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

AstraZeneca

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

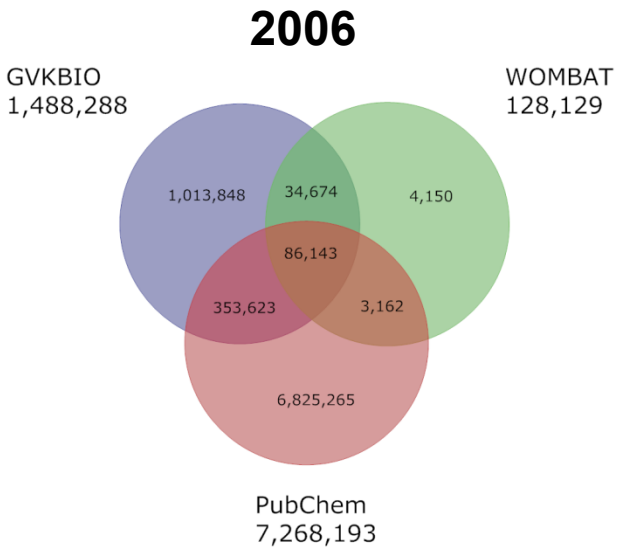
Sorel Muresan[1], Plamen Petrov[1], Christopher Southan[2], Magnus J. Kjellberg[3], Thierry Kogej[1], Christian Tyrchan[3], Peter Varkonyi[1] and Paul Hongxing Xie[1]

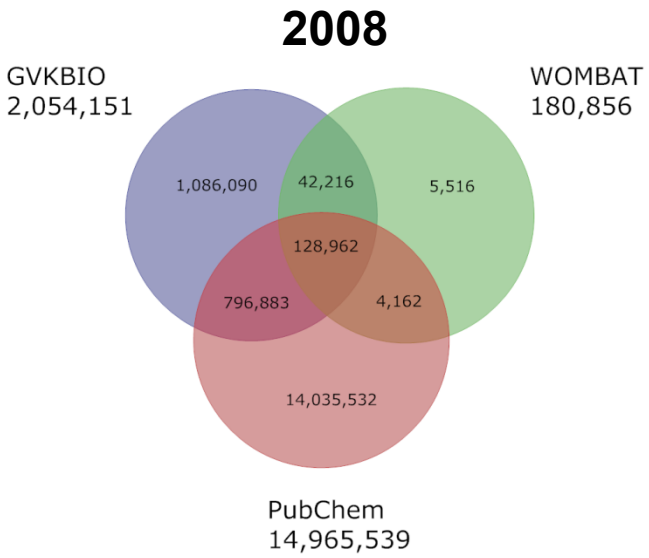
[1] DECS Computational Sciences, AstraZeneca R&D Mölndal, S-431 83 Mölndal, Sweden
[2] R&D Information, AstraZeneca R&D Mölndal, S-431 83 Mölndal, Sweden
[3] CVGI AstraZeneca R&D Mölndal, S-431 83 Mölndal, Sweden

http://dx.doi.org/10.1016/j.drudis.2011.10.005

Driver – explosion in SAR data

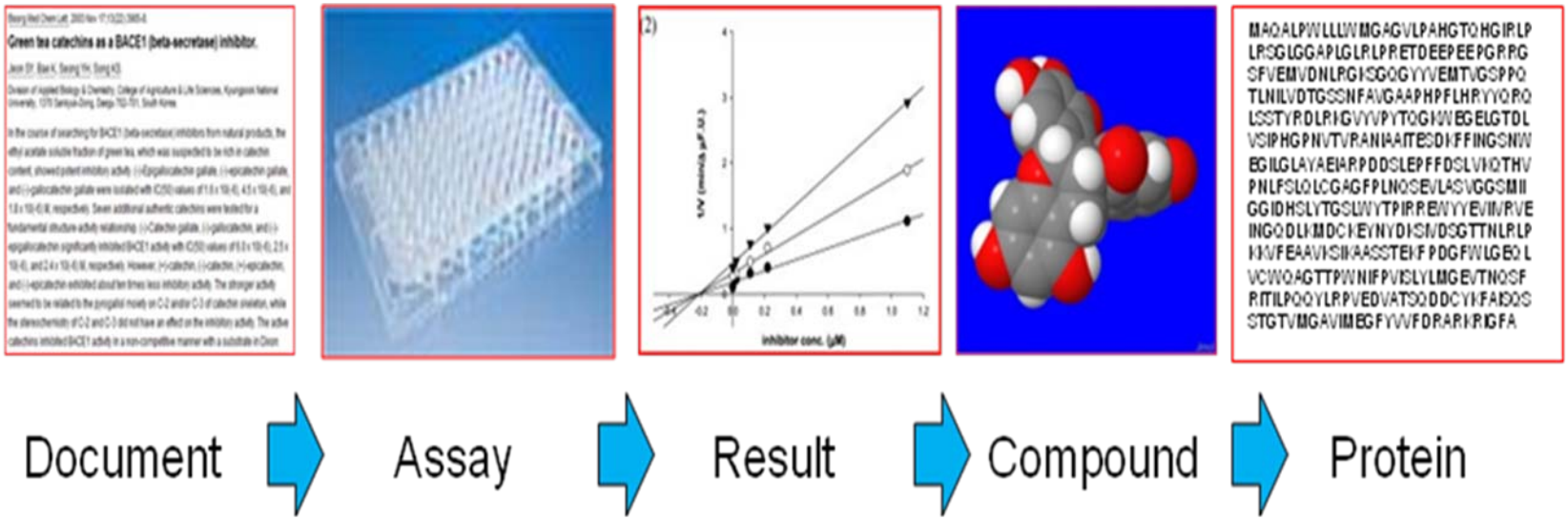
- Chemical information landscape changing fast
- Make every SAR point count, access all available chemistry
- Internal & external datasources





Southan, C.; Varkonyi, P.; Muresan, S., *J. Cheminfo. 2009*

SAR key entities and relationships



Unstructured Data from Documents

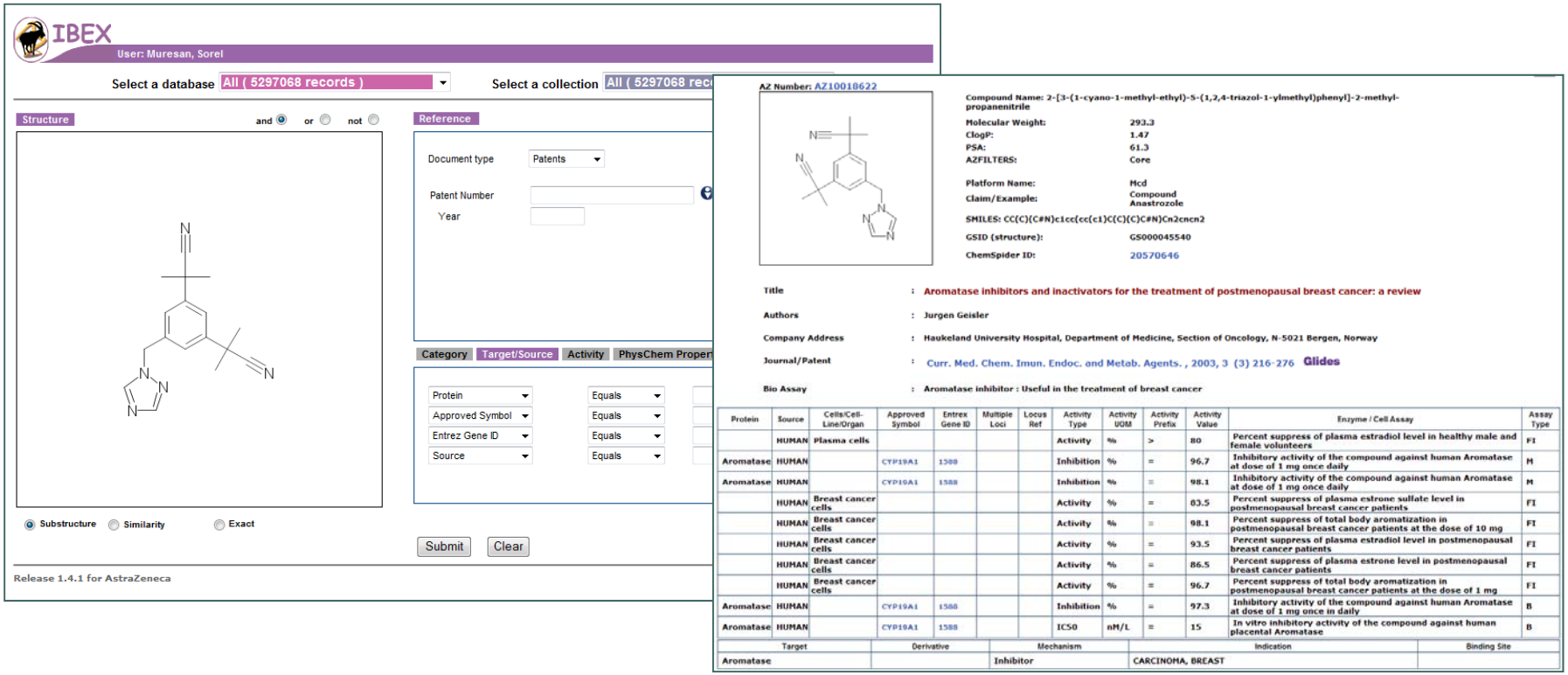
Structured Entries in Relational Databases

Expert Extraction
or
Text Mining

Southan, C.; Boppana, K.; Jagarlapudi, S.; Muresan, S. *J. Cheminfo. 2011*

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.



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.

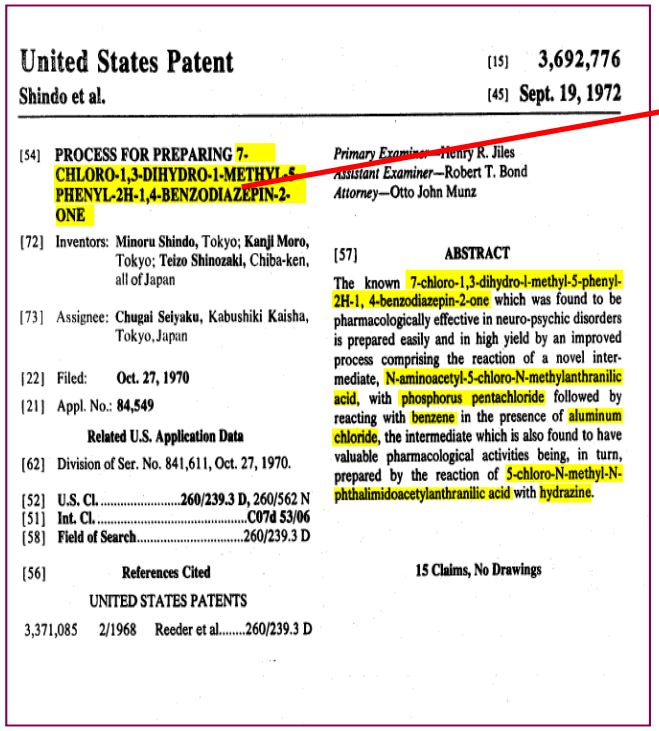
- **ChEMBL**
  - includes drugs, small molecules from the medicinal chemistry or biochemical literature and their targets.

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

Shindo et al.

[15] 3,692,776

[45] Sept. 19, 1972

[54] PROCESS FOR PREPARING 7-CHLORO-1,3-DIHYDRO-1-METHYL-5-PHENYL-2H-1,4-BENZODIAZEPIN-2-ONE

[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

UNITED STATES PATENTS

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

Primary Examiner—Henry R. Jiles

Assistant Examiner—Robert T. Bond

Attorney—Otto John Munz

[57] ABSTRACT

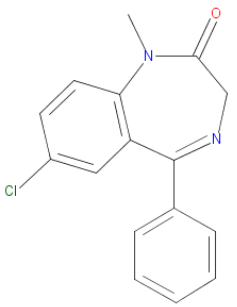
The known 7-chloro-1,3-dihydro-l-methyl-5-phenyl-2H-1, 4-benzodiazepin-2-one which was found to be pharmacologically effective in neuro-psychic disorders is prepared easily and in high yield by an improved process comprising the reaction of a novel intermediate, N-aminoacetyl-5-chloro-N-methylanthranilic acid, with phosphorus pentachloride followed by reacting with benzene in the presence of aluminum chloride, the intermediate which is also found to have valuable pharmacological activities being, in turn, prepared by the reaction of 5-chloro-N-methyl-N-phthalimidoacetylanthranilic acid with hydrazine.

15 Claims, No Drawings

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  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
```

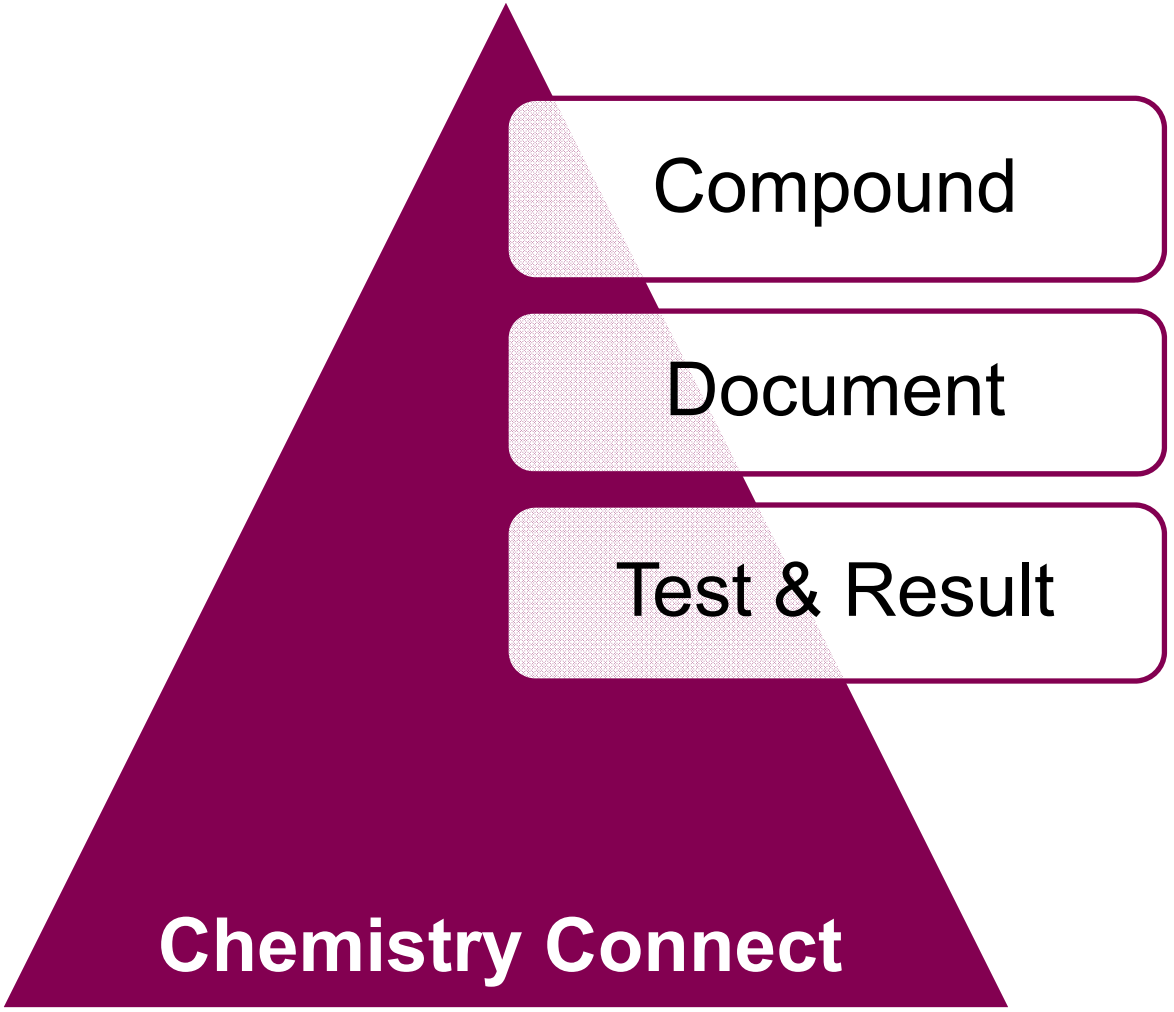
Extracting chemical entities from text

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.

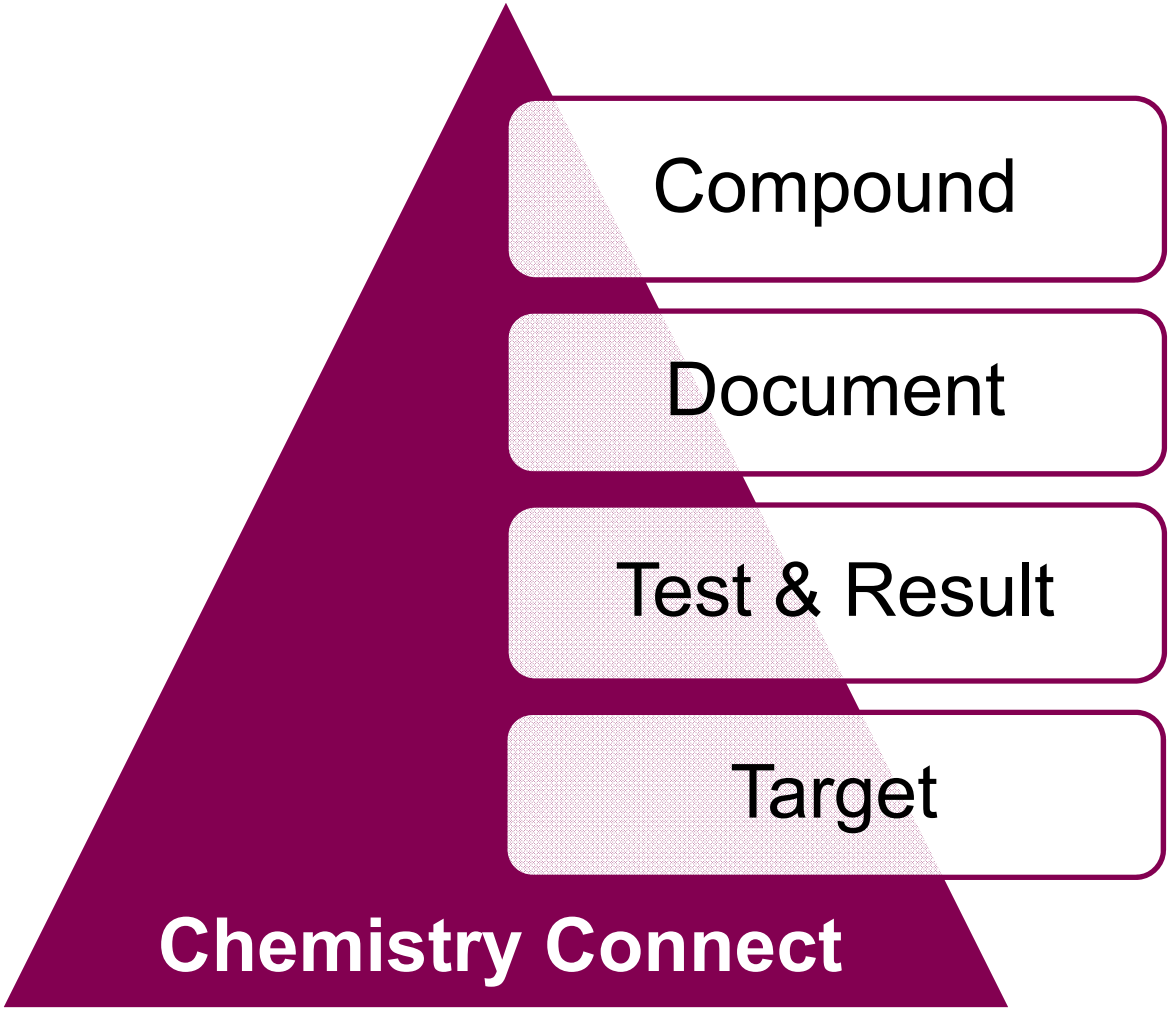
- Automated spelling correction with CaffeineFix from NextMove Software
  - CaffeineFix significantly improves extraction rates (22% increase from D=0 to D=1)
  - name2structure software are complementary (40% of the structures come from single n2s contributions)

Sayle, R.; Xie, P.; Muresan, S., *JCIM 2012*

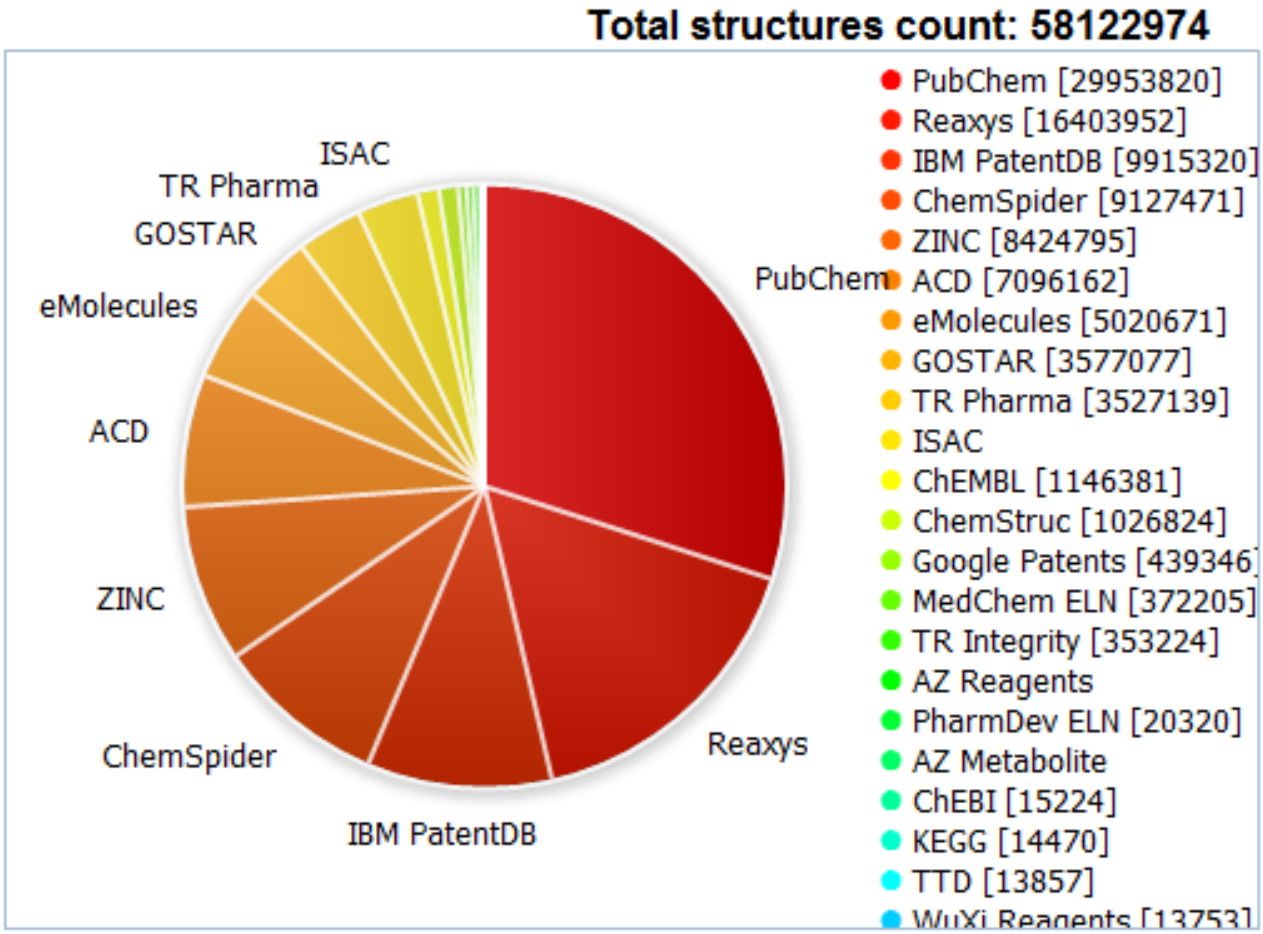
Chemistry Connect



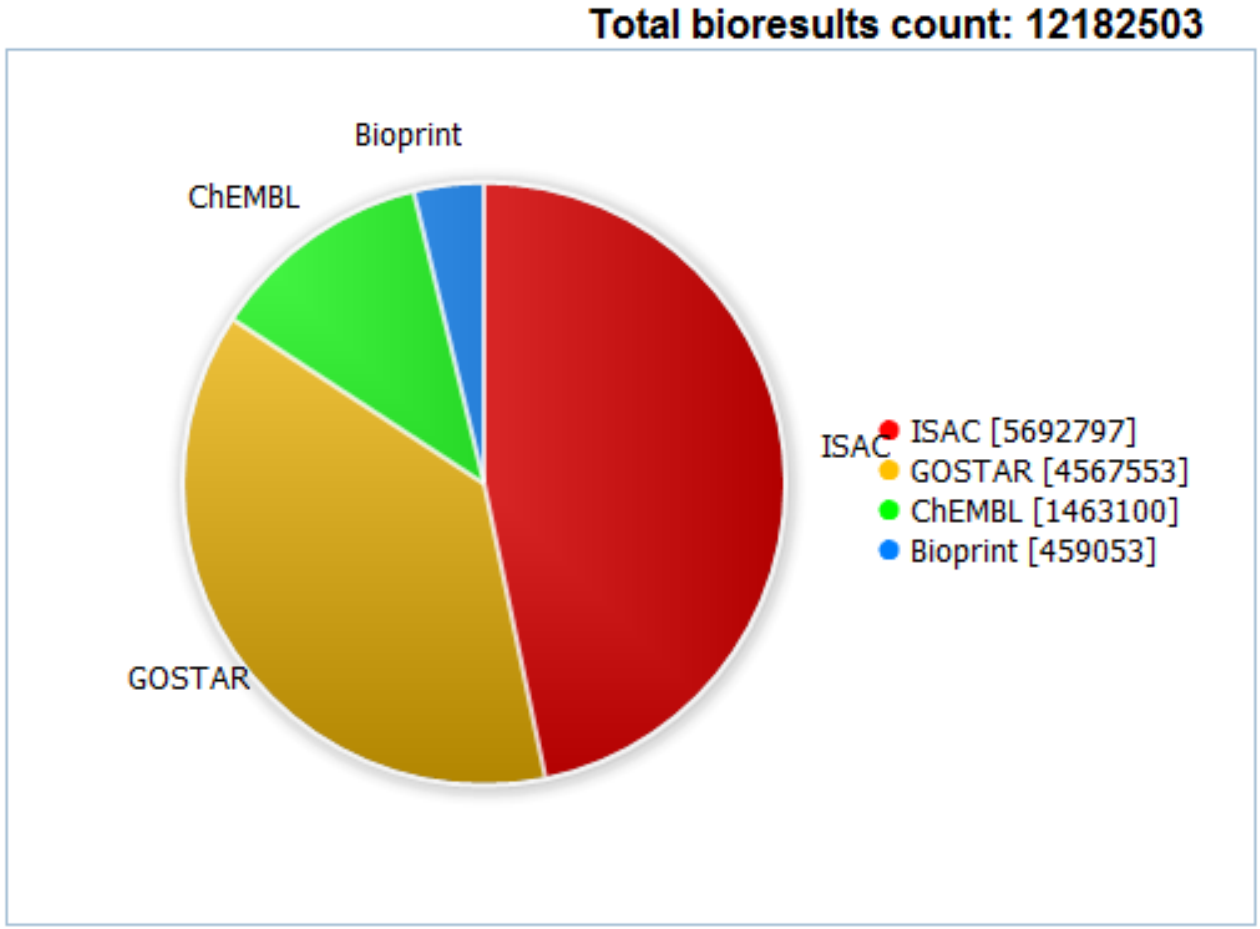
Chemistry Connect



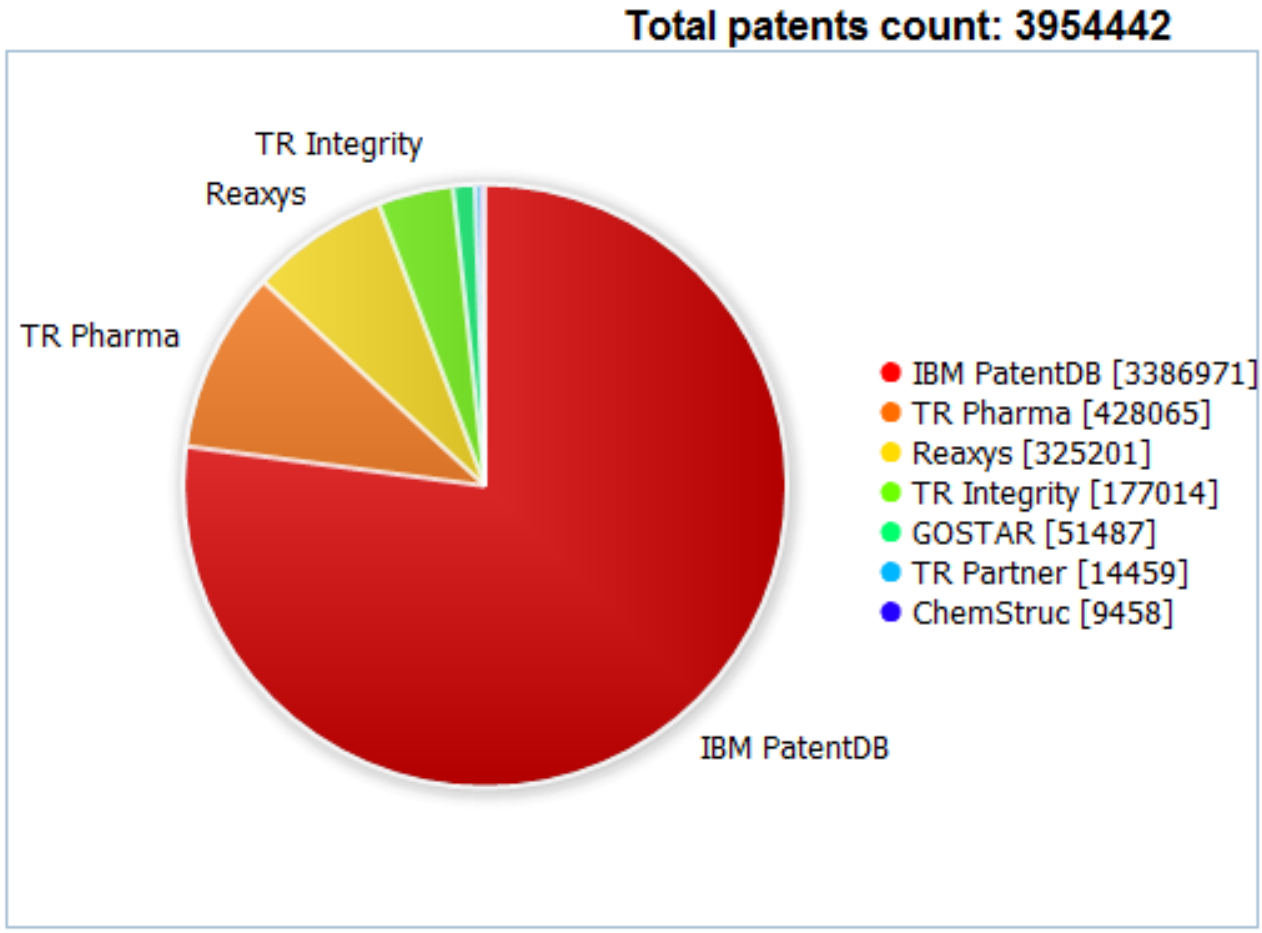
Chemistry Connect



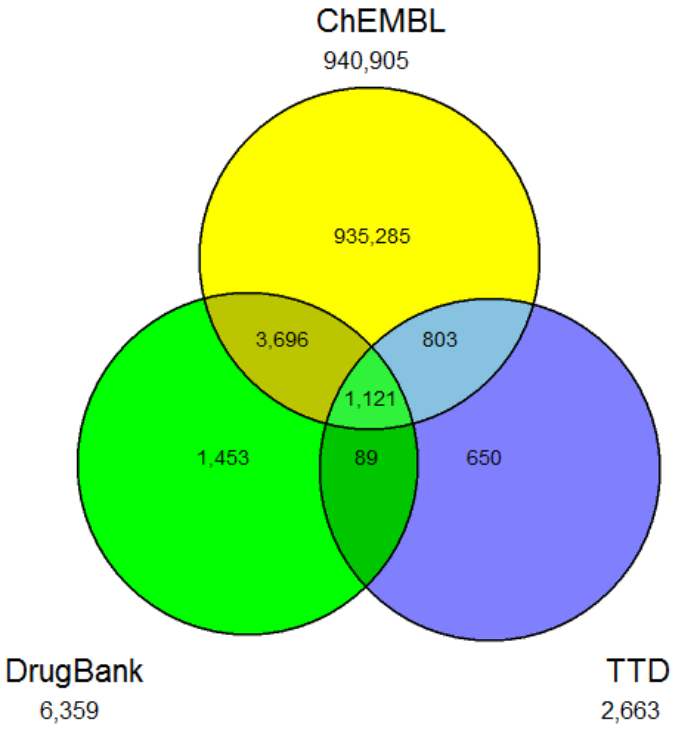
Chemistry Connect

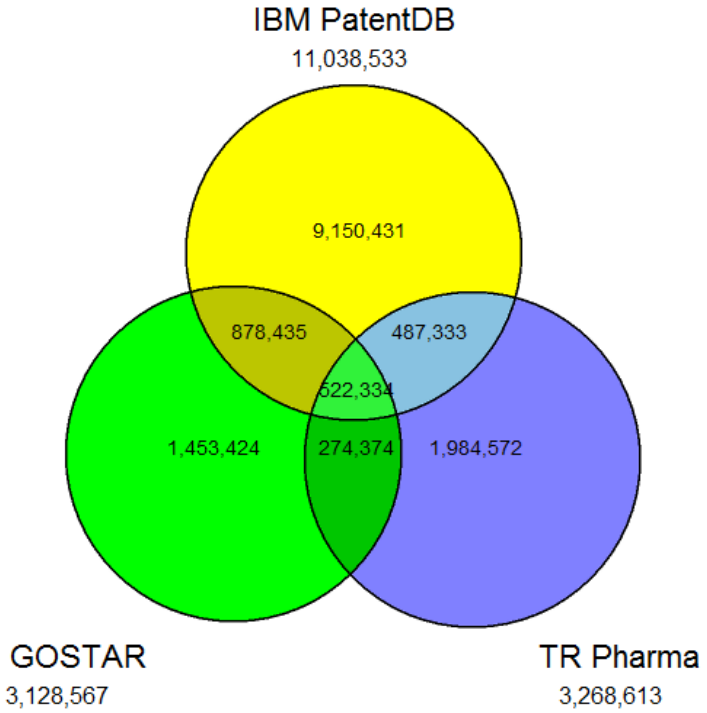


Chemistry Connect



Exact match source comparisons

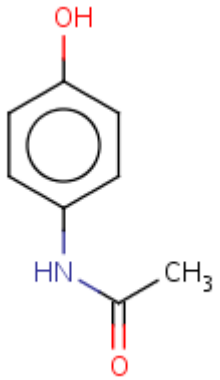




sources that include known drugs

predominantly patent-derived compounds

Finding a common language



Acetaminophen: >1000 synonyms..

[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 sulfate(30%)
acetaminophen sulphate
Acetaminophen Uniserts
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
Amadil
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
Apacet Capsules

Word of the Day : Crowdsourcing

Crowdsourcing in Chemistry Connect

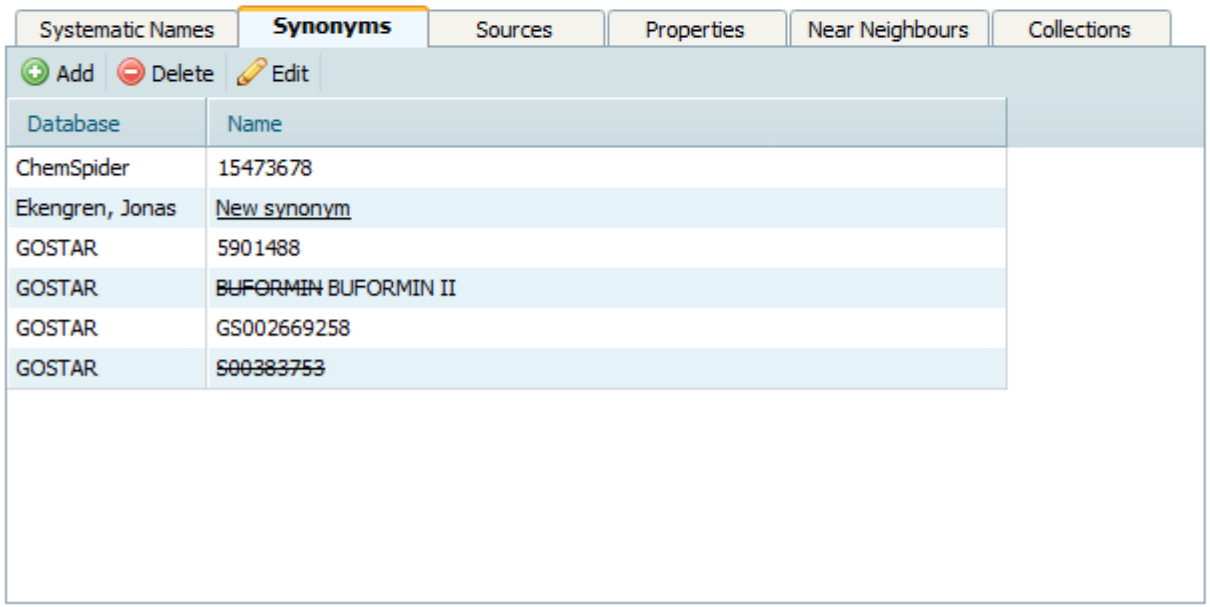
Crowdsourcing will allow users to easily add, remove or update the information in Chemistry Connect. As a first step we have introduced the possibility to curate synonyms but will later expand this to also include curation of structures.

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

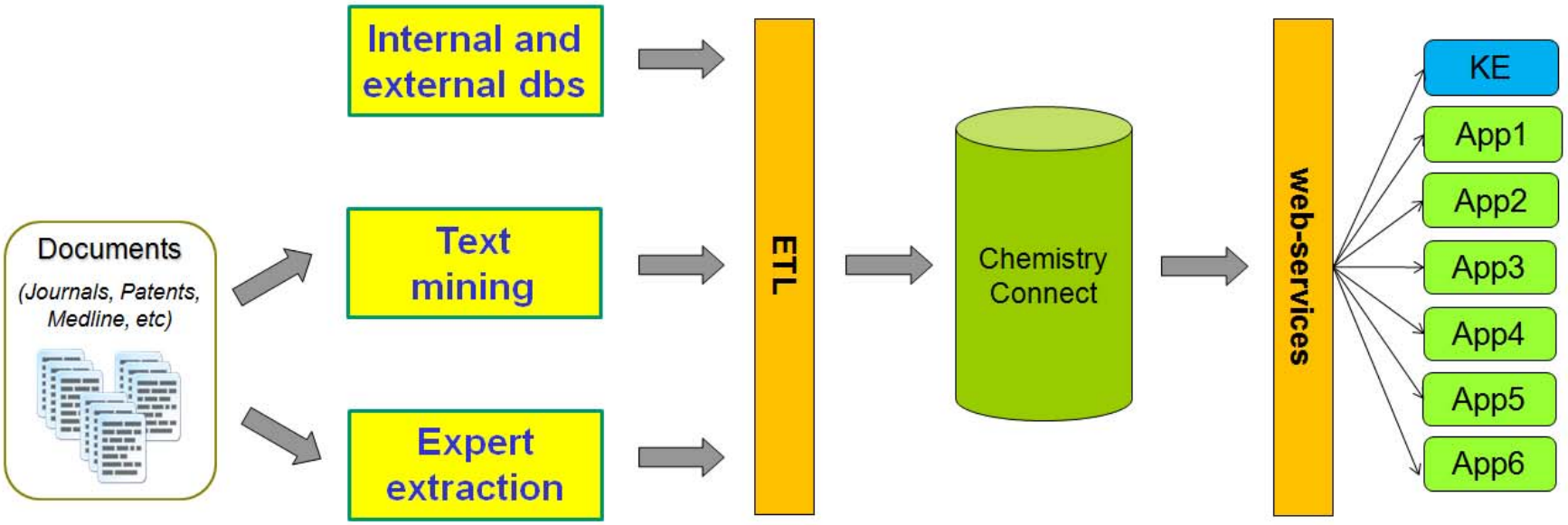
How to curate a synonym

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

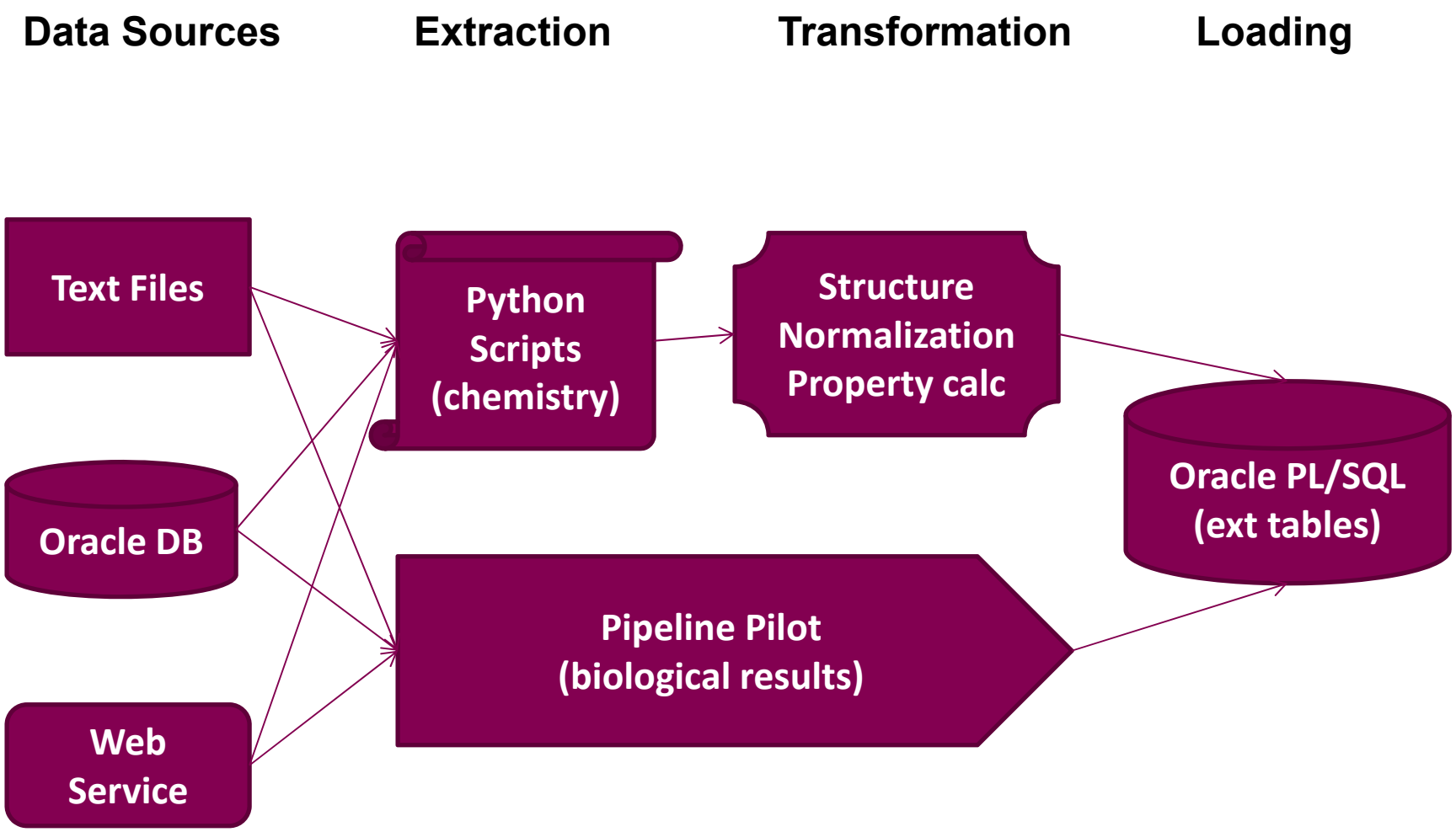
1. Add – will introduce a new synonym. New synonyms are underlined until they are moderated.
2. Delete – remove an incorrect synonym. The value will be shown with a strikethrough.
3. Edit – correct e.g. a misspelling. The old value will be shown with a strikethrough followed by the new value.



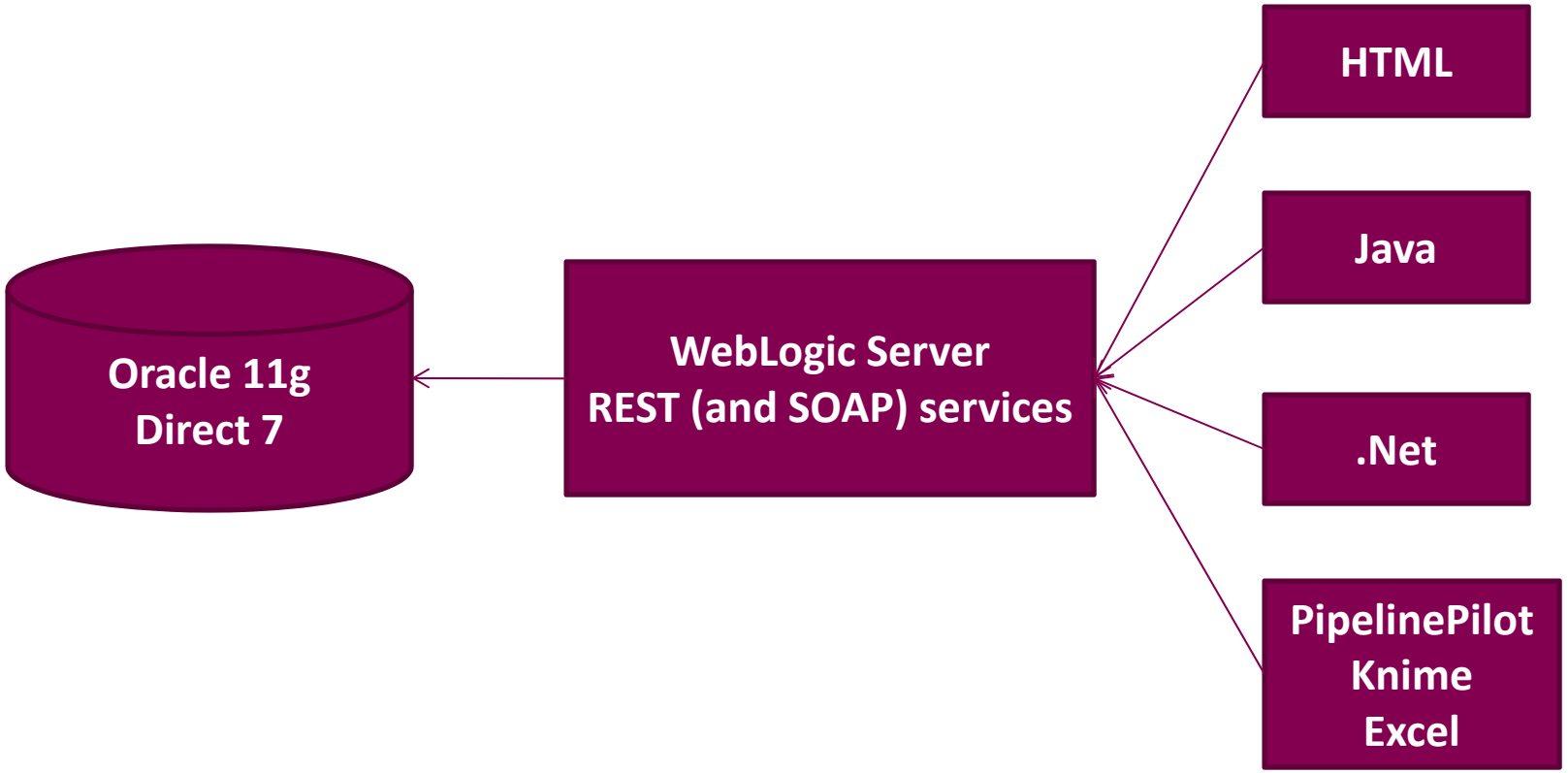
Chemistry Connect



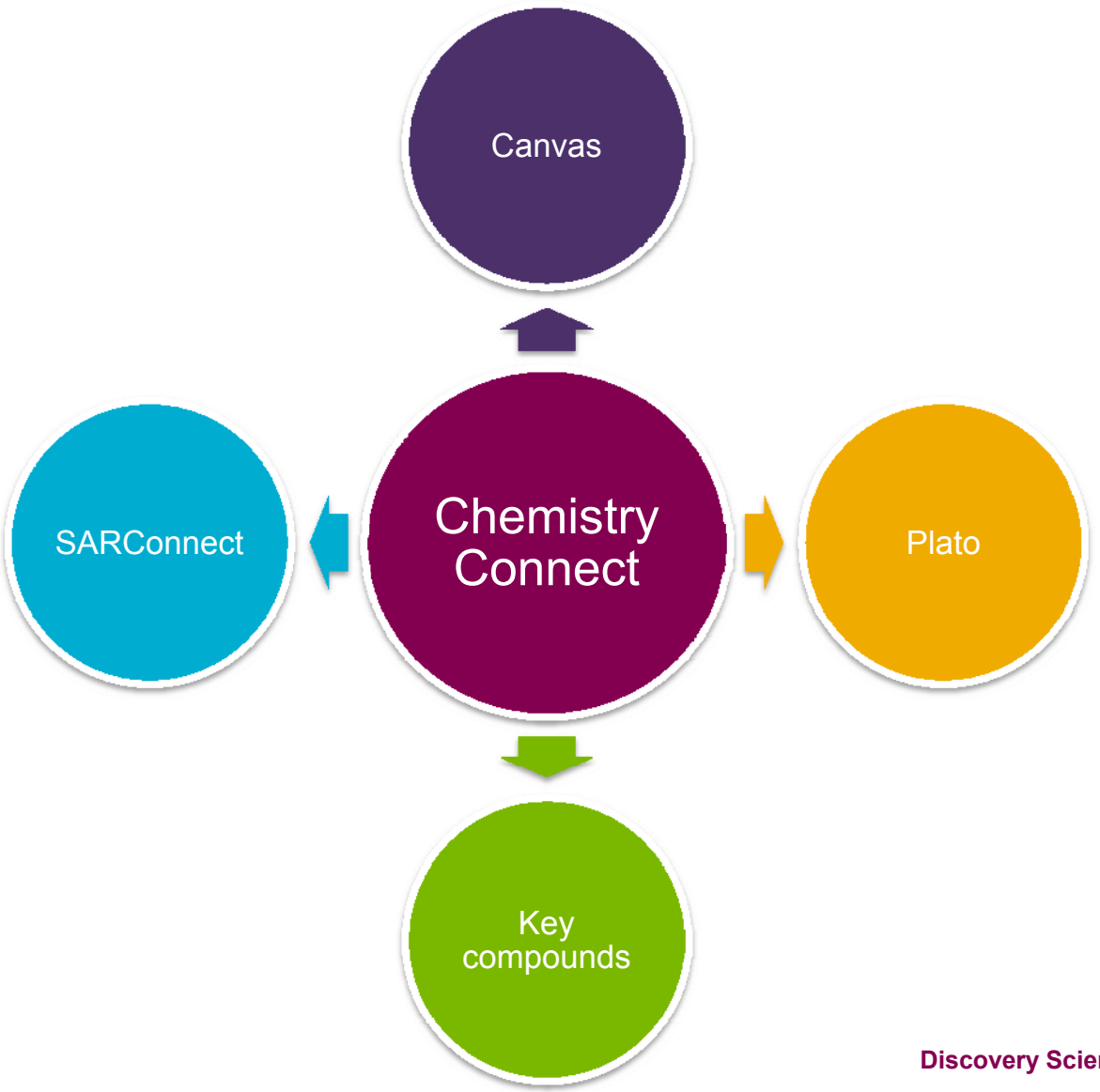
Technical Overview - ETL



Technical Overview - Application

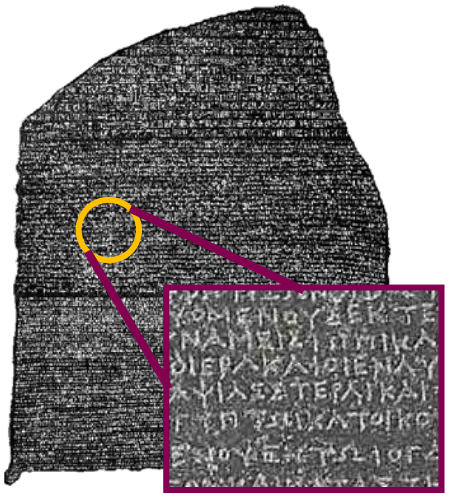


Chemistry Connect Apps

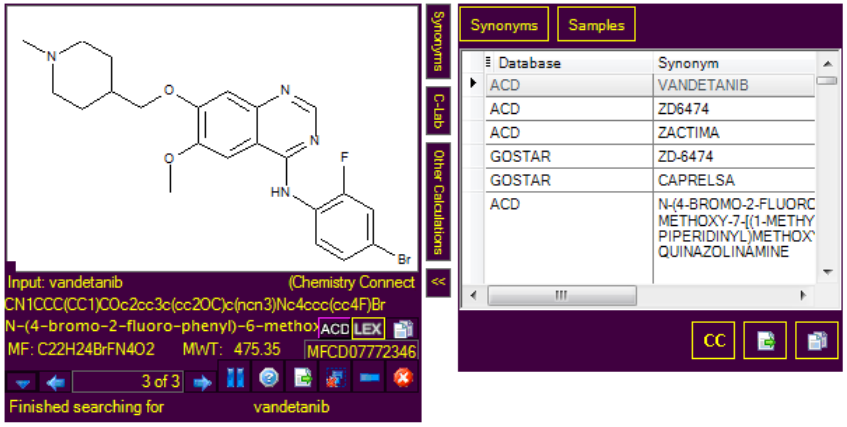


Canvas

is…



196 B.C.



2012 A.D.

…a Rosetta stone for compounds

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…

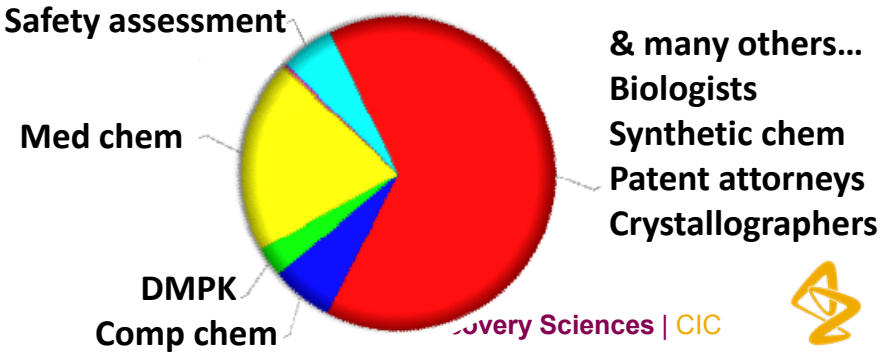
…a compound design tool

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

…really easy to use

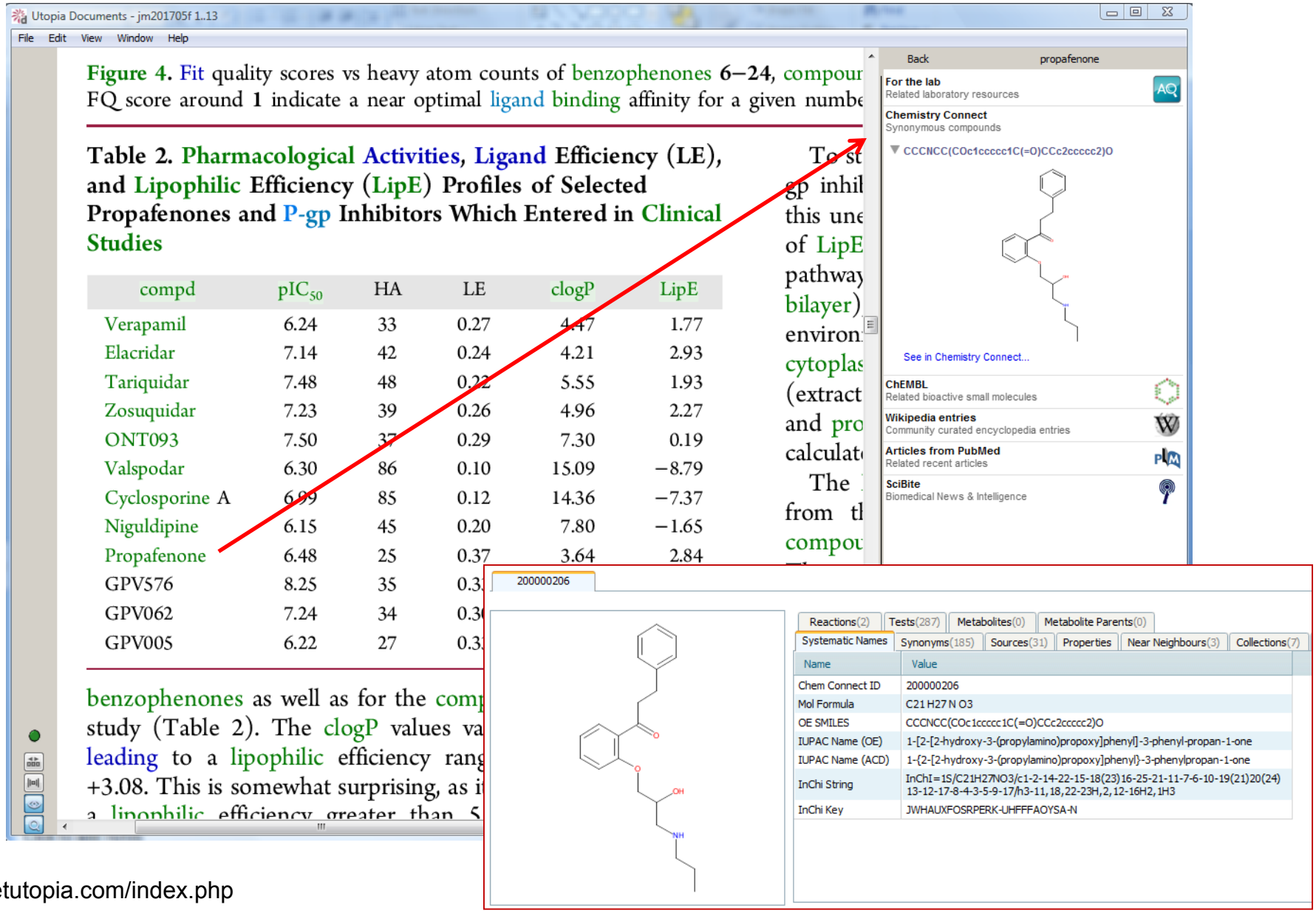
Copy a compound name or structure to the clipboard and let Canvas do the rest of the work

…and in 2011, 1750 AZ scientists did



Jon Winter, Oncology iMed

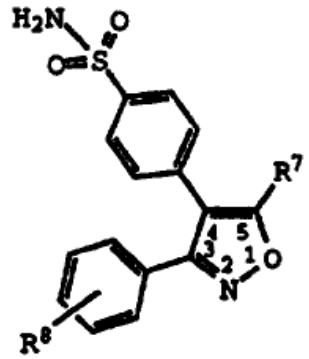
Utopia Documents

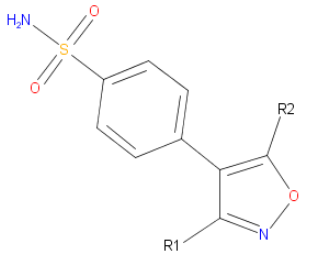


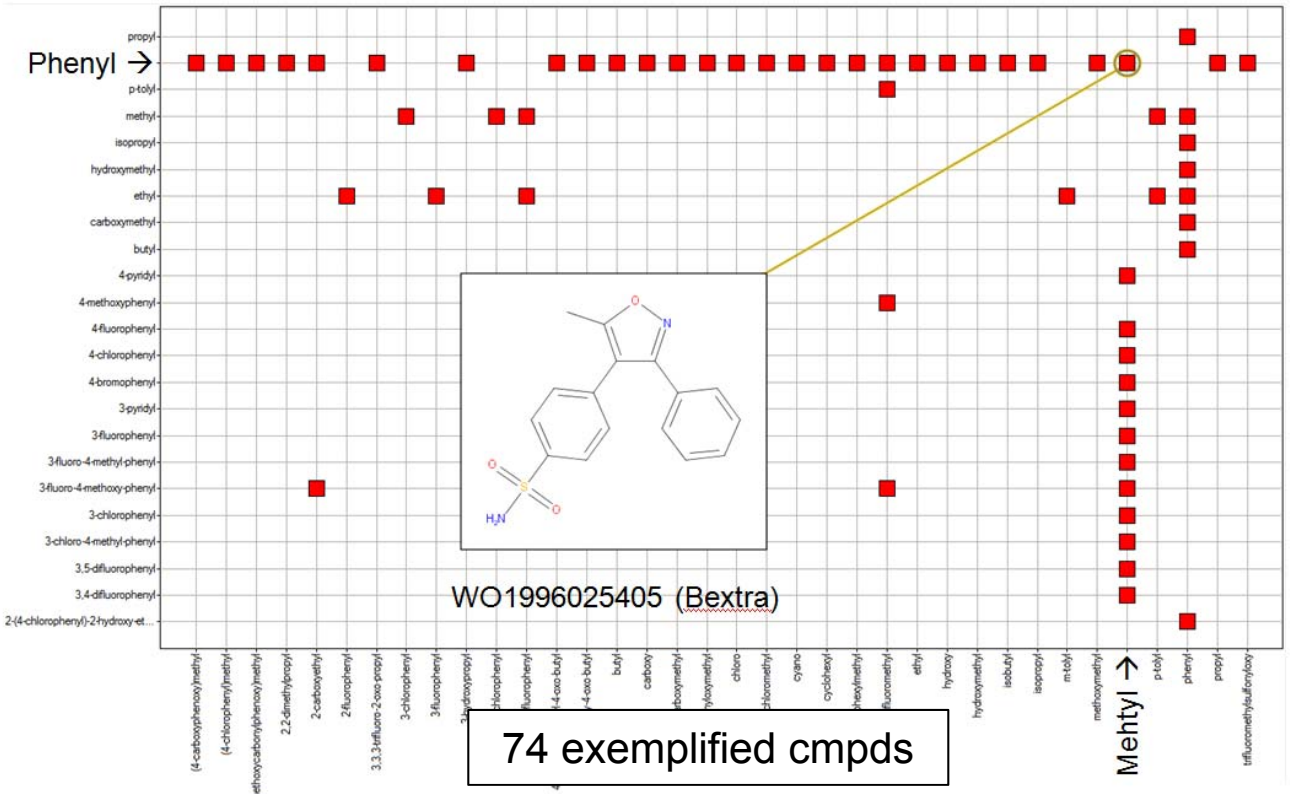
http://getutopia.com/index.php

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?

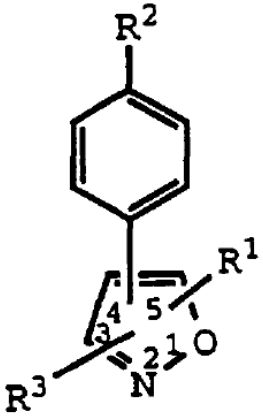


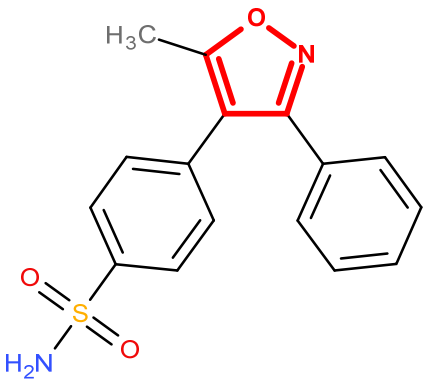


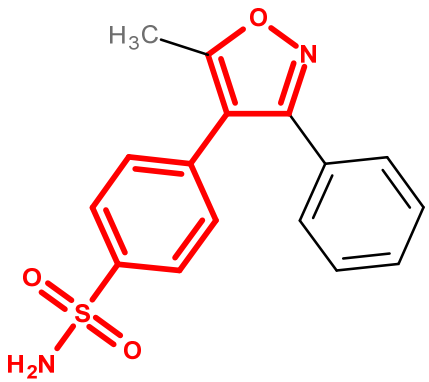


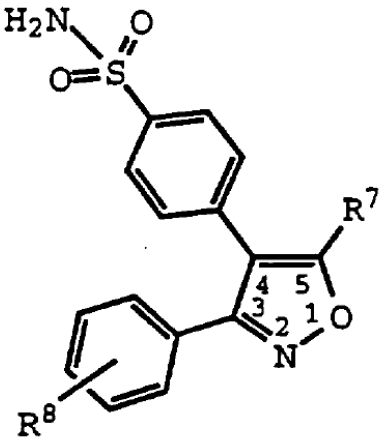
Tyrchan, C. et al *JCIM 2012*

WO1996025405 - Bextra



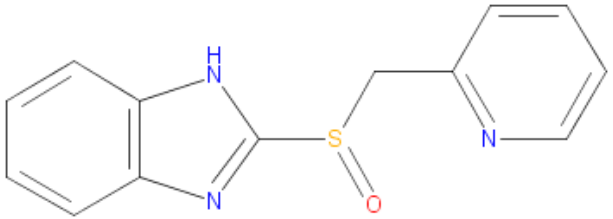


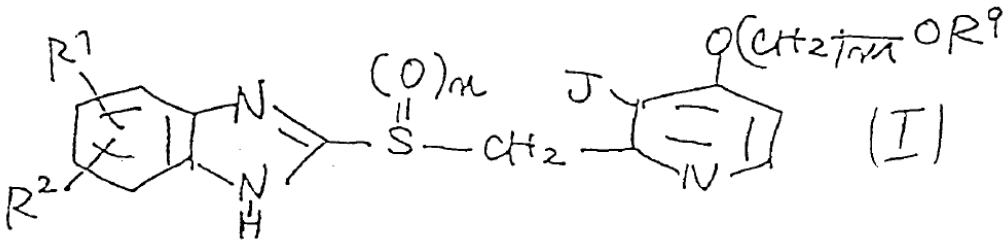


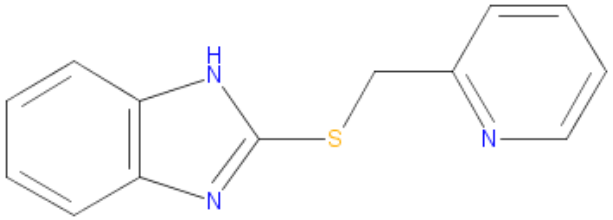


Source	#compounds	Bextra exists	Bextra ranked
GVKBIO	74	Y	1 (broad core) 1 (narrow core)
SureChem	501	Y	1 (broad core) 1 (narrow core)

EP268956 - Aciphex

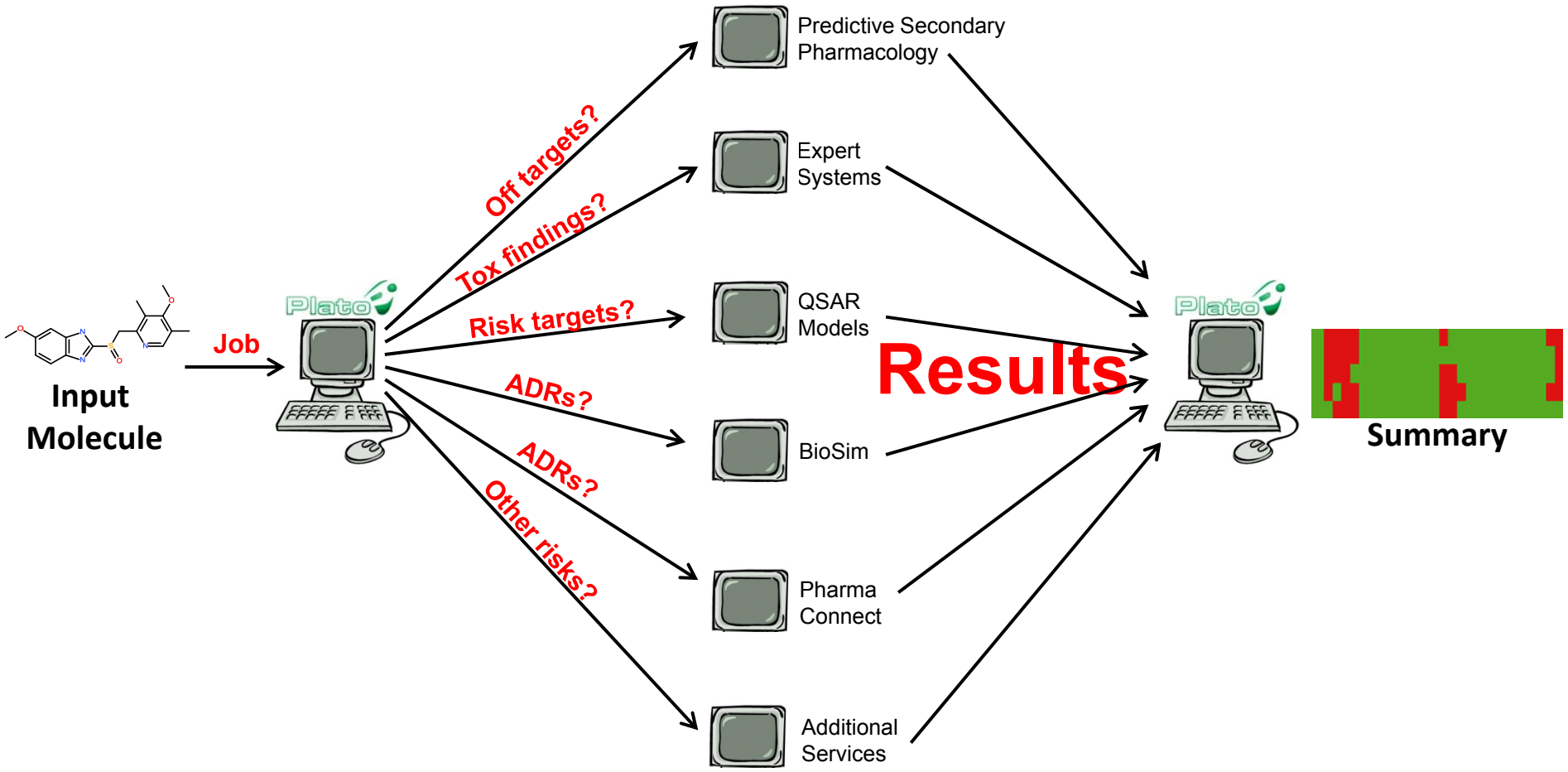






Source	#compounds	Aciphex exists	Aciphex ranked
GVKBIO	27	Y	2 (core1) 1 (core2)
SureChem	168	Y	1 (core1) 1 (core2)

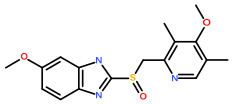
PLATO for Safety/Tox – General Concept



Scott Boyer
Catrin Hasselgren
Lars Carlsson
Tobias Noeske

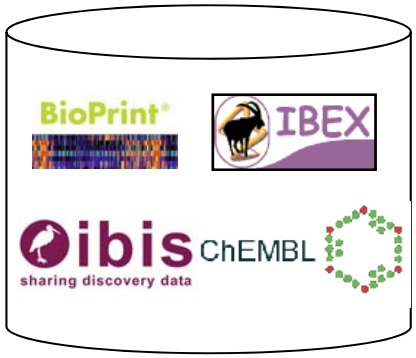
All services in Plato are complementary to find an overall answer to your problem!

Predictive Secondary Pharmacology strategy



Input Molecule

Similarity or substructure search



Chemistry Connect
Compound information: name & structure
Target information: bioactivity data

Compound - Target associations

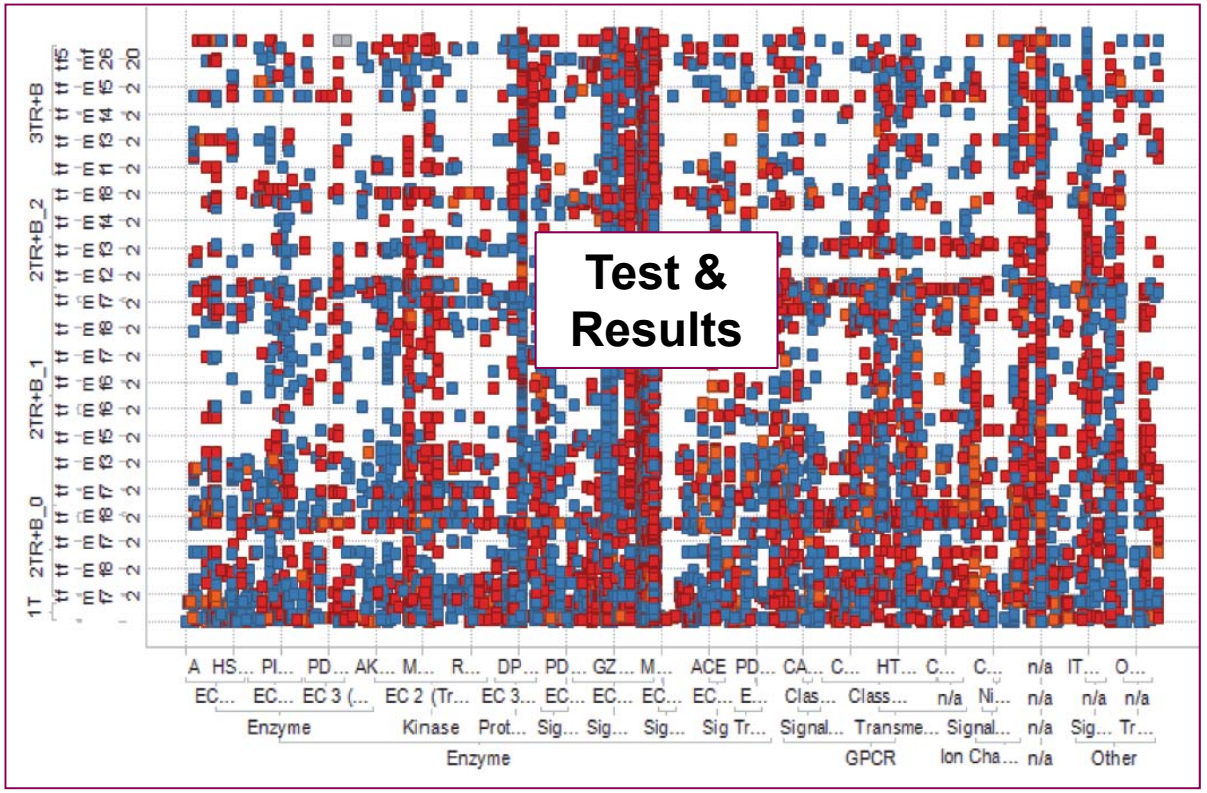
- Potential off-target effects?
- Part of a safety risk assessment

Scott Boyer
Catrin Hasselgren
Lars Carlsson
Tobias Noeske

Similarity concept: Similar compounds bind to similar targets.

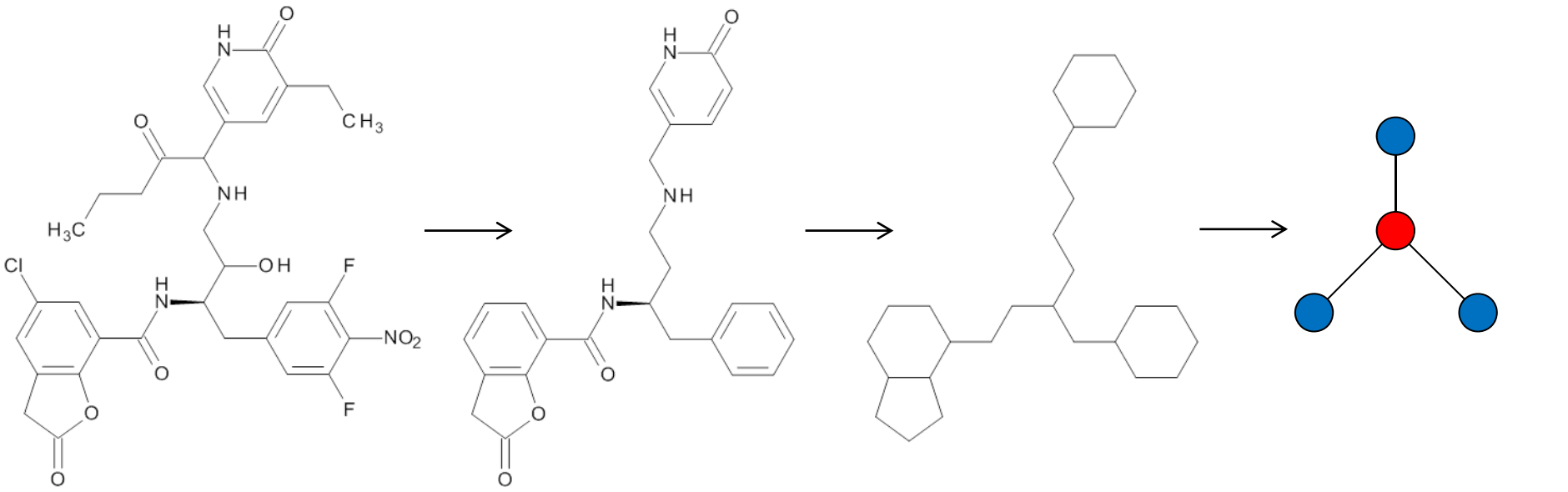
M Johnson et al., *Prog Clin Biol Res* (**1989**), 291:167
P Willet, *Drug Discov Today* (**2006**), 11:1046

SARConnect – navigate SAR landscape



Eriksson, M. et al *Molecular Informatics 2012*

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 Frizzeled Family

Symbol

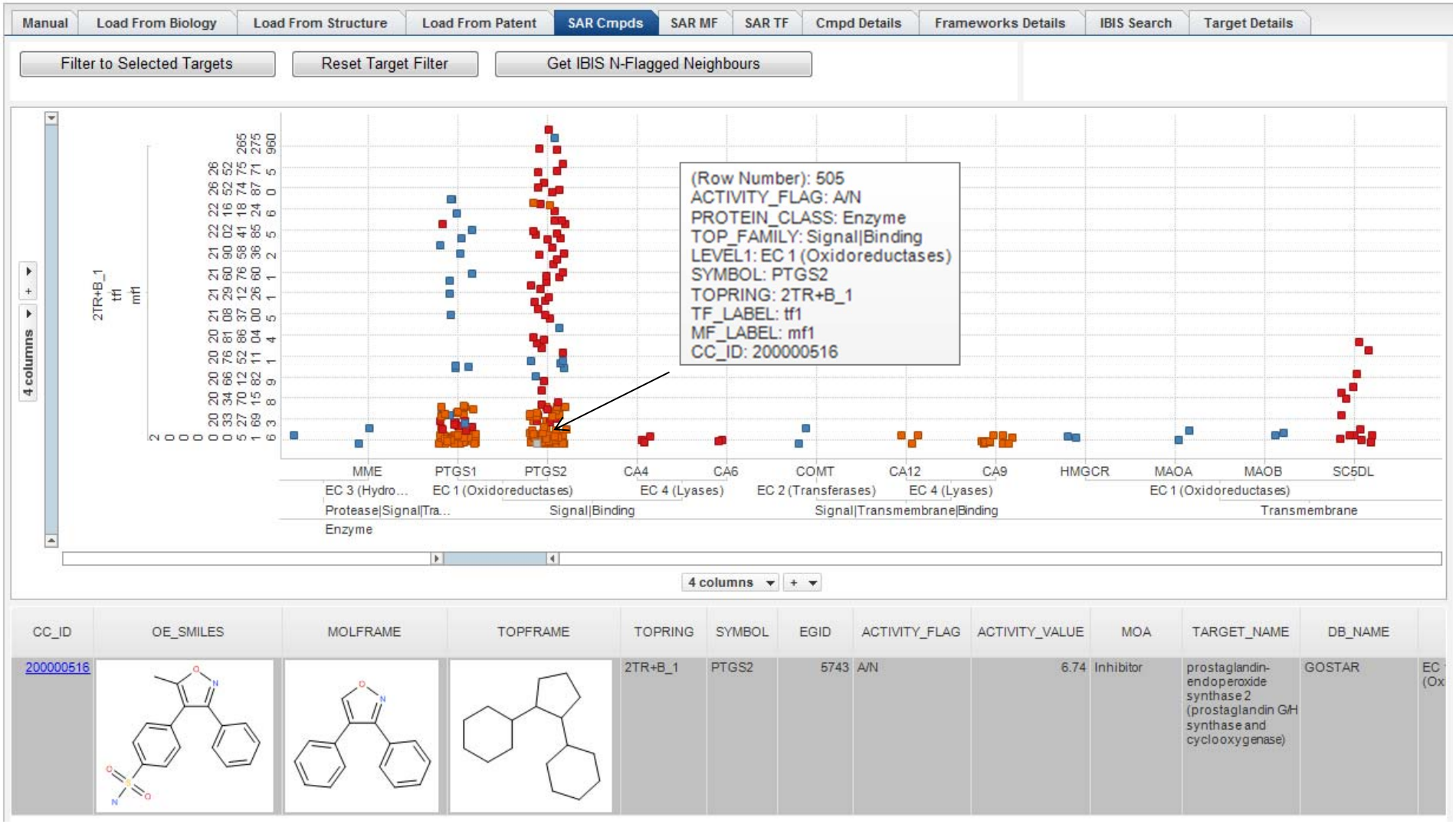
CALCR CALCLR CRHR2 GCCR GIPR GLP1R

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



CC_ID	OE_SMILES	MOLFRAME	TOPFRAME	TOPRING	SYMBOL	EGID	ACTIVITY_FLAG	ACTIVITY_VALUE	MOA	TARGET_NAME	DB_NAME	
200000516				2TR+B_1	PTGS2	5743	A/N	6.74	Inhibitor	prostaglandin-endoperoxide synthase 2 (prostaglandin G/H synthase and cyclooxygenase)	GOSTAR	EC (Ox

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!