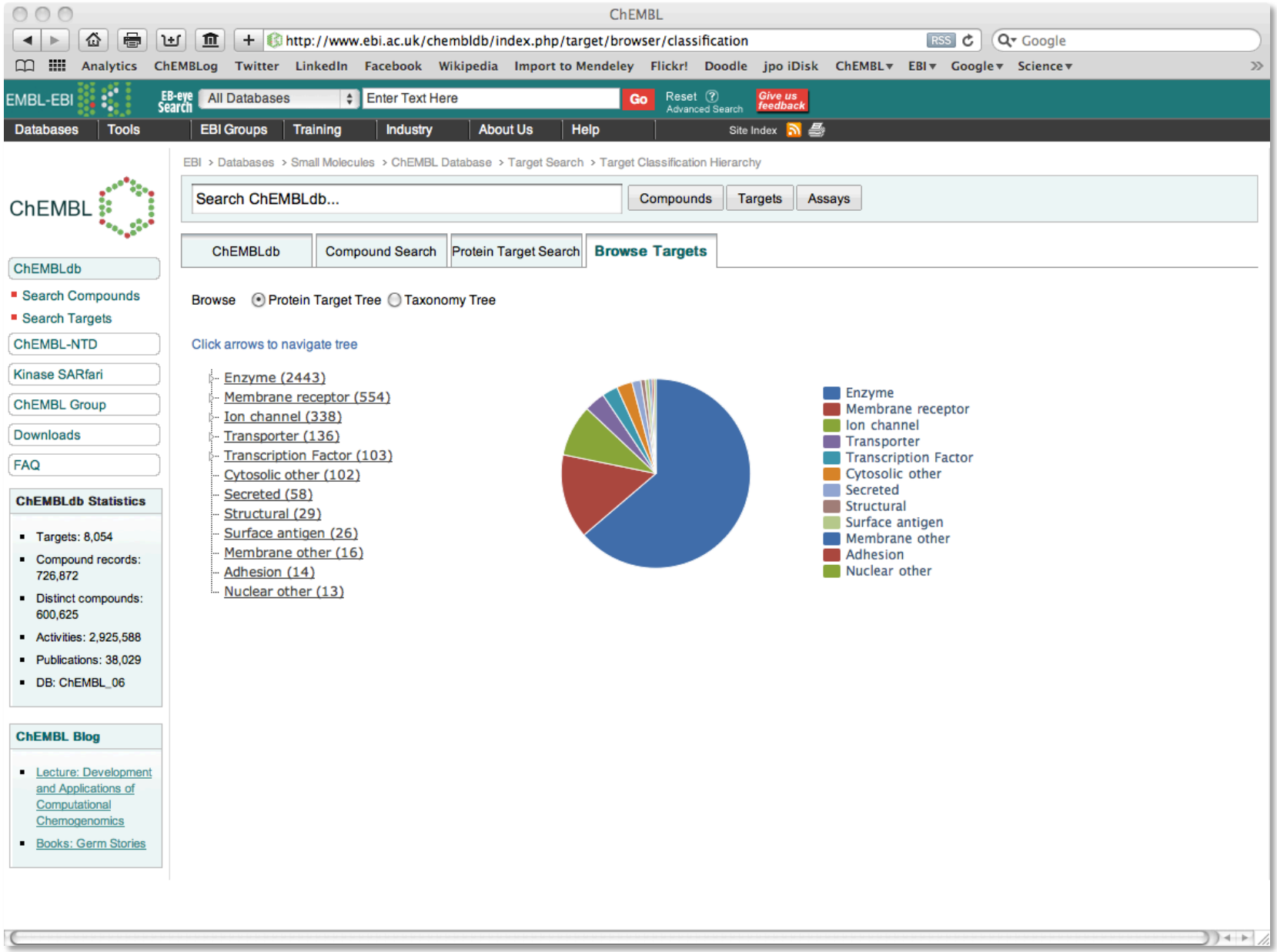


Ligand Efficiency

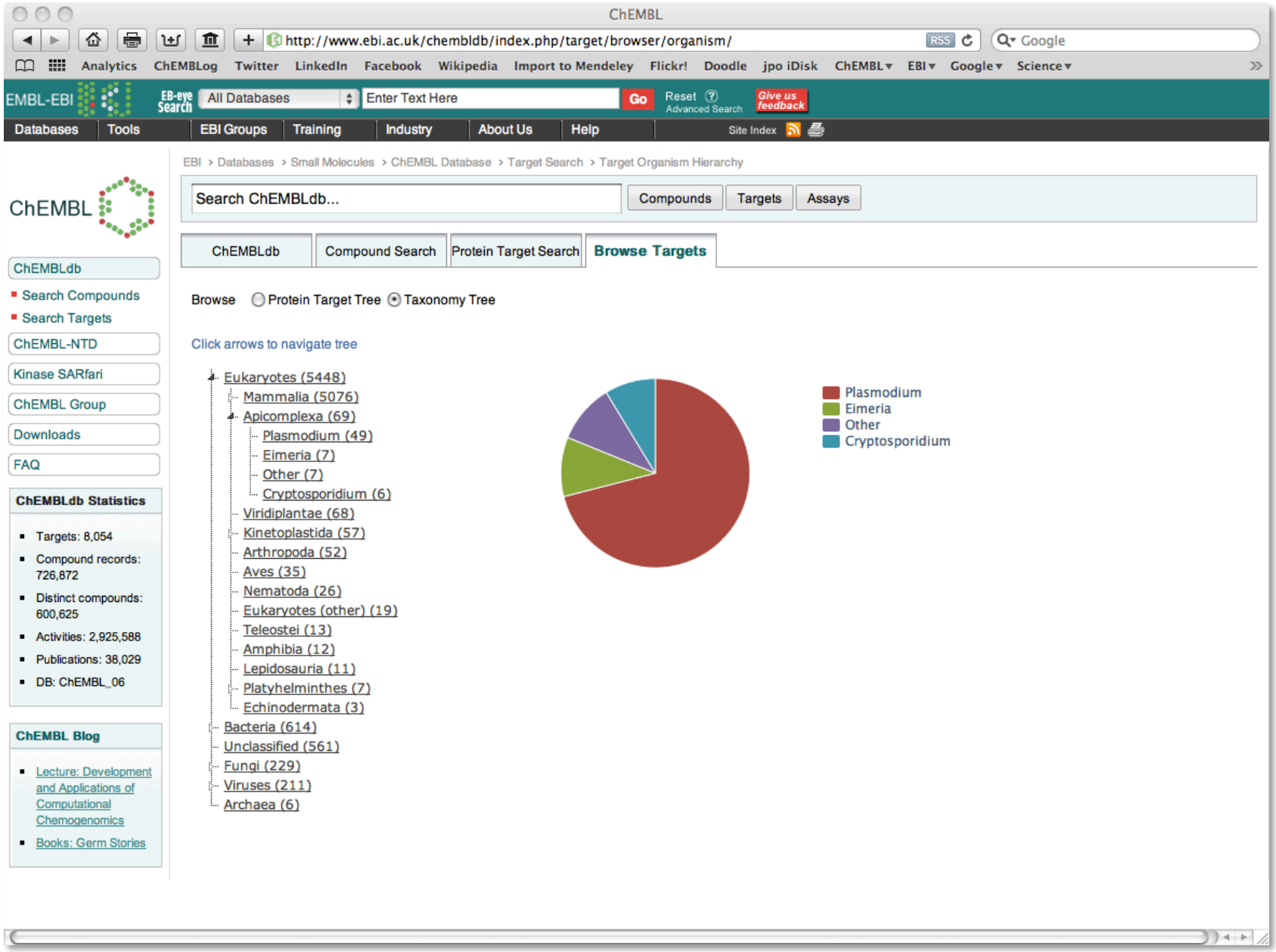


- Ligand efficiency is an objective measure of how much binding energy comes from each atom in a particular interaction
  - Drugs have high ligand efficiency
  - Every atom counts
  - Need to avoid affinity from lipophilicity

Target Class Data

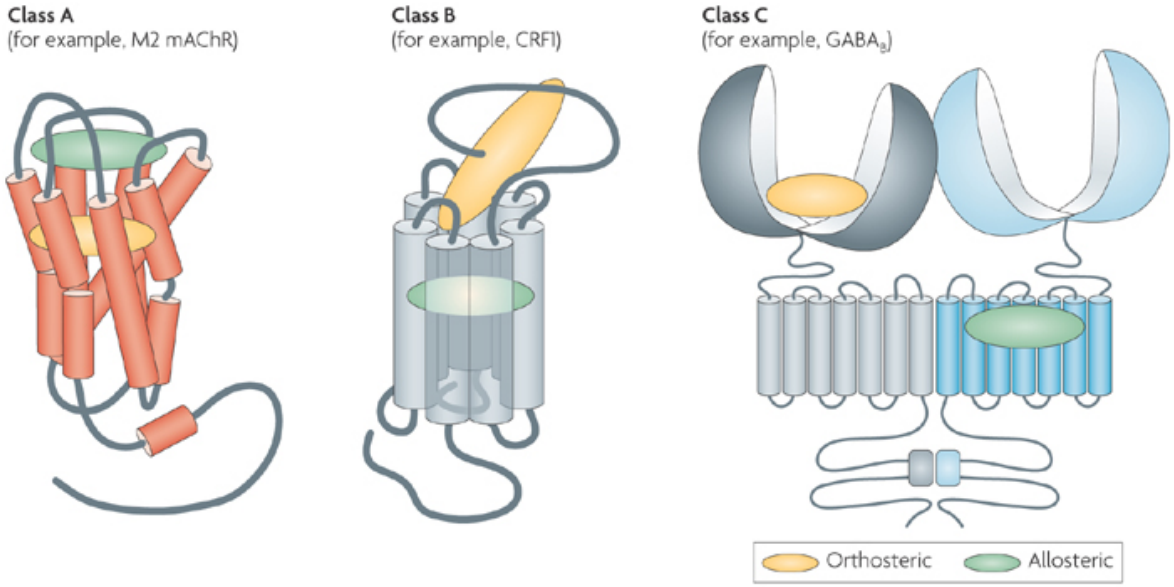


Assay Organism Data



Allosteric Regulators

- Allosteric drugs can have some advantages over orthosteric drugs
  - Selectivity
  - Orthosteric site may be undruggable



Allosteric/Orthosteric sites for GPCRs

Nature Reviews | Drug Discovery

http://www.chemblog.org