

Chemistry Connect

AstraZeneca's cheminformatics platform for large-scale integration of structure and bioactivity data

ICIC 2012 14-17 October Berlin

Sorel Muresan

AstraZeneca R&D Mölndal
Chemistry Innovation Centre, Discovery Sciences

Unrestricted



Chemistry Connect – a team effort

Drug Discovery Today • Volume 16, Numbers 23/24 • December 2011

Making every SAR point count: the development of Chemistry Connect for the large-scale integration of structure and bioactivity data

**Sorel Muresan¹, Plamen Petrov¹, Christopher Southan²,
Magnus J. Kjellberg³, Thierry Kogej¹, Christian Tyrchan³,
Peter Varkonyi¹ and Paul Hongxing Xie¹**

¹ DECS Computational Sciences, AstraZeneca R&D Mölndal, S-431 83 Mölndal, Sweden

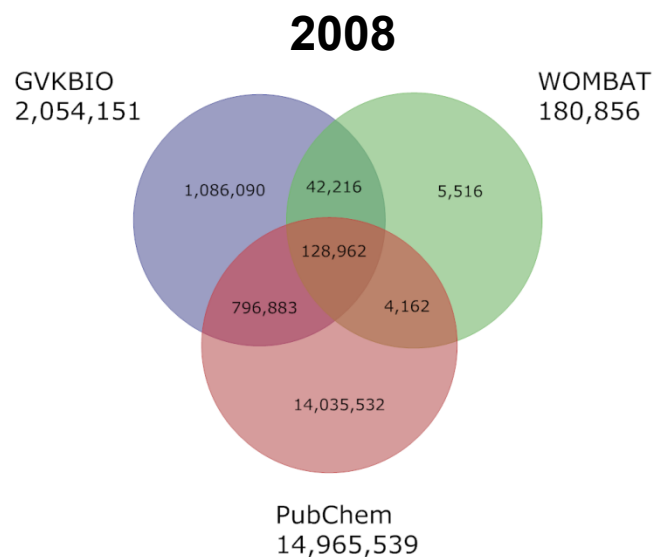
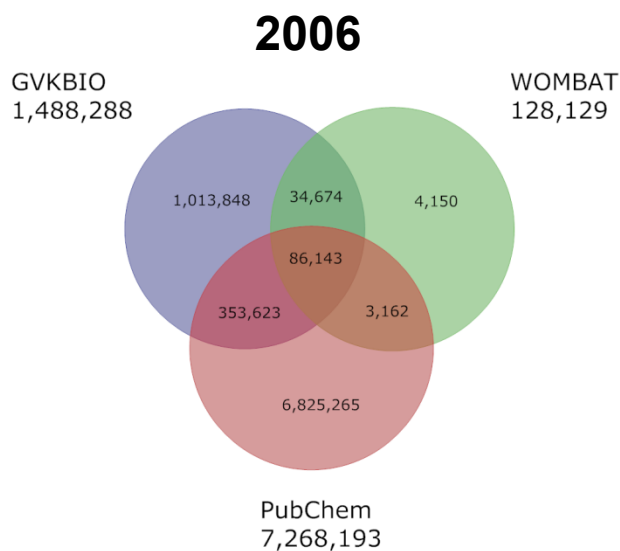
² R&D Information, AstraZeneca R&D Mölndal, S-431 83 Mölndal, Sweden

³ CVGI AstraZeneca R&D Mölndal, S-431 83 Mölndal, Sweden



Driver – explosion in SAR data

- Chemical information landscape changing fast
- Make every SAR point count, access all available chemistry
- Internal & external datasources



SAR key entities and relationships


Biorg Med Chem Lett. 2003 Nov 17;13(22):369-4

Green tea catechins as a BACE1 (beta-secretase) inhibitor.


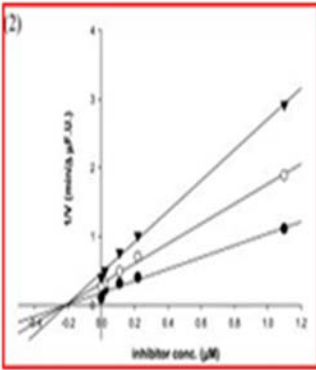
Joon YJ, Baik K, Seung YG, Seung KS

Division of Applied Biology & Chemistry, College of Agriculture & Life Sciences, Kyungpook National University, 1370 Seosuk-Dong, Daegu 702-701, South Korea

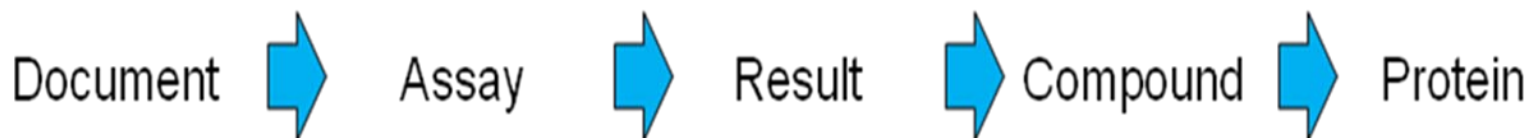
In the course of searching for BACE1 (beta-secretase) inhibitors from natural products, the ethyl acetate soluble fraction of green tea, which was suspected to be rich in catechin content, showed potent inhibition activity. (+)-Epigallocatechin gallate, (-)-epigallocatechin gallate, and (-)-gallocatechin gallate were isolated with IC50 values of 1.8×10^{-6} , 4.5×10^{-6} , and 1.3×10^{-6} M, respectively. Seven additional authentic catechins were tested for a fundamental structure-activity relationship. (-)-Catechin gallate, (-)-gallocatechin, and (-)-epigallocatechin significantly inhibited BACE1 activity with IC50 values of 0.3×10^{-6} , 2.1×10^{-6} , and 2.4×10^{-6} M, respectively. However, (-)-catechin, (-)-catechin, (-)-epigallocatechin, and (-)-epigallocatechin exhibited almost no inhibitory activity. The stronger activity observed to be related to the propyl group on C-2 and/or C-3 of catechin moiety, while the stereochemistry of C-2 and C-3 did not have an effect on the inhibitory activity. The active catechins inhibited BACE1 activity in a non-competitive manner with a substrate in Clever.



2)



```
MAQALPWLLWVMGAGVLPAGHTOHGIRLP  
LRSGIGGAPLGLRPLRETDEEPEEPPGRPG  
SFVEMVDNLRGHSOGGYVEMTVGGSPPO  
TLNILVDTGSSNFVAGAAPFFLHRYYOPO  
LSSTYRDLRIGVYVYPYTOGHVVEGELGTDL  
VSPHGPVNTVRANIAITESDIFFFINGSNW  
EGILGLAYAEIARPPDSDLEPFFDSLVIQTHV  
PNLFSLOLCGAGFPPLNQEVLASVGGSMII  
GGIDHSLYTGSLWYTPIRRENYYEVIMRVE  
INGQDLHMDCHEYNVDISMVDSGTTNLRLP  
KMFEEAAVSIHAAASSTENF PDGFNVLGEO L  
VQWQAGTTPWNIFPVISLYLMGEB/TNOSF  
RITILPQOYLRPVEDVATSDQDDCYHFAKDS  
STGTVMGAVIMEGFYVWFDRARIRIGFA
```



Unstructured Data
from Documents



Structured Entries in
Relational Databases

Expert Extraction
or
Text Mining



Manually extracted SAR data (commercial)

- GOSTAR (*GVKBIO Online Structure Activity Relationship Database*) is a comprehensive database that captures explicit relationships between the three entities of publications, compounds and targets.

IBEX
User: Muresan, Sorel

Select a database **All (5297068 records)** Select a collection **All (5297068 rec**

Structure and or not **Reference**

Document type: Patents

Patent Number:

Year:

Category: **Target/Source** Activity PhysChem Proper

Protein: Equals:

Approved Symbol: Equals:

Entrez Gene ID: Equals:

Source: Equals:

Substructure Similarity Exact

Submit Clear

Release 1.4.1 for AstraZeneca

AZ Number: **AZ10018622**

Chemical Structure:

Compound Name: 2-[[3-(1-cyano-1-methyl-ethyl)-5-(1,2,4-triazol-1-ylmethyl)phenyl]-2-methylpropanenitrile

Molecular Weight: 293.3

ClogP: 1.47

PSA: 61.3

AZFILTERS: Core

Platform Name: Mcd

Claim/Example: Compound Anastrozole

SMILES: CC(C)(C#N)c1cc(cc(c1)C(C)C#N)Cn2cnc2

CSID (structure): CS000045540

ChempSpider ID: 20570646

Title: **Aromatase inhibitors and inactivators for the treatment of postmenopausal breast cancer: a review**

Authors: Jurgen Geisler

Company Address: Haukeland University Hospital, Department of Medicine, Section of Oncology, N-5021 Bergen, Norway

Journal/Patent: *Curr. Med. Chem. Immun. Endoc. and Metab. Agents.*, 2003, 3 (3) 216-276 **Glides**

Bio Assay: **Aromatase inhibitor: Useful in the treatment of breast cancer**

Protein	Source	Cells/Cell-Line/Organ	Approved Symbol	Entrez Gene ID	Multiple Loci	Locus Ref	Activity Type	Activity UOM	Activity Prefix	Activity Value	Enzyme / Cell Assay	Assay Type
	HUMAN	Plasma cells					Activity	%	>	80	Percent suppress of plasma estradiol level in healthy male and female volunteers	F1
Aromatase	HUMAN		CYP19A1	1300			Inhibition	%	=	96.7	Inhibitory activity of the compound against human Aromatase at dose of 1 mg once daily	M
Aromatase	HUMAN		CYP19A1	1588			Inhibition	%	=	98.1	Inhibitory activity of the compound against human Aromatase at dose of 1 mg once daily	M
	HUMAN	Breast cancer cells					Activity	%	=	83.5	Percent suppress of plasma estrone sulfate level in postmenopausal breast cancer patients	F1
	HUMAN	Breast cancer cells					Activity	%	=	98.1	Percent suppress of total body aromatization in postmenopausal breast cancer patients at the dose of 10 mg	F1
	HUMAN	Breast cancer cells					Activity	%	=	93.5	Percent suppress of plasma estradiol level in postmenopausal breast cancer patients	F1
	HUMAN	Breast cancer cells					Activity	%	=	86.5	Percent suppress of plasma estrone level in postmenopausal breast cancer patients	F1
	HUMAN	Breast cancer cells					Activity	%	=	96.7	Percent suppress of total body aromatization in postmenopausal breast cancer patients at the dose of 1 mg	F1
Aromatase	HUMAN		CYP19A1	1508			Inhibition	%	=	97.3	Inhibitory activity of the compound against human Aromatase at dose of 1 mg once in daily	B
Aromatase	HUMAN		CYP19A1	1588			IC50	nM/L	=	15	In vitro inhibitory activity of the compound against human placental Aromatase	B
Target	Derivative		Mechanism			Indication			Binding Site			
Aromatase			Inhibitor			CARCINOMA, BREAST						



SAR data (public)

- **PubChem**

- the NCBI public informatics backbone for the NIH Molecular Libraries Initiative focused on small molecules as systems biology probes and potential therapeutic agents.

- **ChEMBL**

- includes drugs, small molecules from the medicinal chemistry or biochemical literature and their targets.



Extracting chemical entities from text

Collaboration with IBM Research Almaden to apply text analytics technology to analyze intellectual property and scientific literature

- 10 million full text patents
- 11 million structures
- 17% out of 58M parent structures in Chemistry Connect



Chemical Named Entity Recognition (NER)

United States Patent [15] **3,692,776**
Shindo et al. [45] **Sept. 19, 1972**

[54] **PROCESS FOR PREPARING 7-CHLORO-1,3-DIHYDRO-1-METHYL-5-PHENYL-2H-1,4-BENZODIAZEPIN-2-ONE** *Primary Examiner—Henry R. Jiles*
Assistant Examiner—Robert T. Bond
Attorney—Otto John Munz

[72] Inventors: Minoru Shindo, Tokyo; Kanji Moro, Tokyo; Teizo Shinozaki, Chiba-ken, all of Japan

[73] Assignee: Chugai Seiyaku, Kabushiki Kaisha, Tokyo, Japan

[22] Filed: Oct. 27, 1970

[21] Appl. No.: 84,549

Related U.S. Application Data

[62] Division of Ser. No. 841,611, Oct. 27, 1970.

[52] U.S. Cl.260/239.3 D, 260/562 N

[51] Int. Cl.C07d 53/06

[58] Field of Search.....260/239.3 D

[56] **References Cited** **15 Claims, No Drawings**

UNITED STATES PATENTS

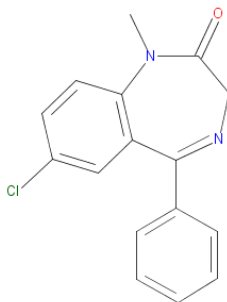
3,371,085 2/1968 Reeder et al.....260/239.3 D

7-CHLORO-1,3-DIHYDRO-1-METHYL-5-PHENYL-2H-1,4-BENZODIAZEPIN-2-ONE

Name-to-Structure software

CN1c2ccc(cc2C(=NCC1=O)c3ccccc3)Cl

ACLJDraw03281108462D



20	22	0	0	0	0	0	0	0	0	0999	V2000
3.8857	-2.1669	0.0000	C	0	0						
4.3272	-3.0642	0.0000	N	0	0						
3.7062	-3.8589	0.0000	C	0	0						
2.7481	-3.5723	0.0000	C	0	0						
2.0156	-4.2601	0.0000	C	0	0						
2.2478	-5.2385	0.0000	C	0	0						
3.2126	-5.5291	0.0000	C	0	0						
3.9384	-4.8372	0.0000	C	0	0						
4.8500	-5.2628	0.0000	C	0	0						
5.7547	-4.8159	0.0000	N	0	0						
5.9724	-3.8346	0.0000	C	0	0						



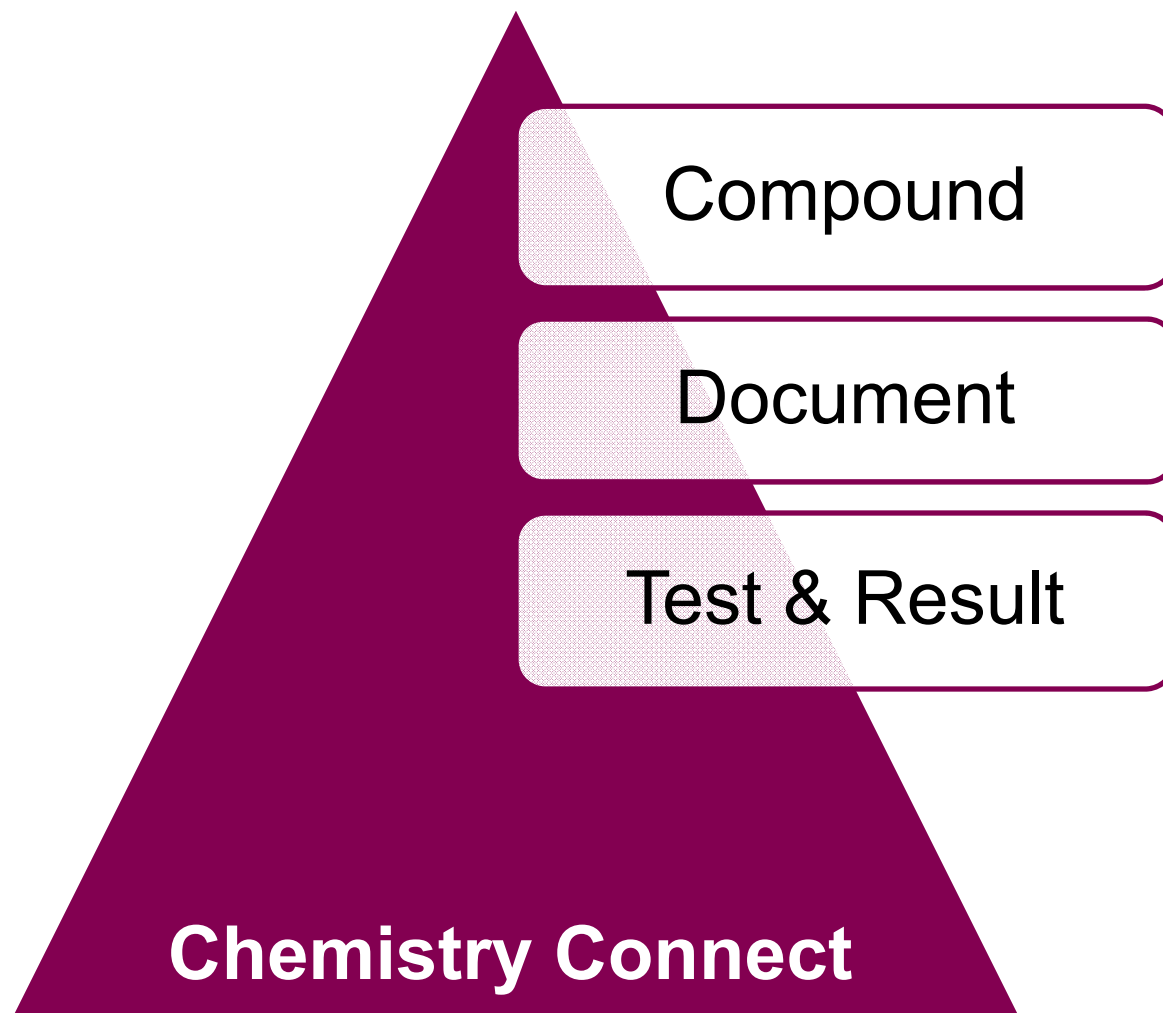
Extracting chemical entities from text

The biggest cause of missing compounds when extracting chemical entities from text is the presence of typographical errors: human errors, OCR failures, hyphenation and multiple line issues, etc.

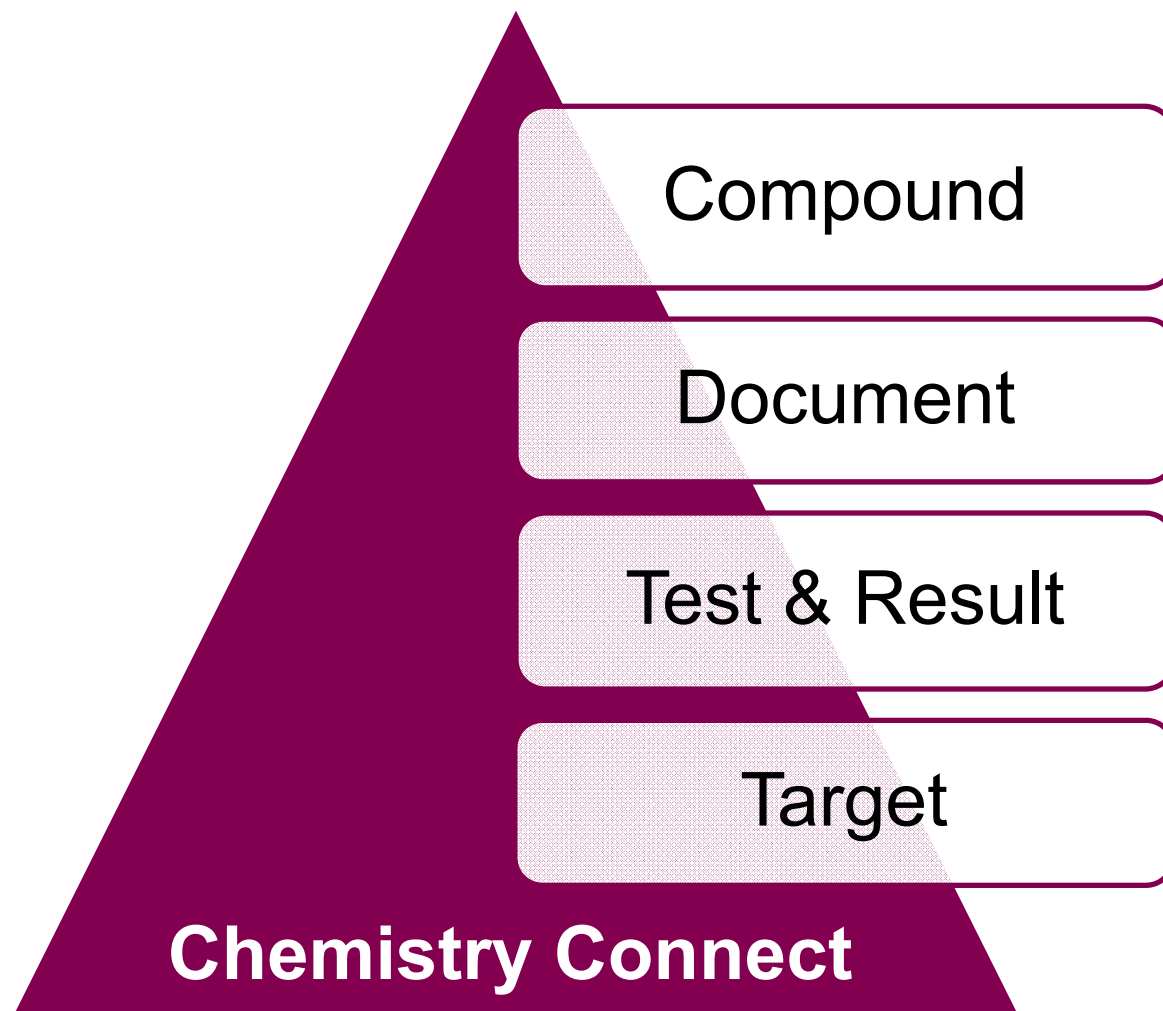
- Automated spelling correction with CaffeineFix from NextMove Software
 - CaffeineFix significantly improves extraction rates (22% increase from D=0 to D=1)
 - name2structure software are complementary (40% of the structures come from single n2s contributions)



Chemistry Connect

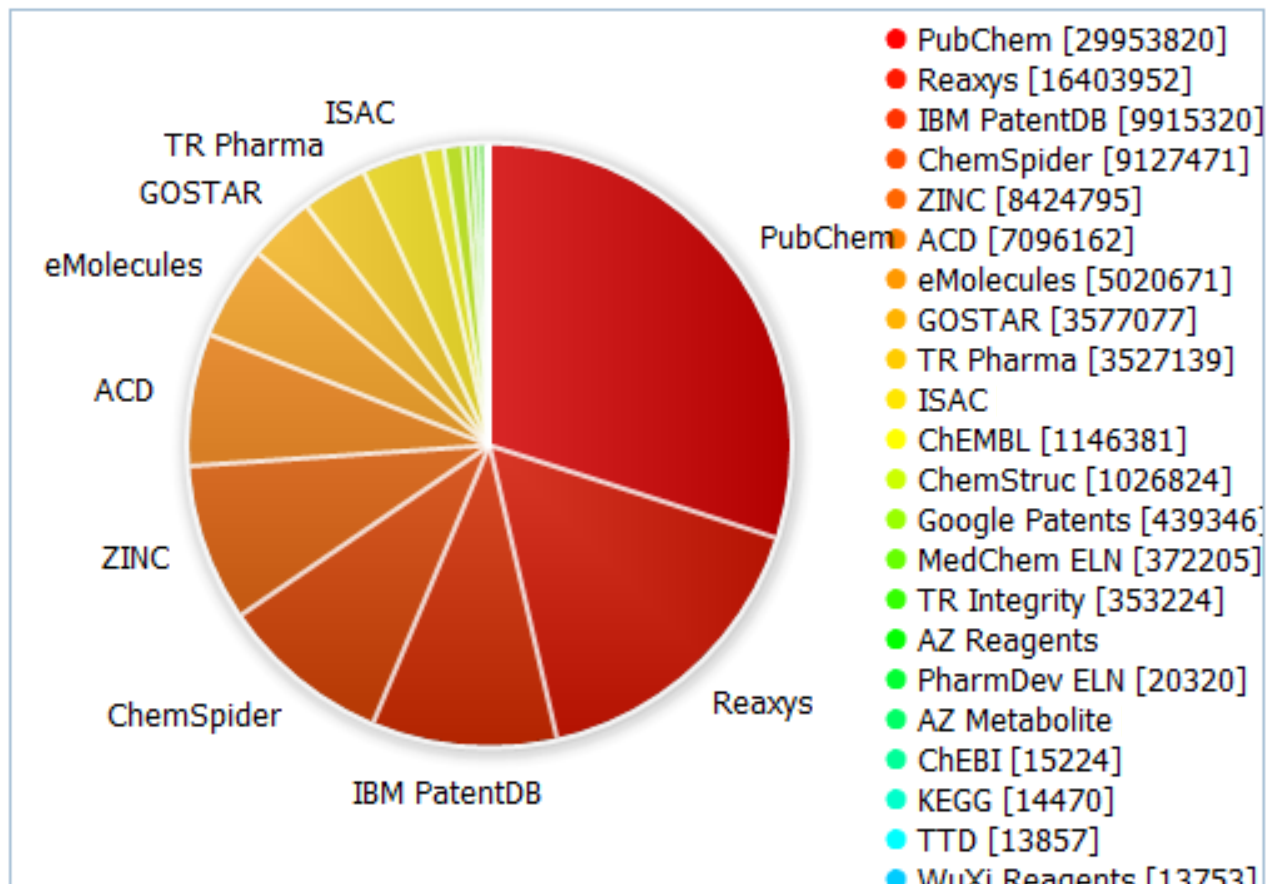


Chemistry Connect



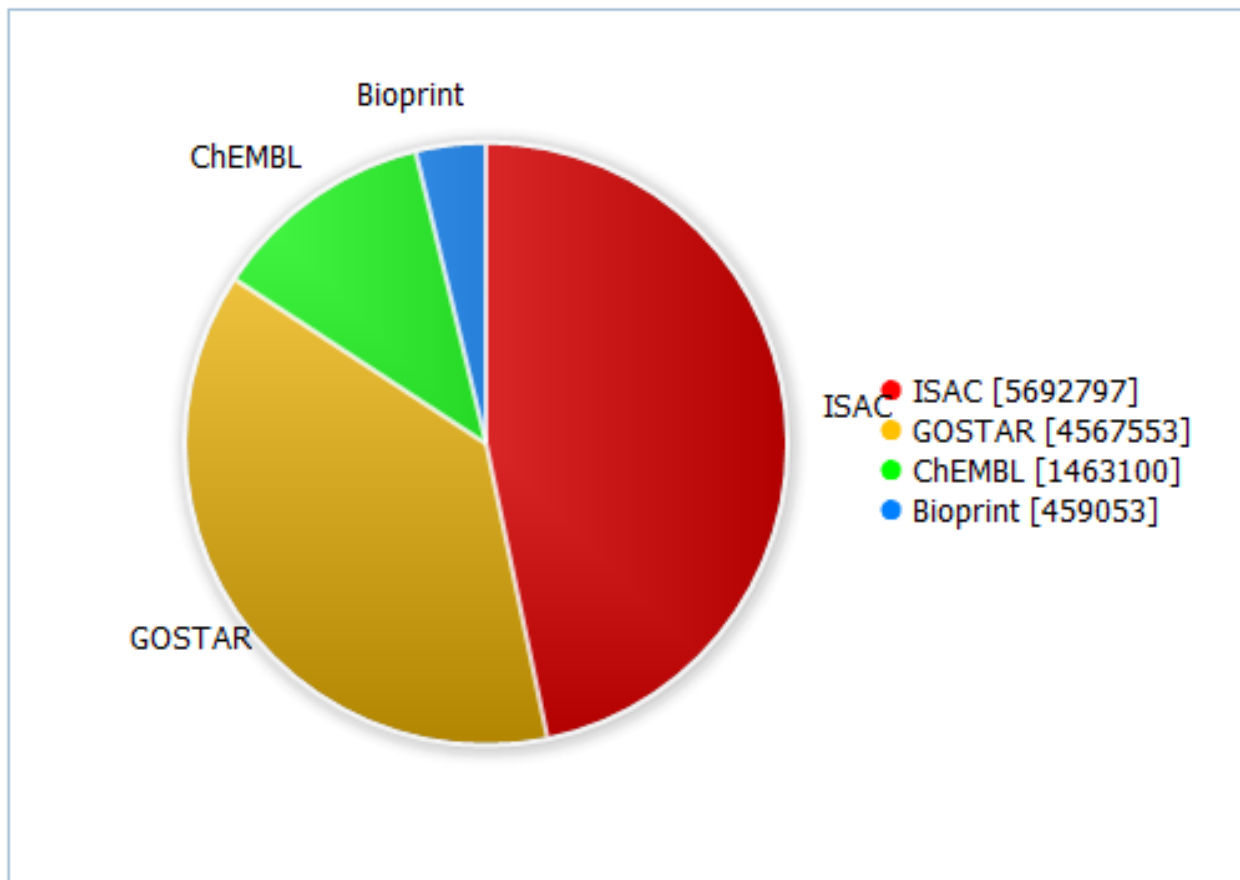
Chemistry Connect

Total structures count: 58122974



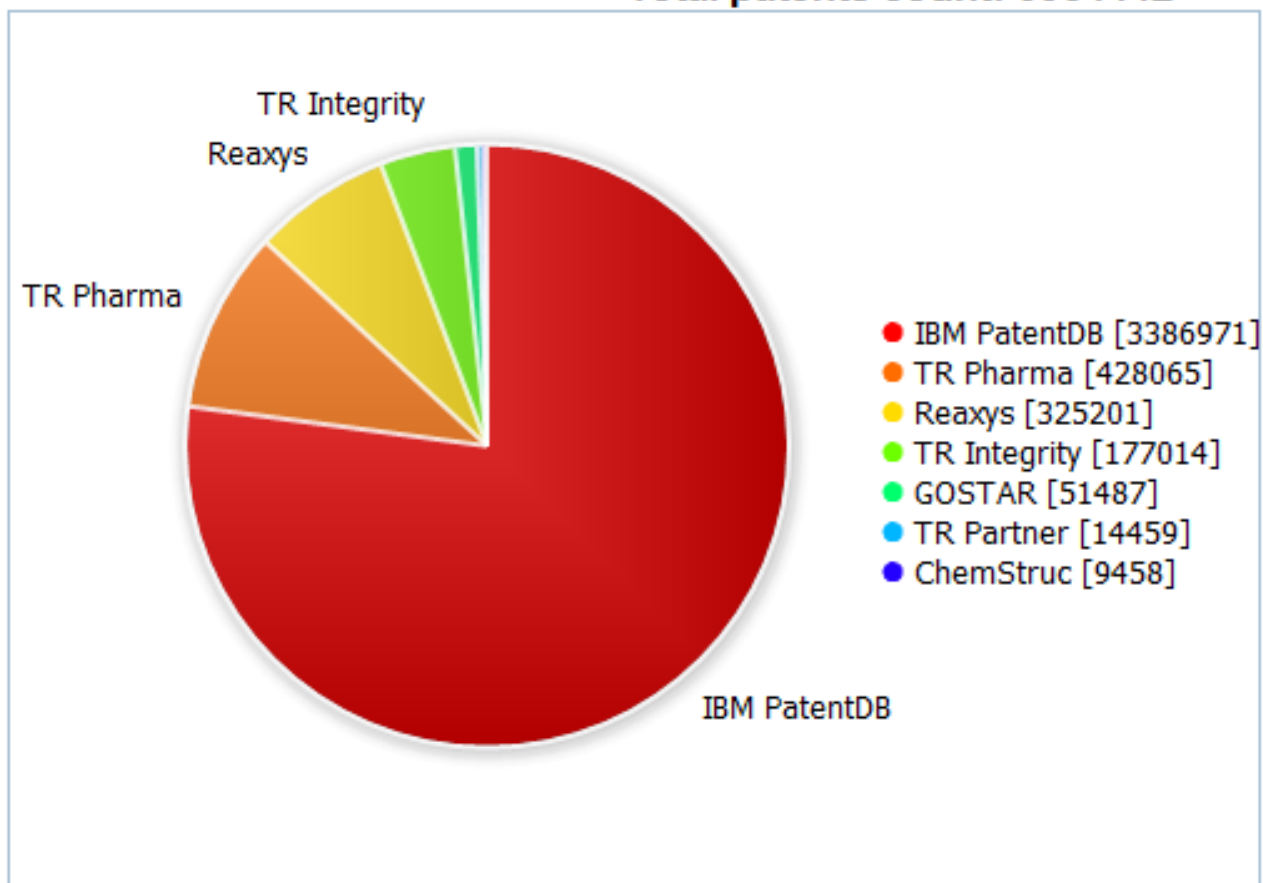
Chemistry Connect

Total bioresults count: 12182503

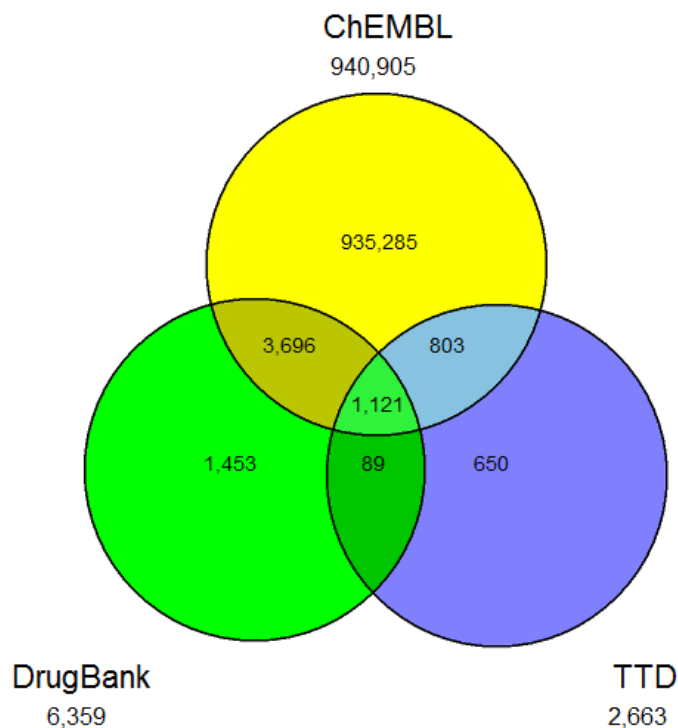


Chemistry Connect

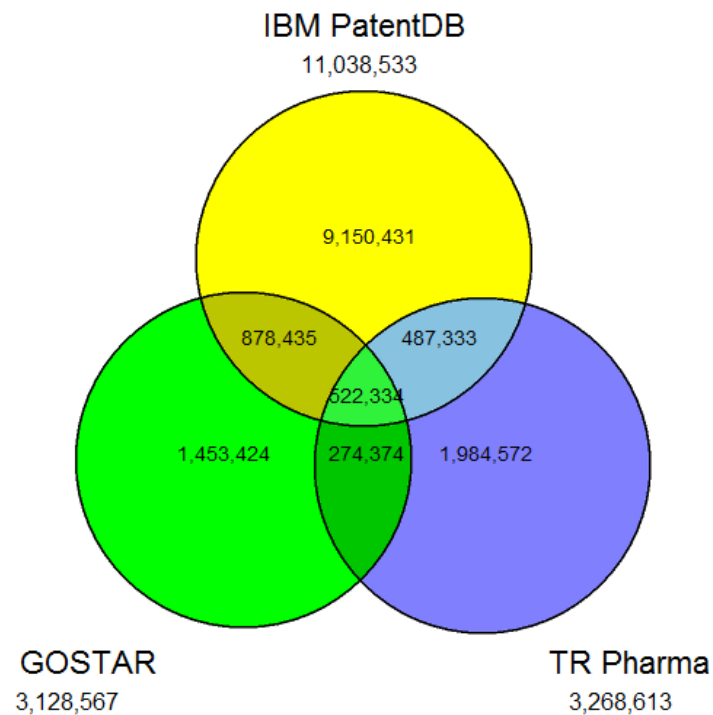
Total patents count: 3954442



Exact match source comparisons



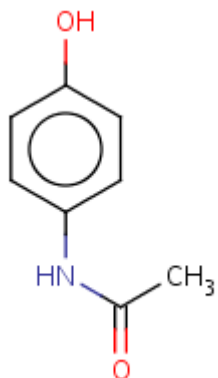
**sources that include
known drugs**



**predominantly patent-
derived compounds**



Finding a common language



[3H]Acetaminophen
10066-90-7
103-90-2
1047-607-00
1169-894-12
16110-10-4
222 AF
222-AF
3-(glutathion-S-yl)acetaminophen
37519-14-5
3-hydroxyacetaminophen
4-(Acetylamino)phenol
4-13-00-01091
4-ACETAMIDOPHENOL
4-Acetaminophenol
4-ACETYLAMINOPHENOL
4'-Hydroxyacetanilide
4-HYDROXYACETANILIDE
4-HYDROXYANILID KYSELINY OCTOVE
4-hydroxyphenolacetamide
644/4046
644/7502
64889-81-2
659/9501
77097-85-9
840-416-00
872-667-00
878-022-04
878-022-09
878-022-14
878-022-19
882-720-04
882-720-07
882-720-10

882-720-13
882-720-16
882-720-20
A F ANACIN
A PER
A.F. ANACIN
AAP
aa-sulfate
AA-sulphate
Abenol
Abensanil
ABROL
ABROLET
AC112578
AC112579
Acamol
Accu-Tap
Acenol
Acenol (pharmaceutical)
Acephen
Acertol
Aceta
Aceta Elixir
Aceta Tablets
Acetaco
Acetagesic
Acetalgin
ACETAMIDE, N-(4-HYDROXYPHENYL)-
ACETAMIDE, N-(P-HYDROXYPHENYL)-
Acetamidophenol
Acetaminofen
Acetaminophen
Acetaminophen (4-hydroxyacetanilide)
Acetaminophen glucuronide(55%)
acetaminophen sulfate

Acetaminophen
Acetaminophen (4-hydroxyacetanilide)
Acetaminophen glucuronide(55%)
acetaminophen sulfate
Acetaminophen sulfata(30%)
acetaminophen sulphate
Acetaminophen Unisers
acetaminophene
Acetamol
ACETANILIDE, 4'-HYDROXY-
Acetavance
Acetofen
ACETOMINOPHEN
Actamin
Actamin Extra
Actamin Super
Actifed Plus
Actimol
Actimol Chewable Tablets
Actimol Children's Suspension
Actimol Infants' Suspension
Actimol Junior Strength Caplets
Actron
Afebrin
Afebryl
Aferadol
AG10223
AG12029
AG124687
AG12800
AG12948
Amadiil
Aminofen
Aminofen Max
Anacin
Anacin-3
Anacin-3 Extra Strength
Anadin dla dzieci
Anaflon
Analter
Anapap
Andox
Anelix
Anexsia
Anexsia 10/660
Anexsia 5/325
Anexsia 7.5/325
Anexsia 7.5/650
Anhiba
Anoquan
Anti-Algos
Antidol
Apacet
vevyn



Acetaminophen:
>1000 synonyms..

Word of the Day : Crowdsourcing

Crowdsourcing in Chemistry Connect

Crowdsourcing will allow users to easily add, remove or update the information in Chemistry Connect. As a first step we have introduced the possibility to curate synonyms but will later expand this to also include curation of structures.

We will soon release this functionality but would very much appreciate your feedback first.

How to curate a synonym

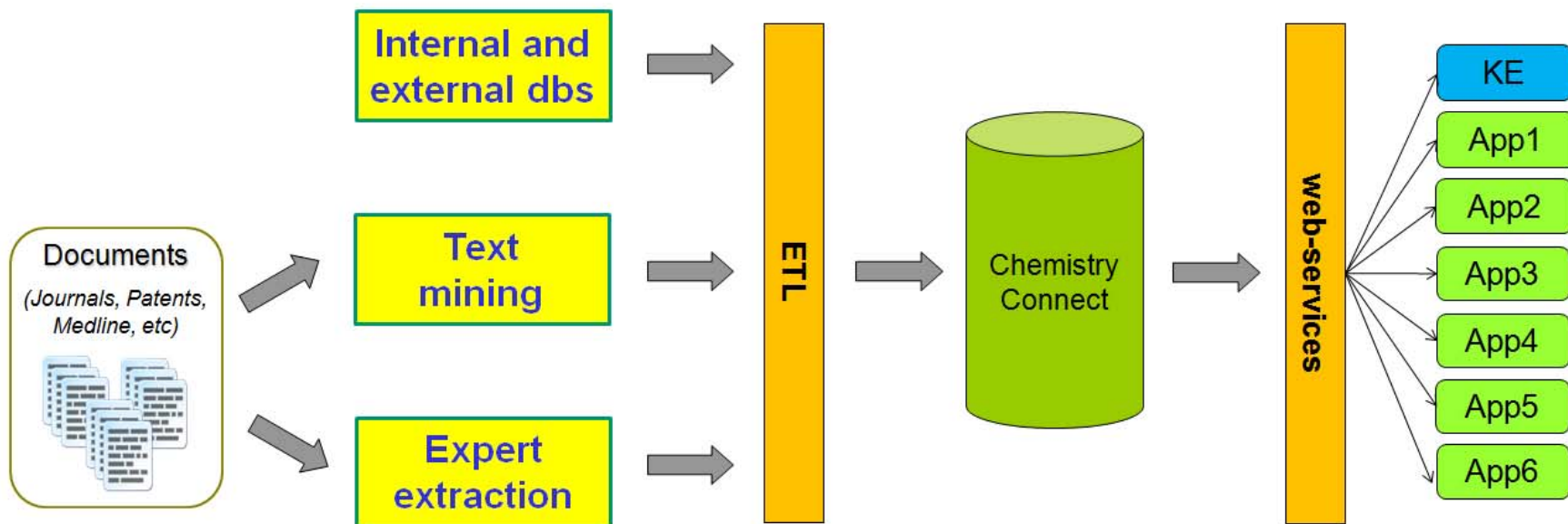
Expand the Chemistry Connect folder, choose any of the three search methods and do a search. Under the Synonyms tab three buttons are shown.

1. Add – will introduce a new synonym. New synonyms are underlined until they are moderated.
2. Delete – remove an incorrect synonym. The value will be shown with a strikethrough.
3. Edit – correct e.g. a misspelling. The old value will be shown with a strikethrough followed by the new value.

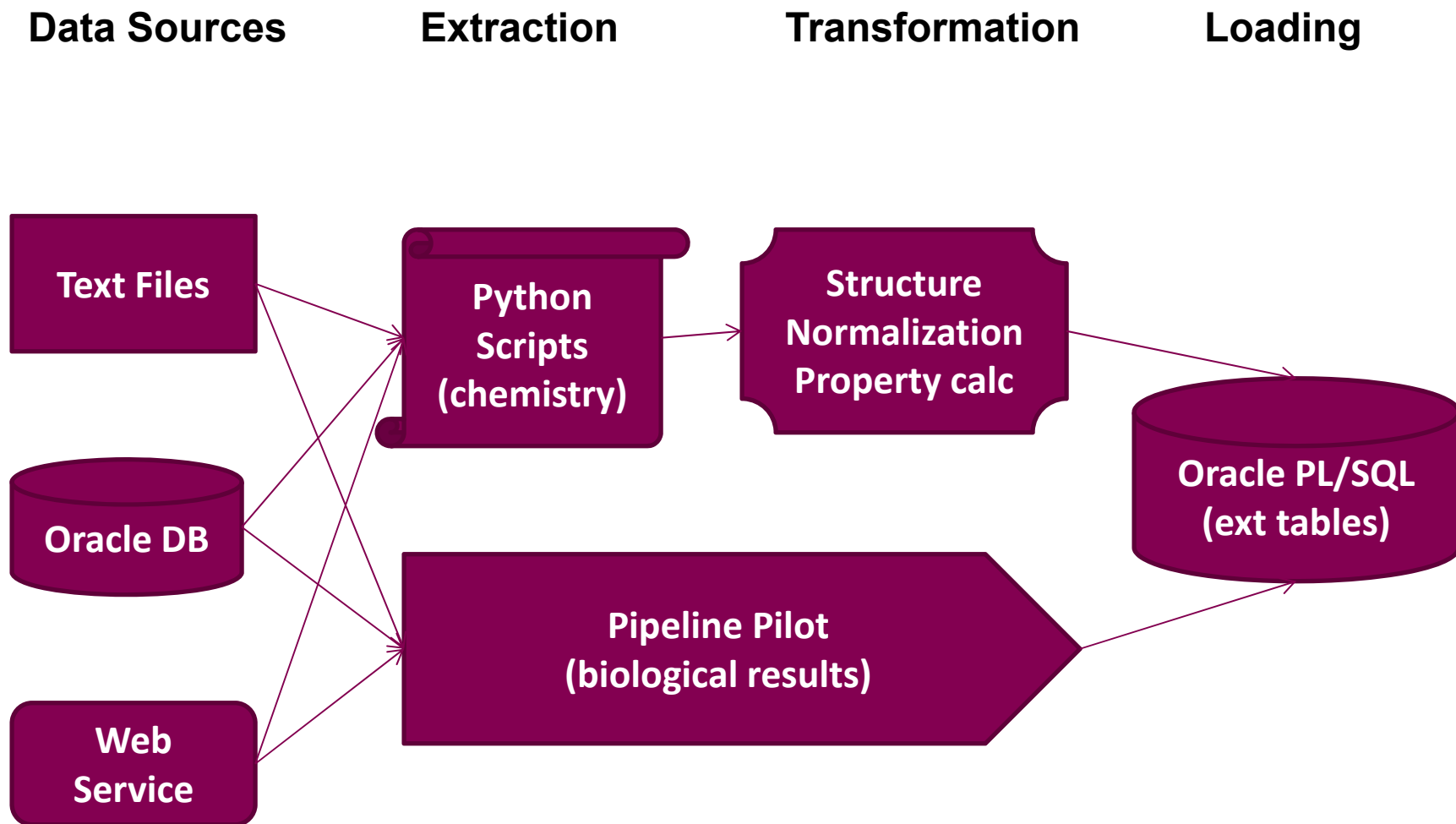
Systematic Names		Synonyms	Sources	Properties	Near Neighbours	Collections
+ Add		- Delete	✎ Edit			
Database	Name					
ChemSpider	15473678					
Ekengren, Jonas	<u>New synonym</u>					
GOSTAR	5901488					
GOSTAR	BUFORMIN BUFORMIN II					
GOSTAR	GS002669258					
GOSTAR	500383753					



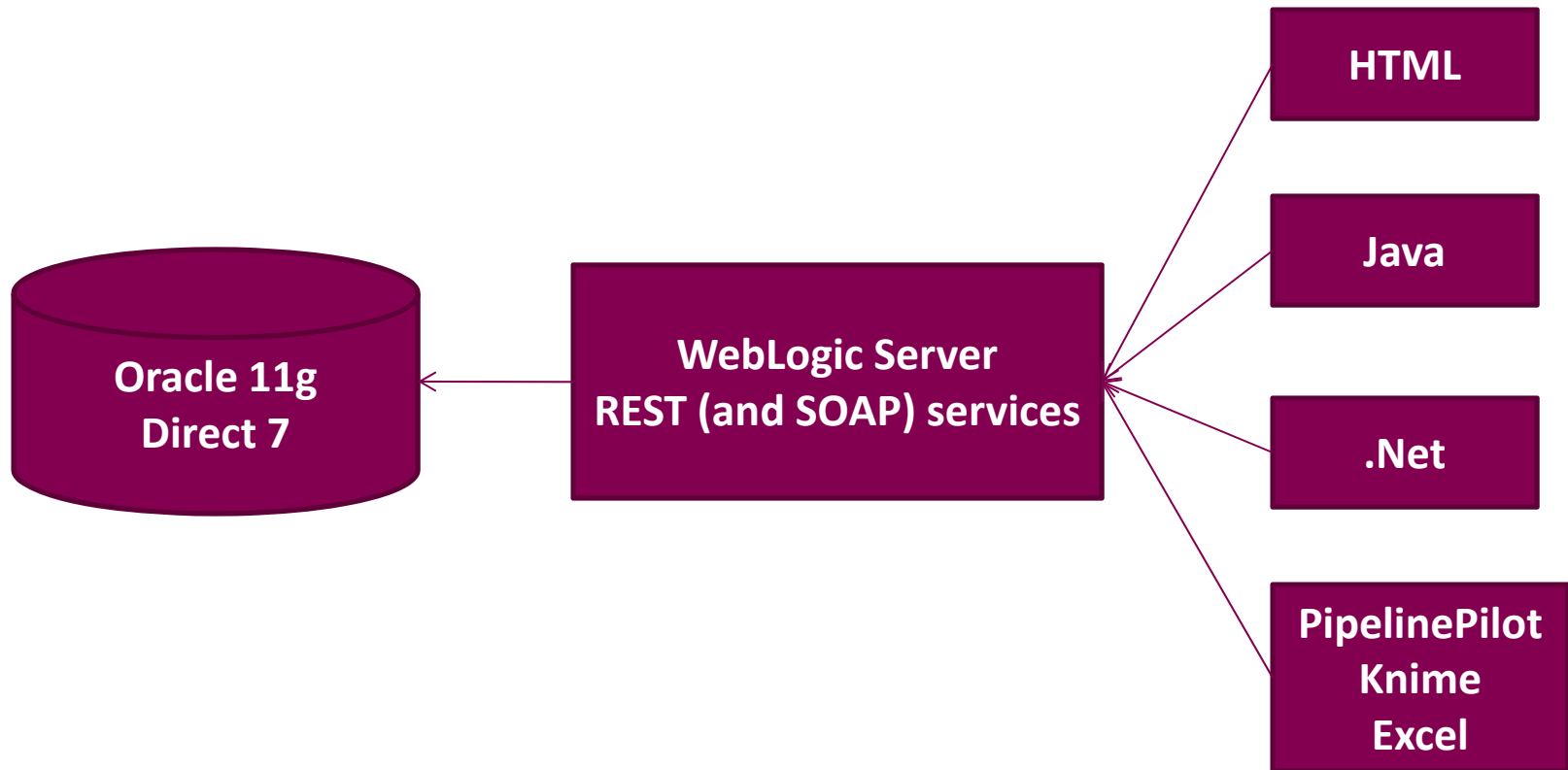
Chemistry Connect



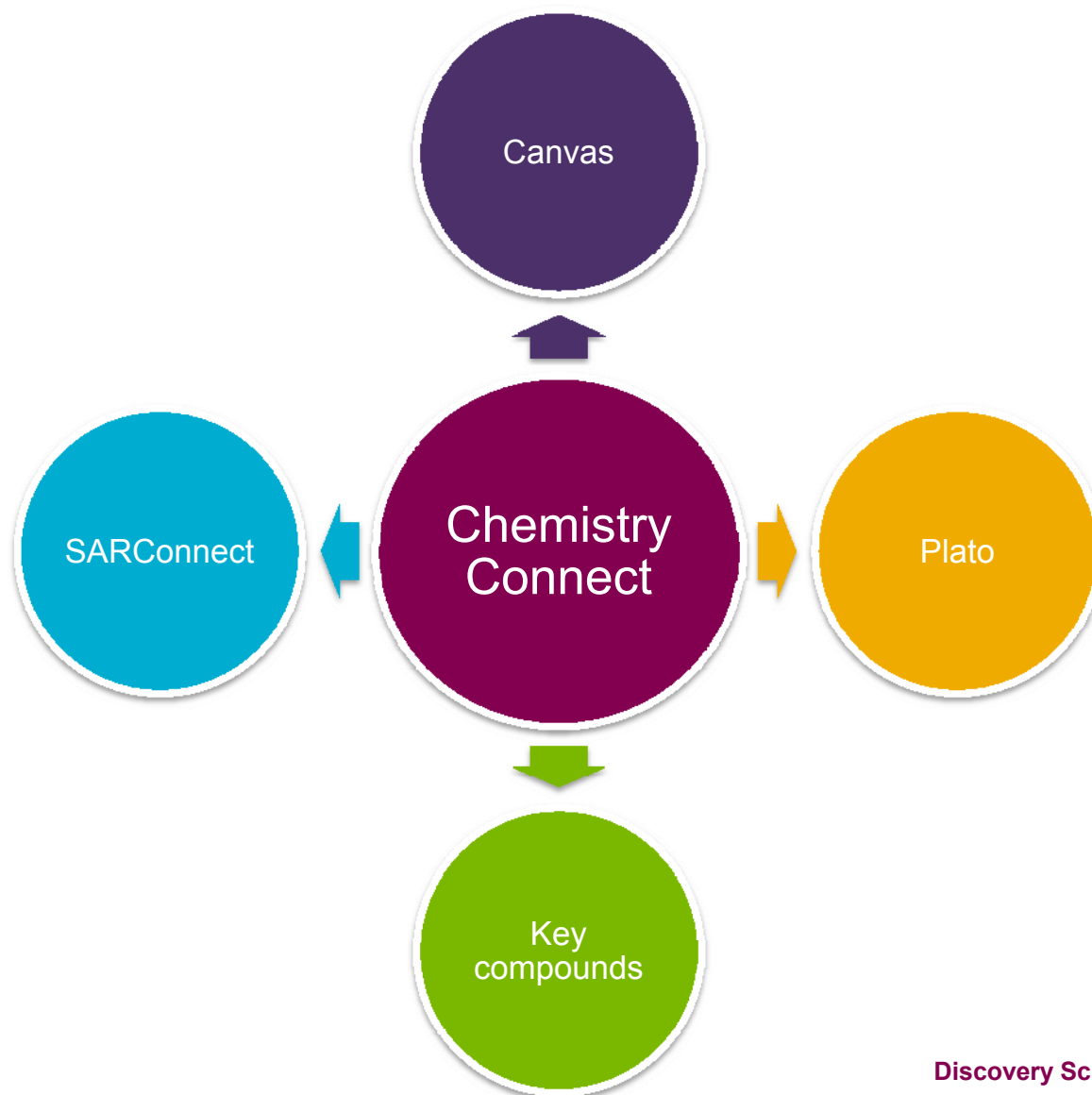
Technical Overview - ETL



Technical Overview - Application



Chemistry Connect Apps



Canvas

is...

...a **Rosetta stone** for compounds



196 B.C.

It automatically translates AZ numbers, SNs, chemical names, structures, SMILES, development IDs, reagent IDs, trade names, legacy Astra & Zeneca IDs...

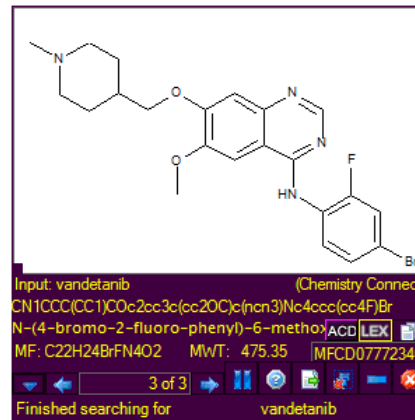
...a **portal** to information

It acts as a springboard to let you access Chemistry Connect, ISAC, IBEX, IBIS data, Compound View, ELN data, AZ Patent Db, IBEX, Integrity...



...a compound **design tool**

It quickly calculates C-lab properties, chemical names, molecular weights, checks novelty...



2012 A.D.

Database	Synonym
ACD	VANDETANIB
ACD	ZD6474
ACD	ZACTIMA
GOSTAR	ZD-6474
GOSTAR	CAPRELSA
ACD	N-(4-BROMO-2-FLUORO-METHOXY-7-(1-METHYLPYPERIDINYL)METHOXY)QUINAZOLINAMINE

...really **easy** to use

Copy a compound name or structure to the clipboard and let Canvas do the rest of the work



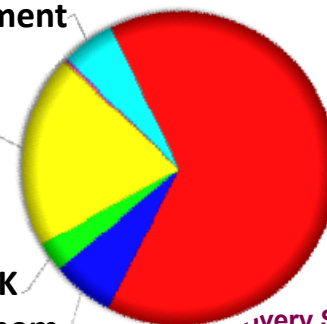
...and in 2011, **1750 AZ scientists** did

Safety assessment

Med chem

DMPK

Comp chem



& many others...

Biologists

Synthetic chem

Patent attorneys

Crystallographers

Discovery Sciences | CIC



Jon Winter, Oncology iMed

Utopia Documents

Utopia Documents - jm201705f 1.13

File Edit View Window Help

Figure 4. Fit quality scores vs heavy atom counts of benzophenones 6–24, compound FQ score around 1 indicate a near optimal ligand binding affinity for a given number

Table 2. Pharmacological Activities, Ligand Efficiency (LE), and Lipophilic Efficiency (LipE) Profiles of Selected Propafenones and P-gp Inhibitors Which Entered in Clinical Studies

compd	pIC ₅₀	HA	LE	clogP	LipE
Verapamil	6.24	33	0.27	4.47	1.77
Elacridar	7.14	42	0.24	4.21	2.93
Tariquidar	7.48	48	0.22	5.55	1.93
Zosuquidar	7.23	39	0.26	4.96	2.27
ONT093	7.50	37	0.29	7.30	0.19
Valspodar	6.30	86	0.10	15.09	-8.79
Cyclosporine A	6.99	85	0.12	14.36	-7.37
Niguldipine	6.15	45	0.20	7.80	-1.65
Propafenone	6.48	25	0.37	3.64	2.84
GPV576	8.25	35	0.3		
GPV062	7.24	34	0.3		
GPV005	6.22	27	0.3		

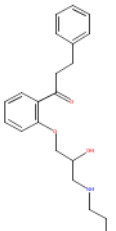
benzophenones as well as for the compound study (Table 2). The clogP values vary leading to a lipophilic efficiency range +3.08. This is somewhat surprising, as it is a lipophilic efficiency greater than 5.

Chemistry Connect
propafenone

For the lab
Related laboratory resources

Chemistry Connect
Synonymous compounds

▼ CCCNCC(COC1CCCCC1C(=O)CCc2ccccc2)O



See in Chemistry Connect...

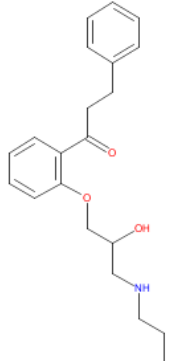
ChEMBL
Related bioactive small molecules

Wikipedia entries
Community curated encyclopedia entries

Articles from PubMed
Related recent articles

SciBite
Biomedical News & Intelligence

200000206

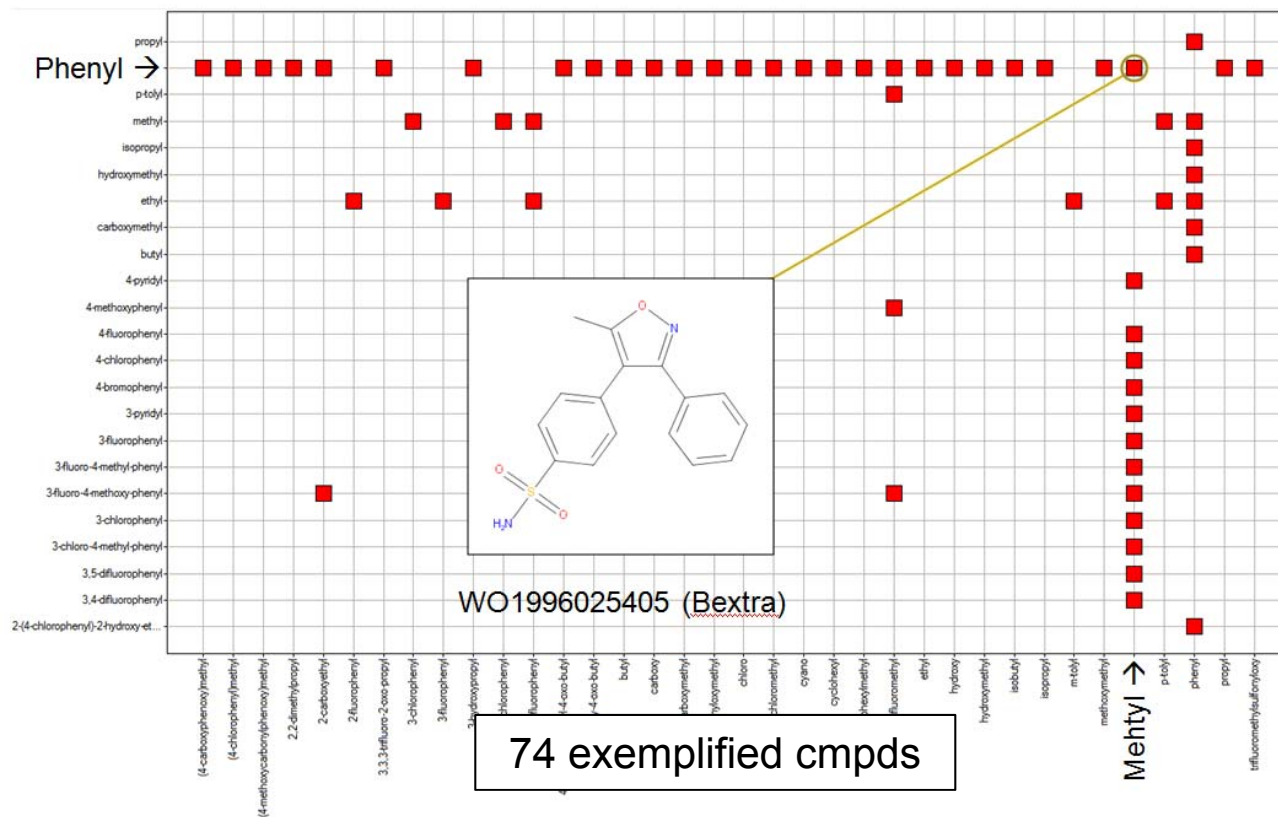
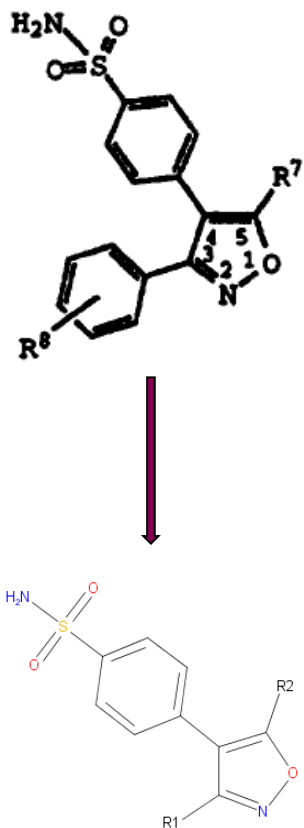


Reactions(2) Tests(287) Metabolites(0) Metabolite Parents(0)

Systematic Names	Synonyms(185)	Sources(31)	Properties	Near Neighbours(3)	Collections(7)
Name	Value				
Chem Connect ID	200000206				
Mol Formula	C21 H27 N O3				
OE SMILES	CCCNCC(COC1CCCCC1C(=O)CCc2ccccc2)O				
IUPAC Name (OE)	1-[2-[2-hydroxy-3-(propylamino)propoxy]phenyl]-3-phenylpropan-1-one				
IUPAC Name (ACD)	1-[2-[2-hydroxy-3-(propylamino)propoxy]phenyl]-3-phenylpropan-1-one				
InChI String	InChI=1S/C21H27NO3/c1-2-14-22-15-18(23)16-25-21-11-7-6-10-19(21)20(24)13-12-17-8-4-3-5-9-17/h3-11,18,22-23H,2,12-16H2,1H3				
InChI Key	JWHAUXFOSRPERK-UHFFFAOYSA-N				

Key compound prediction from patents

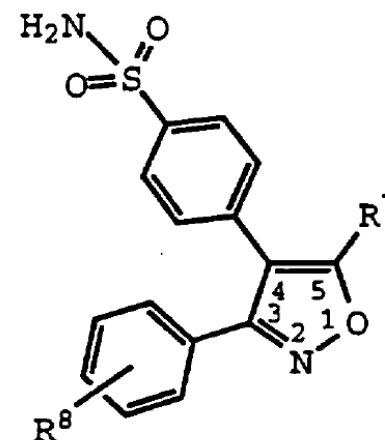
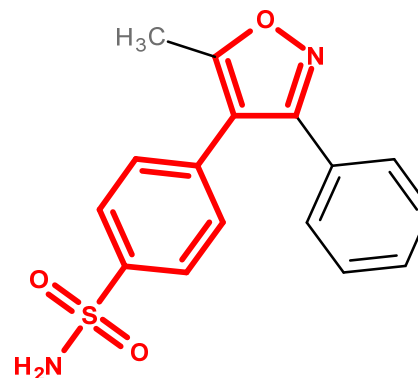
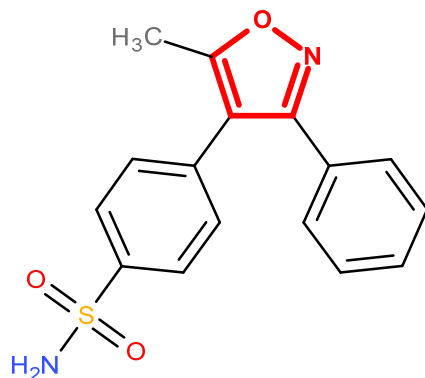
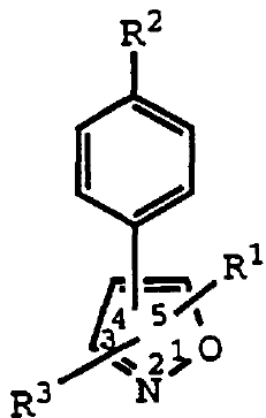
From **WO1996025405** the earliest patent which claims it, can you work out the structure of Bextra (Valdecoxib), the Pfizer NSAID?



WO1996025405 (Bextra)
74 exemplified cmpds



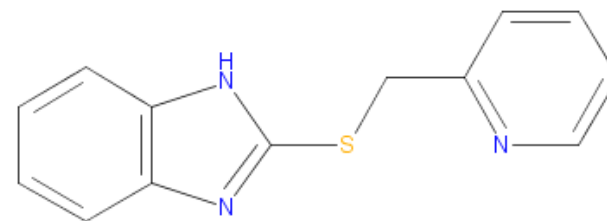
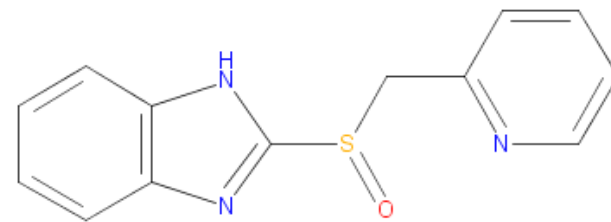
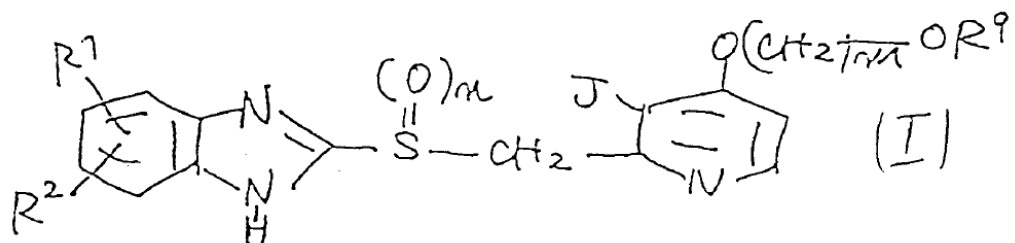
WO1996025405 - Bextra



Source	#compounds	Bextra exists	Bextra ranked
GVKBIO	74	Y	1 (broad core) 1 (narrow core)
SureChem	501	Y	1 (broad core) 1 (narrow core)



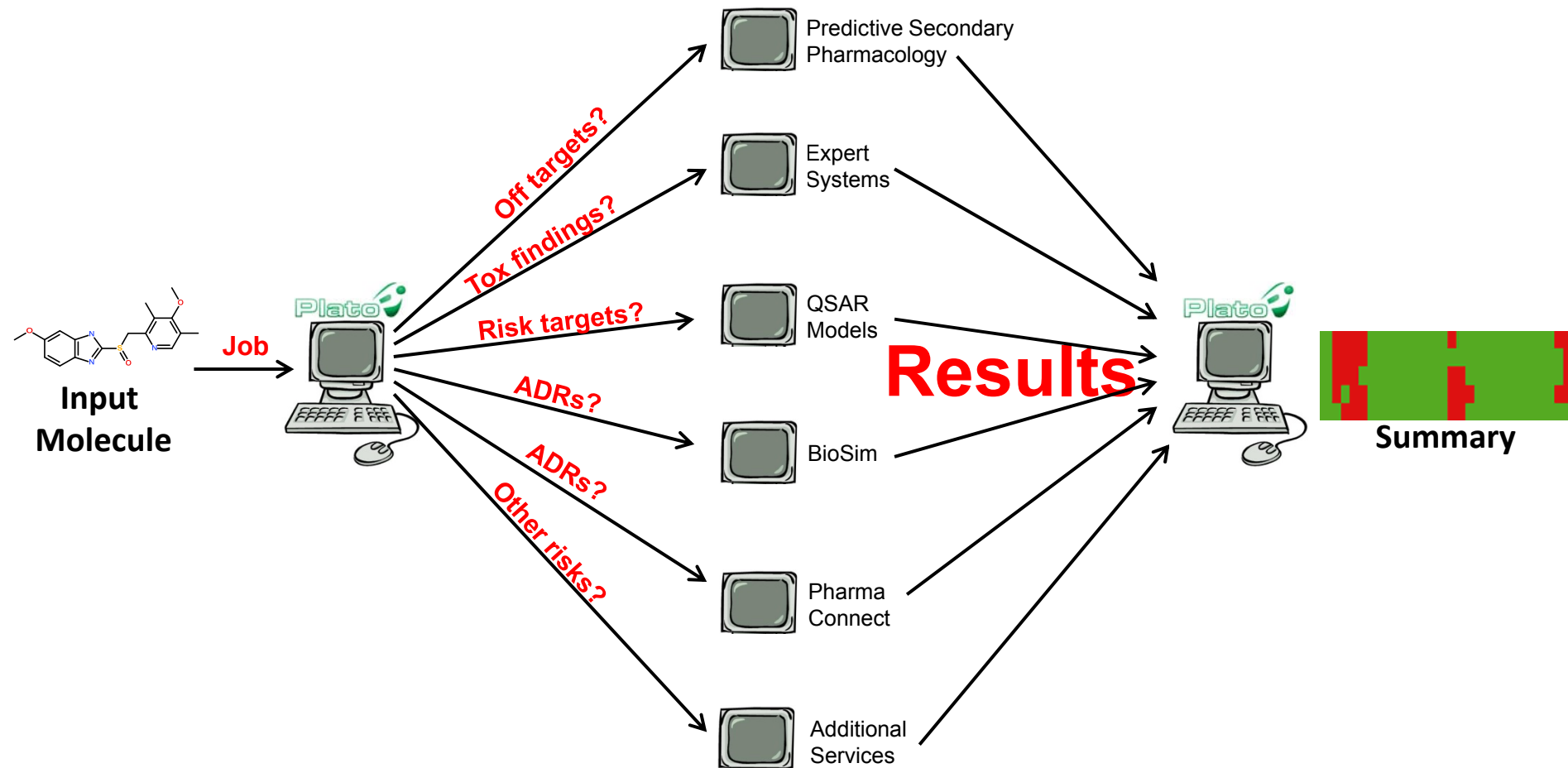
EP268956 - Aciphex



Source	#compounds	Aciphex exists	Aciphex ranked
GVKBIO	27	Y	2 (core1) 1 (core2)
SureChem	168	Y	1 (core1) 1 (core2)



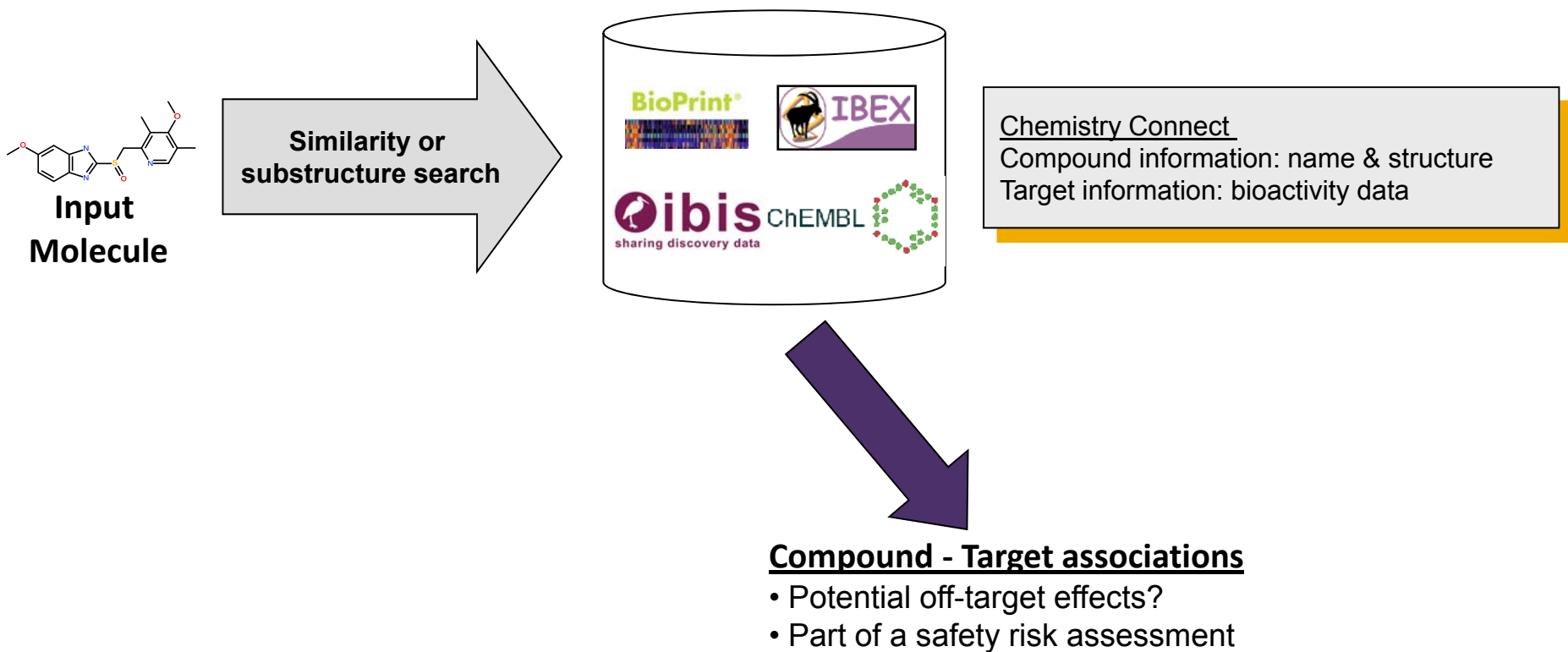
PLATO for Safety/Tox – General Concept



Scott Boyer
Catrin Hasselgren
Lars Carlsson
Tobias Noeske

All services in Plato are complementary to find an overall answer to your problem!

Predictive Secondary Pharmacology strategy



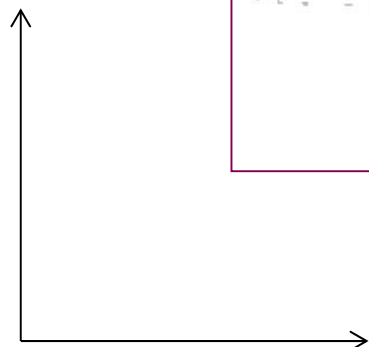
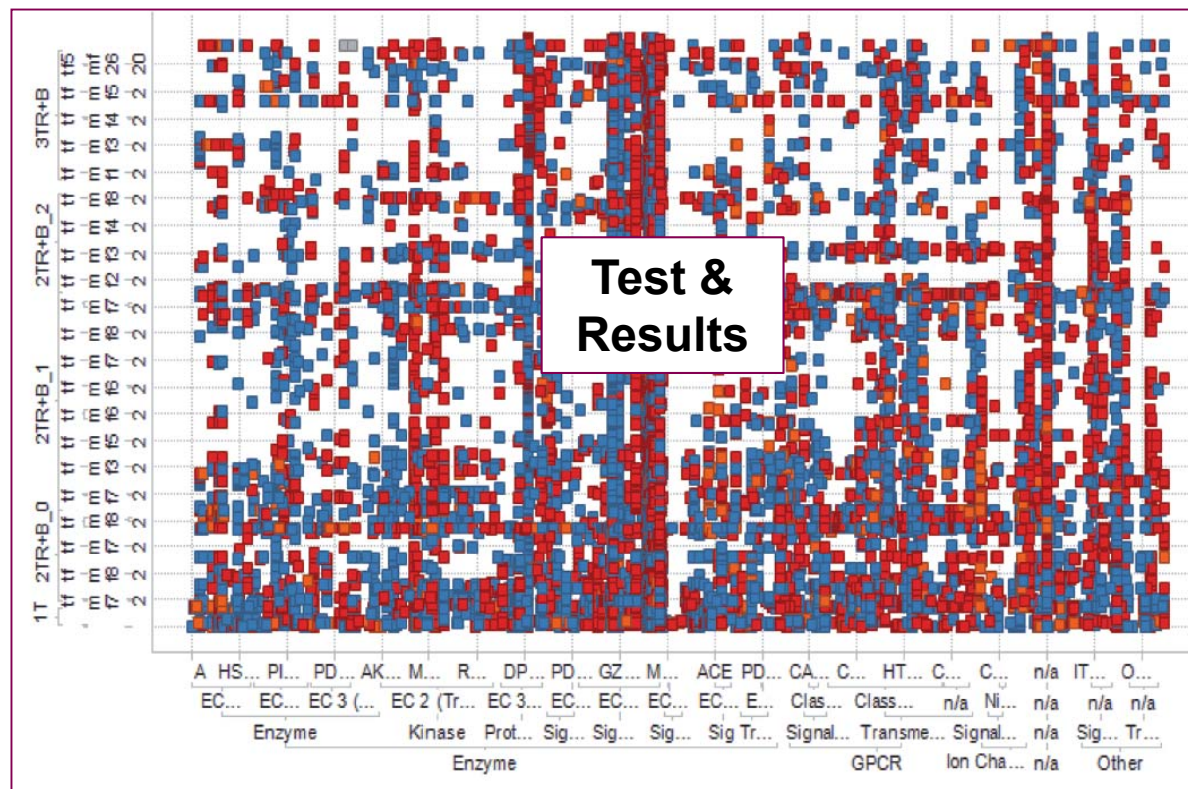
Scott Boyer
Catrin Hasselgren
Lars Carlsson
Tobias Noeske

Similarity concept: Similar compounds bind to similar targets.

M Johnson et al., *Prog Clin Biol Res* (1989), 291:167
P Willet, *Drug Discov Today* (2006), 11:1046

SARConnect – navigate SAR landscape

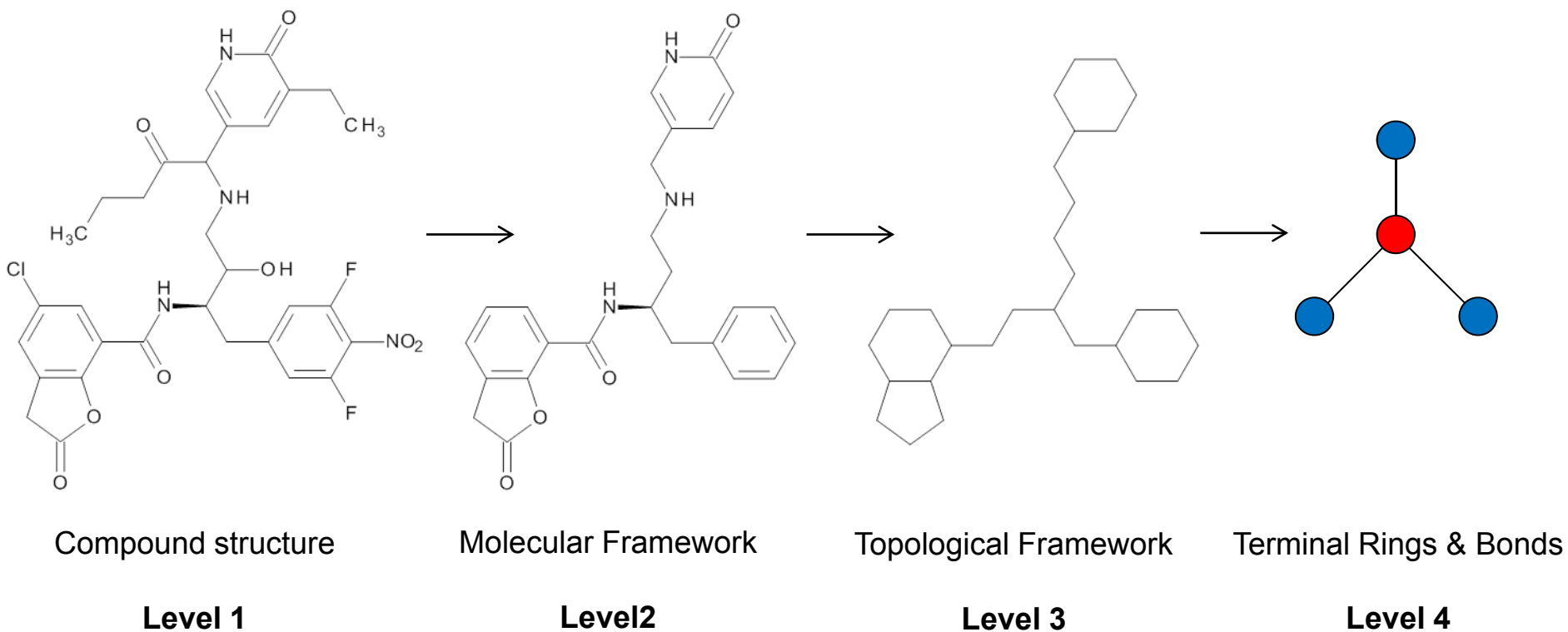
Compound hierarchy



Target hierarchy



SARConnect – structure classification



SARConnect – target classification

Level 1 (Broad target class)

Enzyme NHR **GPCR** Ion Channel Other



Level 2 (Swiss-Prot family class)

GPCR **Signal Transmembrane** Transmembrane



Level 3 (Sub-families)

Class A **Class B** Class C Frizzled Family



Symbol

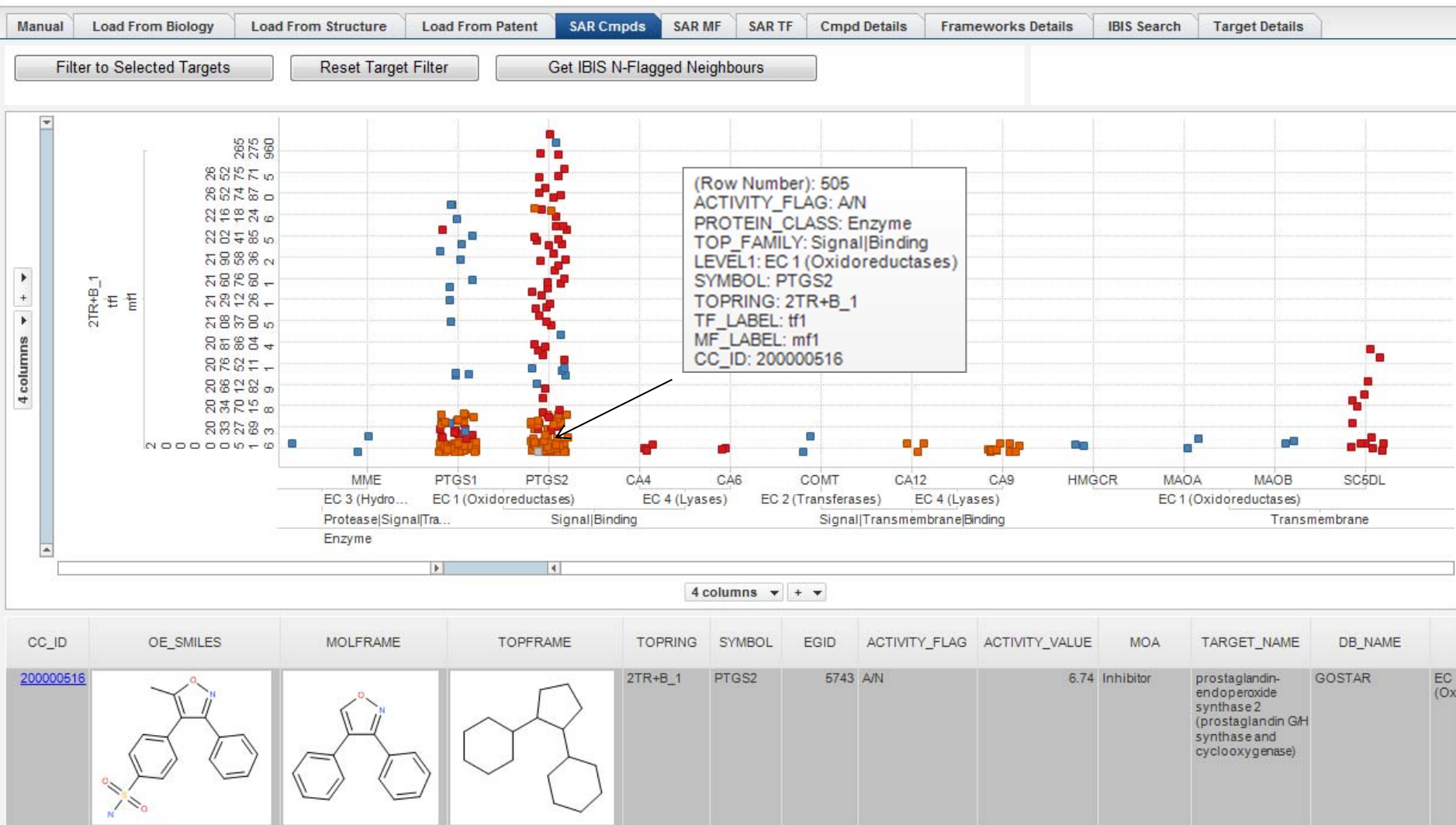
CALCR CALCLR CRHR2 GCCR GIPR GLP1R
PTH2R VIPR1 VIPR2

Table 1. Target statistics for classification in SARConnect.

Target class	Count
G-protein coupled receptor	827
G-protein coupled receptor (Class A)	717
G-protein coupled receptor (Class B)	49
G-protein coupled receptor (Class C)	22
Kinase	608
Nuclear hormone receptor	48
Ion-channel	227
Lipase	40
Phosphatase	180
Protease	575
.. Aspartyl	19
.. Cysteine	153
.. Serine	241
.. Metallo	187
.. Threonine	29
Transporter	538
EC number	4001
PDB entry	4436



SARConnect – navigate SAR landscape



Take-home messages

- Chemistry Connect is enabling AstraZeneca to intensify its exploitation of synergies between internal and external SAR estate and to shorten the time between hypothesis generation during DMTA cycles
- Our Chemical Dictionary of 120 million chemical terms has become a crucial cross-mapping resource between chemistry and the scientific literature
- We cannot wave a magic wand over data quality, provenance issues, drug name space, and the inherent challenges of chemistry representation but Chemistry Connect gives us a unique overview and amelioration options for each source



A Democracy of Ideas (Acknowledgements)

- Plamen Petrov
- Chris Southan
- Paul Xie
- Peter Varkonyi
- Thierry Kogej
- Christian Tyrchan
- Magnus Kjellberg
- Håkan Nilsson
- Mats Eriksson
- Jonas Ekengren
- Ithipol Suriyawongkul
- Niklas Blomberg
- Jon Winter
- John Cumming
- Scott Boyer
- Catrin Hasselgren
- Lars Carlsson
- Tobias Noeske
- and many others...



Thank you!

