The needs for chemistry standards, database tools and data curation at the chemical-biology interface

Antony Williams, Kamel Mansouri, Ann Richard and Chris Grulke*

This work was reviewed by EPA and approved for presentation but does not necessarily reflect official Agency policy.

January 25th, 2016

SLAS 2016, San Diego, CA

Speakers who “experience” data quality

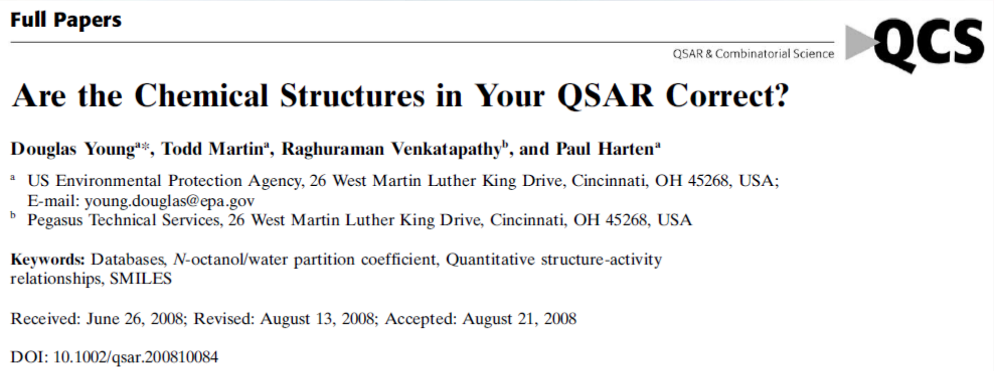
- Ensuring Chemical Structure, Biological Data and Computational Model Quality - Sean Ekins

- Bioassay Variability and Reliability in the Published and Patent Literature - John Overington

- Machine Learning to Optimize Experiments Why Have One Model When You Could Have Thousands? - Alex Clark

- And the rest of you experience it in so many ways…

Why does Quality Matter?



Full Papers

QSAR & Combinatorial Science

QCS

Are the Chemical Structures in Your QSAR Correct?

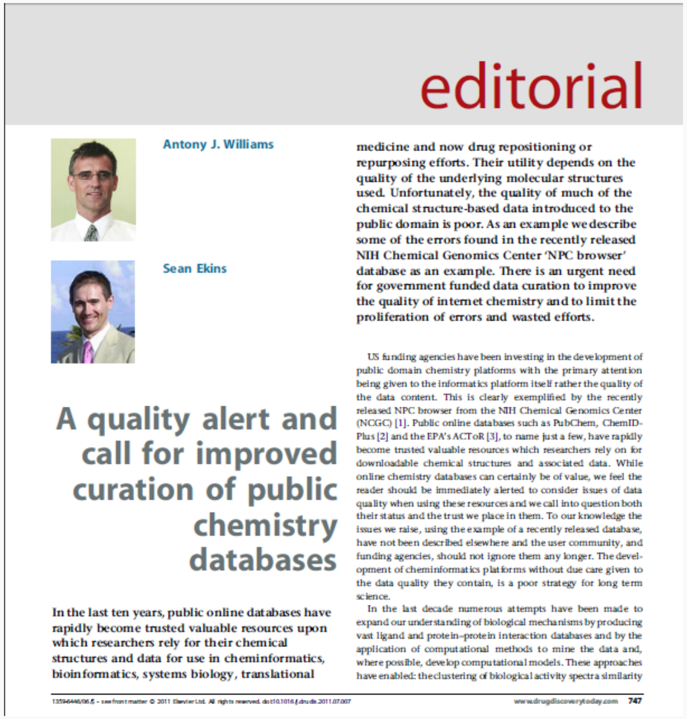
Douglas Young[a]*, Todd Martin[a], Raghuraman Venkatapathy[b], and Paul Harten[a]

[a] US Environmental Protection Agency, 26 West Martin Luther King Drive, Cincinnati, OH 45268, USA; E-mail: young.douglas@epa.gov
[b] Pegasus Technical Services, 26 West Martin Luther King Drive, Cincinnati, OH 45268, USA

Keywords: Databases, *N*-octanol/water partition coefficient, Quantitative structure-activity relationships, SMILES

Received: June 26, 2008; Revised: August 13, 2008; Accepted: August 21, 2008

DOI: 10.1002/qsar.200810084



editorial

Antony J. Williams

Sean Ekins

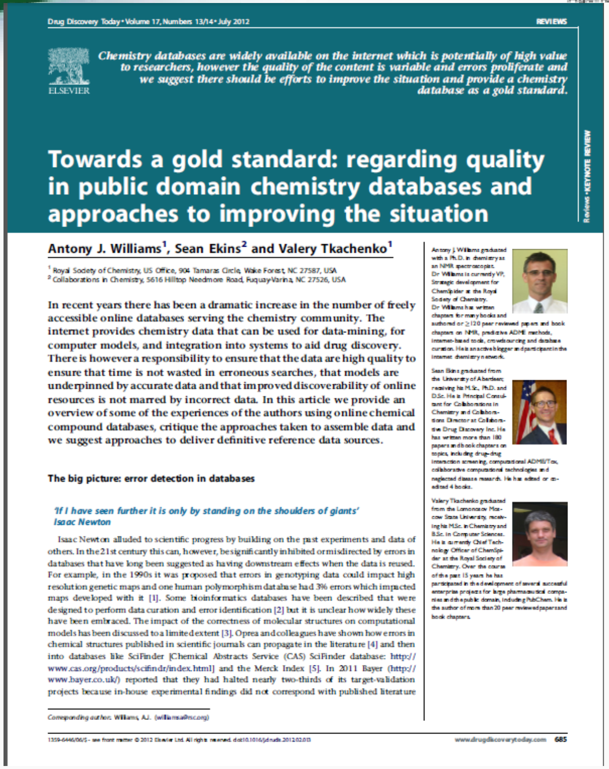
A quality alert and call for improved curation of public chemistry databases

In the last ten years, public online databases have rapidly become trusted valuable resources upon which researchers rely for their chemical structures and data for use in cheminformatics, bioinformatics, systems biology, translational medicine and now drug repositioning or repurposing efforts. Their utility depends on the quality of the underlying molecular structures used. Unfortunately, the quality of much of the chemical structure-based data introduced to the public domain is poor. As an example we describe some of the errors found in the recently released NIH Chemical Genomics Center 'NPC browser' database as an example. There is an urgent need for government funded data curation to improve the quality of internet chemistry and to limit the proliferation of errors and wasted efforts.

US funding agencies have been investing in the development of public domain chemistry platforms with the primary attention being given to the informatics platform itself rather the quality of the data content. This is clearly exemplified by the recently released NPC browser from the NIH Chemical Genomics Center (NCGC) [1]. Public online databases such as PubChem, ChemID-Plus [2] and the EPA's ACToR [3], to name just a few, have rapidly become trusted valuable resources which researchers rely on for downloadable chemical structures and associated data. While online chemistry databases can certainly be of value, we feel the reader should be immediately alerted to consider issues of data quality when using these resources and we call into question both their status and the trust we place in them. To our knowledge the issues we raise, using the example of a recently released database, have not been described elsewhere and the user community, and funding agencies, should not ignore them any longer. The development of cheminformatics platforms without due care given to the data quality they contain, is a poor strategy for long term science.

In the last decade numerous attempts have been made to expand our understanding of biological mechanisms by producing vast ligand and protein–protein interaction databases and by the application of computational methods to mine the data and, where possible, develop computational models. These approaches have enabled: the clustering of biological activity spectra similarity

1359-6446/06/$ - see front matter © 2011 Elsevier Ltd. All rights reserved. doi:10.1016/j.drudis.2011.07.007



ELSEVIER

Chemistry databases are widely available on the internet which is potentially of high value to researchers, however the quality of the content is variable and errors proliferate and we suggest there should be efforts to improve the situation and provide a chemistry database as a gold standard.

Reviews • KEYNOTE REVIEW

Towards a gold standard: regarding quality in public domain chemistry databases and approaches to improving the situation

Antony J. Williams[1], Sean Ekins[2] and Valery Tkachenko[1]

[1] Royal Society of Chemistry, US Office, 904 Tamaras Circle, Wake Forest, NC 27587, USA
[2] Collaborations in Chemistry, 5616 Hilltop Needmore Road, Fuquay-Varina, NC 27526, USA

In recent years there has been a dramatic increase in the number of freely accessible online databases serving the chemistry community. The internet provides chemistry data that can be used for data-mining, for computer models, and integration into systems to aid drug discovery. There is however a responsibility to ensure that the data are high quality to ensure that time is not wasted in erroneous searches, that models are underpinned by accurate data and that improved discoverability of online resources is not marred by incorrect data. In this article we provide an overview of some of the experiences of the authors using online chemical compound databases, critique the approaches taken to assemble data and we suggest approaches to deliver definitive reference data sources.

The big picture: error detection in databases

> *'If I have seen further it is only by standing on the shoulders of giants'*
> *Isaac Newton*

Isaac Newton alluded to scientific progress by building on the past experiments and data of others. In the 21st century this can, however, be significantly inhibited or misdirected by errors in databases that have long been suggested as having downstream effects when the data is reused. For example, in the 1990s it was proposed that errors in genotyping data could impact high resolution genetic maps and one human polymorphism database had 3% errors which impacted maps developed with it [1]. Some bioinformatics databases have been described that were designed to perform data curation and error identification [2] but it is unclear how widely these have been embraced. The impact of the correctness of molecular structures on computational models has been discussed to a limited extent [3]. Oprea and colleagues have shown how errors in chemical structures published in scientific journals can propagate in the literature [4] and then into databases like SciFinder [Chemical Abstracts Service (CAS) SciFinder database: http://www.cas.org/products/scifindr/index.html] and the Merck Index [5]. In 2011 Bayer (http://www.bayer.co.uk/) reported that they had halted nearly two-thirds of its target-validation projects because in-house experimental findings did not correspond with published literature

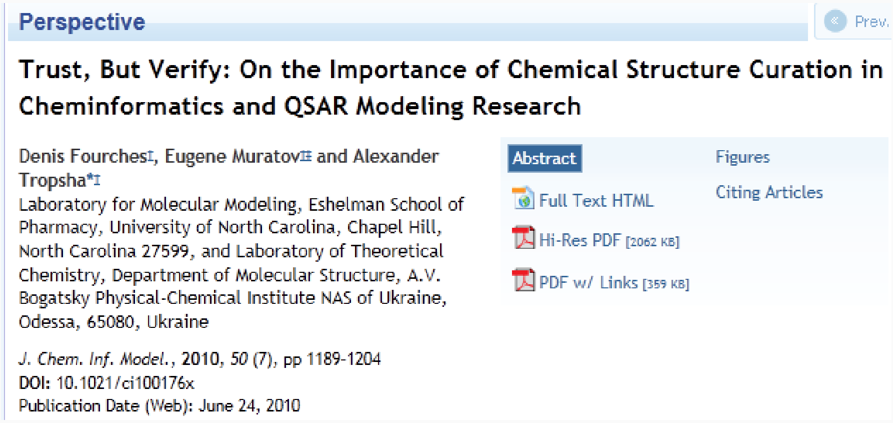
Antony J. Williams graduated with a Ph.D. in chemistry as an NMR spectroscopist. Dr Williams is currently VP, Strategic development for ChemSpider at the Royal Society of Chemistry. Dr Williams has written chapters for many books and authored or ≥120 peer reviewed papers and book chapters on NMR, predictive ADME methods, internet-based tools, crowdsourcing and database curation. He is an active blogger and participant in the internet chemistry network.

Sean Ekins graduated from the University of Aberdeen; receiving his M.Sc., Ph.D. and D.Sc. He is Principal Consultant for Collaborations in Chemistry and Collaborations Director at Collaborative Drug Discovery Inc. He has written more than 180 papers and book chapters on topics, including drug–drug interaction screening, computational ADME/Tox, collaborative computational technologies and neglected disease research. He has edited or co-edited 4 books.

Valery Tkachenko graduated from the Lomonosov Moscow State University, receiving his M.Sc. in Chemistry and B.Sc. in Computer Sciences. He is currently Chief Technology Officer of ChemSpider at the Royal Society of Chemistry. Over the course of the past 15 years he has participated in the development of several successful enterprise projects for large pharmaceutical companies and the public domain, including PubChem. He is the author of more than 20 peer reviewed papers and book chapters.

Corresponding author: Williams, A.J. (williamsa@rsc.org)

1359-6446/06/$ - see front matter © 2012 Elsevier Ltd. All rights reserved. doi:10.1016/j.drudis.2012.02.013



Perspective

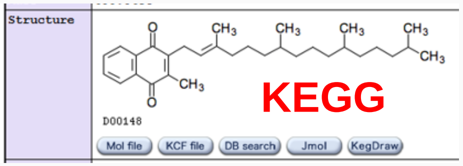
« Prev.

Trust, But Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research

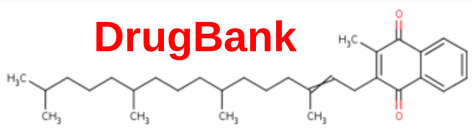
Denis Fourches[†], Eugene Muratov[‡] and Alexander Tropsha[*†]
Laboratory for Molecular Modeling, Eshelman School of Pharmacy, University of North Carolina, Chapel Hill, North Carolina 27599, and Laboratory of Theoretical Chemistry, Department of Molecular Structure, A.V. Bogatsky Physical-Chemical Institute NAS of Ukraine, Odessa, 65080, Ukraine

J. Chem. Inf. Model., 2010, *50* (7), pp 1189-1204
DOI: 10.1021/ci100176x
Publication Date (Web): June 24, 2010

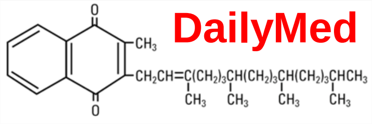
What is the Correct Structure of Vitamin K?



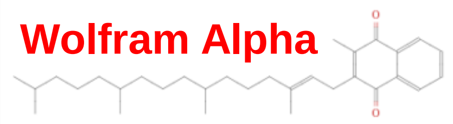
KEGG



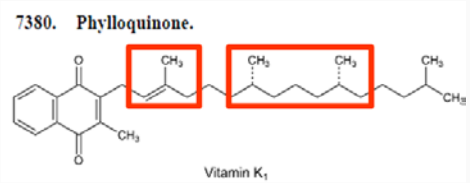
DrugBank



DailyMed



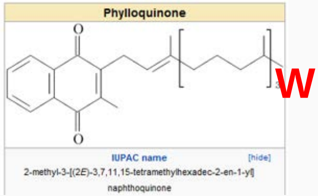
Wolfram Alpha



7380. Phylloquinone.

Vitamin K_1

Merck Index

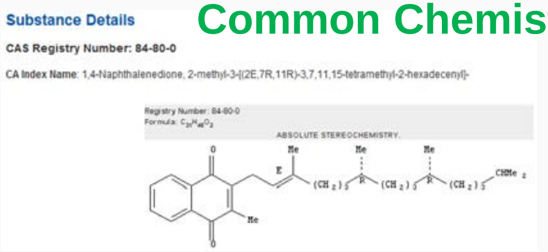


Phylloquinone

IUPAC name [hide]

2-methyl-3-[(2*E*)-3,7,11,15-tetramethylhexadec-2-en-1-yl]naphthoquinone

Wikipedia



Substance Details

CAS Registry Number: 84-80-0

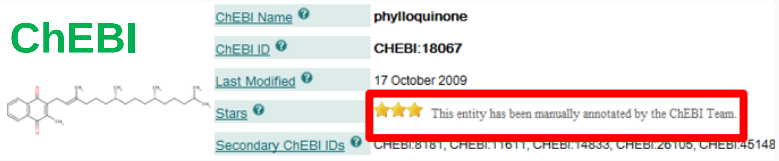
CA Index Name: 1,4-Naphthalenedione, 2-methyl-3-[(2E,7R,11R)-3,7,11,15-tetramethyl-2-hexadecenyl]-

Registry Number: 84-80-0
Formula: $C_{31}H_{46}O_2$

ABSOLUTE STEREOCHEMISTRY.

Common Chemistry

ChEBI



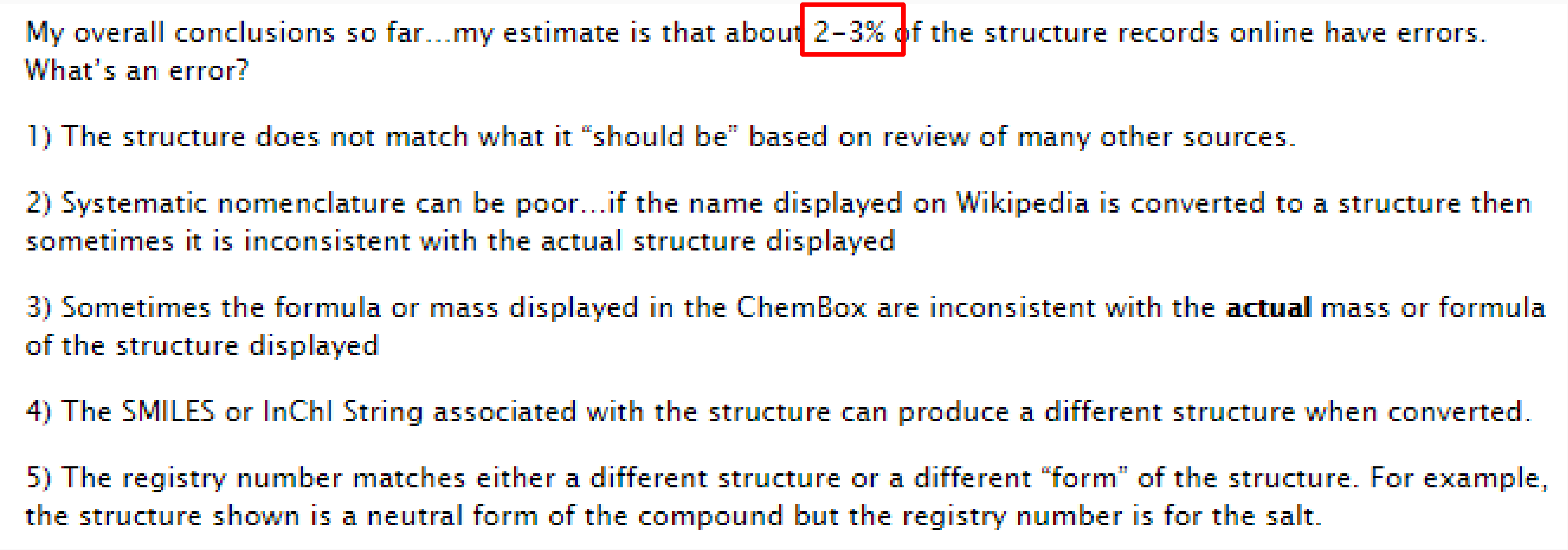
ChEBI Name	phylloquinone
ChEBI ID	CHEBI:18067
Last Modified	17 October 2009
Stars	★★★ This entity has been manually annotated by the ChEBI Team.
Secondary ChEBI IDs	CHEBI:8181, CHEBI:11611, CHEBI:14833, CHEBI:26106, CHEBI:45148

Experiences over the years

« Welcome to the ChemConnector Blog and the Art of Chemunication

Sign up to Receive ChemConnector Via Feedburner »

Dedicating Christmas Time to the Cause of Curating Wikipedia



My overall conclusions so far…my estimate is that about 2-3% of the structure records online have errors. What's an error?

1) The structure does not match what it "should be" based on review of many other sources.

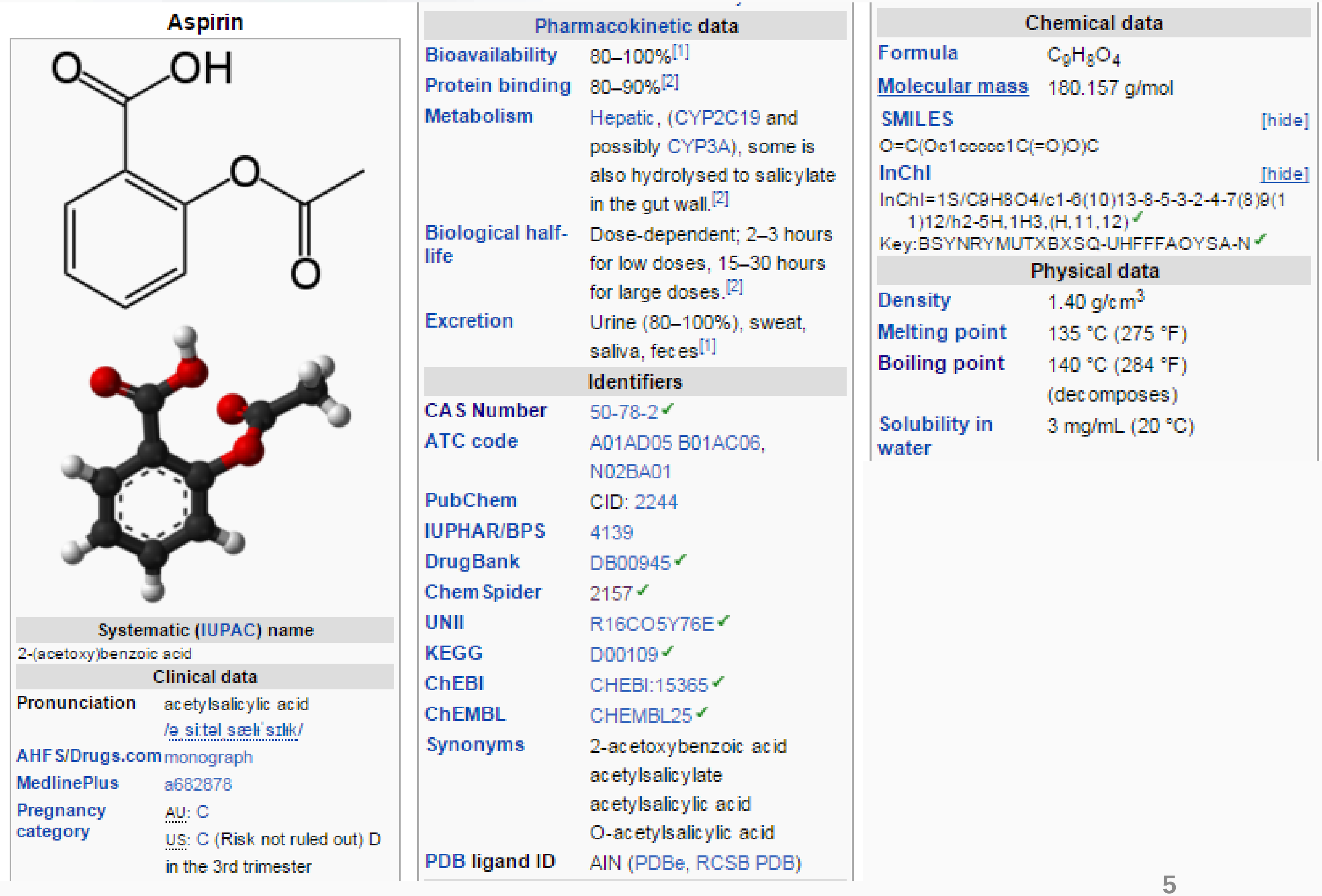
2) Systematic nomenclature can be poor…if the name displayed on Wikipedia is converted to a structure then sometimes it is inconsistent with the actual structure displayed

3) Sometimes the formula or mass displayed in the ChemBox are inconsistent with the **actual** mass or formula of the structure displayed

4) The SMILES or InChI String associated with the structure can produce a different structure when converted.

5) The registry number matches either a different structure or a different "form" of the structure. For example, the structure shown is a neutral form of the compound but the registry number is for the salt.

Experiences over the years



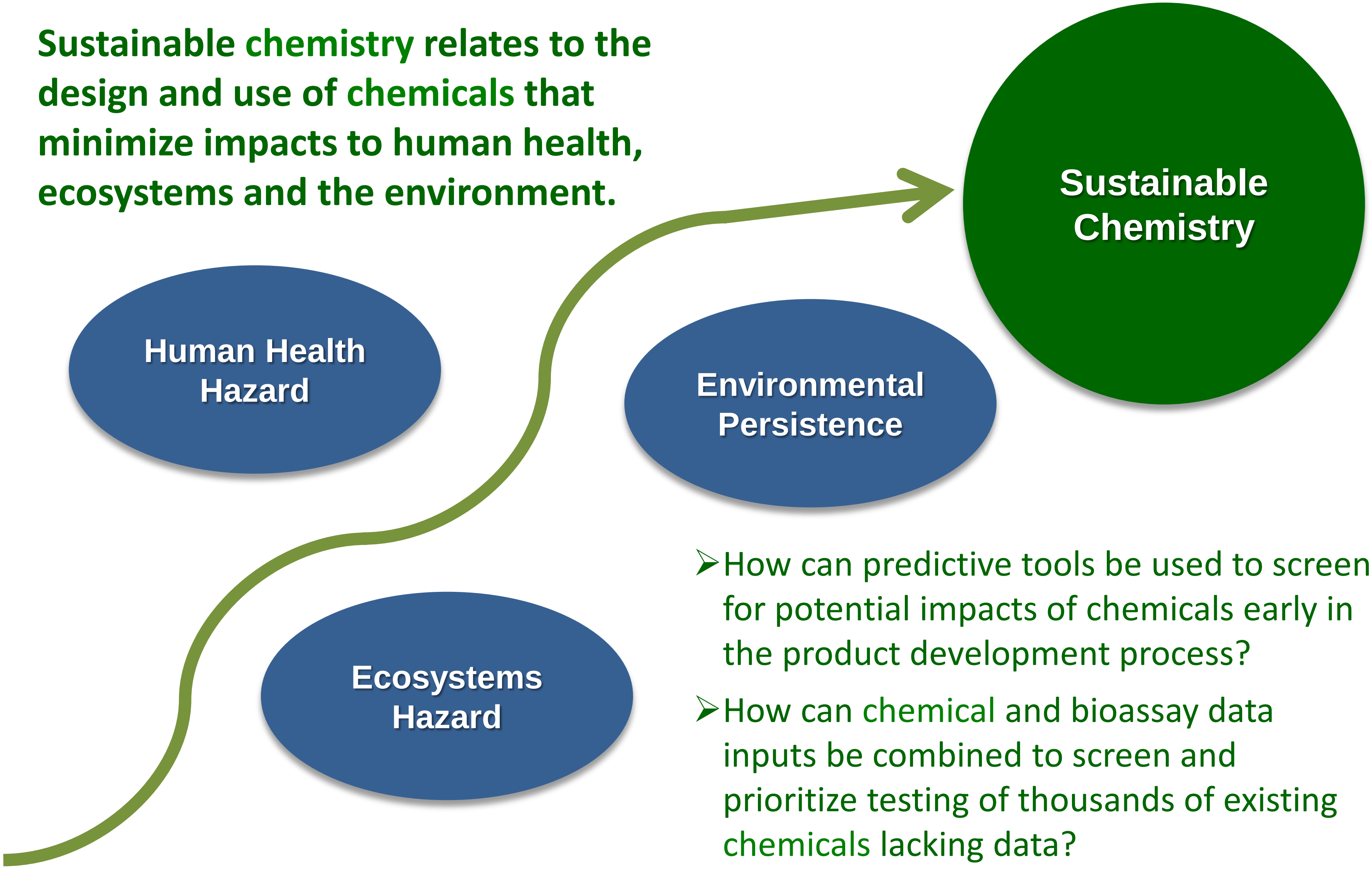
Aspirin

Systematic (IUPAC) name	
2-(acetoxy)benzoic acid	
Clinical data	
Pronunciation	acetylsalicylic acid /ə siːtəl sælɪˈsɪlɪk/
AHFS/Drugs.com	monograph
MedlinePlus	a682878
Pregnancy category	AU: C US: C (Risk not ruled out) D in the 3rd trimester

Pharmacokinetic data	
Bioavailability	80–100%[1]
Protein binding	80–90%[2]
Metabolism	Hepatic, (CYP2C19 and possibly CYP3A), some is also hydrolysed to salicylate in the gut wall.[2]
Biological half-life	Dose-dependent; 2–3 hours for low doses, 15–30 hours for large doses.[2]
Excretion	Urine (80–100%), sweat, saliva, feces[1]
Identifiers	
CAS Number	50-78-2 ✓
ATC code	A01AD05 B01AC06, N02BA01
PubChem	CID: 2244
IUPHAR/BPS	4139
DrugBank	DB00945 ✓
ChemSpider	2157 ✓
UNII	R16CO5Y76E ✓
KEGG	D00109 ✓
ChEBI	CHEBI:15365 ✓
ChEMBL	CHEMBL25 ✓
Synonyms	2-acetoxybenzoic acid acetylsalicylate acetylsalicylic acid O-acetylsalicylic acid
PDB ligand ID	AIN (PDBe, RCSB PDB)

Chemical data	
Formula	$C_9H_8O_4$
Molecular mass	180.157 g/mol
SMILES	[hide]
O=C(Oc1ccccc1C(=O)O)C	
InChI	[hide]
InChI=1S/C9H8O4/c1-6(10)13-8-5-3-2-4-7(8)9(11)12/h2-5H,1H3,(H,11,12) ✓ Key:BSYNRYMUTXBXSQ-UHFFFAOYSA-N ✓	
Physical data	
Density	1.40 g/cm^3
Melting point	135 °C (275 °F)
Boiling point	140 °C (284 °F) (decomposes)
Solubility in water	3 mg/mL (20 °C)

Sustainable chemistry relates to the design and use of chemicals that minimize impacts to human health, ecosystems and the environment.

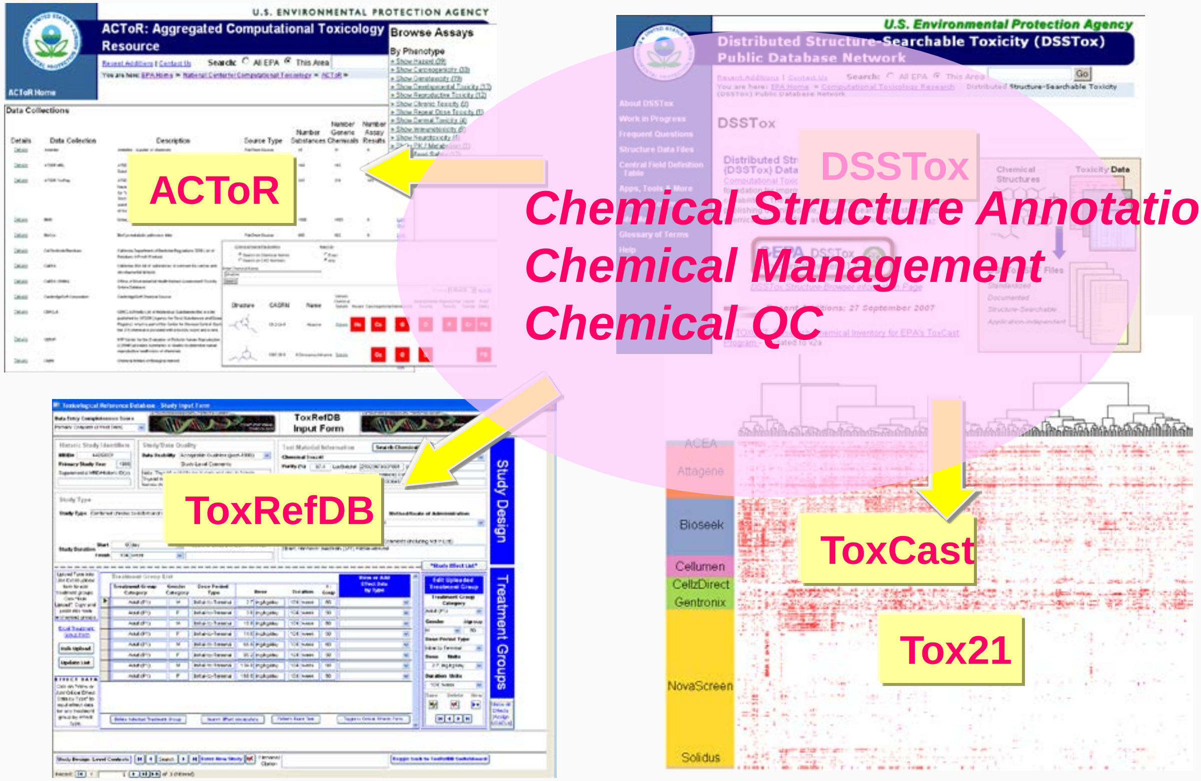


- How can predictive tools be used to screen for potential impacts of chemicals early in the product development process?
- How can chemical and bioassay data inputs be combined to screen and prioritize testing of thousands of existing chemicals lacking data?
- How do chemical transformations impact the hazard potential and environmental persistence?

NCCT Publicly Available Data

http://www.epa.gov/ncct

EPA
United States Environmental Protection Agency



ACToR

DSSTox

Chemical Structure Annotation

Chemical Management

Chemical QC

ToxRefDB

ToxCast

Tox21

U.S. ENVIRONMENTAL PROTECTION AGENCY
ACToR: Aggregated Computational Toxicology Resource
Browse Assays
By Phenotype
ACToR Home
Data Collections
U.S. Environmental Protection Agency
Distributed Structure-Searchable Toxicity (DSSTox) Public Database Network
DSSTox
ToxRefDB Input Form
Study Design
Treatment Groups
ACEA
Attagene
Bioseek
Cellumen
CellzDirect
Gentronix
NovaScreen
Solidus

Chemistry as a Data Foundation

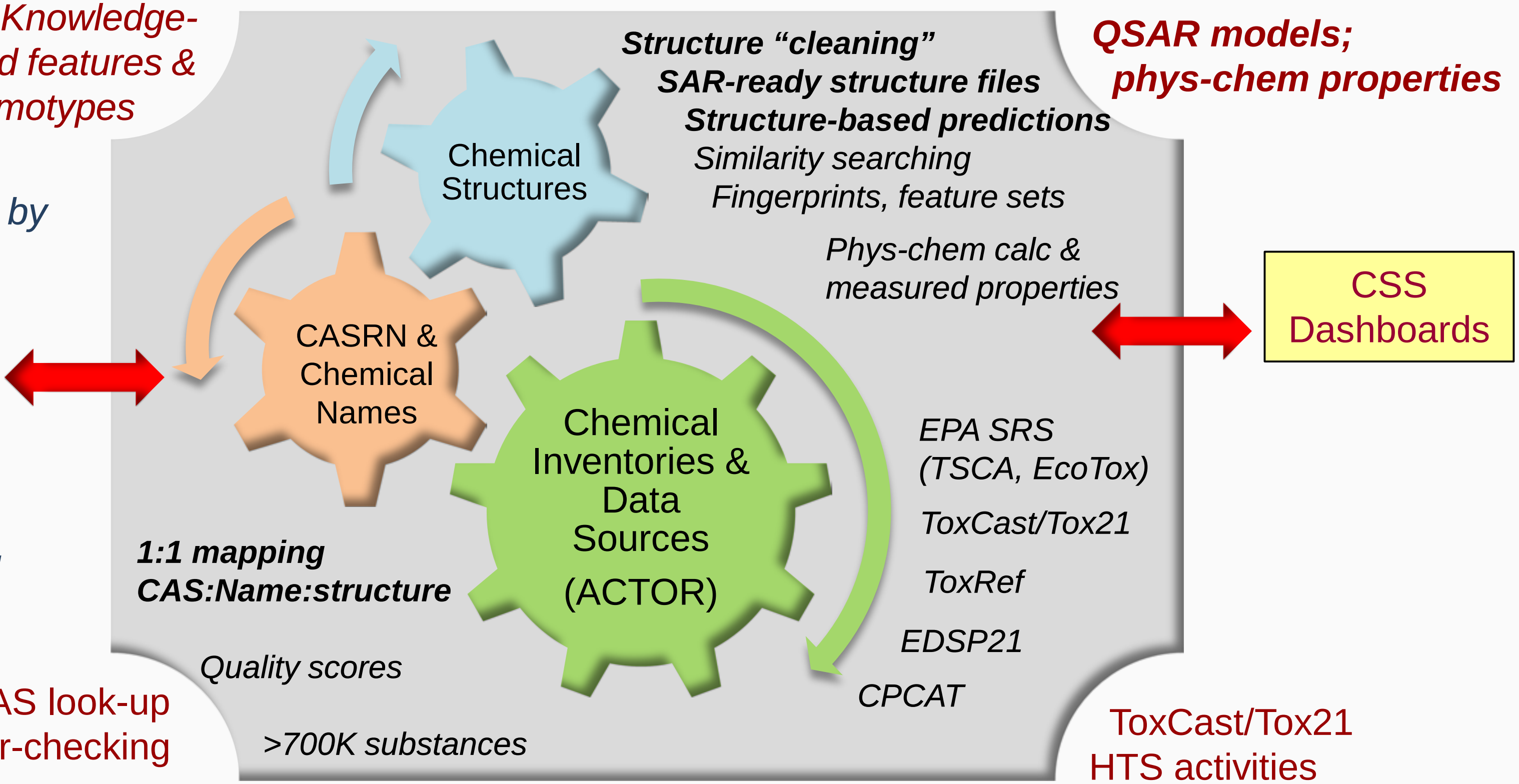
MOA & Knowledge-informed features & chemotypes

Chemical search by
CAS-names
SMILES
Structures

Structure file downloads

PK/ADME model inputs & outputs

CAS look-up
New list error-checking



Chemical Structures

CASRN & Chemical Names

Chemical Inventories & Data Sources

(ACTOR)

Structure "cleaning"
SAR-ready structure files
Structure-based predictions
Similarity searching
Fingerprints, feature sets

Phys-chem calc &
measured properties

EPA SRS
(TSCA, EcoTox)
ToxCast/Tox21
ToxRef
EDSP21
CPCAT

1:1 mapping
CAS:Name:structure

Quality scores

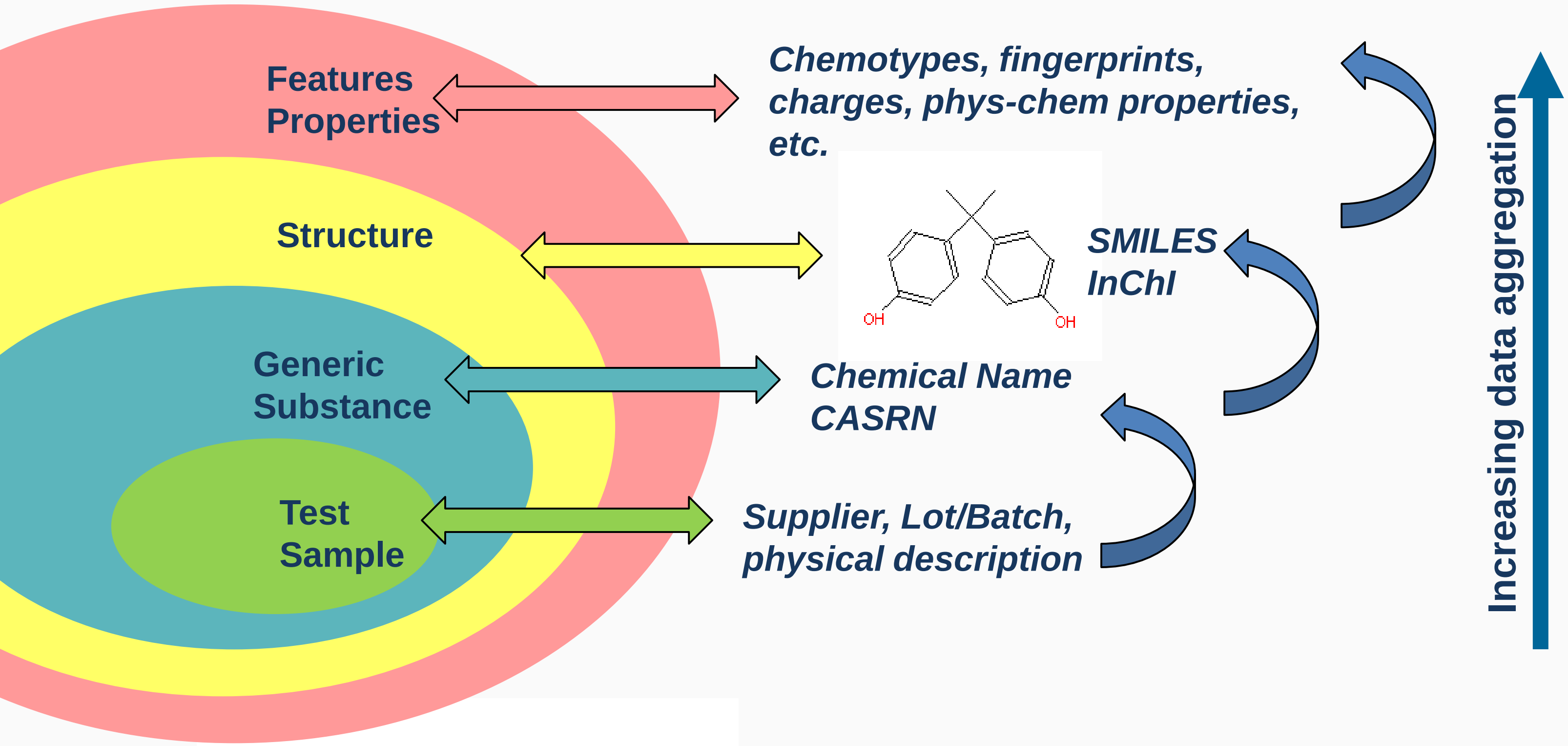
>700K substances

QSAR models;
phys-chem properties

CSS
Dashboards

ToxCast/Tox21
HTS activities

Chemical Elements to Quality Data Integration – Chemical “Representations”



Features Properties

Chemotypes, fingerprints, charges, phys-chem properties, etc.

Structure

SMILES
InChI

Generic Substance

Chemical Name
CASRN

Test Sample

Supplier, Lot/Batch, physical description

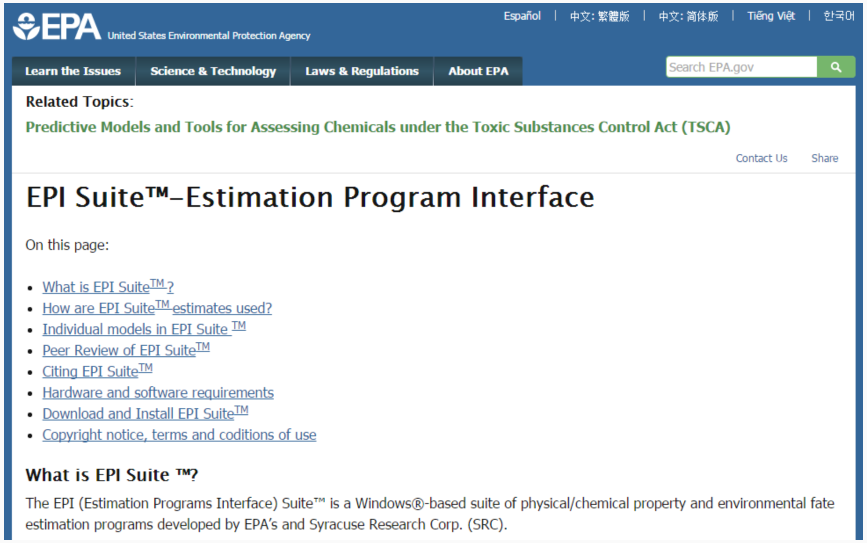
Increasing data aggregation

Diverse quality in databases

- Our challenges in assembling a database
  - Sourcing high quality data – sorting wheat from chaff
  - How to mesh data together – based on structure? On name? On CAS number? Other identifier?
  - Checking self-consistency of data?
  - Structure validation versus property value validation – very different challenges

iCSS Chemistry Dashboard

- Deliver a web-based platform hosting prediction models – both in-house and “community models”
- Initial set of models to be based on EPISuite’s PHYSPROP datasets – logP, BP, MP, Wsol etc.
- A question of “data quality” - a project to review data initiated.


EPA United States Environmental Protection Agency
Español | 中文: 繁體版 | 中文: 简体版 | Tiếng Việt | 한국어
Learn the Issues
Science & Technology
Laws & Regulations
About EPA
Search EPA.gov
Contact Us
Share

Related Topics:

Predictive Models and Tools for Assessing Chemicals under the Toxic Substances Control Act (TSCA)

EPI Suite™–Estimation Program Interface

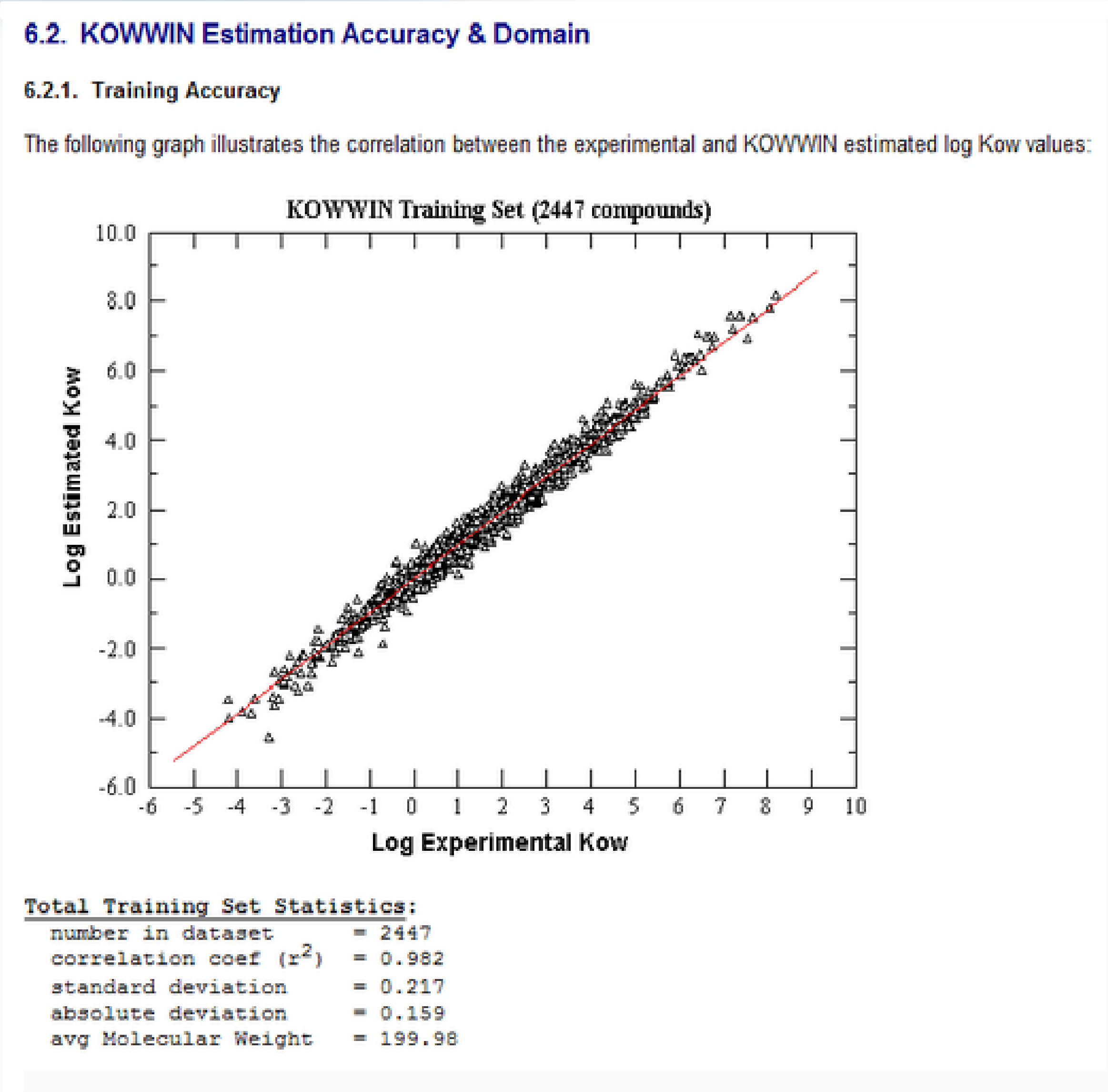
On this page:

- What is EPI Suite™?
- How are EPI Suite™ estimates used?
- Individual models in EPI Suite ™
- Peer Review of EPI Suite™
- Citing EPI Suite™
- Hardware and software requirements
- Download and Install EPI Suite™
- Copyright notice, terms and coditions of use

What is EPI Suite ™?

The EPI (Estimation Programs Interface) Suite™ is a Windows®-based suite of physical/chemical property and environmental fate estimation programs developed by EPA’s and Syracuse Research Corp. (SRC).

KOWWIN: 2447 Training Compounds



6.2. KOWWIN Estimation Accuracy & Domain

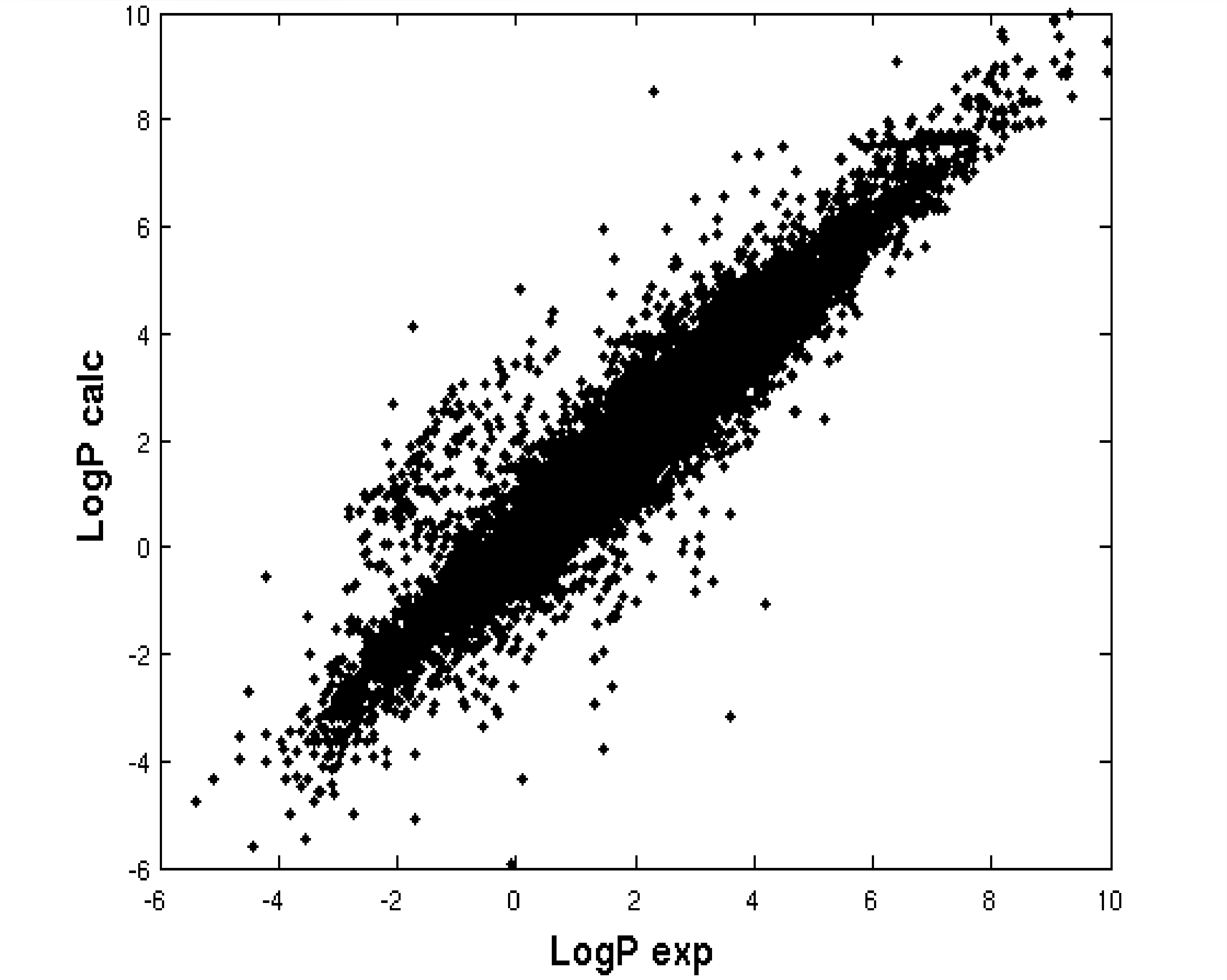
6.2.1. Training Accuracy

The following graph illustrates the correlation between the experimental and KOWWIN estimated log Kow values:

```
Total Training Set Statistics:
   number in dataset          = 2447
   correlation coef (r2)      = 0.982
   standard deviation         = 0.217
   absolute deviation         = 0.159
   avg Molecular Weight       = 199.98
```

Overall Predicted vs. Experimental

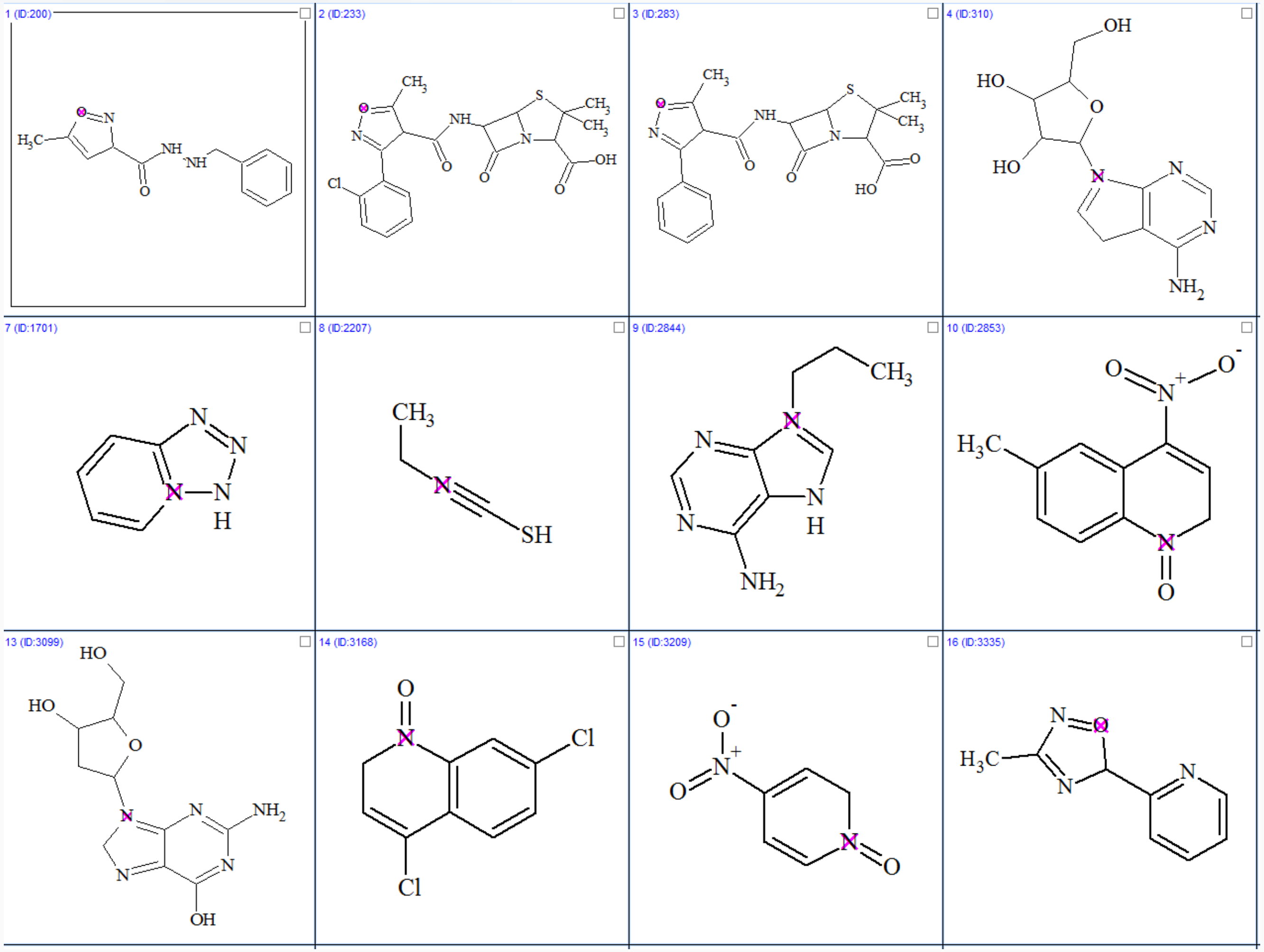
(>15k chemicals)



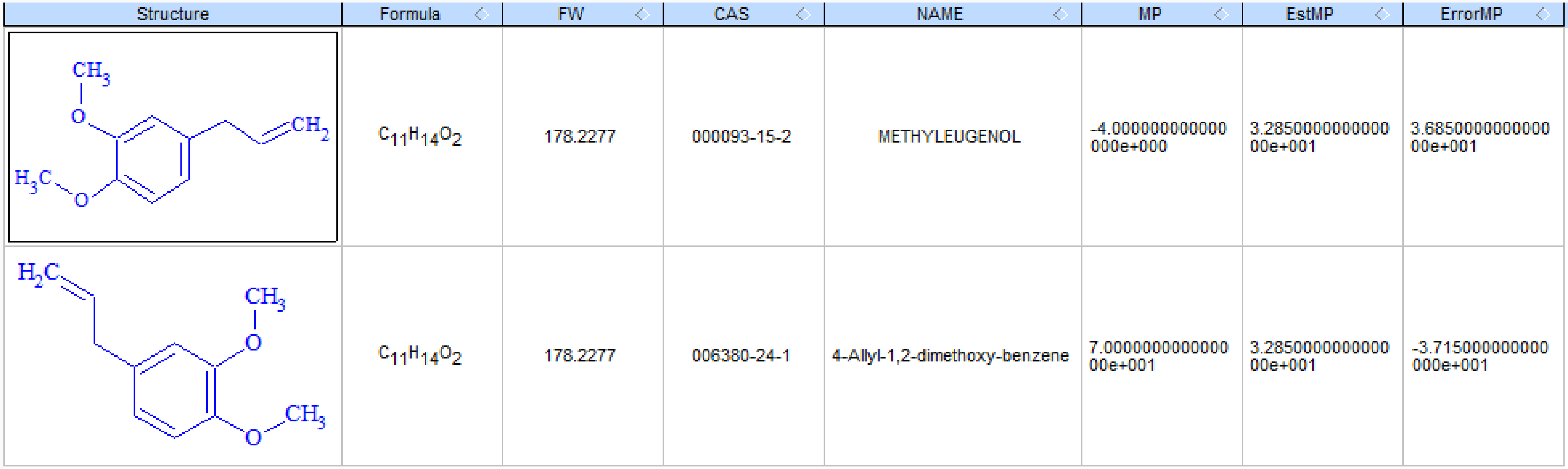
Original EPISuite Data

- Sourced the SDF files online as downloads
- Basic Manual review searching for errors:
  - hypervalency, charge imbalance, undefined stereo
  - deduplication
  - mismatches between identifiers
    - CAS Numbers not matching structure
    - Names not matching structure
    - Collisions between identifiers

Incorrect valences

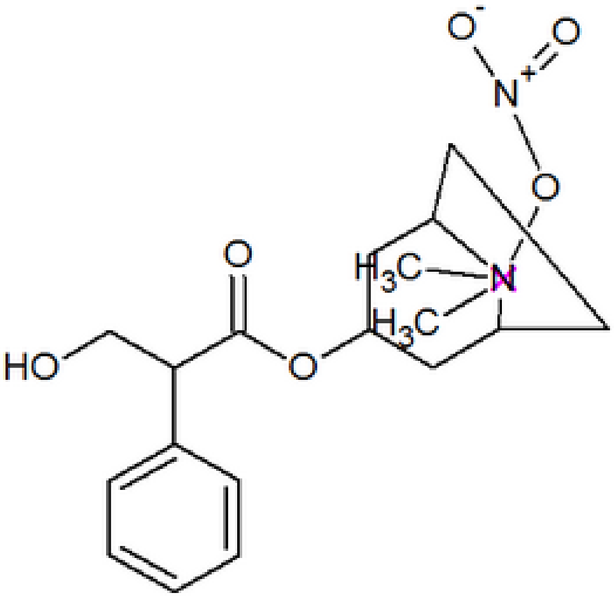


Differences in Values (MP)

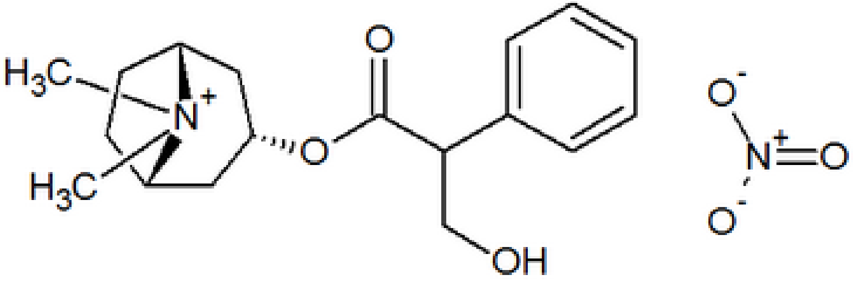


Structure	Formula	FW	CAS	NAME	MP	EstMP	ErrorMP
	$C_{11}H_{14}O_2$	178.2277	000093-15-2	METHYLEUGENOL	-4.000000000000000e+000	3.285000000000000e+001	3.685000000000000e+001
	$C_{11}H_{14}O_2$	178.2277	006380-24-1	4-Allyl-1,2-dimethoxy-benzene	7.000000000000000e+001	3.285000000000000e+001	-3.715000000000000e+001

Salts



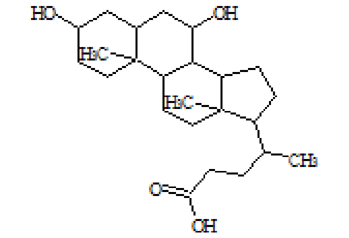
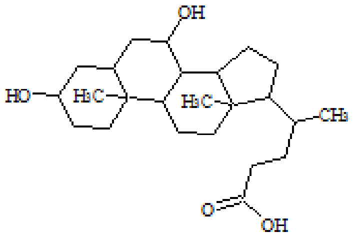
PHYSPROP database



Appropriate Depiction

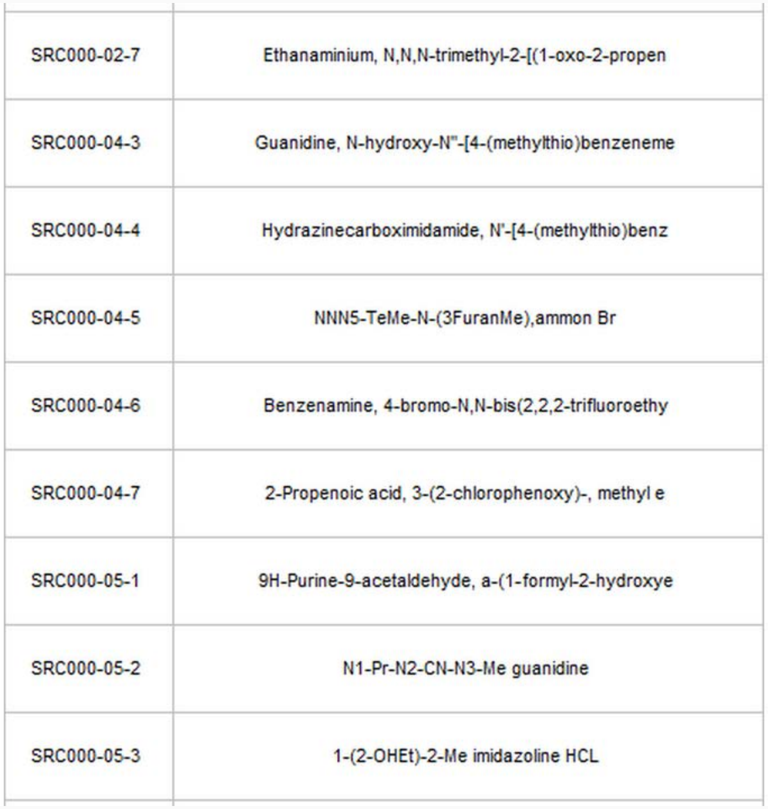
Collisions in Records

- Same structure depictions (Molfiles) - different CAS, different names, different SMILES

	$C_{24}H_{40}O_4$	392.5720	000083-49-8	HYODEOXYCHOLIC ACID	3.0800000000000 00e+000
	$C_{24}H_{40}O_4$	392.5720	000128-13-2	URSODEOXYCHOLIC ACID	3.0000000000000 00e+000

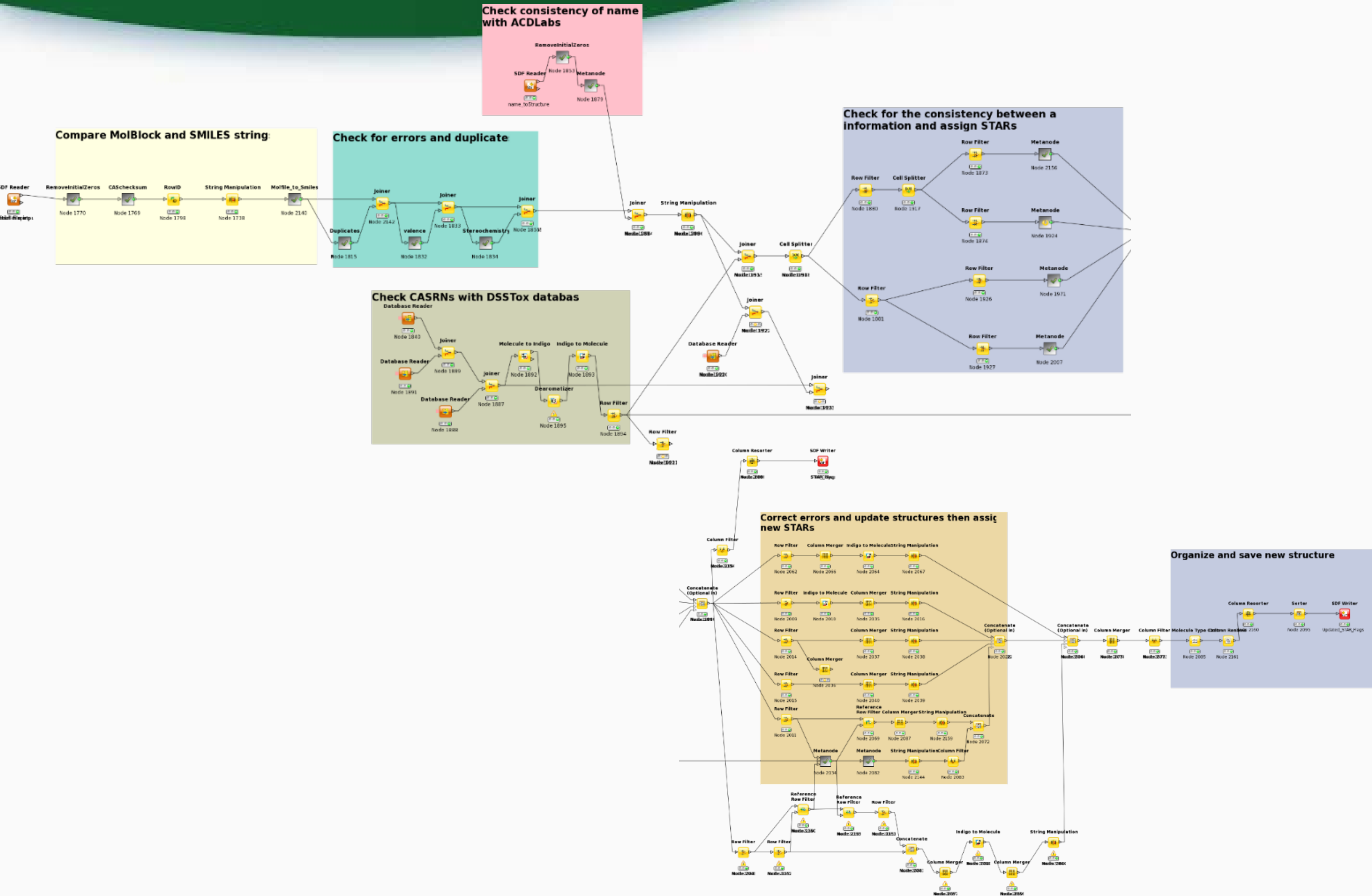
Identifiers

- Many chemical names are truncated
- Many chemicals don’t have CAS Numbers



SRC000-02-7	Ethanaminium, N,N,N-trimethyl-2-[(1-oxo-2-propen
SRC000-04-3	Guanidine, N-hydroxy-N"-[4-(methylthio)benzeneme
SRC000-04-4	Hydrazinecarboximidamide, N'-[4-(methylthio)benz
SRC000-04-5	NNN5-TeMe-N-(3FuranMe),ammon Br
SRC000-04-6	Benzenamine, 4-bromo-N,N-bis(2,2,2-trifluoroethy
SRC000-04-7	2-Propenoic acid, 3-(2-chlorophenoxy)-, methyl e
SRC000-05-1	9H-Purine-9-acetaldehyde, a-(1-formyl-2-hydroxye
SRC000-05-2	N1-Pr-N2-CN-N3-Me guanidine
SRC000-05-3	1-(2-OHEt)-2-Me imidazoline HCL

KNIME workflow to evaluate dataset



Statistics and data quality flags inserted

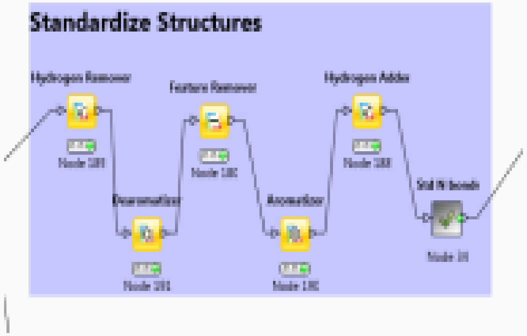
15,809 chemicals in the KOWWIN dataset

- CAS Checksum: 12163 valid, 3646 invalid
- Invalid names: 555
- Invalid SMILES 133
- Valence errors: 322 Molfile, 3782 SMILES
- Duplicates check:
  - 31 DUPLICATES MOLFILE
  - 626 DUPLICATES SMILES
  - 531 DUPLICATES NAMES
- SMILES vs. Molfiles (structure check)
  - 1279 differ in stereochemistry
  - 362 “Covalent Halogens”
  - 191 differ as tautomers
  - 436 are different compounds

Quality flags inserted into data

- 4 Stars ENHANCED: 4 levels of consistency with stereo information
- 4 Stars: 4 levels of consistency, stereo ignored.
- 3 Stars Plus: 3 out of 4 levels. The 4th is a tautomer.
- 3 Stars ENHANCED: 4 levels of consistency with stereo information
- 3 Stars: 3 levels of consistency, stereo ignored.
- 2 Stars PLUS : 2 out of 4 levels. The 3rd is a tautomer.
- 1 Star - What's left.

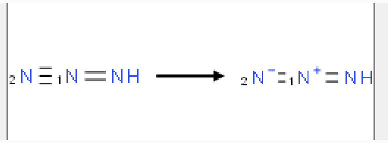
Standardization of structures
Make QSAR ready file for modeling

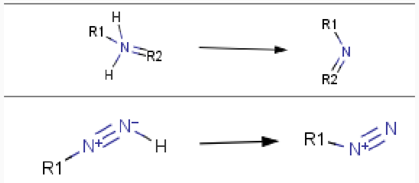


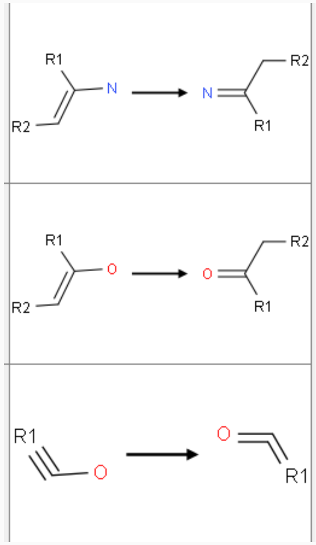
- Explicit hydrogen removal
- Removal of chirality info, isotopes and pseudo-atoms
- Aromatization + add explicit hydrogen atoms
- Standardize Nitro groups
- Other tautomerization/mesomerization checking
- Neutralize (when possible)

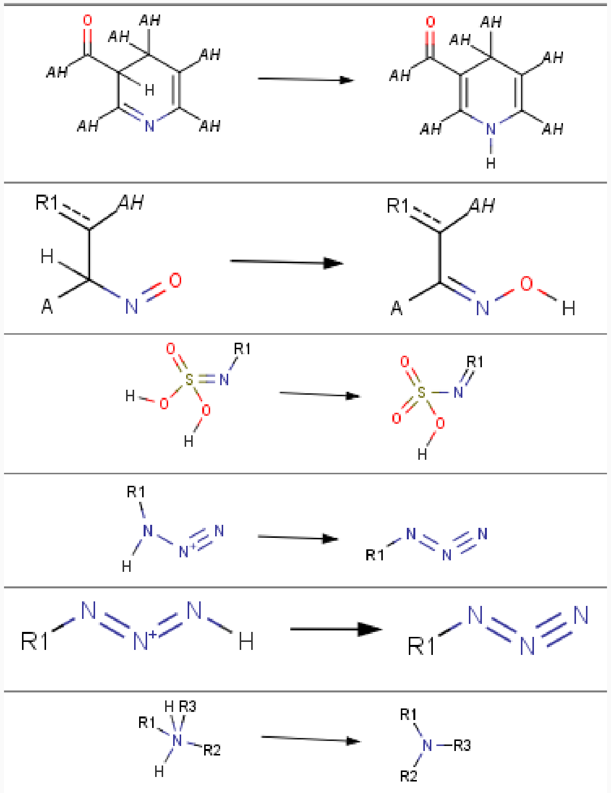
Mesomerization/tautomerization

- Azide mesomers
- Exo-enol tautomers
- Enamine-Imine tautomers
- Ynol-ketene tautomers
- ….

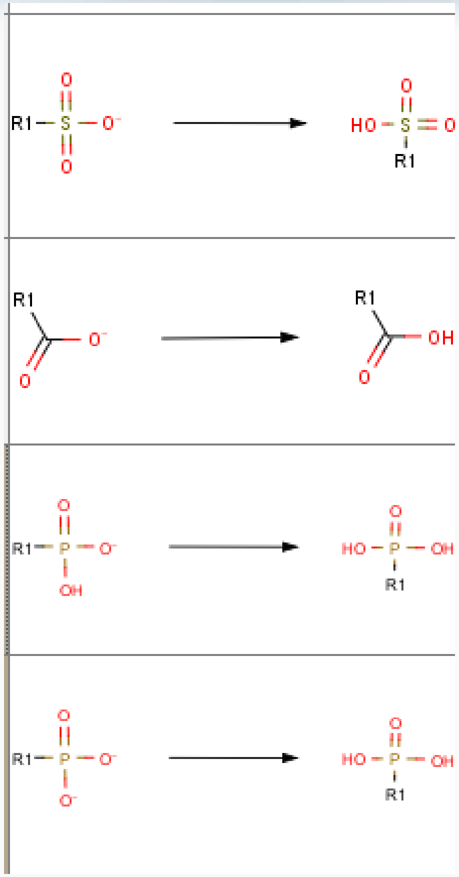


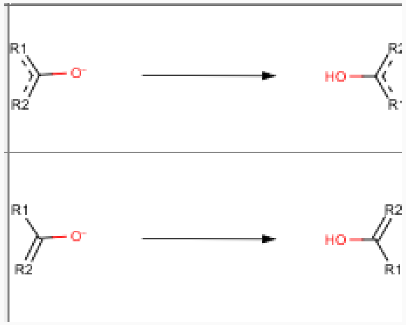


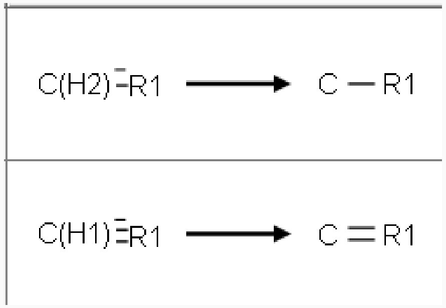


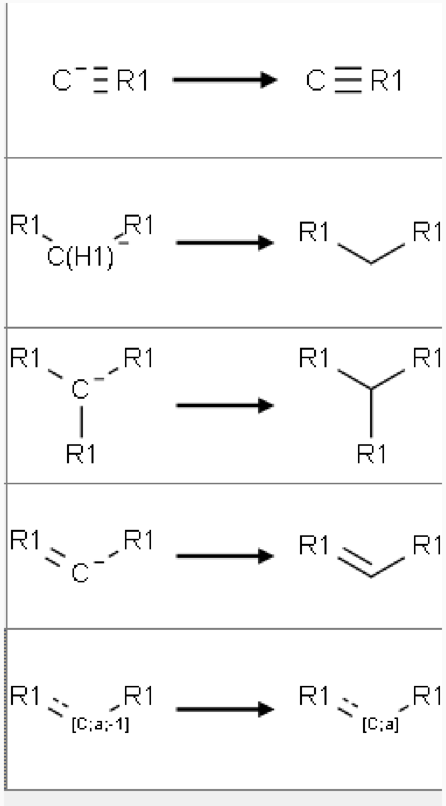


Neutralize Structures





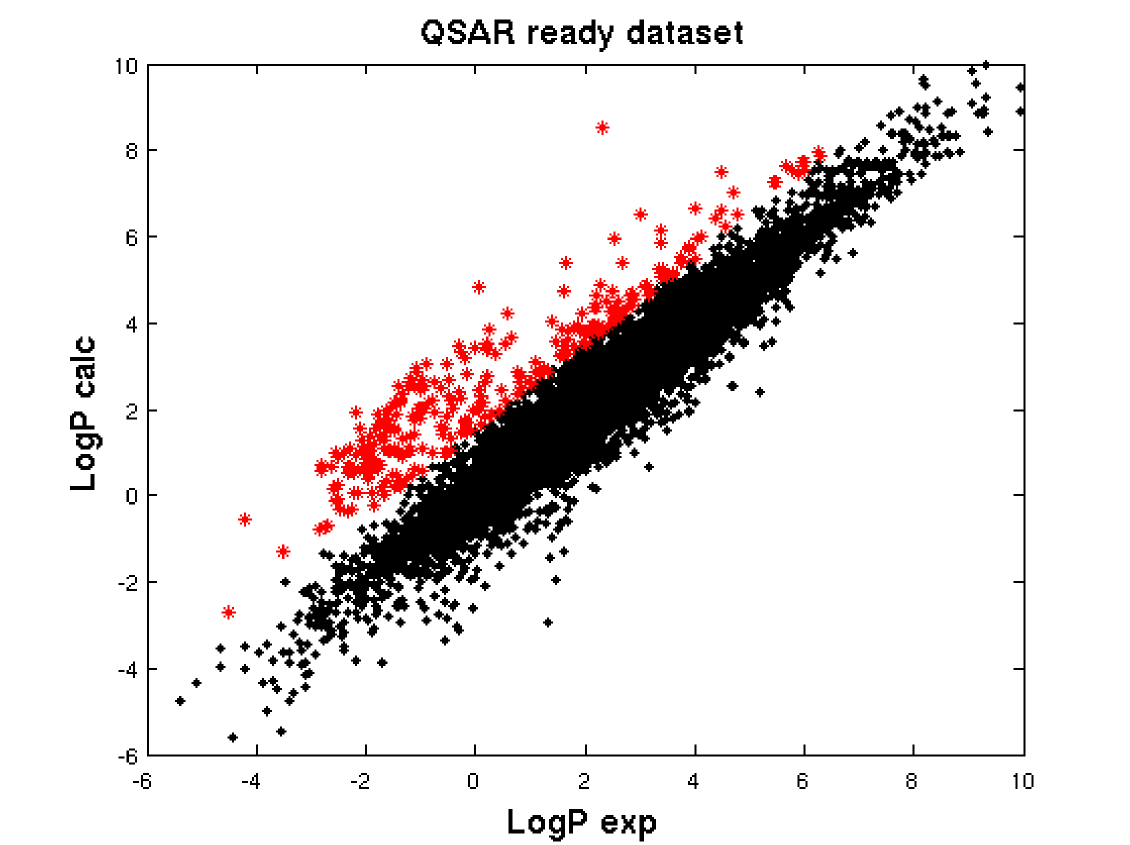


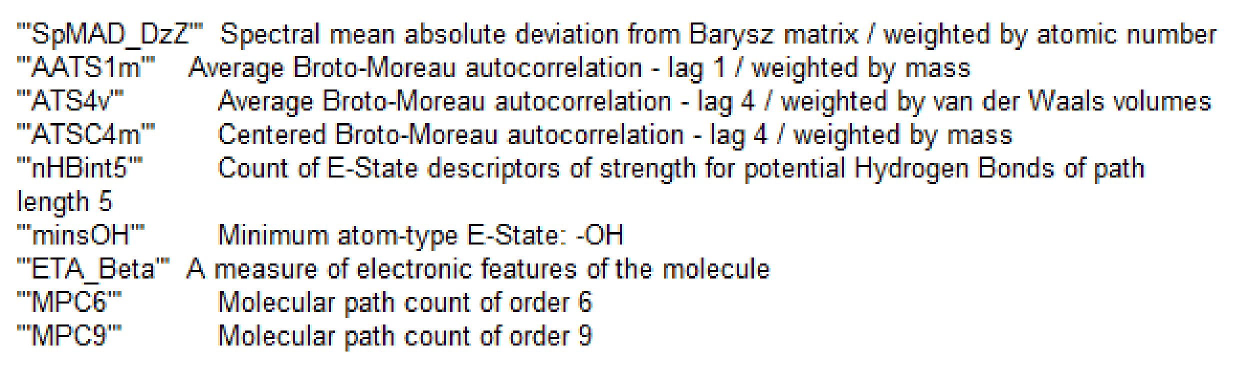


Building the Models

- Interested in “impact of quality vs. quantity of data on prediction models”
- Is it *worth* the effort to clean and enhance data underpinning the models?
- QSAR ready forms for modeling – standardize tautomers, remove stereochemistry, no salts
- Compare 3 STAR and BETTER models with entire dataset

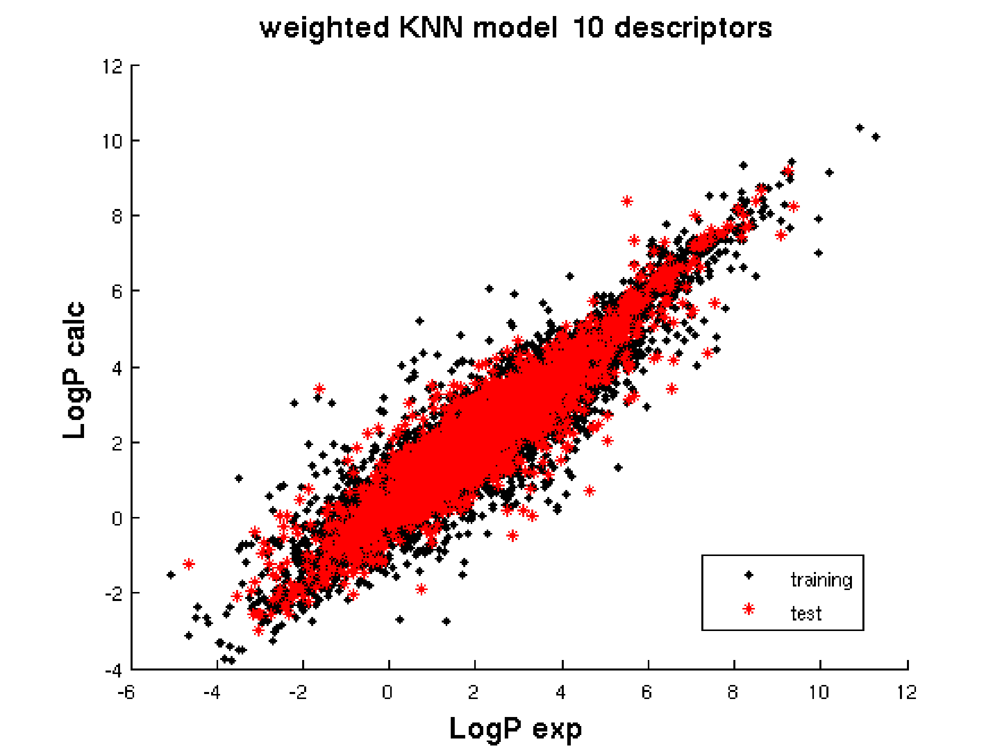
What about the “cluster”?





"'SpMAD_DzZ'"	Spectral mean absolute deviation from Barysz matrix / weighted by atomic number
"'AATS1m'"	Average Broto-Moreau autocorrelation - lag 1 / weighted by mass
"'ATS4v'"	Average Broto-Moreau autocorrelation - lag 4 / weighted by van der Waals volumes
"'ATSC4m'"	Centered Broto-Moreau autocorrelation - lag 4 / weighted by mass
"'nHBint5'"	Count of E-State descriptors of strength for potential Hydrogen Bonds of path length 5
"'minsOH'"	Minimum atom-type E-State: -OH
"'ETA_Beta'"	A measure of electronic features of the molecule
"'MPC6'"	Molecular path count of order 6
"'MPC9'"	Molecular path count of order 9

Weighted kNN Model, 10 descriptors



Weighted 5-nearest neighbors

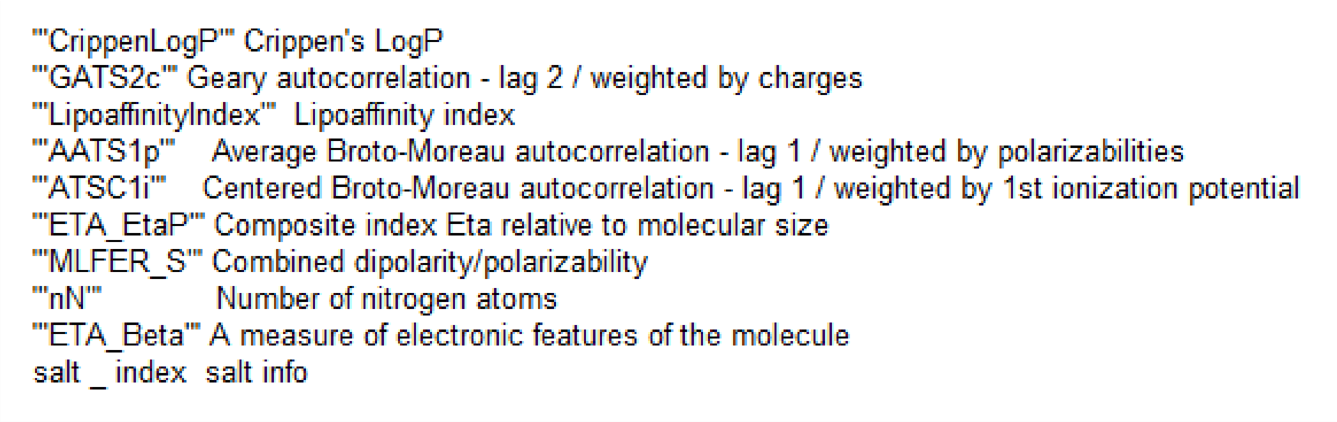
Training: 11251 chemicals

Test set: 2798 chemicals

5 fold cross validation:

R2: 0.87

RMSE: 0.67



"'CrippenLogP'" Crippen's LogP

"'GATS2c'" Geary autocorrelation - lag 2 / weighted by charges

"'LipoaffinityIndex'" Lipoaffinity index

"'AATS1p'" Average Broto-Moreau autocorrelation - lag 1 / weighted by polarizabilities

"'ATSC1i'" Centered Broto-Moreau autocorrelation - lag 1 / weighted by 1st ionization potential

"'ETA_EtaP'" Composite index Eta relative to molecular size

"'MLFER_S'" Combined dipolarity/polarizability

"'nN'" Number of nitrogen atoms

"'ETA_Beta'" A measure of electronic features of the molecule

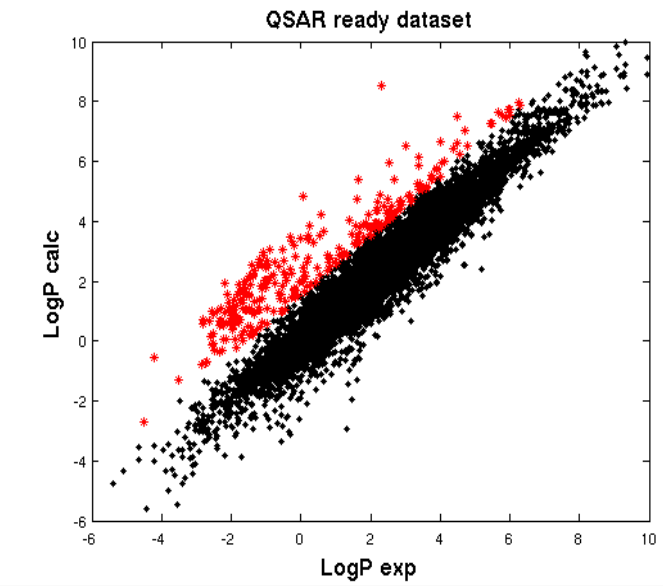
salt _ index salt info

Global vs Local Models???

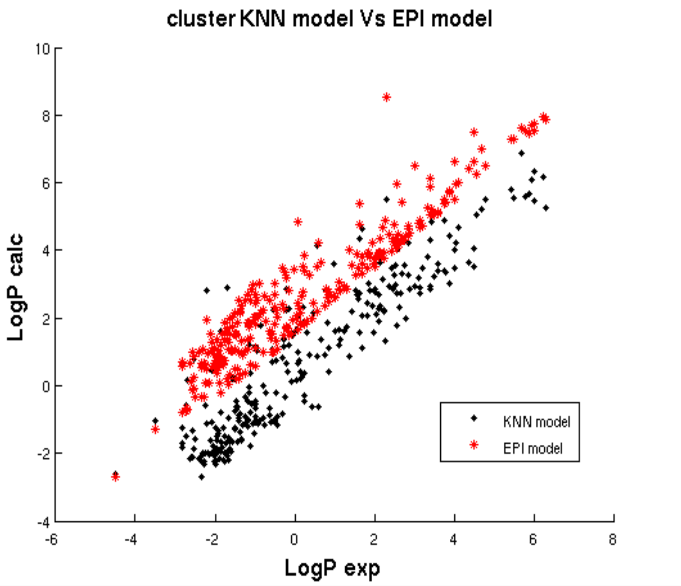
280 compound cluster

Applicability Domain of the Original EPISuite was an issue…

EPISuite Predicted vs. Experimental



EPISuite vs kNN Comparison of Cluster

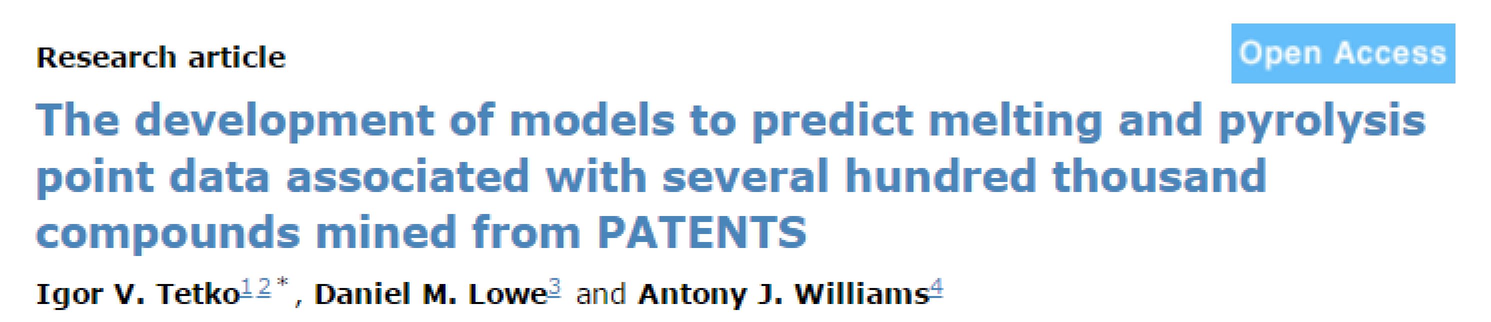


Bringing together PhysChem Data

- Additional Experimental Data being assembled
  - *PubChem*
  - *eChemPortal*
  - *Open Data Sources*

Open Data Set - example

- 200,000 melting points and 13,000 pyrolysis data points



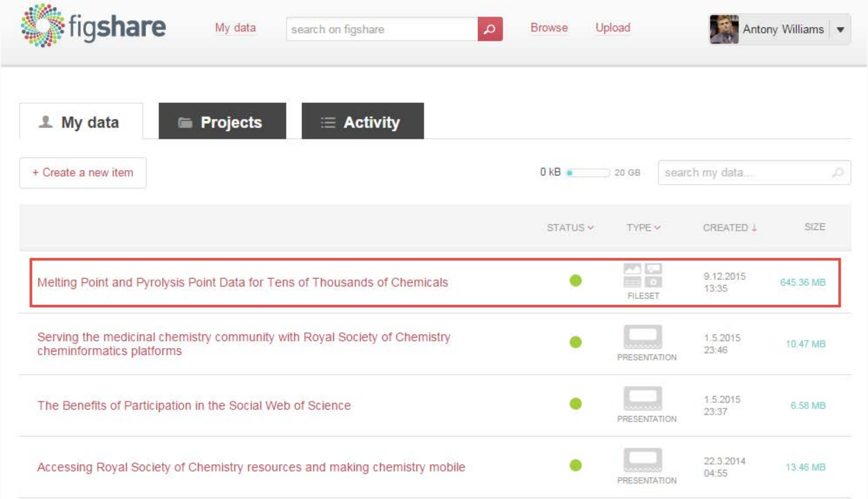
Research article

Open Access

The development of models to predict melting and pyrolysis point data associated with several hundred thousand compounds mined from PATENTS

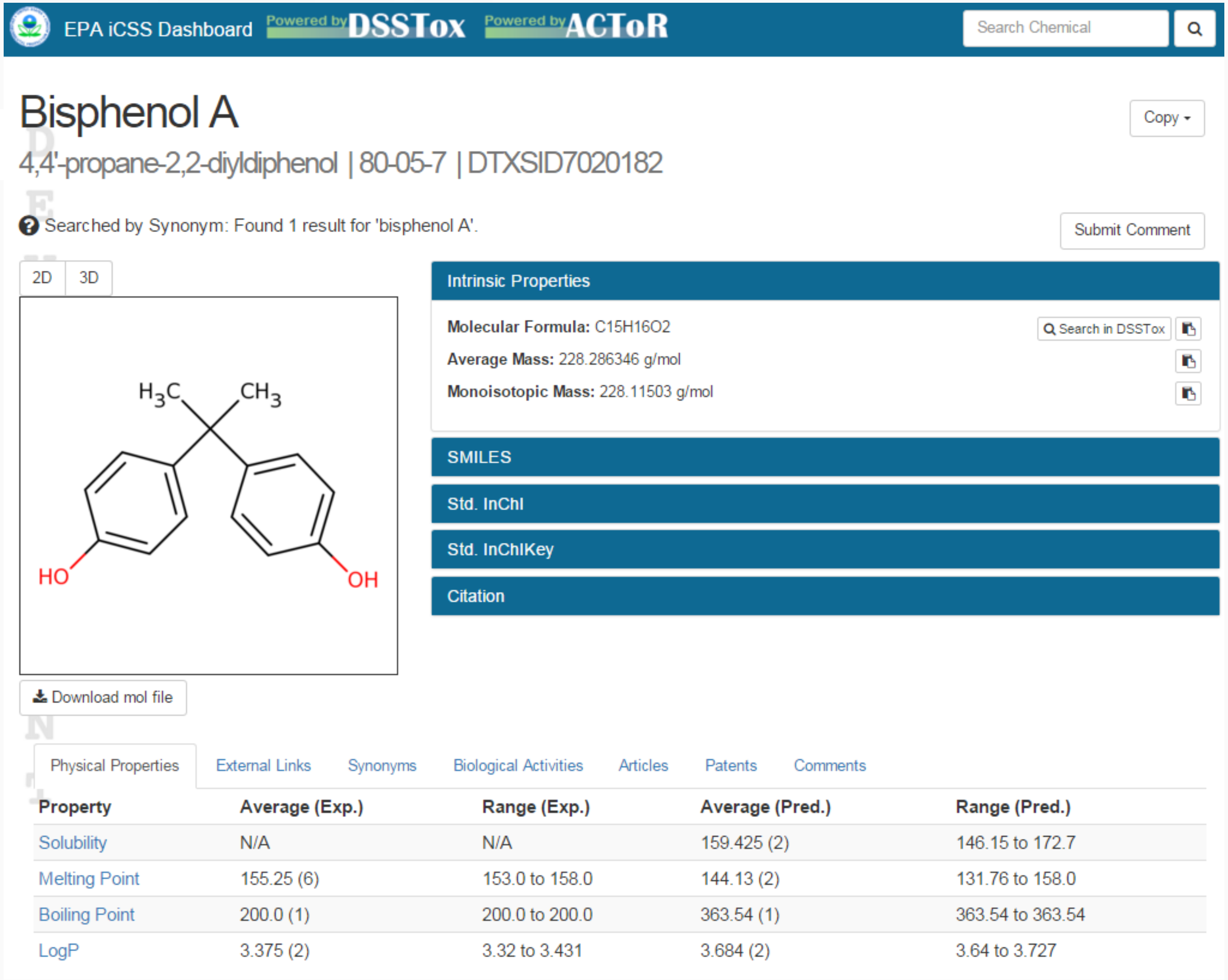
Igor V. Tetko[1,2,*], Daniel M. Lowe[3] and Antony J. Williams[4]

- Modeling this much data is a challenge!
- Accessibility to data?
  - – Figshare.com
  - – DataDryad.com
  - – Institutional Repositories
  - – Publishers?



	STATUS	TYPE	CREATED	SIZE
Melting Point and Pyrolysis Point Data for Tens of Thousands of Chemicals	●	FILESET	9.12.2015 13:35	645.36 MB
Serving the medicinal chemistry community with Royal Society of Chemistry cheminformatics platforms	●	PRESENTATION	1.5.2015 23:46	10.47 MB
The Benefits of Participation in the Social Web of Science	●	PRESENTATION	1.5.2015 23:37	6.58 MB
Accessing Royal Society of Chemistry resources and making chemistry mobile	●	PRESENTATION	22.3.2014 04:55	13.46 MB

iCSS Chemistry Dashboard Releasing in March 2016



Bisphenol A

4,4'-propane-2,2-diyldiphenol | 80-05-7 | DTXSID7020182

Searched by Synonym: Found 1 result for 'bisphenol A'.

Intrinsic Properties

Molecular Formula: C15H16O2

Average Mass: 228.286346 g/mol

Monoisotopic Mass: 228.11503 g/mol

SMILES

Std. InChI

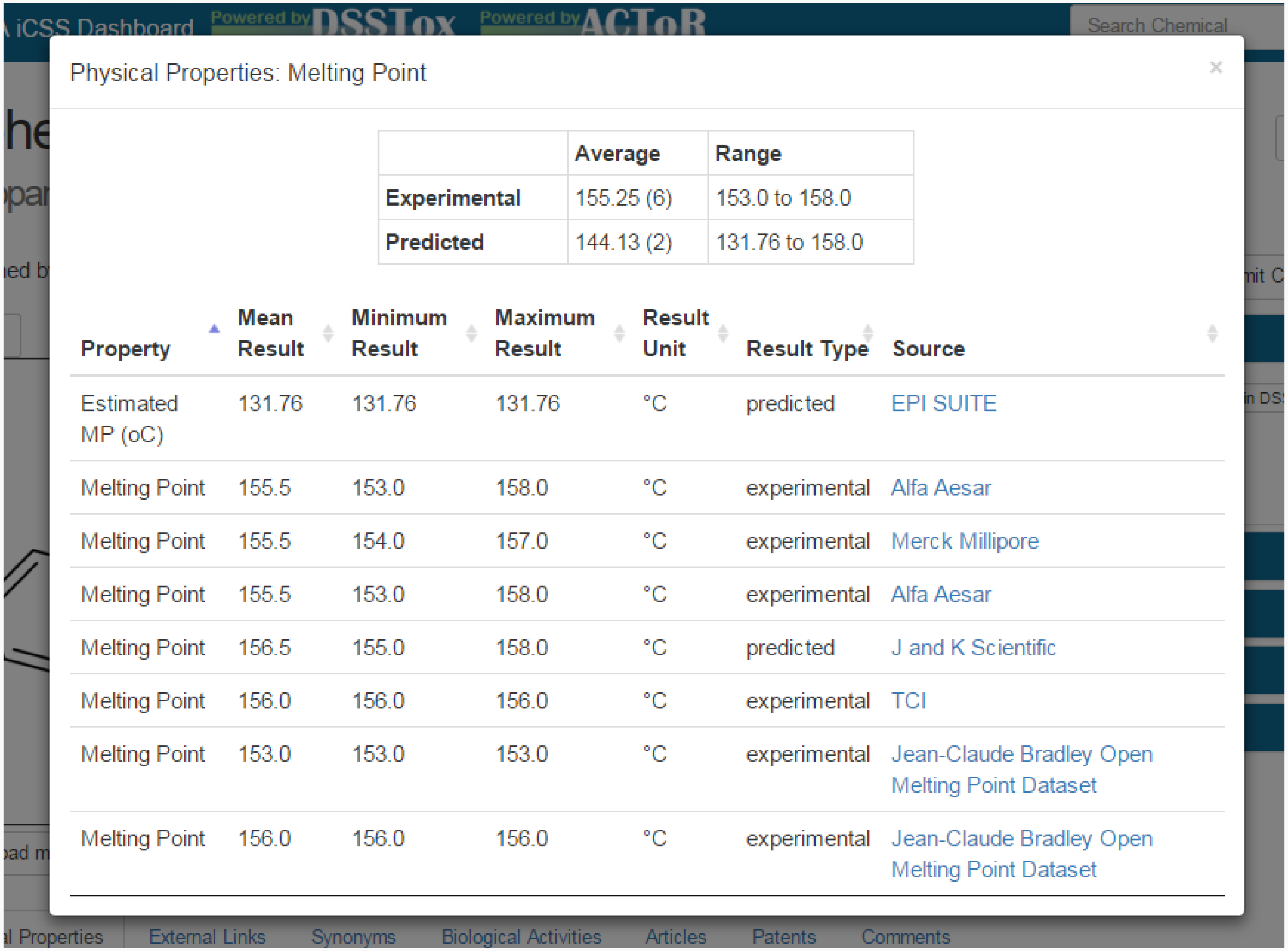
Std. InChIKey

Citation

Physical Properties | External Links | Synonyms | Biological Activities | Articles | Patents | Comments

Property	Average (Exp.)	Range (Exp.)	Average (Pred.)	Range (Pred.)
Solubility	N/A	N/A	159.425 (2)	146.15 to 172.7
Melting Point	155.25 (6)	153.0 to 158.0	144.13 (2)	131.76 to 158.0
Boiling Point	200.0 (1)	200.0 to 200.0	363.54 (1)	363.54 to 363.54
LogP	3.375 (2)	3.32 to 3.431	3.684 (2)	3.64 to 3.727

iCSS Chemistry Dashboard Releasing in March 2016



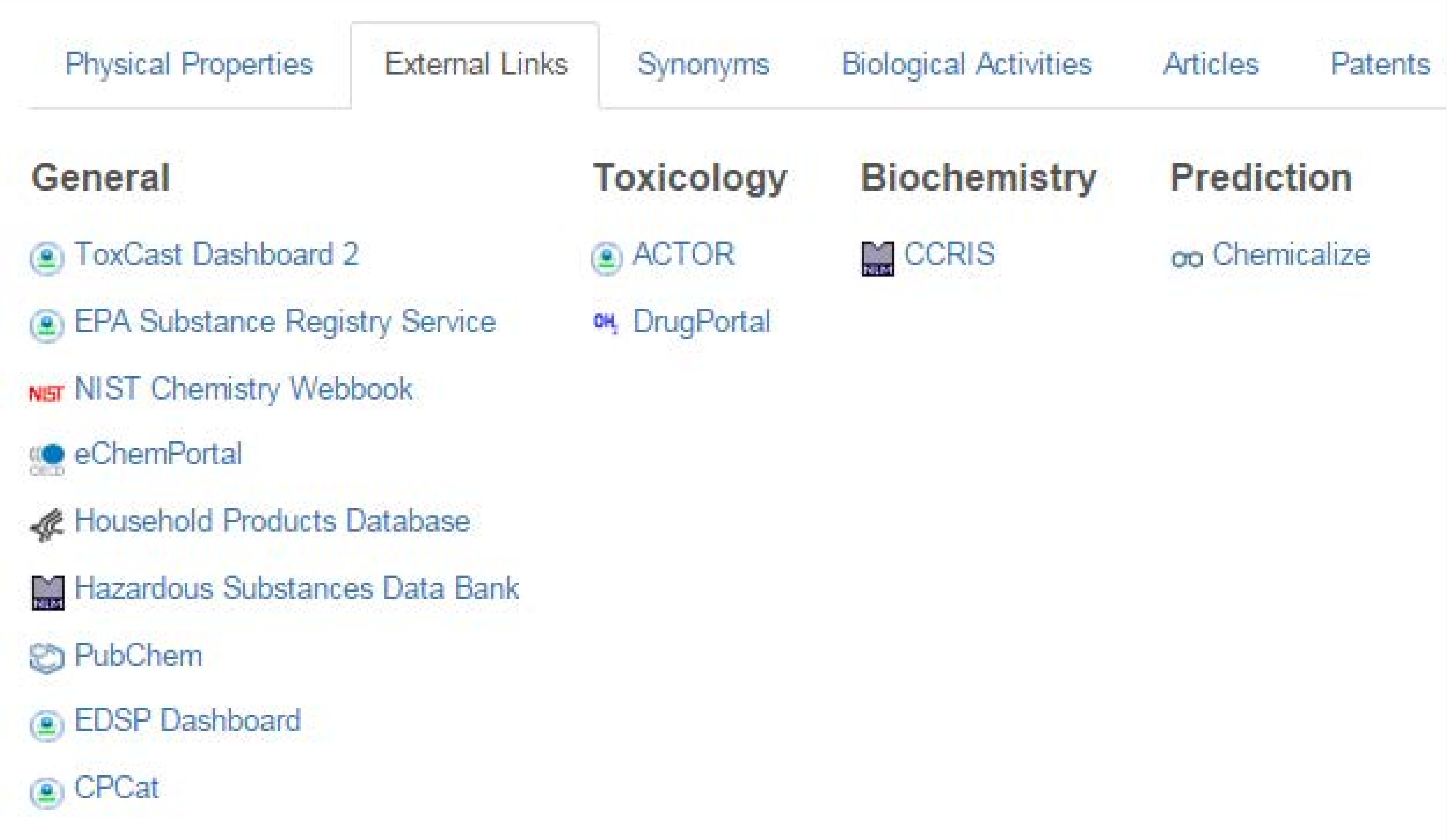
Physical Properties: Melting Point

	Average	Range
Experimental	155.25 (6)	153.0 to 158.0
Predicted	144.13 (2)	131.76 to 158.0

Property	Mean Result	Minimum Result	Maximum Result	Result Unit	Result Type	Source
Estimated MP (oC)	131.76	131.76	131.76	°C	predicted	EPI SUITE
Melting Point	155.5	153.0	158.0	°C	experimental	Alfa Aesar
Melting Point	155.5	154.0	157.0	°C	experimental	Merck Millipore
Melting Point	155.5	153.0	158.0	°C	experimental	Alfa Aesar
Melting Point	156.5	155.0	158.0	°C	predicted	J and K Scientific
Melting Point	156.0	156.0	156.0	°C	experimental	TCI
Melting Point	153.0	153.0	153.0	°C	experimental	Jean-Claude Bradley Open Melting Point Dataset
Melting Point	156.0	156.0	156.0	°C	experimental	Jean-Claude Bradley Open Melting Point Dataset

iCSS Chemistry Dashboard Releasing in March 2016

- Links external resources: EPA, NIH, property predictors



Physical Properties | External Links | Synonyms | Biological Activities | Articles | Patents

General

- ToxCast Dashboard 2
- EPA Substance Registry Service
- NIST Chemistry Webbook
- eChemPortal
- Household Products Database
- Hazardous Substances Data Bank
- PubChem
- EDSP Dashboard
- CPCat

Toxicology

- ACTOR
- DrugPortal

Biochemistry

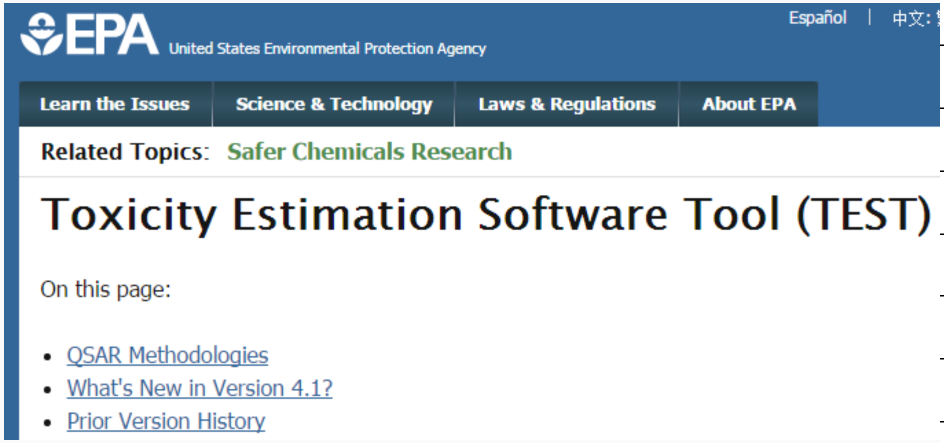
- CCRIS

Prediction

- Chemicalize

…Integrate Toxicity Estimation Software Tool

http://www.epa.gov/chemical-research/toxicity-estimation-software-tool-test

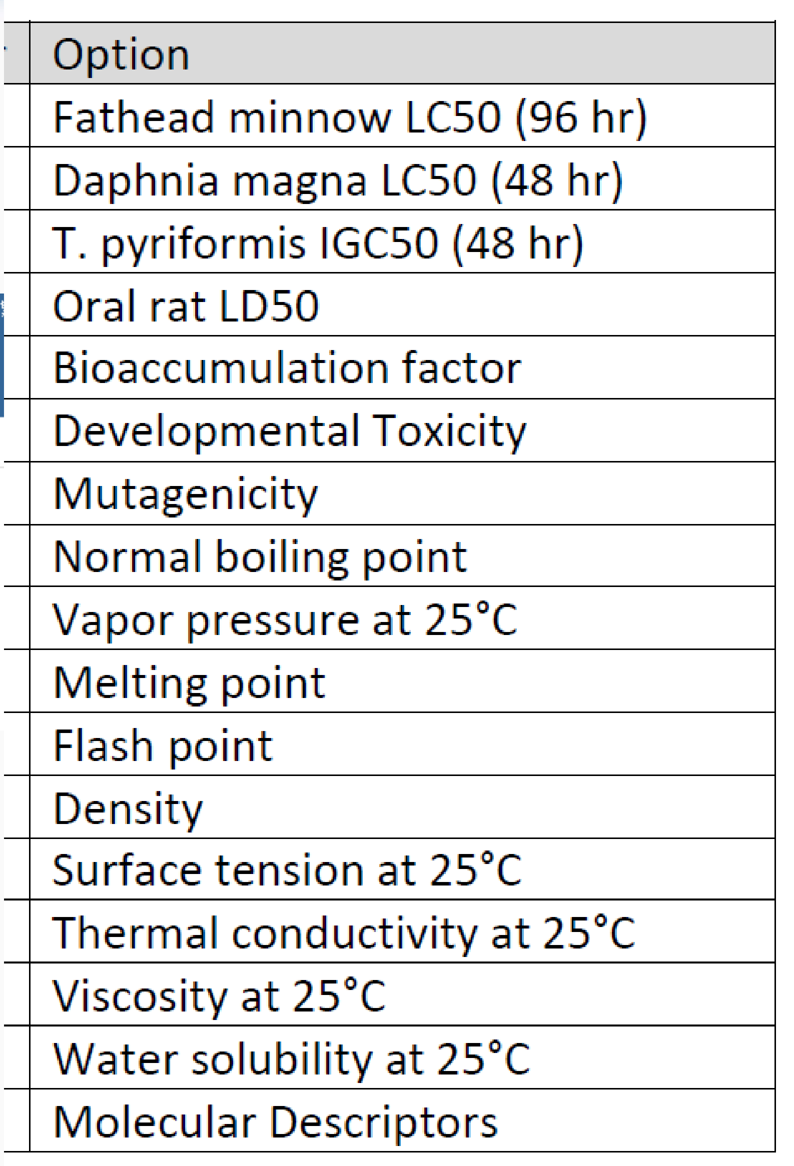


Related Topics: Safer Chemicals Research

Toxicity Estimation Software Tool (TEST)

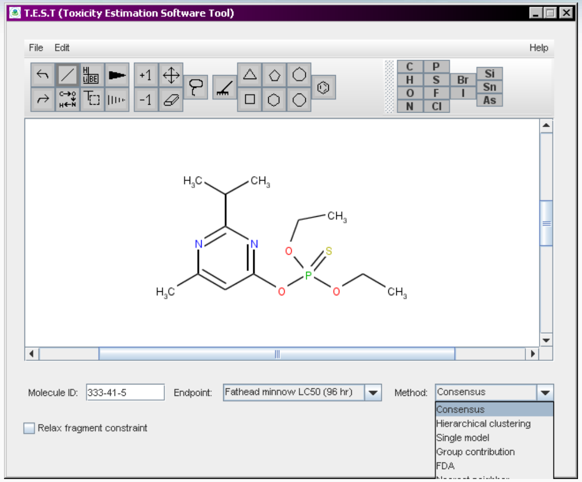
On this page:

- QSAR Methodologies
- What's New in Version 4.1?
- Prior Version History



Option
Fathead minnow LC50 (96 hr)
Daphnia magna LC50 (48 hr)
T. pyriformis IGC50 (48 hr)
Oral rat LD50
Bioaccumulation factor
Developmental Toxicity
Mutagenicity
Normal boiling point
Vapor pressure at 25°C
Melting point
Flash point
Density
Surface tension at 25°C
Thermal conductivity at 25°C
Viscosity at 25°C
Water solubility at 25°C
Molecular Descriptors

Toxicity Estimation Software Tool



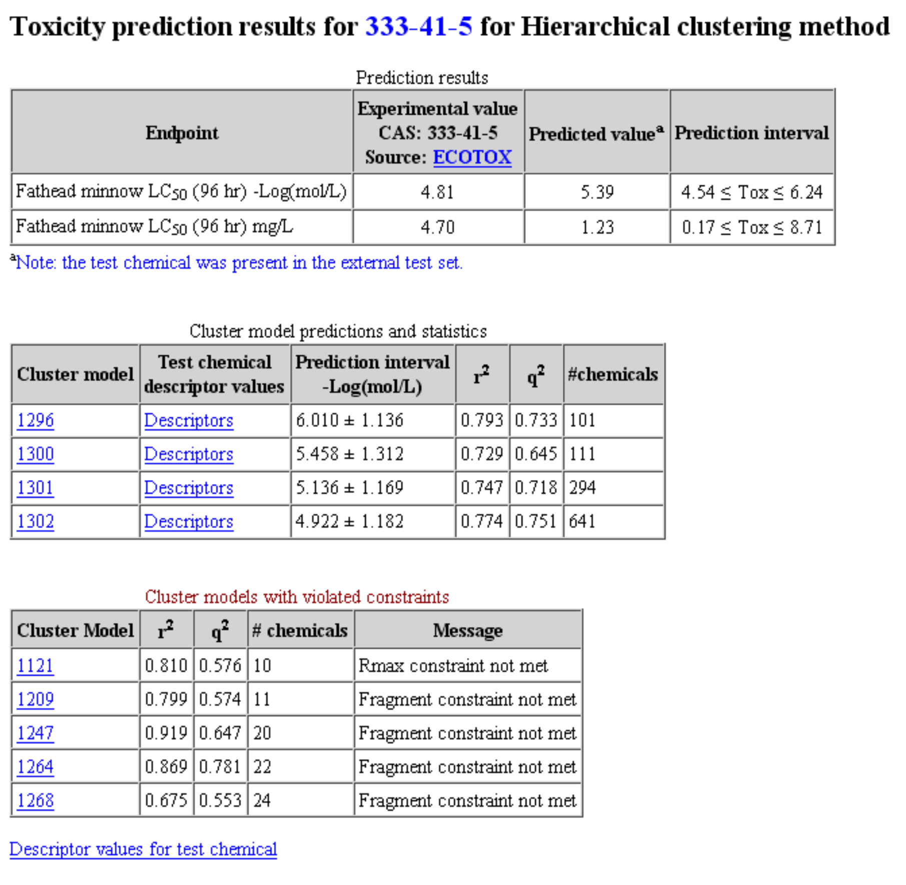
Molecule ID: 333-41-5

Endpoint: Fathead minnow LC50 (96 hr)

Method: Consensus

- Consensus
- Hierarchical clustering
- Single model
- Group contribution
- FDA

□ Relax fragment constraint



Toxicity prediction results for 333-41-5 for Hierarchical clustering method

Prediction results

Endpoint	Experimental value CAS: 333-41-5 Source: ECOTOX	Predicted value[a]	Prediction interval
Fathead minnow LC_{50} (96 hr) -Log(mol/L)	4.81	5.39	4.54 ≤ Tox ≤ 6.24
Fathead minnow LC_{50} (96 hr) mg/L	4.70	1.23	0.17 ≤ Tox ≤ 8.71

[a] Note: the test chemical was present in the external test set.

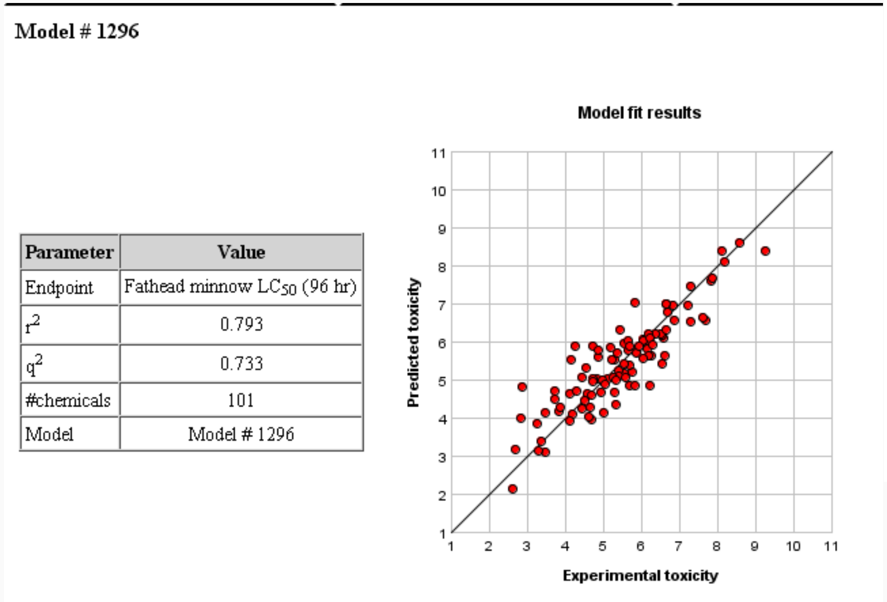
Cluster model predictions and statistics

Cluster model	Test chemical descriptor values	Prediction interval -Log(mol/L)	r^2	q^2	#chemicals
1296	Descriptors	6.010 ± 1.136	0.793	0.733	101
1300	Descriptors	5.458 ± 1.312	0.729	0.645	111
1301	Descriptors	5.136 ± 1.169	0.747	0.718	294
1302	Descriptors	4.922 ± 1.182	0.774	0.751	641

Cluster models with violated constraints

Cluster Model	r^2	q^2	# chemicals	Message
1121	0.810	0.576	10	Rmax constraint not met
1209	0.799	0.574	11	Fragment constraint not met
1247	0.919	0.647	20	Fragment constraint not met
1264	0.869	0.781	22	Fragment constraint not met
1268	0.675	0.553	24	Fragment constraint not met

Descriptor values for test chemical



Model # 1296

Parameter	Value
Endpoint	Fathead minnow LC_{50} (96 hr)
r^2	0.793
q^2	0.733
#chemicals	101
Model	Model # 1296

Conclusion

- Improved chemistry support and data will impact every NCCT project
- Creating high quality data is difficult & iterative
- Public domain and open data is valuable *but* requires validation
- Modern cheminformatics approaches can help validate and prepare data for modeling
- **There are increasing efforts to source and release high quality, open data to the world**

Acknowledgements

Contributors to the iCSS Chemistry Dashboard

- Jeff Edwards
- Jeremy Fitzpatrick
- Jordan Foster
- Chris Grulke
- Jason Harris
- Dave Lyons
- Kamel Mansouri
- Ann Richard
- Aria Smith
- Jennifer Smith
- Indira Thillainadarajah