The advantages, and cautions, of applying AI to the development of prediction models

Antony Williams

AAAS Conference, Denver CO. February 2024

US-EPA Office of Research and Development

- Center for Computational Toxicology and Exposure (CCTE)
- Some of our primary efforts in AI are in modeling (QSAR/QSPR QSUR) based on measurement, harvesting, assembly and curation of streams of data
- We develop and deliver to the community
  - curated chemistry data streams to support our applications and models
  - prediction models, web-based applications and data streams to support others

AI is not new, but it is modernizing...

- We have benefited from AI modeling approaches for decades
- Approaches continue to advance with the promise of improved performance
- Many challenges remain the same – quality of data, data overfitting, and interpretability of the models
- The latest form of modernization is “Large Language Models”

Data, Model and Tool Development

- There are many tools developed by our cheminformatics team and across other centers in EPA. I will represent **ours** only…
- We have production level public-facing tools, proof-of-concept public-facing tools, and many tools in development…
- We focus on FAIR data releasing it to the community and making it available on Public APIs

Example Problems We Apply Models to

- Property prediction – e.g., water solubility, vapor pressure
- Fate and Transport – e.g., bioaccumulation, bioconcentration
- Bioactivity – e.g., endocrine disruption, steroidogenesis
- Analytical Chemistry – e.g., what techniques are chemicals amenable to - GC/MS? LC/MS? +ve/-ve ion?

OECD Principles for Modeling

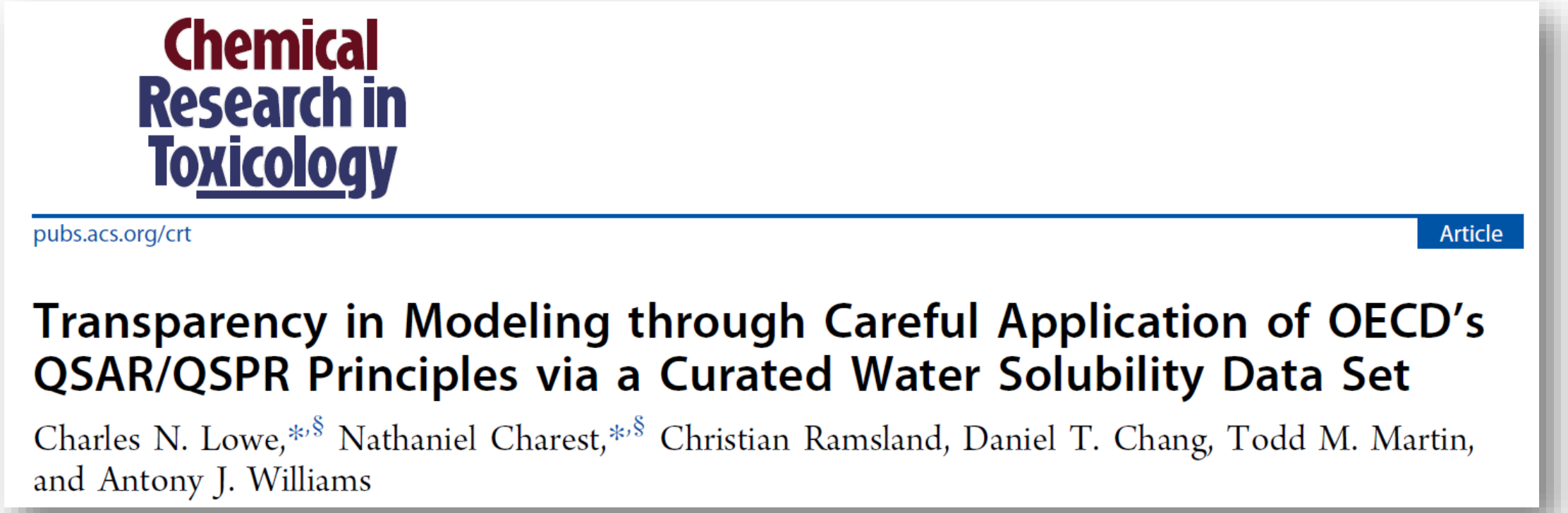
https://www.oecd.org/chemicalsafety/risk-assessment/37849783.pdf

- To facilitate the consideration of a (Q)SAR model for regulatory purposes, it should be associated with the following information:
  1) a defined endpoint
  2) an unambiguous algorithm
  3) a defined domain of applicability
  4) appropriate measures of goodness-of–fit, robustness and predictivity
  5) a mechanistic interpretation, if possible
- These principles have been around a long time…

These principles were agreed by OECD member countries at the 37th Joint Meeting of the Chemicals Committee and Working Party on Chemicals, Pesticides and Biotechnology in November 2004.

Our approaches to building models

- Exemplified through our recent water solubility work



Chemical Research in Toxicology

pubs.acs.org/crt

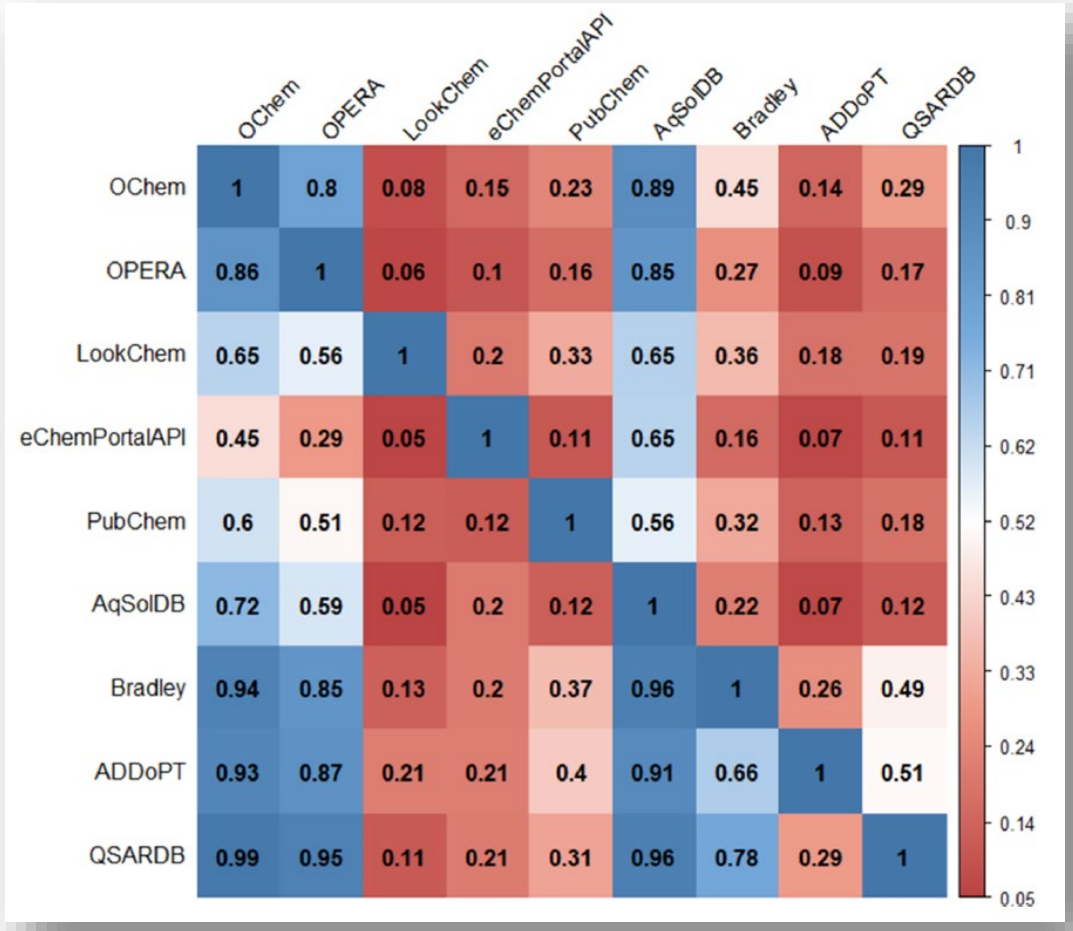
Article

Transparency in Modeling through Careful Application of OECD's QSAR/QSPR Principles via a Curated Water Solubility Data Set

Charles N. Lowe,*,§ Nathaniel Charest,*,§ Christian Ramsland, Daniel T. Chang, Todd M. Martin, and Antony J. Williams

Data assembly and curation

- ~10,200 unique chemical structures with experimental data
- Great care is needed in data assembly for modeling
- For water solubility large collections based on standard methods provide threshold measurements >X, <Y
- The same data is shared between public platforms



Lots of descriptors to choose from

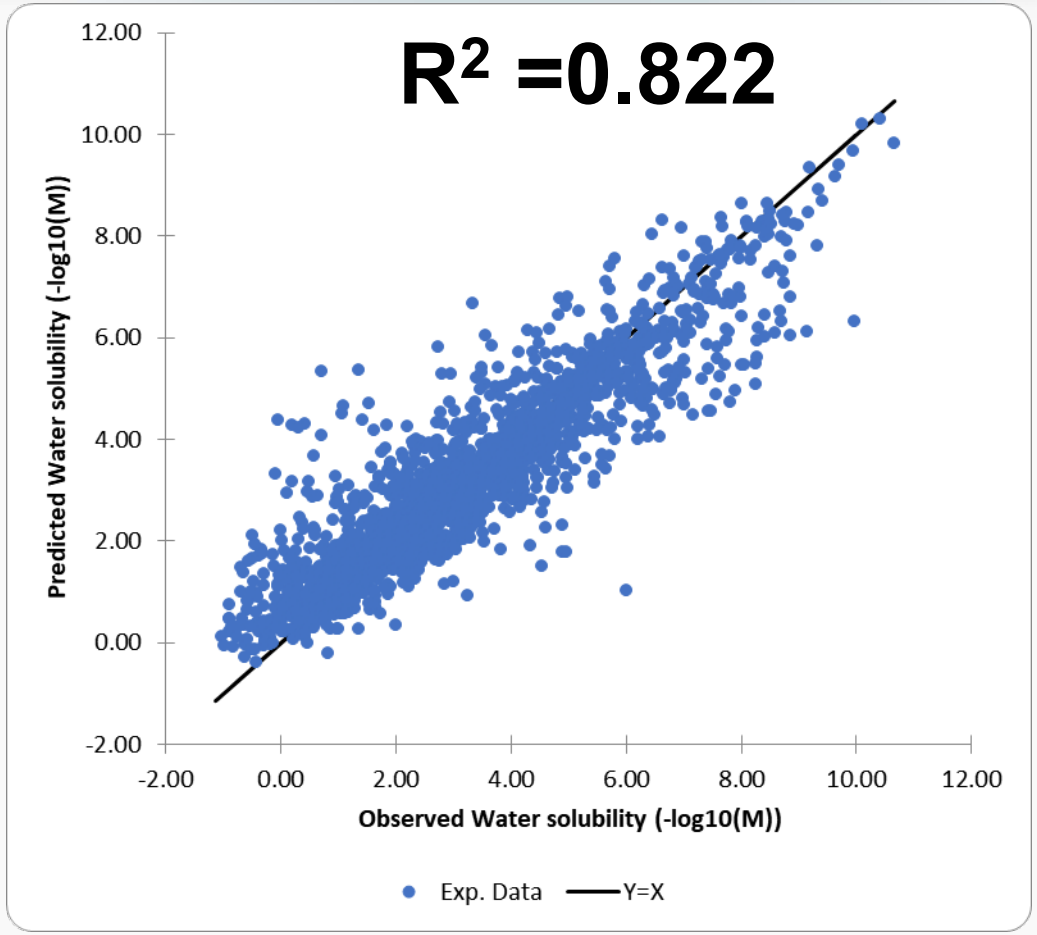
- Many Descriptors to choose: commercial and open source
- We use Padel, Mordred and TEST descriptors (open)
- Example: http://www.yapcwsoft.com/dd/padeldescriptor/

PaDEL-Descriptor

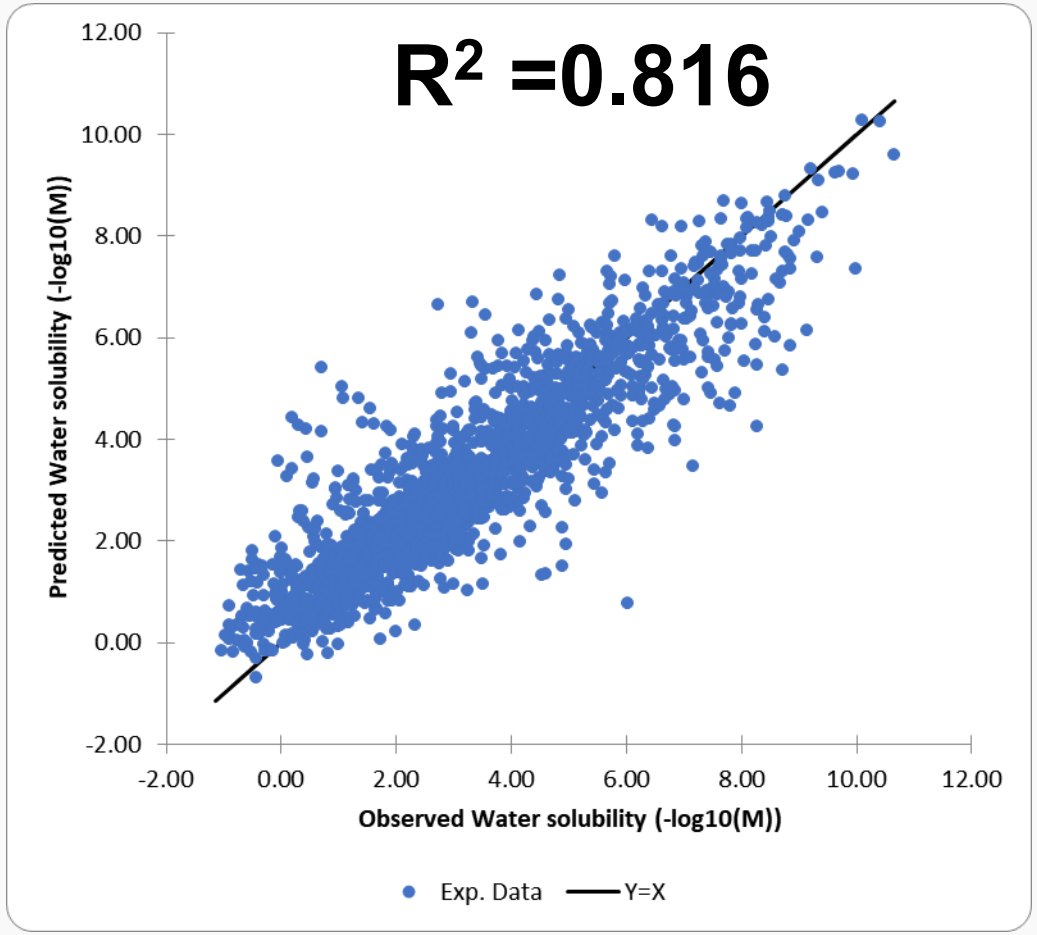
Description

A software to calculate molecular descriptors and fingerprints. The software currently calculates 1875 descriptors (1444 1D, 2D descriptors and 431 3D descriptors) and 12 types of fingerprints (total 16092 bits). The descriptors and fingerprints are calculated using The Chemistry Development Kit with additional descriptors and fingerprints such as atom type electrotopological state descriptors, Crippen's logP and MR, extended topochemical atom (ETA) descriptors, McGowan volume, molecular linear free energy relation descriptors, ring counts, count of chemical substructures identified by Laggner, and binary fingerprints and count of chemical substructures identified by Klekota and Roth.

Feature Selection and Variables can help **mechanistic understanding**



Without Feature Selection – 427 variables



With Feature Selection – **19 variables**

Todd Martin, SERMACS, 2023

Descriptors

Descriptor	Definition
XLOGP	Wang **octanol water partition coefficient**
XLOGP2	Wang octanol water partition coefficient squared
BEHm3	Highest eigenvalue n. 3 of Burden matrix / weighted by atomic masses
eim	Mean information content on the edge magnitude
Qs	Molecular and group **polarity** index
BEHm4	Highest eigenvalue n. 4 of Burden matrix / weighted by atomic masses
BEHm2	Highest eigenvalue n. 2 of Burden matrix / weighted by atomic masses
BEHm7	Highest eigenvalue n. 7 of Burden matrix / weighted by atomic masses
BEHm6	Highest eigenvalue n. 6 of Burden matrix / weighted by atomic masses
MATS1v	Moran autocorrelation - lag 1 / weighted by atomic van der Waals volumes
MATS1p	Moran autocorrelation - lag 1 / weighted by atomic polarizabilities
ATS4m	Broto-Moreau autocorrelation of a topological structure - lag 4 / weighted by atomic masses
BEHm5	Highest eigenvalue n. 5 of Burden matrix / weighted by atomic masses
Hmax	Maximum **hydrogen E-State** value in molecule.
piID	Conventional bond order id number
ieadjem	Mean information content on the edge adjacency equality
Mv	Mean atomic van der Waals volume (scaled on Carbon atom)
SHHBd	Sum of E-State indices for hydrogen bond donors
ide	Total information content on the distance equality

Delivering Modeled Data and Models via Free-Access Cheminformatics Tools

- Many of our tools are publicly available
  - CompTox Chemicals Dashboard (https://comptox.epa.gov/dashboard)
    - Predicted properties (ACD/Labs, OPERA, TEST)
    - Bioactivity models: Estrogen receptor, Androgen receptor
  - Proof-of-Concept cheminformatics modules (https://www.epa.gov/chemical-research/cheminformatics)
    - Hazard Profiling of chemicals
    - Single and batch prediction of physprop and toxicity endpoints
- Chemical curation is into the DSSTox database

Assembling data is easy. Curation is hard

https://pubs.acs.org/doi/10.1021/acs.jcim.2c00268

- It is very easy to harvest and download massive amounts of data. FAIRness has expanded access…

RETURN TO ISSUE | < PREV **CHEMICAL INFORMATION** NEXT >

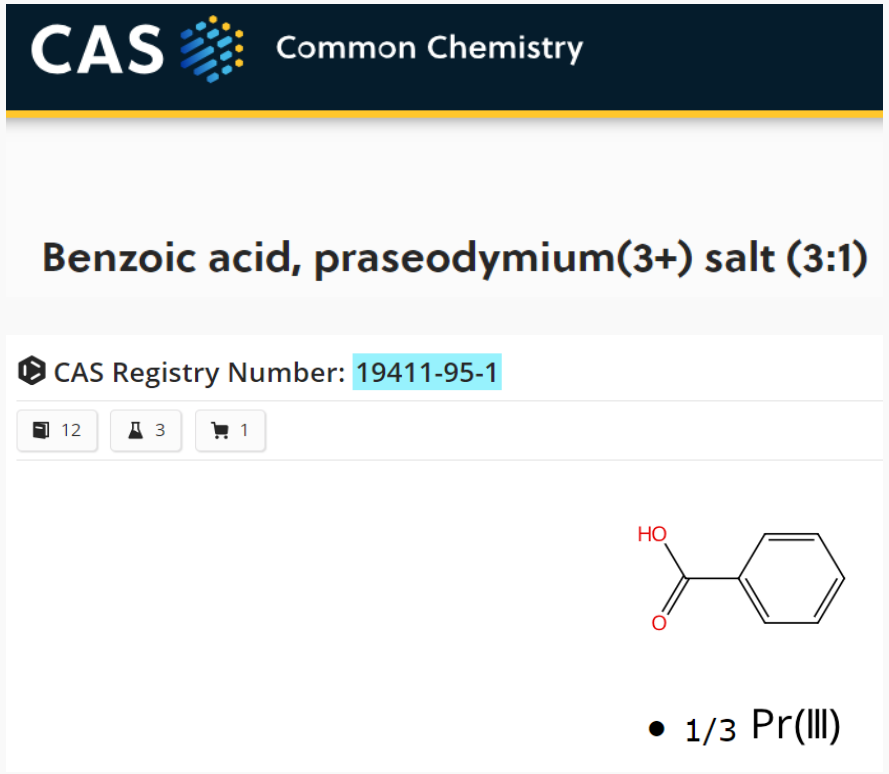
CAS Common Chemistry in 2021: Expanding Access to Trusted Chemical Information for the Scientific Community

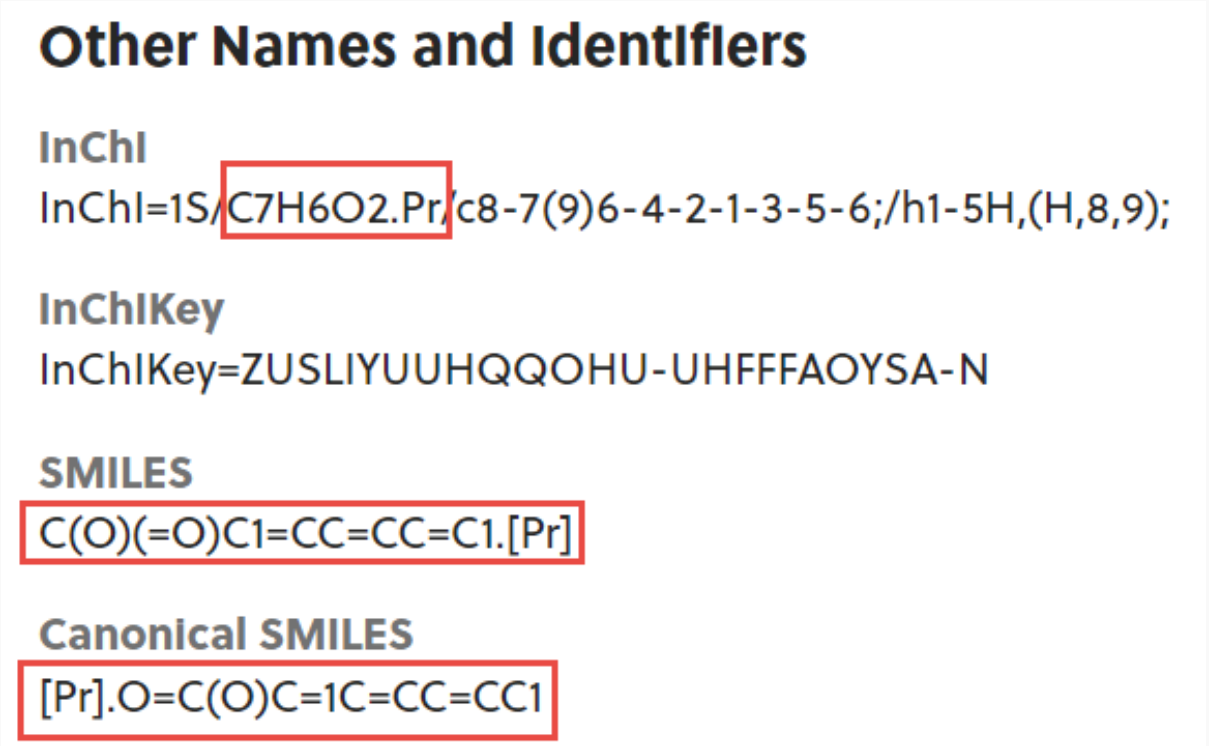
Andrea Jacobs*, Dustin Williams, Katherine Hickey, Nathan Patrick, Antony J. Williams, Stuart Chalk, Leah McEwen, Egon Willighagen, Martin Walker, Evan Bolton, Gabriel Sinclair, and Adam Sanford

- Open API and downloadable dataset – contributing CASRNs, Names and Structures to Open Chemistry

Stoichiometry is important

- SIMPLE example…1 to 3 stoichiometry



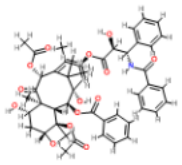


- 1000s of structures with bad stoichiometry into the wild

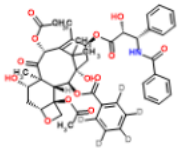
Data Quality issues proliferate
Taxol skeletons (105 CS/202 PubChem)

EPA United States Environmental Protection Agency

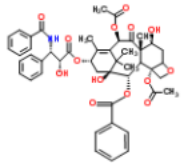
ChemSpider
Search and share chemistry



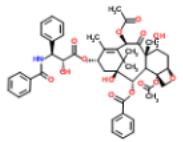
NSC-LSC-1



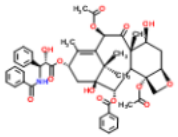
(2alpha,5beta,7beta,10beta,13alpha)-4,10-Diacetoxy-13-{[(2R,3S)-3-(benzoylamino)-2-hydroxy-3-phenylpropanoyl]oxy}-1,7-dihydroxy-



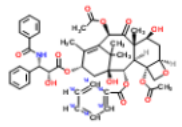
(2alpha,3xi,7beta,10beta,13alpha)-4,10-Diacetoxy-13-{[(2R,3S)-3-(benzoylamino)-2-hydroxy-3-phenylpropanoyl]oxy}-1,7-dihydroxy-



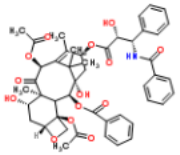
(2alpha,3xi,5beta,7alpha,10beta,13alpha)-4,10-Diacetoxy-13-{[(2R,3S)-3-(benzoylamino)-2-hydroxy-3-phenylpropanoyl]oxy}-1,7-dihydroxy-



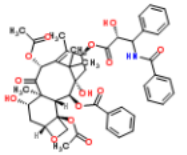
(2alpha,3xi,5beta,7beta,10beta,13alpha)-4,10-Diacetoxy-13-{[(2S,3S)-3-(benzoylamino)-2-hydroxy-3-phenylpropanoyl]oxy}-1,7-dihydroxy-



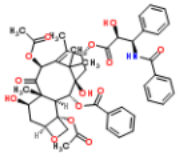
(2alpha,5beta,7beta,10beta,13alpha)-4,10-Diacetoxy-13-{[(2R,3S)-3-(benzoylamino)-2-hydroxy-3-phenylpropanoyl]oxy}-1,7-dihydroxy-



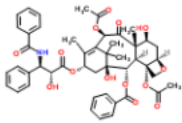
(2alpha,3xi,5beta,7beta,10alpha,13alpha)-4,10-Diacetoxy-13-{[(2R,3S)-3-(benzoylamino)-2-hydroxy-3-phenylpropanoyl]oxy}-1,7-dihydroxy-



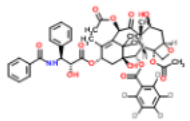
(2alpha,5beta,7beta,10beta,13alpha)-4,10-Diacetoxy-13-{[(2R)-3-(benzoylamino)-2-hydroxy-3-phenylpropanoyl]oxy}-1,7-dihydroxy-



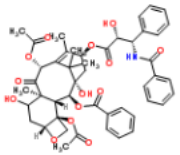
(1beta,2beta,3beta,4alpha,5alpha,7alpha,8alpha Diacetoxy-13-{[(2S,3R)-3-(benzoylamino)-2-hydroxy-3-phenylpropanoyl]oxy}-1,7-dihydroxy-



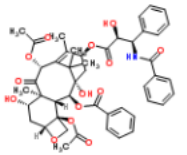
(2alpha,5beta,7beta,10beta,13alpha)-4,10-Diacetoxy-13-{[(2R,3R)-3-(benzoylamino)-2-hydroxy-3-phenylpropanoyl]oxy}-1,7-dihydroxy-



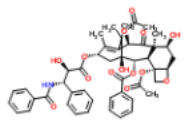
(2alpha,3beta,5beta,7beta,10beta,13alpha)-4,10-Diacetoxy-13-{[(2R,3S)-3-(benzoylamino)-2-hydroxy-3-phenylpropanoyl]oxy}-1,7-dihydroxy-



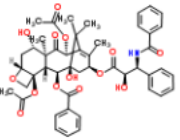
(2alpha,5beta,10beta,13alpha)-4,10-Diacetoxy-13-{[(2R,3S)-3-(benzoylamino)-2-hydroxy-3-phenylpropanoyl]oxy}-1,7-dihydroxy-9-oxo-



(2alpha,5beta,7beta,10beta,13alpha)-4,10-Diacetoxy-13-{[(2S,3R)-3-(benzoylamino)-2-hydroxy-3-phenylpropanoyl]oxy}-1,7-dihydroxy-



(2alpha,5beta,7beta,8alpha,10beta,13alpha)-4,1 Diacetoxy-13-{[(2R,3S)-3-(benzoylamino)-2-hydroxy-3-phenylpropanoyl]oxy}-1,7-dihydroxy-



(1b,2a,5b,7b,10b,13a)-4,10-bis(acetyloxy)-13-{[(2R,3S)-3-(benzoylamino)-2-hydroxy-3-phenylpropanoyl]oxy}-1,7-dihydroxy-9-oxo-

Accessing DSSTox chemistry: CompTox Chemicals Dashboard

- A **publicly accessible website** delivering:
  - 1.2M chemicals with related property data
  - Related substances: transformation products, mono/polymer
  - Experimental/predicted physicochemical property data
  - Experimental Human and Ecological hazard data
  - Integration to “biological assay data” (ToxCast/Tox21)
  - Information regarding chemicals in consumer products
  - Links to other agency websites and public data resources
  - “Batch searching” for tens to thousands of chemicals

Curating Chemistry into the DSSTox Database

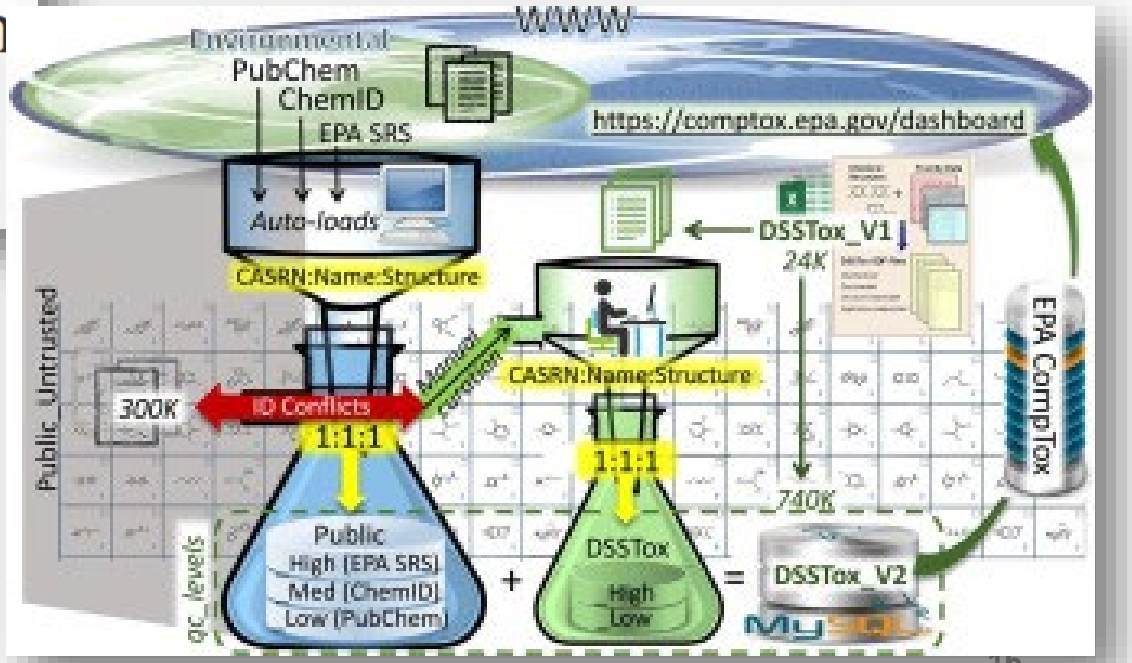
ELSEVIER

Computational Toxicology

Volume 12, November 2019, 100096

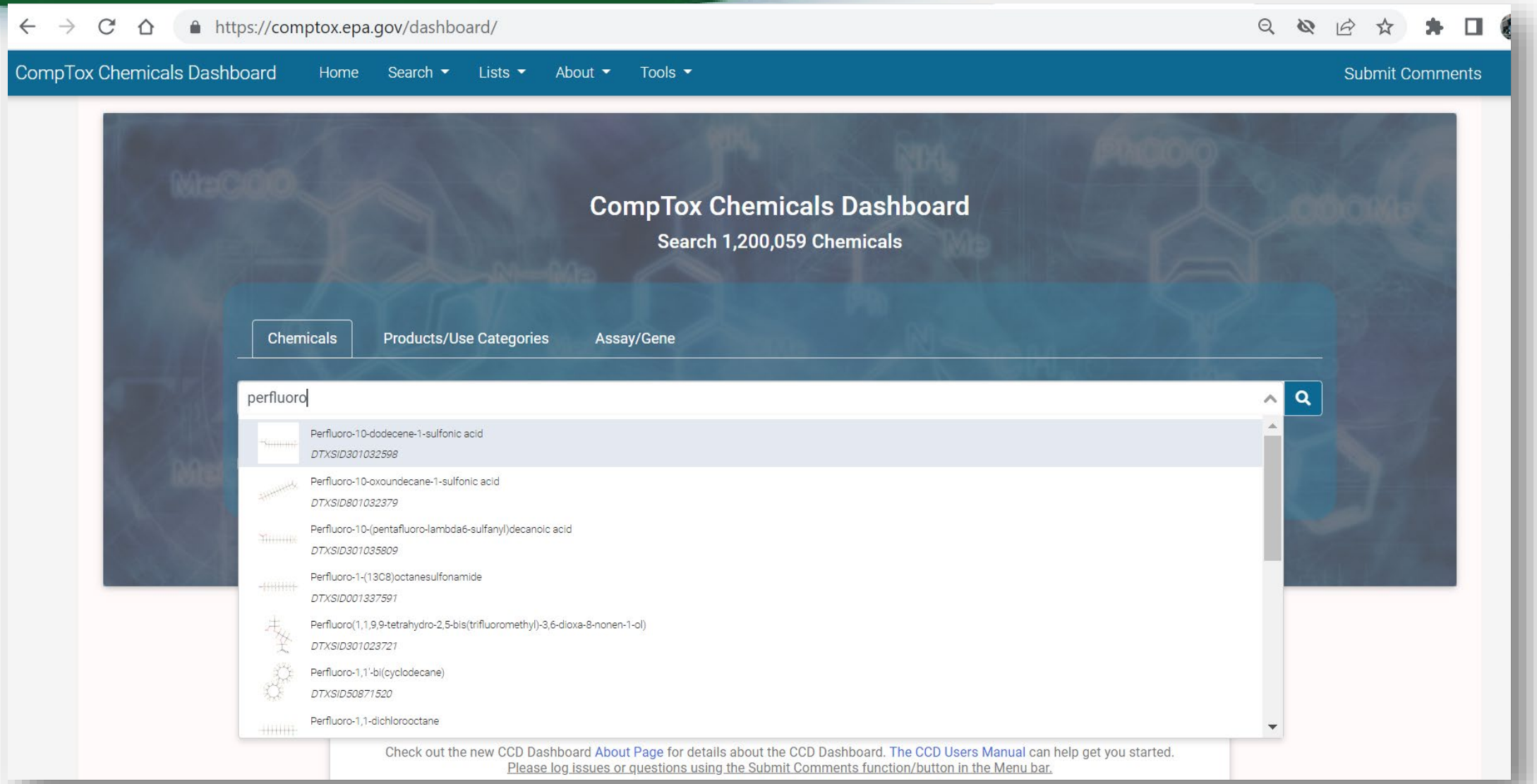
EPA's DSSTox database: History of development of a curated chemistry resource supporting computational toxicology research

Christopher M. Grulke [a], Antony J. Williams [a], Inthirany Thillanadarajah [b], Ann M. Richard [a]



CompTox Chemicals Dashboard

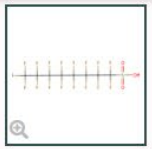
https://comptox.epa.gov/dashboard



1 of ~1.2M Chemical Pages

Experimental/Predicted Properties

CompTox Chemicals Dashboard | Home | Search | Lists | About | Tools | Submit Comments | Search all data



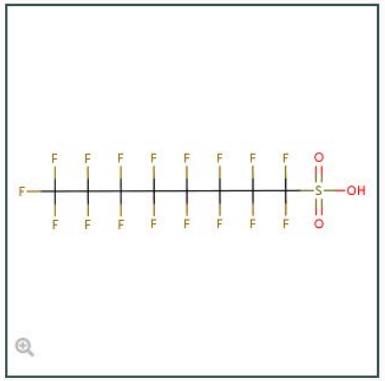
Perfluorooctanesulfonic acid

1763-23-1 | DTXSID3031864

Searched by DTXSID

- Details
- Executive Summary
- Properties
- Env. Fate/Transport
- Hazard
- Safety > GHS Data
- ADME > IVIVE
- Exposure
- Bioactivity
- Similar Compounds
- GenRA
- Related Substances
- Synonyms
- Literature
- Links
- Comments

Chemical Details



Wikipedia

Perfluorooctanesulfonic acid (**PFOS**) (conjugate base **perfluorooctanesulfonate**) is a chemical compound having an eight-carbon fluorocarbon chain and a sulfonic acid functional group and thus a perfluorosulfonic acid. It is an anthropogenic (man-made) fluorosurfactant, now regarded as a global pollutant. PFOS was the key ingredient in Scotchgard, a fabric protector made by 3M, and related stain repellents. The acronym "PFOS" refers to the parent sulfonic acid and to various salts

Read more

Quality Control Notes

Intrinsic Properties

Molecular Formula: $C_8HF_{17}O_3S$ | MOL FILE | FIND ALL CHEMICALS

Average Mass: 500.13 g/mol | ISOTOPE MASS DISTRIBUTION

Monoisotopic Mass: 499.937494 g/mol

Structural Identifiers

Linked Substances

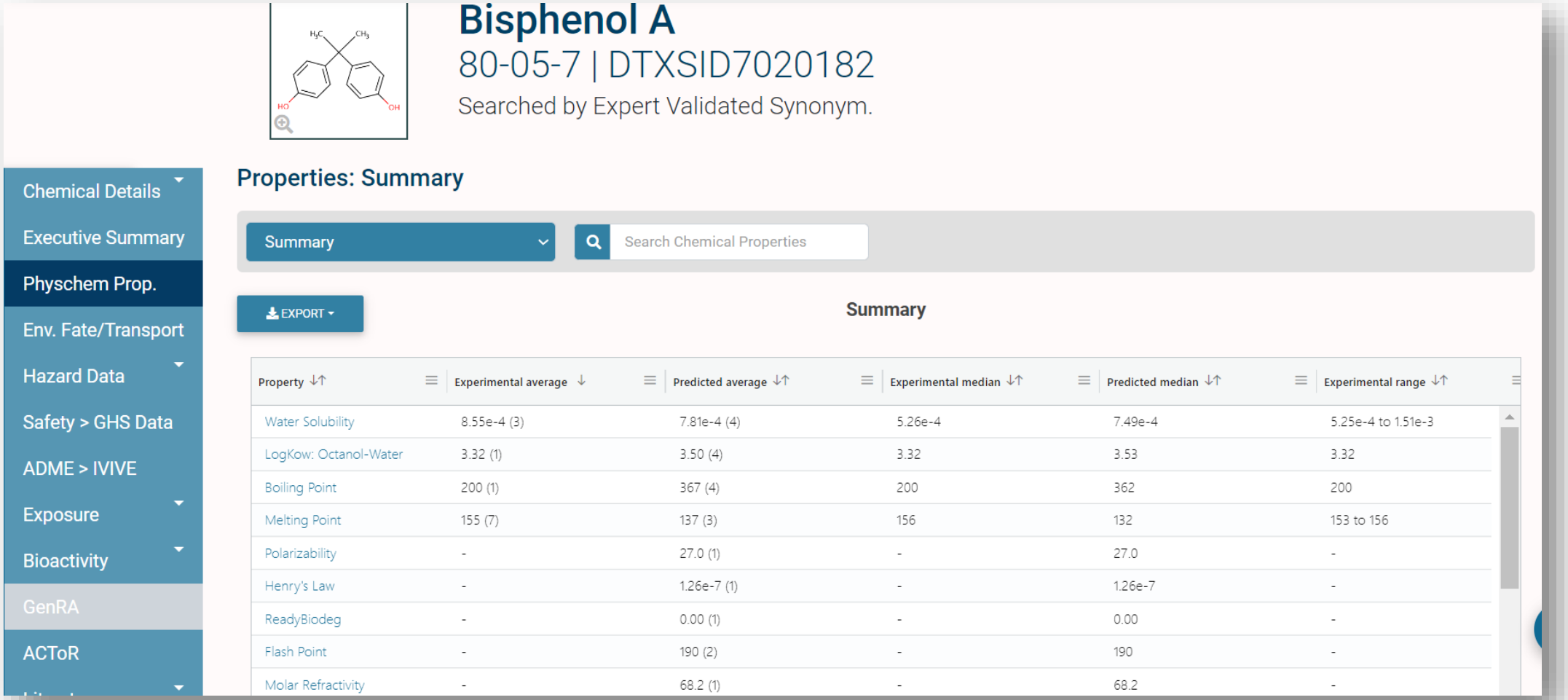
Presence in Lists

Record Information

QSAR Modeled Data are available

- We build models then apply then to our curated datasets for release, PLUS deliver the models for realtime use
- Data can be harvested one chemical at a time or 1000s using the batch search (https://pubs.acs.org/doi/10.1021/acs.jcim.0c01273)

Modeled values are not enough!



Where is all the calculation detail? Are predictions in applicability domain etc?

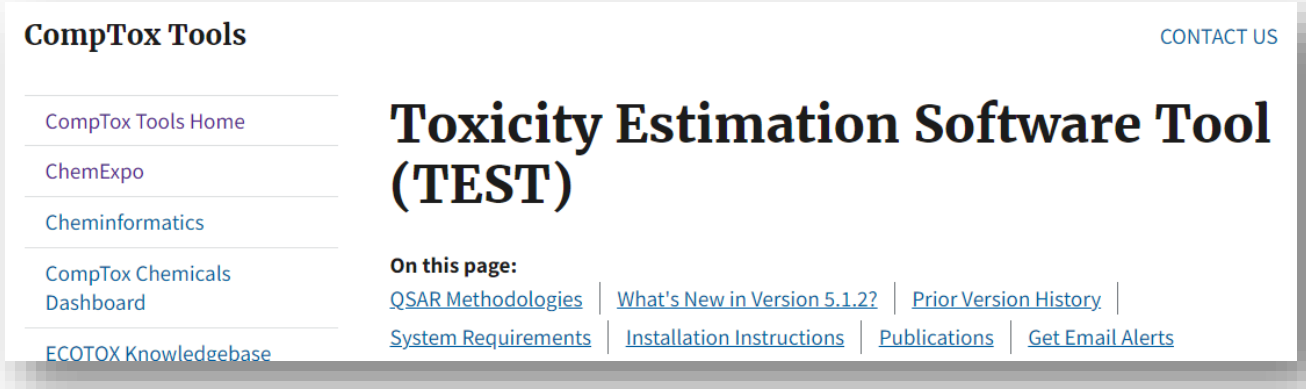
- For OPERA and TEST models we have all the details
  - OPERA https://jcheminf.biomedcentral.com/articles/10.1186/s13321-018-0263-1


RESEARCH ARTICLE Open Access

OPERA models for predicting physicochemical properties and environmental fate endpoints

Kamel Mansouri[1,2,3*], Chris M. Grulke[1], Richard S. Judson[1] and Antony J. Williams[1]

  - TEST https://www.epa.gov/comptox-tools/toxicity-estimation-software-tool-test


CompTox Tools

CONTACT US

CompTox Tools Home
ChemExpo
Cheminformatics
CompTox Chemicals Dashboard
ECOTOX Knowledgebase

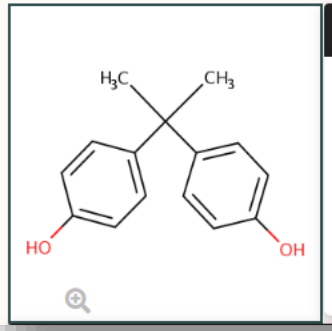
Toxicity Estimation Software Tool (TEST)

On this page:
QSAR Methodologies | What's New in Version 5.1.2? | Prior Version History | System Requirements | Installation Instructions | Publications | Get Email Alerts

OPERA Model Details

OPERA Model Calculation Details: LogKow: Octanol-Water

PRINT PDF



Chemical Identifiers

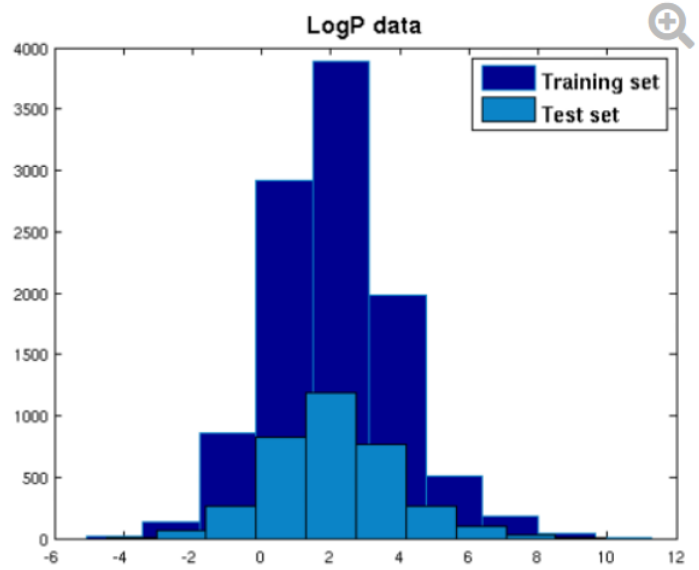
Preferred Name: Bisphenol A
DTXSID: DTXSID7020182
DTXCID: DTXCID30182
CASRN: 80-05-7

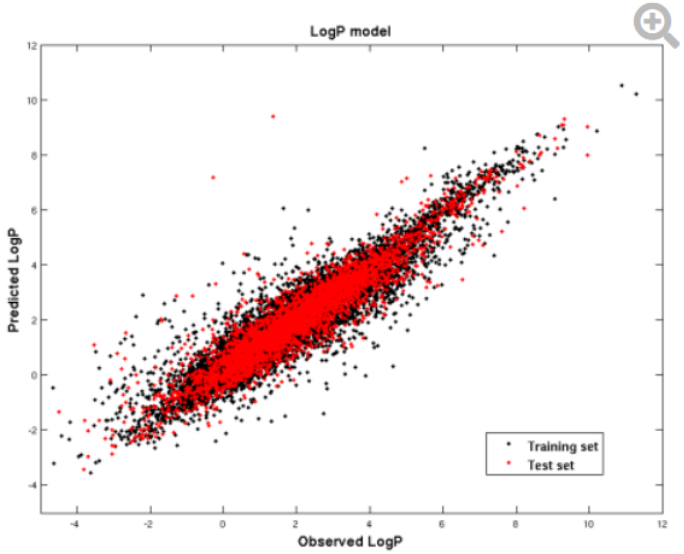
Model Results

Predicted value: 3.32
Experimental value: 3.32
Global applicability domain: Inside
Local applicability domain index: 1.00
Confidence level: 0.938
Opera Version: OPERA 2.6

OPERA Model Details

Model Performance





QMRF

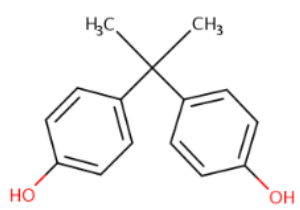
Weighted KNN model

5-fold CV (75%)		Training (75%)		Test (25%)	
Q2	RMSE	RMSE	R2	RMSE	R2
0.850	0.690	0.670	0.860	0.790	0.860

OPERA Model Details

Nearest Neighbors from the Training Set

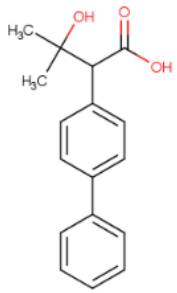
1



Bisphenol A

Measured : 3.32
Predicted : 3.52

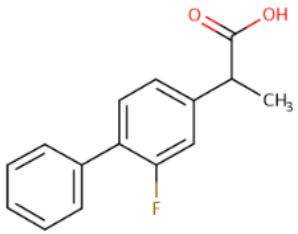
2



BUTANOIC ACID,2-(4-BIPH...

Measured : 3.25
Predicted : 3.75

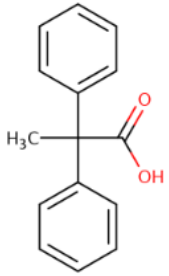
3



Flurbiprofen

Measured : 4.16
Predicted : 3.33

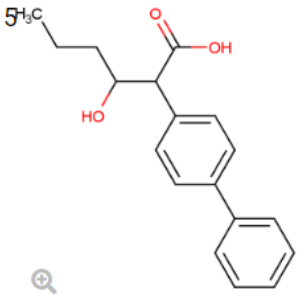
4



2,2-Diphenylpropionic acid

Measured : 2.69
Predicted : 3.43

5



3-OH-2-(4-BIPHENYLYL)H...

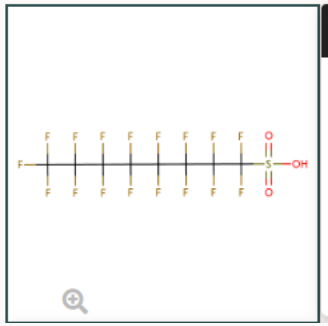
Measured : 3.75
Predicted : 3.78

Why this detail is required

- Predicted Fish Biotrans. Half-Life (Km) of PFOS is 2.7 days

OPERA Model Calculation Details: Fish Biotrans. Half-Life (Km)

PRINT PDF



Chemical Identifiers

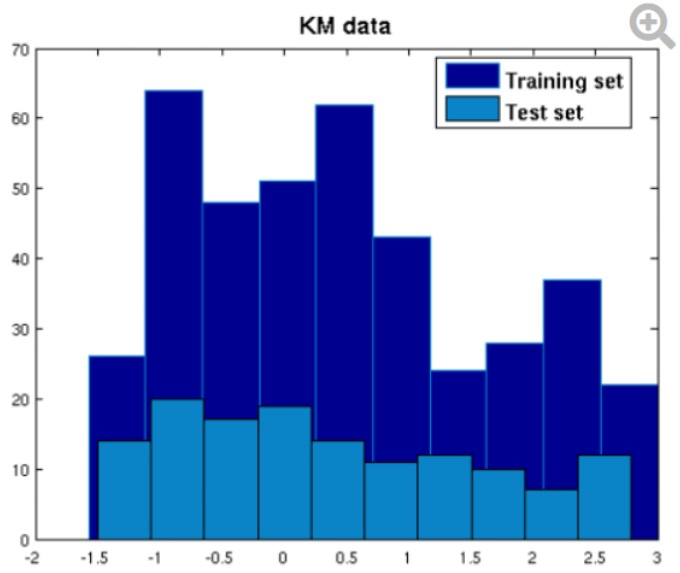
Preferred Name: Perfluorooctanesulfonic acid
DTXSID: DTXSID3031864
DTXCID: DTXCID1011864
CASRN: 1763-23-1

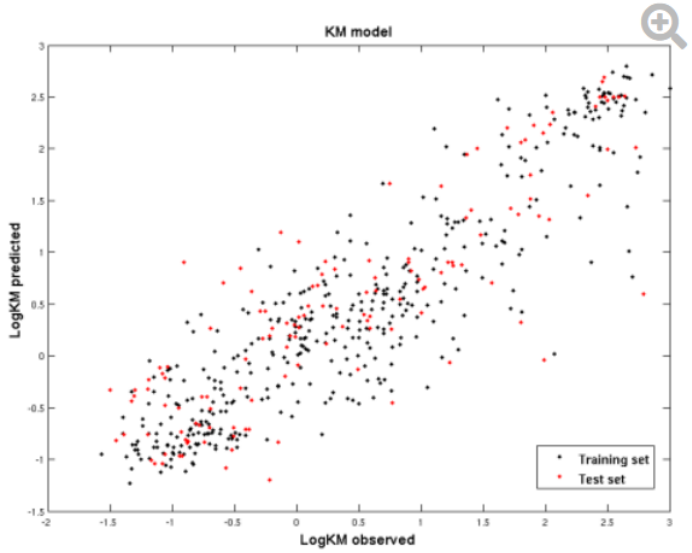
Model Results

Predicted value: 2.69 days
Global applicability domain: Outside
Local applicability domain index: 0.174
Confidence level: 0.145
Opera Version: OPERA 2.6

Why this detail is required

Model Performance





QMRF

Weighted KNN model

5-fold CV (75%)		Training (75%)		Test (25%)	
Q2	RMSE	RMSE	R2	R2	RMSE
0.830	0.490	0.500	0.820	0.730	0.620

QMRF details

https://comptox.epa.gov/dashboard-api/ccdapp1/qmrfdata/file/by-modelid/28

QMRF identifier (JRC Inventory): **Q17-66-0019**

QMRF Title: **OPERA-model for biotransformation rate constant**

Printing Date: **Oct 17, 2017**

1.QSAR identifier

1.1.QSAR identifier (title):

OPERA-model for biotransforma

1.2.Other related models:

No related models

1.3.Software coding the model:

OPERA V1.5

OPERA (OPEn (quantitative) str
open source command line appl
physicochemical properties and
descriptors. It is available for do
license.

Kamel Mansouri (mansourikame
https://github.com/kmansouri/OP

3.7.Endpoint data quality and variability:

The original data collected from the PHYSPROP database (631 chemicals) have undergone a se
the chemical structures and remove duplicates,
erroneous entries. This procedure also included
ensure only good quality data is used for the de
model (548 chemicals).

Then, QSAR-ready structures were generated by
standardizing all chemical structures and remov
and metallo-organic chemicals (541 chemicals).
workflows that were developed for the purpose
standardization of the data are available in the p
Section 2.7].

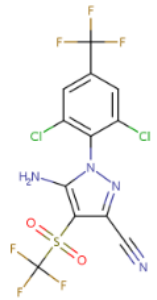
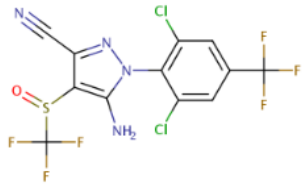
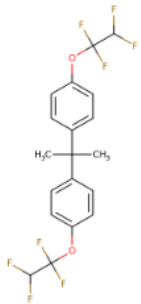
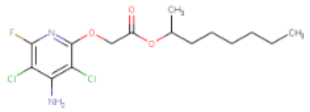
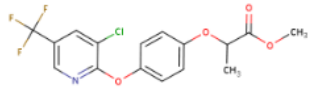
The curated outlier-free experimental data (541
chemicals) was divided into training and validati
machine learning and modeling steps.

4.4.Descriptor selection:

PaDEL software was used to calculate 1440 molecular descriptors. A first filter was applied in order to remove descriptors with missing values, constant and near constant (standard deviation of 0.25 as a threshold) and highly correlated descriptors (96% as a threshold). The remaining 837 descriptors were used in a feature selection procedure to select a minimum number of variables encoding the most relevant structural information to the modeled endpoint. This step consisted of coupling Genetic Algorithms (GA) with the weighted kNN algorithm and was applied in 5 fold cross validation on the training set (405 chemicals). This procedure was run for 200 consecutive independent runs maximizing Q^2 in cross-validation and minimizing the number of descriptors. The number of k neighbors is optimized within the range of 3 to 7. The descriptors were then ranked based on their frequency of selection during the GA runs. The best model showed an optimal compromise between the simplicity (minimum number of descriptors) and performance (Q^2 in cross-validation) to ensure transparency and facilitate the mechanistic interpretation as required by OECD principles 2 and 5. More details in paper [ref2 Section 2.7].

Why this detail is required

Nearest Neighbors from the Training Set

1	2	3	4	5
				
Fipronil sulfone	Fipronil	1,1'-Isopropylidenebis(4-(1...	Fluroxypyr-meptyl	Haloxyfop-methyl
Measured : 1.05 Predicted : -0.25	Measured : -0.21 Predicted : 0.56	Measured : 2.79 Predicted : 0.21	Measured : -0.61 Predicted : 0.05	Measured : -1.08 Predicted : 0.16

TEST Prediction Reports

Predicted Water solubility at 25 Â°C for 80-05-7 from Consensus method

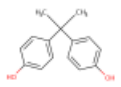
Prediction results

Endpoint	Experimental value (CAS= 80-05-7) Source: EPI Suite v 4.00	Predicted value[a]
Water solubility at 25 Â°C -Log10(mol/L)	3.28	2.90
Water solubility at 25 Â°C mg/L	120.10	284.38

[a]Note: the test chemical was present in the training set. The prediction *does not* represent an external prediction.

Individual Predictions

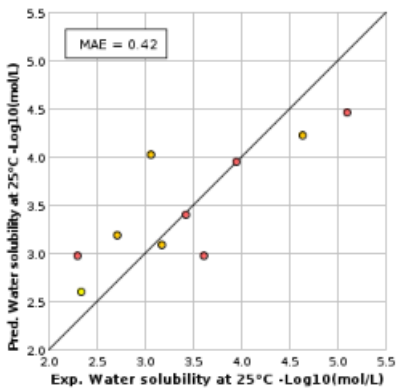
Method	Predicted value -Log10(mol/L)
Hierarchical clustering	3.58
Group contribution	3.10
Nearest neighbor	2.04



Predictions for the test chemical and for the most similar chemicals in the e

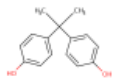
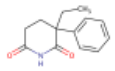
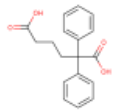
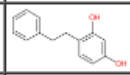
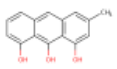
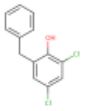
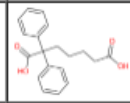
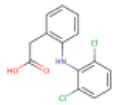
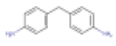
If the predicted value matches the experimental values for similar chemicals in the test set (and t

Prediction results (colors defined in table below)

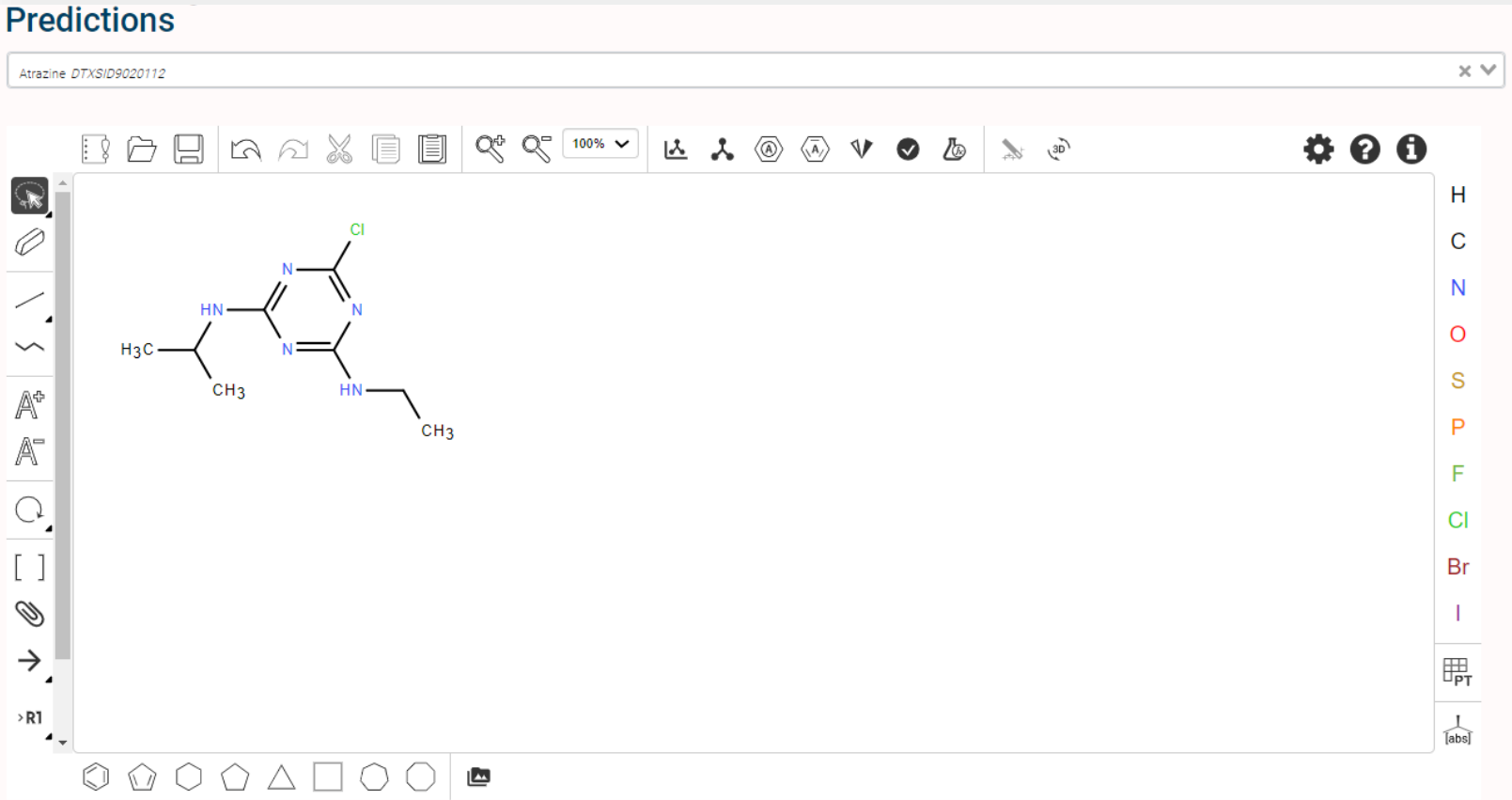


Chemicals	MAE*
Entire set	0.58
Similarity coefficient ≥ 0.5	0.42

*Mean absolute error in -Log10(mol/L)

CAS	Structure	Similarity Coefficient	Experimental value -Log10(mol/L)	Predicted value -Log10(mol/L)
80-05-7 (test chemical)			3.28	2.90
77-21-4		0.72	2.34	2.60
103387-44-6		0.70	3.17	3.09
94-77-9		0.69	2.72	3.19
491-59-8		0.61	3.06	4.03
19578-81-5		0.60	4.64	4.23
110048-81-2		0.60	3.43	3.40
15307-86-5		0.59	5.10	4.47
101-77-9		0.58	2.30	2.97

Access to Real Time Predictions



- Toxicological properties
  - 96 hour fathead minnow LC50
  - 48 hour D. magna LC50
  - 48 hour T. pyriformis IGC50
  - Oral rat LD50
  - Bioconcentration factor
  - Developmental toxicity
  - Ames mutagenicity
  - Estrogen Receptor RBA
  - Estrogen Receptor Binding
- Physical properties
  - Normal boiling point
  - Melting point
  - Flash point
  - Vapor pressure
  - Density
  - Surface tension
  - Thermal conductivity
  - Viscosity
  - Water solubility

Multiple modeling approaches plus Consensus

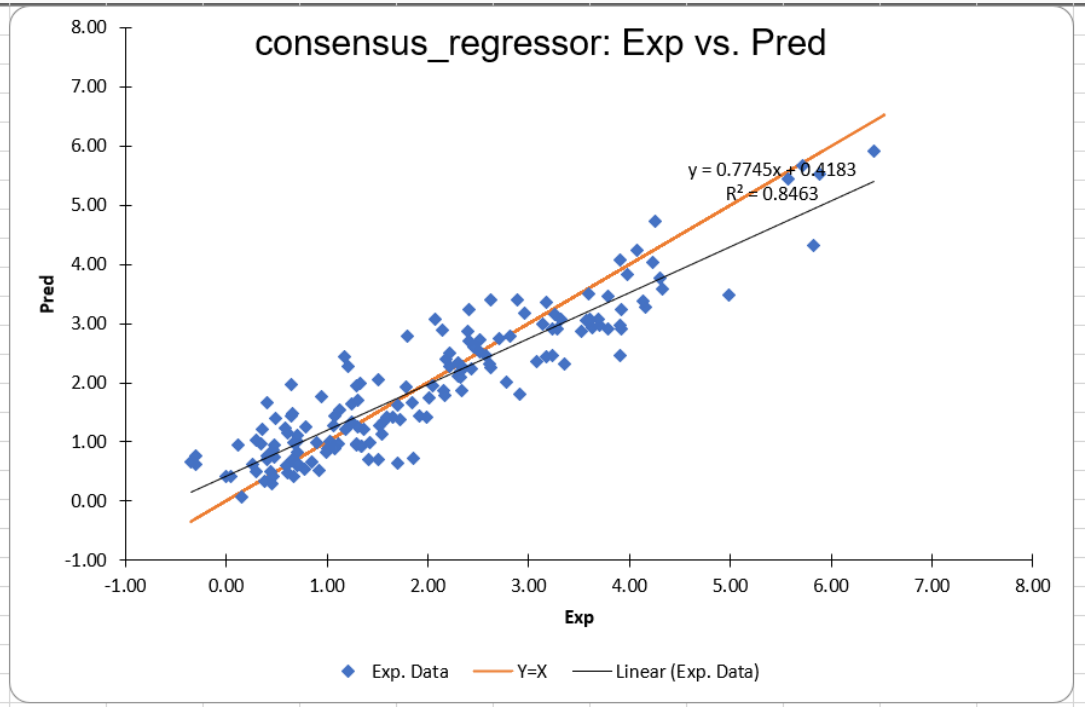
Property	Experimental Value	Consensus	Hierarchical clustering	Single model	Group contribution	Nearest neighbor
96 hour fathead minnow LC50	4.158 -Log10(mol/L) 14.993 mg/L	4.001 -Log10(mol/L) 21.524 mg/L	4.067 -Log10(mol/L) 18.489 mg/L	3.997 -Log10(mol/L) 21.697 mg/L	3.876 -Log10(mol/L) 28.721 mg/L	4.064 -Log10(mol/L) 18.630 mg/L
48 hour D. magna LC50	3.601 -Log10(mol/L) 54.062 mg/L	3.854 -Log10(mol/L) 30.189 mg/L	3.725 -Log10(mol/L) 40.679 mg/L	4.354 -Log10(mol/L) 9.555 mg/L	3.817 -Log10(mol/L) 32.852 mg/L	3.521 -Log10(mol/L) 65.046 mg/L
48 hour T. pyriformis IGC50			3.525 -Log10(mol/L) 64.465 mg/L			
Oral rat LD50	2.506 -Log10(mol/kg) 672.807 mg/kg	2.197 -Log10(mol/kg) 1370.798 mg/kg	2.295 -Log10(mol/kg) 1093.260 mg/kg			2.099 -Log10(mol/kg) 1718.794 mg/kg
Bioconcentration factor	0.557 Log10 3.606	0.950 Log10 8.910	0.837 Log10 6.873	0.985 Log10 9.650	1.004 Log10 10.102	0.974 Log10 9.409
Developmental toxicity		true	true	true		false
Ames mutagenicity	false	false	false			false
Estrogen Receptor RBA						
Estrogen Receptor Binding	false	false	false	false	false	
Normal boiling point		312.2 °C	310.7 °C		341.5 °C	284.3 °C
Melting point	173.0 °C	163.6 °C	180.6 °C		106.2 °C	204.0 °C

Rows: **18** Total Rows: **18**

Coming Soon:
Excel report for models for each data set

- Cover sheet with model metadata
- Training and test set statistics
- Training and test set statistics
- Prediction results for each method

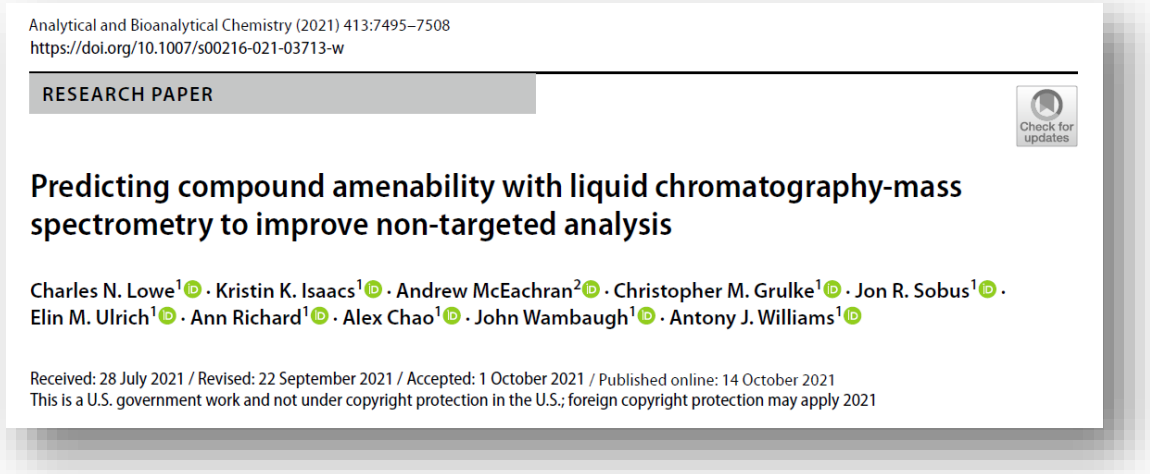
	Canonical QSAR Ready Smiles	Exp	Pre	Erro
2	C1=C2C=CC=CC2=CC2C=CC=CC=21	3.26	3.16	0.10
3	CCCCCCC1=CC=C(C=C1)C1=CC=C(C=C1)C#N	3.52	2.87	0.65
4	BrC1=CC=C(C=C1)C1=CC=C(Br)C=C1	3.92	3.24	0.68
5	NC1=C2C(=O)C3=C(C=CC=C3)C(=O)C2=C(O)C=C1OC1=CC=C(Br)C=C1	2.05	1.95	0.10
6	C1COCCN1	0.45	0.30	0.15
7	CC1CC(C)(C)CC(=O)C=1	0.30	1.02	0.72
8	OC1C=CC(CC2C=CC=CC=2)=CC=1	1.33	2.00	0.67
9	CC(C)OC1=CC(=CC=C1)NC(=O)C1C=CC=CC=1C(F)(F)F	1.29	1.95	0.66
10	NC1=C2C=CC=CC2=C(C=C1)S(O)(=O)=O	0.78	0.54	0.24
11	OC(C1C=CC=CC=1)C1C=CC(Cl)=CC=1	1.79	1.93	0.14
12	OC1=CC(=CC2=CC(NC(=O)NC3=CC=C4C(O)=CC(=CC4=C3)S(O)(=O)=O)=CC=C12)S(O)(=O)=O	0.26	0.63	0.37
13	CC1C=CC=CC=1O	1.03	1.01	0.02
14	ClC1=CC(C2C=CC=CC=2Cl)=C(Cl)C=C1	4.23	4.04	0.19
15	ClC1=C2C(=O)C3=C(C=CC=C3)C(=O)C2=CC=C1	2.30	2.11	0.19
16	CC1=C(C=CC=C1Cl)[N+]([O-])=O	1.99	1.41	0.58
17	ClC(Cl)Cl	1.11	0.98	0.13
18	[O-][N+](=O)C1=CC=CC(Cl)=C1Cl	2.16	1.87	0.29
19	OC(=O)C1C=C(Cl)C(Cl)=CC=1	0.36	1.21	0.85
20	NC(N)=S	0.30	0.51	0.21
21	BrC1=CC(Br)=C(OC2=C(Br)C(Br)=C(Br)C(Br)=C2Br)C(Br)=C1	5.72	5.68	0.04
22	COCN(C(=O)CCl)C1C(CC)=CC=CC=1CC	1.12	1.54	0.42
23	CC(C)(CO)CO	0.67	0.42	0.25
24	CC(C)(C)C1=CC(N2N=C3C=CC(Cl)=CC3=N2)=C(O)C(=C1)C(C)(C)C	3.57	3.05	0.52
25	BrC1=CC(Br)=C(OC2=C(Br)C=CC=C2Br)C=C1	5.88	5.52	0.36
26	CCC(=O)NC1C=C(Cl)C(Cl)=CC=1	1.84	1.67	0.17
27	OC(C1C=CC(Cl)=CC=1)(C1C=CC(Cl)=CC=1)C(Cl)(Cl)Cl	3.79	2.92	0.87
28	COC1=C(Cl)C(Cl)=C(Cl)C(Cl)=C1Cl	4.16	3.27	0.89
29	FC(F)(Cl)C(F)(Cl)Cl	1.70	1.62	0.08



Summary sheet | Training | Test | svm_regressor_1.1 | rf_regressor_1.1 | xgb_regressor_1.0 | consensus_regressor

Where do we use predictions like this?

- Models are used in many places in our computational toxicology research
- They are used in the analytical labs to help guide non-targeted analysis



Analytical and Bioanalytical Chemistry (2021) 413:7495–7508
https://doi.org/10.1007/s00216-021-03713-w

RESEARCH PAPER

Check for updates

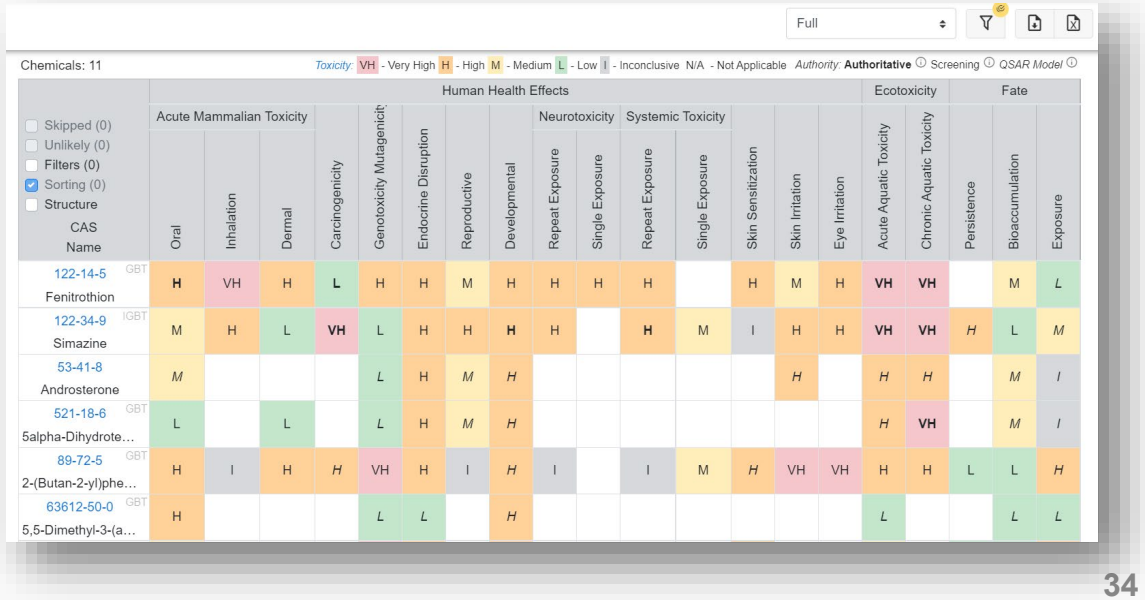
Predicting compound amenability with liquid chromatography-mass spectrometry to improve non-targeted analysis

Charles N. Lowe[1] · Kristin K. Isaacs[1] · Andrew McEachran[2] · Christopher M. Grulke[1] · Jon R. Sobus[1] · Elin M. Ulrich[1] · Ann Richard[1] · Alex Chao[1] · John Wambaugh[1] · Antony J. Williams[1]

Received: 28 July 2021 / Revised: 22 September 2021 / Accepted: 1 October 2021 / Published online: 14 October 2021

Where do we use predictions like this?

- Models are used in many places in our computational toxicology research
- They are used in the analytical labs to help guide non-targeted analysis
- By stakeholders for Hazard profiling of chemicals



Where do we use predictions like this?

- Models are used in many places in our computational toxicology research
- They are used in the analytical labs to help guide non-targeted analysis
- By stakeholders for Hazard profiling of chemicals
- Predictions for breakdown products in the environment

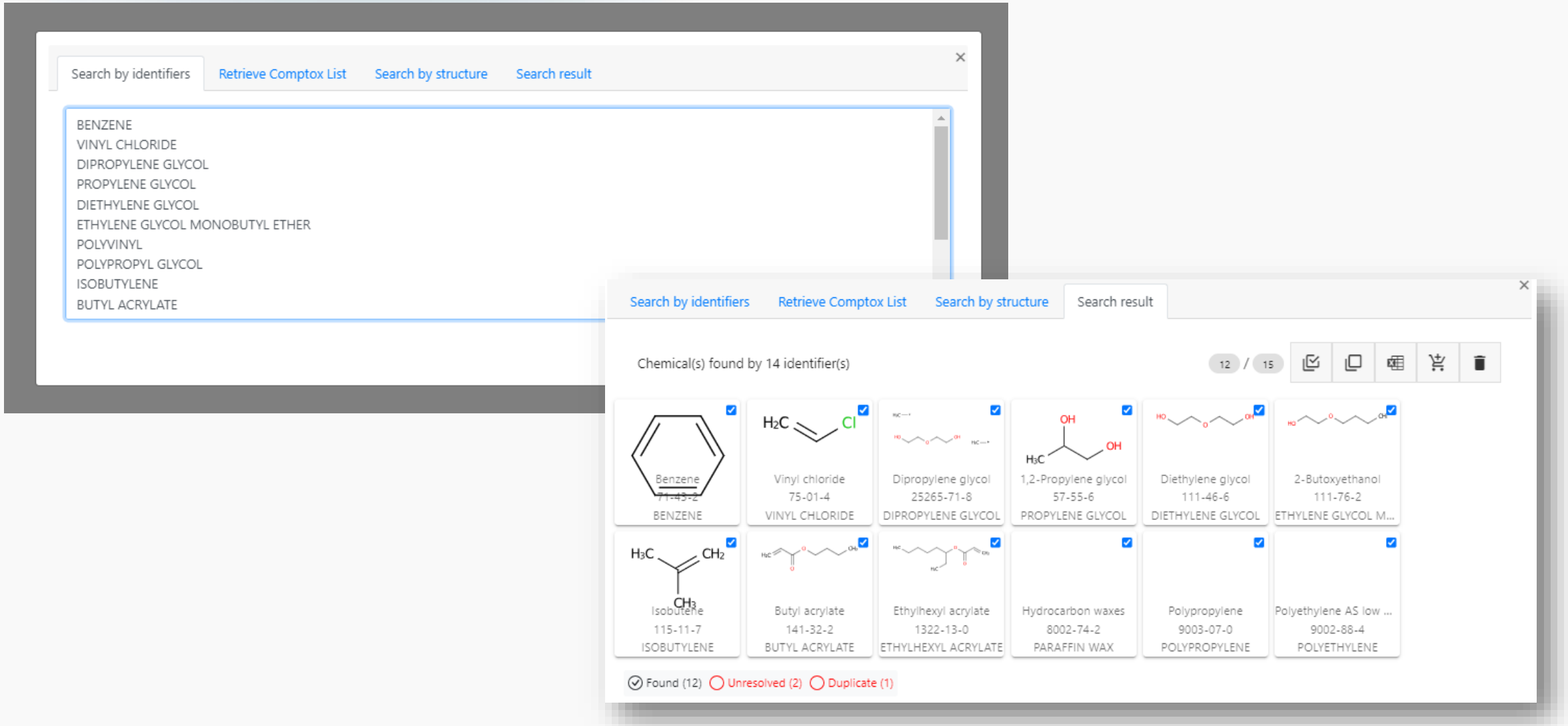
Where can our tools be applied

- Emergency Response utility is obvious…
- Consider East Palestine

https://www.cleveland19.com/2023/02/14/ntsb-announces-preliminary-malfunction-that-caused-east-palestine-train-derailment/

POLYPROPYLENE
POLYETHYLENE
Residue lube oil
VINYL CHLORIDE
DIPROPYLENE GLYCOL
PROPYLENE GLYCOL
DIETHYLENE GLYCOL
COMBUSTIBLE LIQ., NOS (ETHYLENE GLYCOL MONOBUTYL ETHER)
SEMOLINA
COMBUSTIBLE LIQ., NOS (ETHYLHEXYL ACRYLATE)
POLYVINYL
PETROLEUM LUBEOIL
POLYPROPYL GLYCOL
ISOBUTYLENE
BUTYL ACRYLATES, STABILIZED
PETRO OIL, NEC
ADDITIVES, FUEL
BALLS,CTN,M EDCL
SHEET STEEL
VEGTABLE, FROZEN
BENZENE
PARAFFIN WAX
FLAKES, POWDER
HYDRAULIC CEMENT
AUTOS PASSENGER
MALT LIQUORS

Hazard Comparison Profiling



Hazard Comparison Profiling

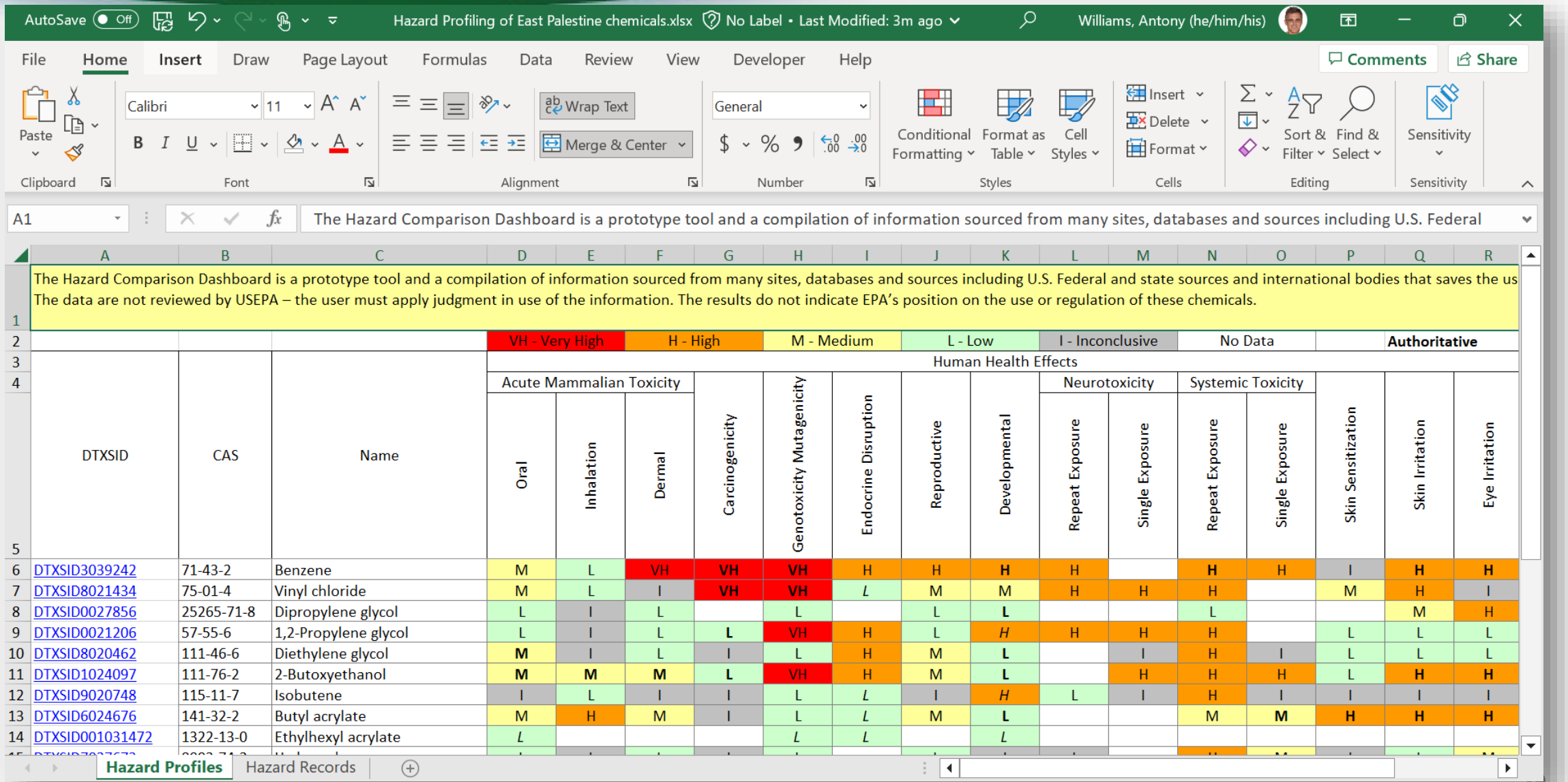
Chemicals: 12

Toxicity: VH - Very High H - High M - Medium L - Low I - Inconclusive N/A - Not Applicable *Authority:* **Authoritative** Screening *QSAR Model*

Skipped (0)
Unlikely (0)
Filters (0)
Sorting (0)
Structure

CAS	Name		Human Health Effects															Ecotoxicity		Fate		
			Acute Mammalian Toxicity							Neurotoxicity		Systemic Toxicity										
			Oral	Inhalation	Dermal	Carcinogenicity	Genotoxicity Mutagenicit	Endocrine Disruption	Reproductive	Developmental	Repeat Exposure	Single Exposure	Repeat Exposure	Single Exposure	Skin Sensitization	Skin Irritation	Eye Irritation	Acute Aquatic Toxicity	Chronic Aquatic Toxicity	Persistence	Bioaccumulation	Exposure
71-43-2	Benzene	AIGBT	M	L	VH	**VH**	**VH**	H	H	**H**	H		**H**	H	I	**H**	**H**	**H**	**M**	*M*	H	*H*
75-01-4	Vinyl chloride	AIGBT	M	L	I	**VH**	**VH**	*L*	M	M	H	H	H		M	H	I	M	**VH**	*M*	L	*VH*
25265-71-8	Dipropylene glycol	GBT	L	I	L		L		L	**L**			L			M	H	L			L	
57-55-6	1,2-Propylene gly…	AIGBT	L	I	L	**L**	VH	H	L	*H*	H	H	H		L	L	L	**L**	**L**	*L*	L	*VH*
111-46-6	Diethylene glycol	GBTM	**M**	I	L	I	L	H	M	**L**		I	H	I	L	L	L	**L**	L	*L*	L	*VH*
111-76-2	2-Butoxyethanol	AIGBT	**M**	**M**	**M**	**L**	VH	H	M	**L**		H	H	H	L	**H**	**H**	**L**	L	*L*	L	*VH*
115-11-7	Isobutene	GBTM	I	L	I	I	L	*L*	I	*H*	L	I	H	I	I	I	I	L		*M*	L	*VH*
141-32-2	Butyl acrylate	GBTM	M	H	M	I	L	*L*	M	**L**			M	**M**	**H**	**H**	**H**	H	H	*L*	L	*H*
1322-13-0	Ethylhexyl acrylate		*L*				*L*	*L*		*L*								*VH*			*I*	
8002-74-2	Hydrocarbon waxes	GBT	L	I	L	I	L		L	I	I		H	M	I	L	M	L	I		L	
9003-07-0	Polypropylene	GBT				**I**												L			L	
9002-88-4	Polyethylene AS I…	GBT		M		**I**												L			L	

Data to Excel in <60s



AutoSave Off | Hazard Profiling of East Palestine chemicals.xlsx | No Label • Last Modified: 3m ago | Williams, Antony (he/him/his)

File Home Insert Draw Page Layout Formulas Data Review View Developer Help | Comments | Share

Paste | Calibri | 11 | Wrap Text | General | Merge & Center | Conditional Formatting | Format as Table | Cell Styles | Insert | Delete | Format | Sort & Filter | Find & Select | Sensitivity

Clipboard | Font | Alignment | Number | Styles | Cells | Editing | Sensitivity

A1 | The Hazard Comparison Dashboard is a prototype tool and a compilation of information sourced from many sites, databases and sources including U.S. Federal

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	The Hazard Comparison Dashboard is a prototype tool and a compilation of information sourced from many sites, databases and sources including U.S. Federal and state sources and international bodies that saves the us The data are not reviewed by USEPA – the user must apply judgment in use of the information. The results do not indicate EPA's position on the use or regulation of these chemicals.																	
2				VH - Very High		H - High		M - Medium		L - Low		I - Inconclusive		No Data			Authoritative	
3				Human Health Effects														
4				Acute Mammalian Toxicity								Neurotoxicity		Systemic Toxicity				
5	DTXSID	CAS	Name	Oral	Inhalation	Dermal	Carcinogenicity	Genotoxicity Mutagenicity	Endocrine Disruption	Reproductive	Developmental	Repeat Exposure	Single Exposure	Repeat Exposure	Single Exposure	Skin Sensitization	Skin Irritation	Eye Irritation
6	DTXSID3039242	71-43-2	Benzene	M	L	VH	VH	VH	H	H	H	H		H	H	I	H	H
7	DTXSID8021434	75-01-4	Vinyl chloride	M	L	I	VH	VH	*L*	M	M	H	H	H		M	H	I
8	DTXSID0027856	25265-71-8	Dipropylene glycol	L	I	L		L		L	L			L			M	H
9	DTXSID0021206	57-55-6	1,2-Propylene glycol	L	I	L	L	VH	H	L	*H*	H	H	H		L	L	L
10	DTXSID8020462	111-46-6	Diethylene glycol	M	I	L	I	L	H	M	L		I	H	I	L	L	L
11	DTXSID1024097	111-76-2	2-Butoxyethanol	M	M	M	L	VH	H	M	L		H	H	H	L	H	H
12	DTXSID9020748	115-11-7	Isobutene	I	L	I	I	L	*L*	I	*H*	L	I	H	I	I	I	I
13	DTXSID6024676	141-32-2	Butyl acrylate	M	H	M	I	L	*L*	M	L			M	M	H	H	H
14	DTXSID001031472	1322-13-0	Ethylhexyl acrylate	*L*				*L*	*L*		*L*							

Hazard Profiles | Hazard Records

Perfect Example of FAIR Data and APIs

- We owe a lot to FAIR data and availability of information
- We curate a lot of our chemistry data using public resources such as PubChem, ChEBI, Common Chemistry and others
- The availability of Public APIs takes things to another level!
- We have been using the PubChem API to harvest data so we can build new applications, like the Safety Module

Cheminformatics Safety Module
Integrate multiple data streams…

The CompTox API is now public
https://api-ccte.epa.gov/docs/index.html

Computational Toxicology and Exposure Data APIs

EPA's computational toxicology research efforts evaluate the potential health effects of thousands of chemicals. The process of evaluating potential health effects involves generating data that investigates the potential harm, or hazard of a chemical, the degree of exposure to chemicals as well as the unique chemical characteristics.

The APIs provided by EPA enable users to extract specific data from various databases and integrate them into their applications. These data are also available for download on our Data Download page

As part of EPA's commitment to share data, all of the computational toxicology data is publicly available for anyone to access and use. EPA's computational toxicology data is considered "open data", and thus all of the data are free of all copyright restrictions, and fully and freely available for both non-commercial and commercial use.

Limited Access APIs

An API Key is needed to access these APIs. Each user will need a specific key for each application. Please send an email to request an API key.

Chemical APIs

Access APIs for searching chemicals, files for chemical structures, and chemical details.

Documentation

Hazard APIs

Access APIs for human and ecotoxicology data.

Documentation

Bioactivity APIs

Access APIs for chemical bioactivity data.

Documentation

Conclusions

- Modeling is essential to our research efforts – we have models covering dozens of endpoints and continuing to expand
- Careful data assembly and curation is required
- OECD Modeling principles are guiding our future modeling
- CompTox Chemicals Dashboard provides public access to our various curated data streams, pre-predicted data and realtime prediction

Acknowledgments

- Modeling Team – Dan Chang, Nate Charest, Charlie Lowe, Todd Martin
- Valery Tkachenko – Cheminformatics Module
- Our DSSTox curation team
- Our SCDCD colleagues and DevOps team