http://www.orcid.org/0000-0002-2668-4821

CompTox Chemicals Dashboard providing access to experimental and predicted environmental fate and transport data

Antony Williams[1], *Chris Grulke*[1], *Kamel Mansouri*[2] *and Todd Martin*[3]

1) *NCCT, U.S. Environmental Protection Agency, RTP, NC*
2) *Integrated Laboratory Systems, Research Triangle Park, NC.*
3) *NRMRL, U.S. Environmental Protection Agency, Cincinnati, OH*

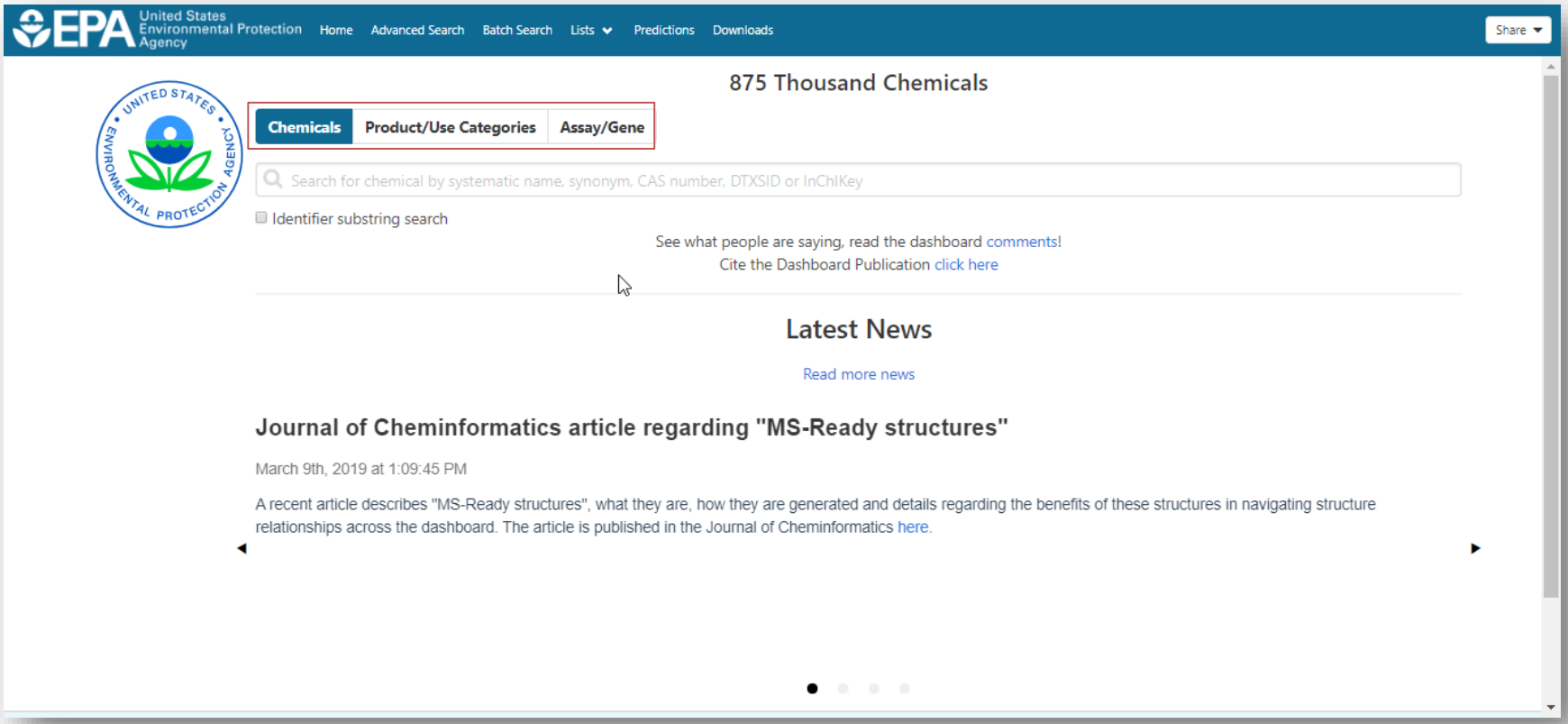
Fall 2019
ACS Fall Meeting, San Diego

Overview

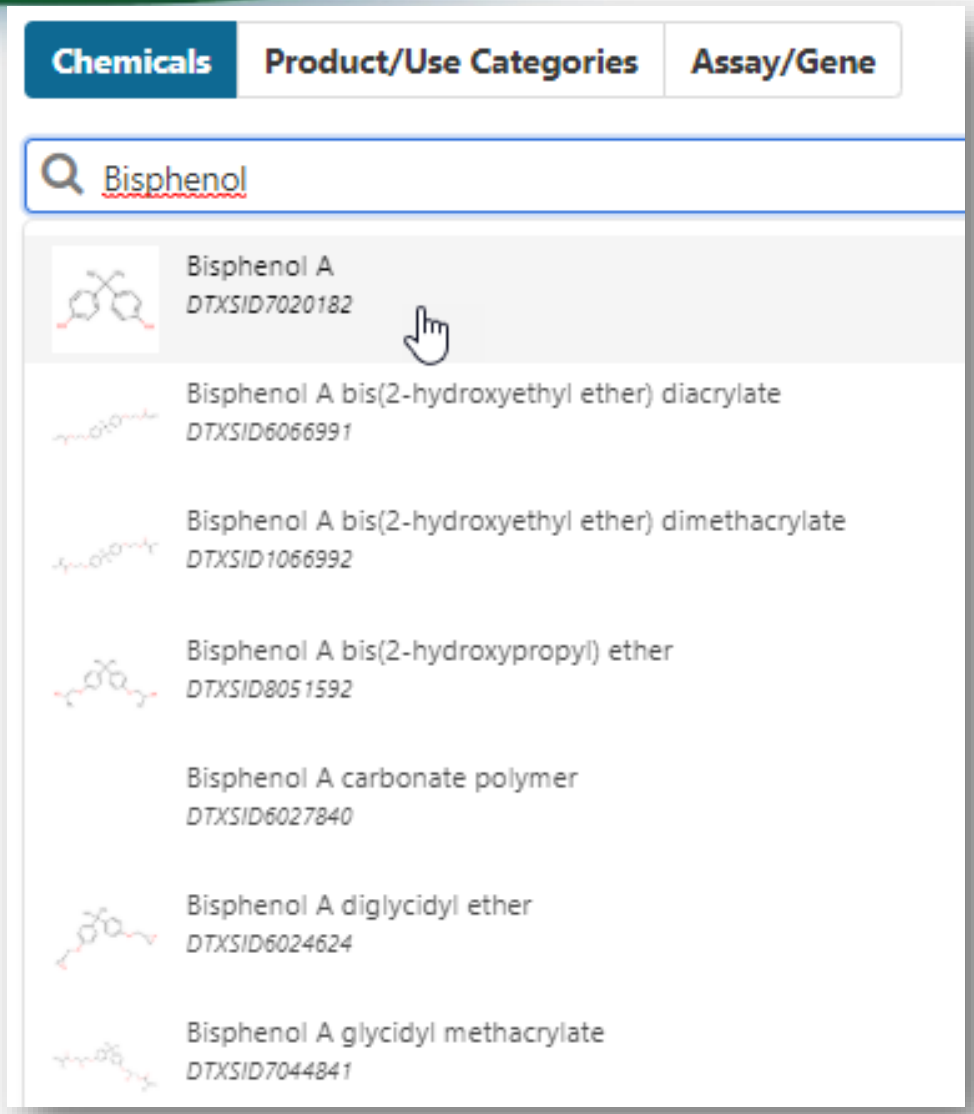
- The CompTox Chemicals Dashboard - web-based database of 875k substances
- Associated data including:
  - *In vivo* hazard data
  - *In vitro* bioactivity screening data
  - Link farm to tens of public resources
- Includes experimental and predicted physchem and experimental fate and transport data
- Access to real-time predictions
- A quick overview of capabilities…

CompTox Chemicals Dashboard

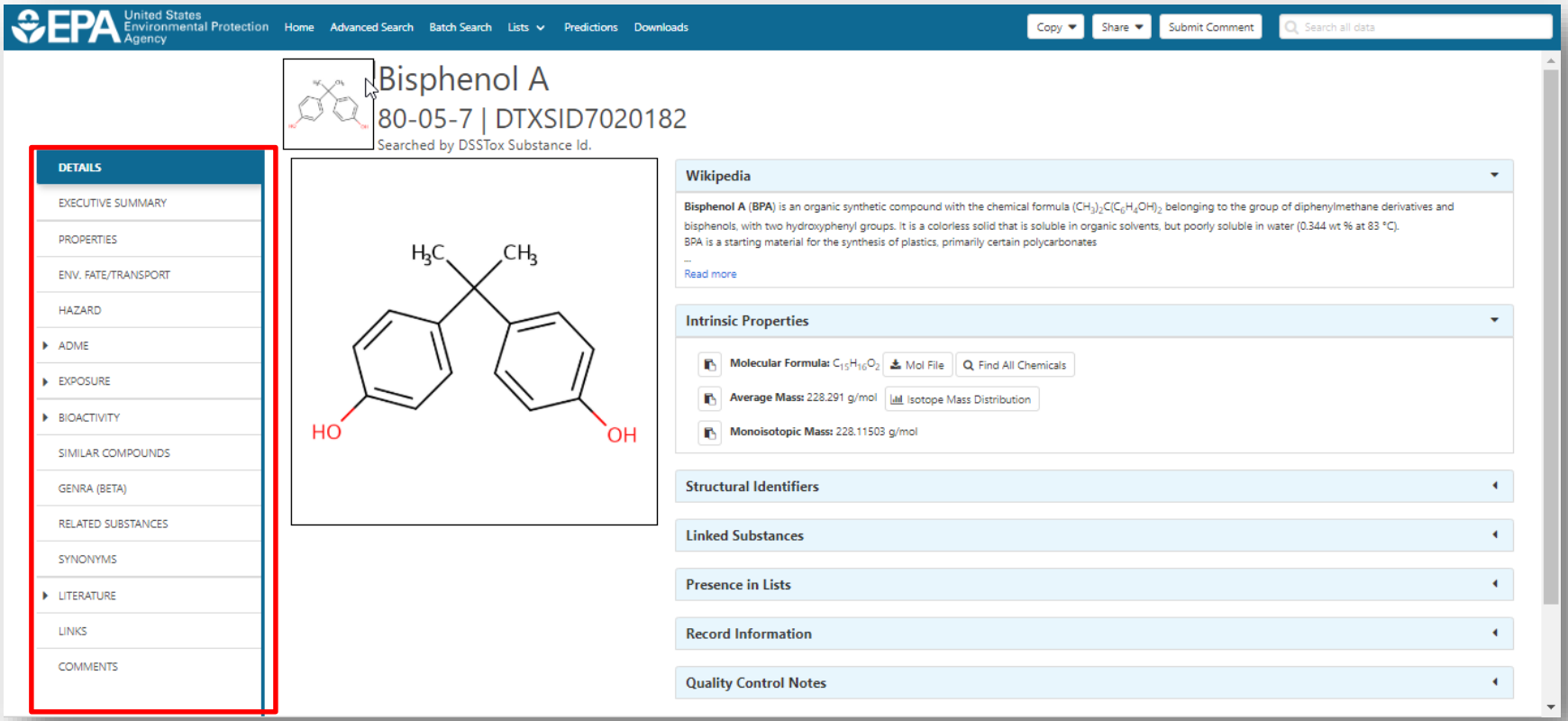
https://comptox.epa.gov/dashboard



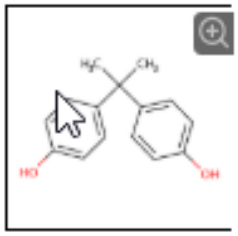
BASIC Search



Detailed Chemical Pages



Experimental and Predicted Data



Bisphenol A

80-05-7 | DTXSID7020182

Searched by DSSTox Substance Id.

Property

Summary

Download Columns

Property	Experimental average	Predicted average
LogP: Octanol-Water	3.32 (1)	3.29
Melting Point	155 (7)	139
Boiling Point	200 (1)	363
Water Solubility	5.26e-4 (1)	9.62e-4
Vapor Pressure	-	8.37e-7
Flash Point	-	190

Prediction models and transparency

EPA United States Environmental Protection Agency

Predicted

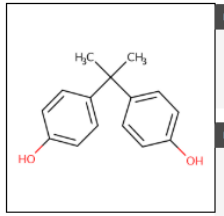
Download Predicted Data

Source	Result	Calculation Details	QMRF
EPISUITE	3.64	Not Available	Not Available
NICEATM	2.40	Not Available	Available
ACD/Labs Consensus	3.63	Not Available	Not Available
ACD/Labs	3.43	Not Available	Not Available
OPERA	3.35	OPERA Model Report [Inside AD]	Available

OPERA Models: LogP: Octanol-Water

Bisphenol A

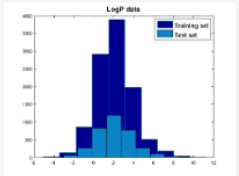
80-05-7 | DTXSID7020182

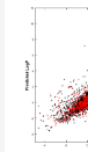


Model Results

Predicted value: 3.35
Global applicability domain: Inside
Local applicability domain index: 0.877
Confidence level: 0.813

Model Performance



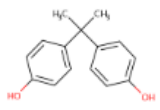


QMRF

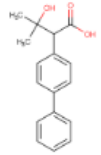
Weighted KNN model

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

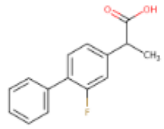
Nearest Neighbors from the Training Set



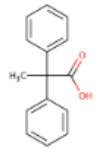
Bisphenol A
Measured:3.32
Predicted: 3.35076



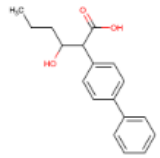
BUTANOIC ACID,2-(4-BIPHENYLYL)-3-HYDROXY-3-METHY
Measured:3.25
Predicted: 3.39062



Flurbiprofen
Measured:4.16
Predicted: 3.94445



2,2-Diphenylpropionic acid
Measured:2.69
Predicted: 2.84603



3-OH-2-(4-BIPHENYLYL)HEXANOIC ACID
Measured:3.75
Predicted: 3.70322

Curated data

EPA United States Environmental Protection Agency

An automated curation procedure for addressing chemical errors and inconsistencies in public datasets used in QSAR modelling

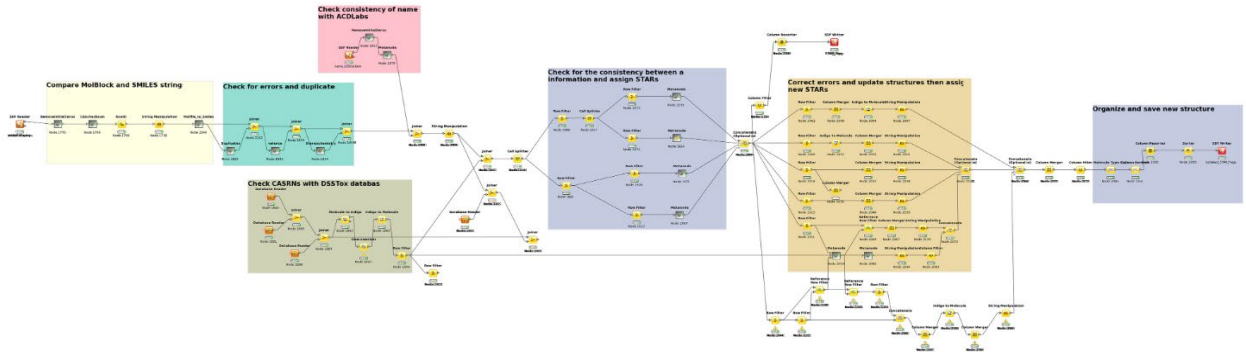
K. Mansouri, C. M. Grulke, A. M. Richard, R. S. Judson & A. J. Williams

To cite this article: K. Mansouri, C. M. Grulke, A. M. Richard, R. S. Judson & A. J. Williams (2016) An automated curation procedure for addressing chemical errors and inconsistencies in public datasets used in QSAR modelling, SAR and QSAR in Environmental Research, 27:11, 911-937, DOI: 10.1080/1062936X.2016.1253611

To link to this article: https://doi.org/10.1080/1062936X.2016.1253611

Figure 8 of 14

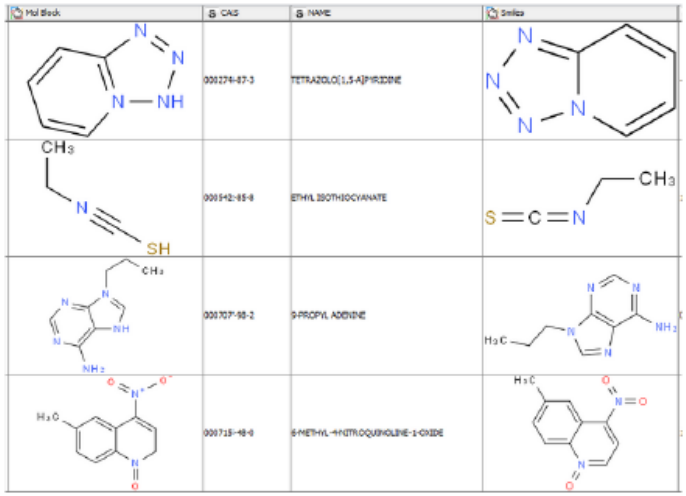
Figure 8. The KNIME curation workflow developed based on the log *P* PHYSPROP dataset and generalized for application to other datasets.



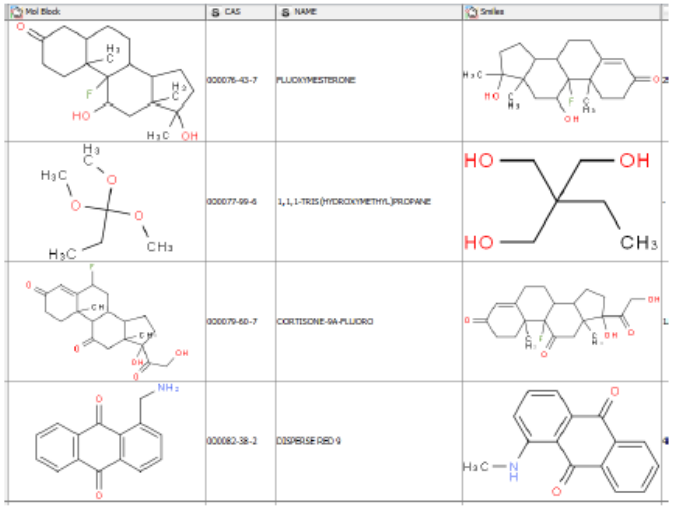
Public PHYSPROP Dataset

EPA United States Environmental Protection Agency

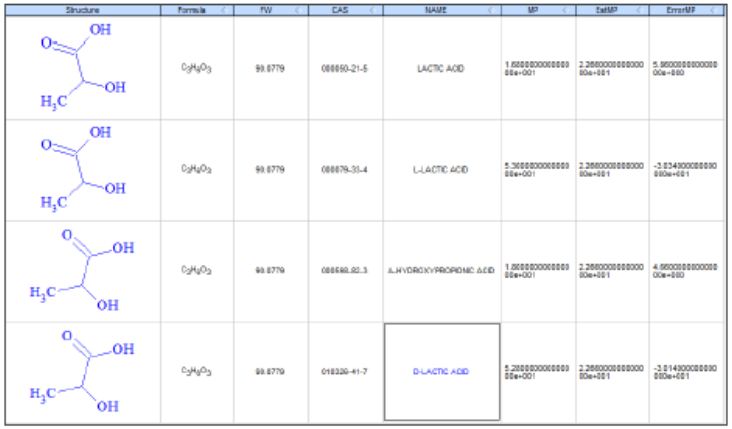
Valence Errors



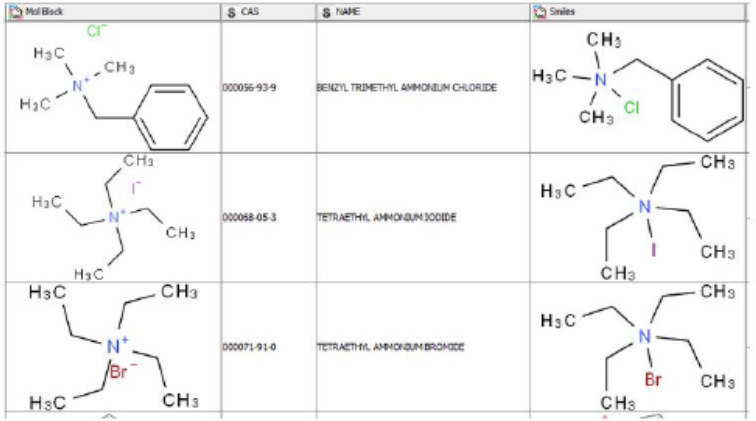
Mismatching structures



Duplicate Structures



Covalent Halogens



LogP dataset: 15,809 structures

- CAS Checksum: 12163 valid, 3646 invalid **(>23%)**
- Invalid names: 555
- Invalid SMILES 133
- Valence errors: 322 Molfile, 3782 SMILES **(>24%)**
- Duplicates check:
  - –31 DUPLICATE MOLFILES
  - –626 DUPLICATE SMILES
  - –531 DUPLICATE NAMES
- SMILES vs. Molfiles (structure check)
  - –1279 differ in stereochemistry **(~8%)**
  - –362 “Covalent Halogens”
  - –191 differ as tautomers
  - –436 are different compounds **(~3%)**

OPERA Predicted Properties

Mansouri *et al. J Cheminform (2018) 10:10*
https://doi.org/10.1186/s13321-018-0263-1

Journal of Cheminformatics

RESEARCH ARTICLE Open Access

CrossMark

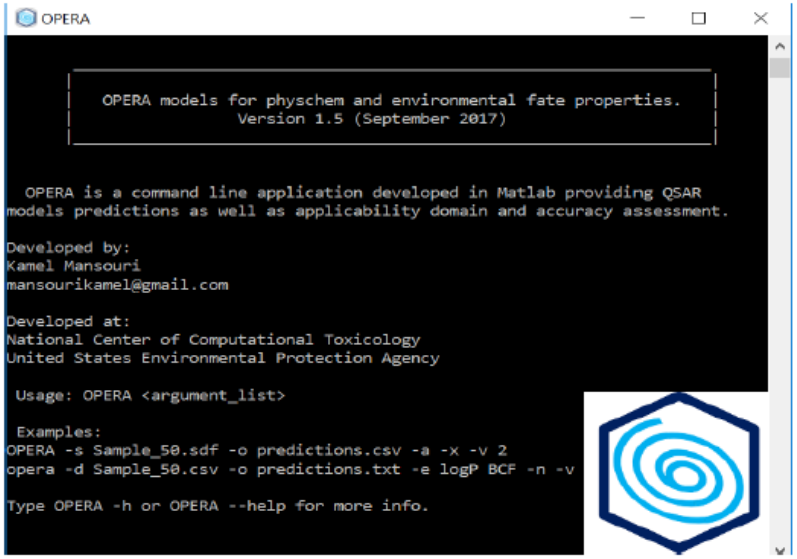
OPERA models for predicting physicochemical properties and environmental fate endpoints

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

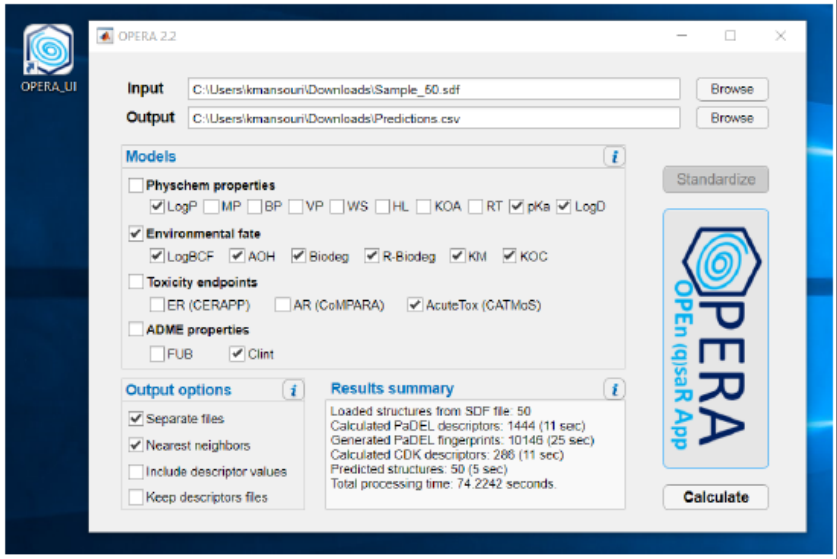
OPERA Models: https://github.com/kmansouri/OPERA

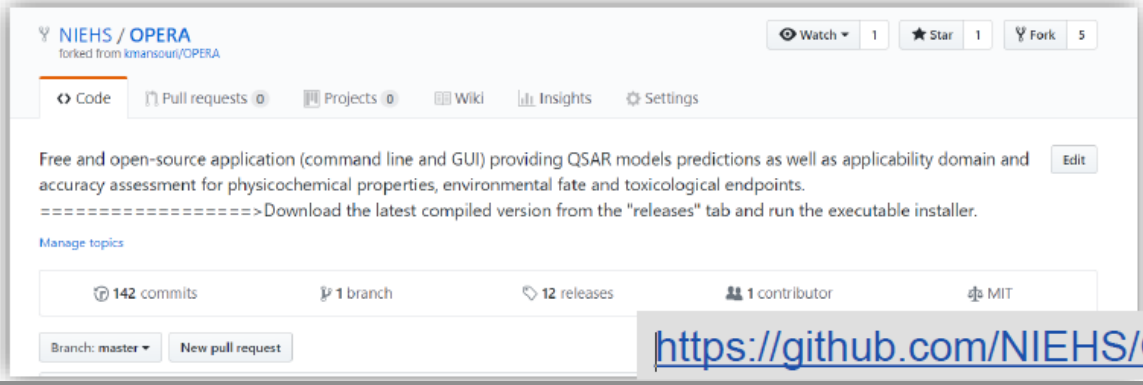
OPERA Standalone Application

Command line



Graphical User Interface





https://github.com/NIEHS/OPERA

Open Source
https://github.com/kmansouri/OPERA

README.md

OPERA

OPERA is a free and open-source/open-data suite of QSAR models providing predictions on physicochemical properties, environmental fate and toxcicity endpoints as well as additional information including applicability domain and accuracy assessment. All models were built on curated data and standardized QSAR-ready chemical structures. OPERA is available in command line and user-friendly graphical interface for Windows and Linux operating systems. It can be installed as a standalone desktop application or embedded in a different tool/workflow.

References:

[1] Mansouri K. et al. J Cheminform (2018) https://doi.org/10.1186/s13321-018-0263-1.

[2] Mansouri, K. et al. SAR and QSAR in Env. Res. (2016). https://doi.org/10.1080/1062936X.2016.1253611

[3] Williams A. J. et al. J Cheminform (2017) https://doi.org/10.1186/s13321-017-0247-6

[4] The CompTox Chemistry Dashboard (https://comptox.epa.gov/dashboard)

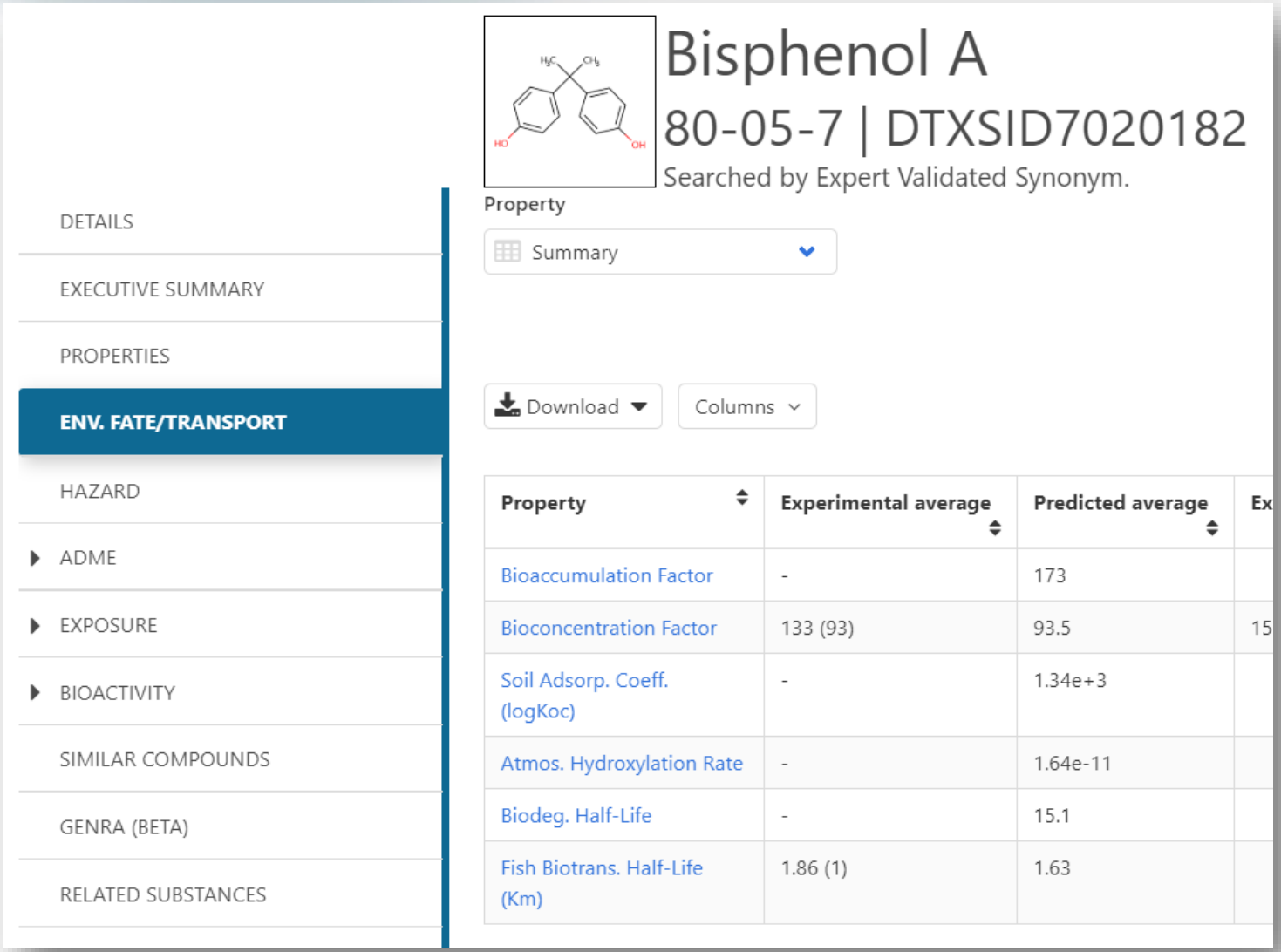
[5] JRC QSAR Model Database https://qsardb.jrc.ec.europa.eu/qmrf/endpoint

Models:

```
*  Latest version OPERA v2.2:

      + Molecular descriptors:
- PaDEL (2.21) (https://doi.org/10.1002/jcc.21707 )
- CDK (2.0) (https://doi.org/10.1186/s13321-017-0220-4)
```

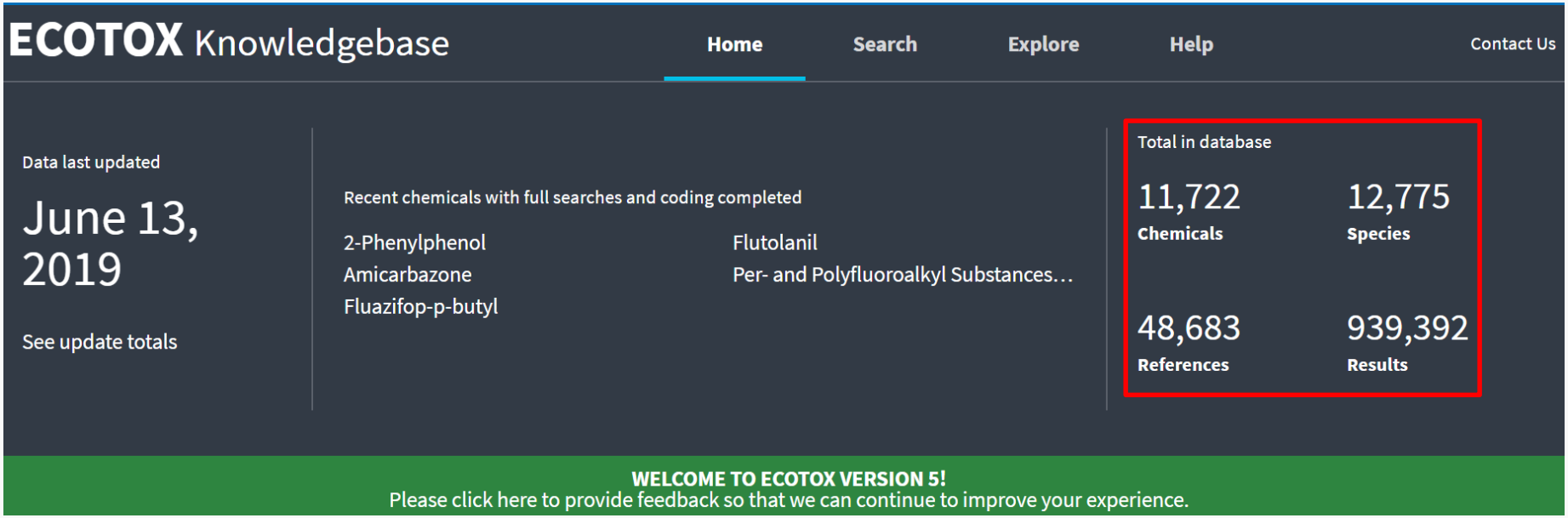
Environmental Fate and Transport



Environmental Fate and Transport

Source	Result	Experimental Details
ECOTOX: aquatic	150	Species: Navicula incerta; Response Site: Not reported
ECOTOX: aquatic	150	Species: Navicula incerta; Response Site: Not reported
ECOTOX: aquatic	100	Species: Navicula incerta; Response Site: Not reported
ECOTOX: aquatic	100	Species: Navicula incerta; Response Site: Not reported
PhysPropNCCT	43.7	
ECOTOX: aquatic	38.4	Species: Oncorhynchus mykiss; Response Site: Liver
ECOTOX: aquatic	25.0	Species: Oncorhynchus mykiss; Response Site: Liver
ECOTOX: aquatic	22.0	Species: Oncorhynchus mykiss; Response Site: Liver
ECOTOX: aquatic	10.8	Species: Oncorhynchus mykiss; Response Site: Liver
ECOTOX: aquatic	8.70	Species: Oncorhynchus mykiss; Response Site: Liver

ECOTOX data


ECOTOX Knowledgebase
Home
Search
Explore
Help
Contact Us
Data last updated
June 13, 2019
See update totals
Recent chemicals with full searches and coding completed
2-Phenylphenol
Amicarbazone
Fluazifop-p-butyl
Flutolanil
Per- and Polyfluoroalkyl Substances…
Total in database
11,722
Chemicals
12,775
Species
48,683
References
939,392
Results
WELCOME TO ECOTOX VERSION 5!
Please click here to provide feedback so that we can continue to improve your experience.

Other prediction modules
TEST Desktop Software

EPA United States Environmental Protection Agency

Toxicity Estimation Software Tool (TEST)

On this page:

- QSAR Methodologies
- What's New in Version 4.2.1?
- Prior Version History
- System Requirements
- Installation Instructions
- Publications
- Get Email Alerts

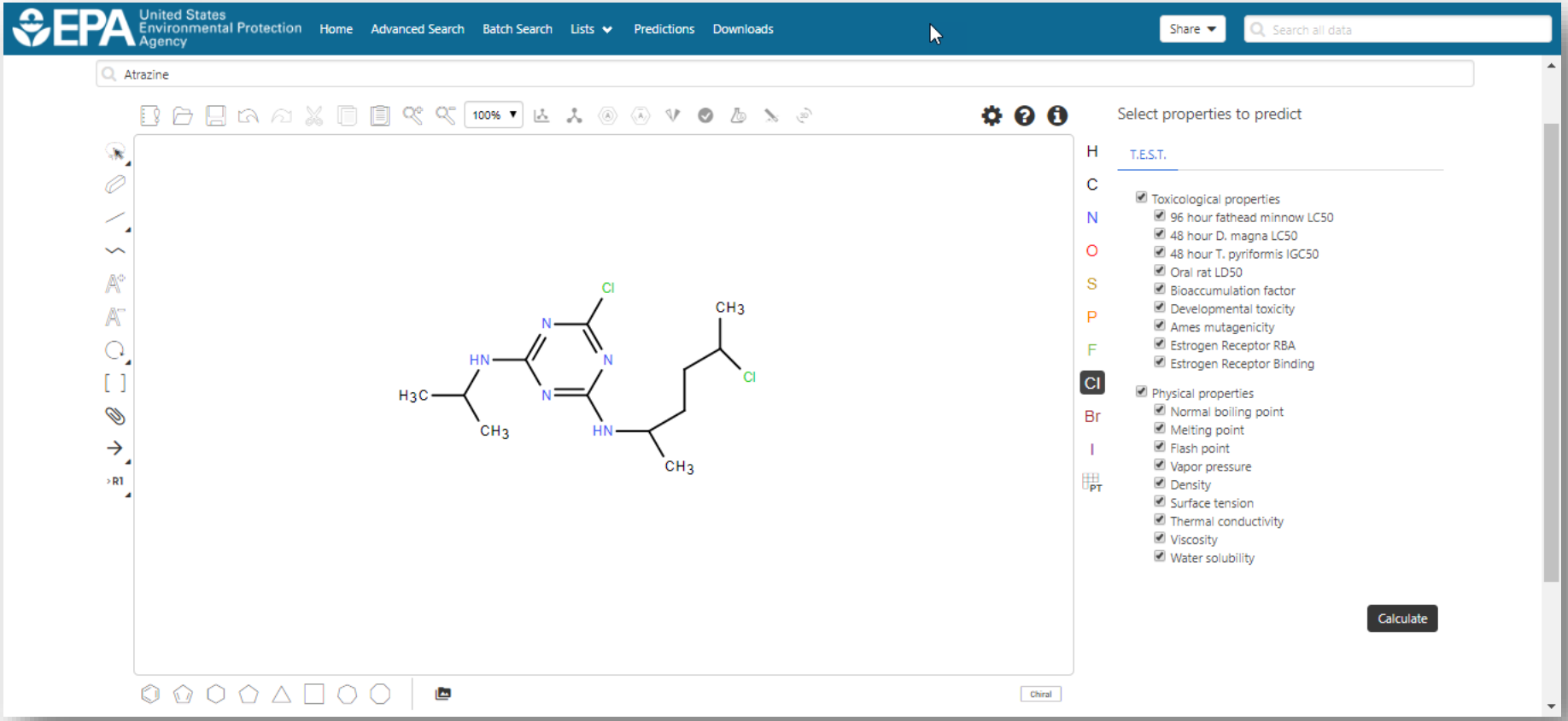
The Toxicity Estimation Software Tool (TEST) was developed to allow users to easily estimate the toxicity of chemicals using Quantitative Structure Activity Relationships (QSARs) methodologies. QSARs are mathematical models used to predict measures of toxicity from the physical characteristics of the structure of chemicals (known as molecular descriptors). Simple QSAR models calculate the toxicity of chemicals using a simple linear function of molecular descriptors:

Ask a Technical Expert

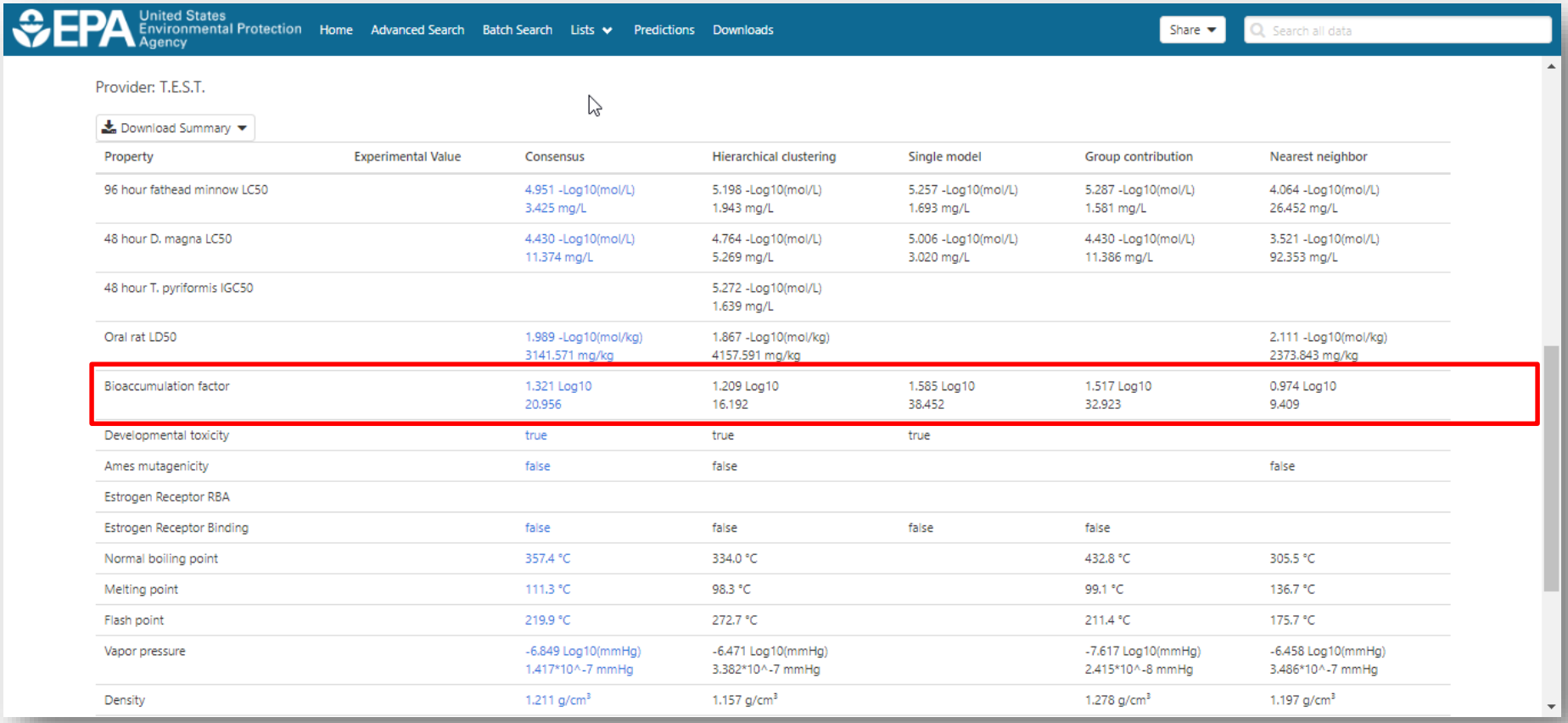
Got a question about our research model? Want to give us feedback? Contact a technical expert about TEST.

Real-Time Predictions Based on TEST

Real-Time Predictions



TEST Predictions
Detailed calculation reports



Property	Experimental Value	Consensus	Hierarchical clustering	Single model	Group contribution	Nearest neighbor
96 hour fathead minnow LC50		4.951 -Log10(mol/L) 3.425 mg/L	5.198 -Log10(mol/L) 1.943 mg/L	5.257 -Log10(mol/L) 1.693 mg/L	5.287 -Log10(mol/L) 1.581 mg/L	4.064 -Log10(mol/L) 26.452 mg/L
48 hour D. magna LC50		4.430 -Log10(mol/L) 11.374 mg/L	4.764 -Log10(mol/L) 5.269 mg/L	5.006 -Log10(mol/L) 3.020 mg/L	4.430 -Log10(mol/L) 11.386 mg/L	3.521 -Log10(mol/L) 92.353 mg/L
48 hour T. pyriformis IGC50			5.272 -Log10(mol/L) 1.639 mg/L			
Oral rat LD50		1.989 -Log10(mol/kg) 3141.571 mg/kg	1.867 -Log10(mol/kg) 4157.591 mg/kg			2.111 -Log10(mol/kg) 2373.843 mg/kg
Bioaccumulation factor		1.321 Log10 20.956	1.209 Log10 16.192	1.585 Log10 38.452	1.517 Log10 32.923	0.974 Log10 9.409
Developmental toxicity		true	true	true		
Ames mutagenicity		false	false			false
Estrogen Receptor RBA						
Estrogen Receptor Binding		false	false	false	false	
Normal boiling point		357.4 °C	334.0 °C		432.8 °C	305.5 °C
Melting point		111.3 °C	98.3 °C		99.1 °C	136.7 °C
Flash point		219.9 °C	272.7 °C		211.4 °C	175.7 °C
Vapor pressure		-6.849 Log10(mmHg) 1.417*10^-7 mmHg	-6.471 Log10(mmHg) 3.382*10^-7 mmHg		-7.617 Log10(mmHg) 2.415*10^-8 mmHg	-6.458 Log10(mmHg) 3.486*10^-7 mmHg
Density		1.211 g/cm³	1.157 g/cm³		1.278 g/cm³	1.197 g/cm³

TEST Predictions
Detailed calculation reports

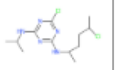
Predicted Vapor pressure at 25°C for ClC=1N=C(N=C(N1)NC(C)CCC(Cl)C)NC(C)C from Consensus method

Prediction results

Endpoint	Experimental value	Predicted value
Vapor pressure at 25°C Log10(mmHg)	N/A	-6.85
Vapor pressure at 25°C mmHg	N/A	1.42E-07

Individual Predictions

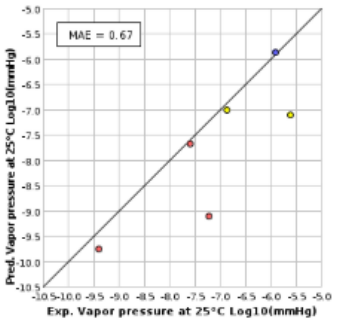
Method	Predicted value Log10(mmHg)
Hierarchical clustering	-6.47
Group contribution	-7.62
Nearest neighbor	-6.46



Predictions for the test chemical and for the most similar chemicals in the external tes

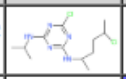
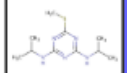
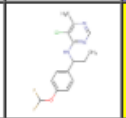
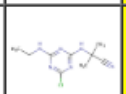
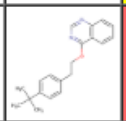
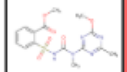
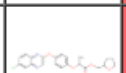
If the predicted value matches the experimental values for similar chemicals in the test set (and the similar che

Prediction results (colors defined in table below)



Chemicals	MAE*
Entire set	0.47
Similarity coefficient ≥ 0.5	0.67

*Mean absolute error in Log10(mmHg)

CAS	Structure	Similarity Coefficient	Experimental value Log10(mmHg)	Predicted value Log10(mmHg)
ClC=1N=C(N=C(N1)NC(C)CCC(Cl)C)NC(C)C (test chemical)			N/A	-6.85
7287-19-6		0.83	-5.91	-5.86
130339-07-0		0.77	-5.62	-7.11
21725-46-2		0.76	-6.86	-7.01
120928-09-8		0.58	-7.59	-7.67
101200-48-0		0.56	-9.41	-9.76
119738-06-6		0.55	-7.23	-9.11

Built on TEST Web Services

https://www.epa.gov/sites/production/files/2018-08/documents/webtest_users_guide.pdf

https://comptox.epa.gov/dashboard/web-test/

User's Guide for WebTEST (version 1.0) (Web-services Toxicity Estimation Software Tool)

A Web-Ser
from Mole

Batch Searching

- Given a set of chemicals how can data be harvested?
- OPERA and TEST predictions have been generated for structures in the database

- How can the dashboard be used to harvest data for hundreds to thousands of chemicals…

Example list of chemicals - Opioids

ELSEVIER

Trends in Environmental Analytical Chemistry

Volume 20, October 2018, e00059

Opioid occurrence in environmental water samples—A review

Marina Celia Campos-Mañas [a], Imma Ferrer [b], E.Michael Thurman [b], Ana Agüera [a]

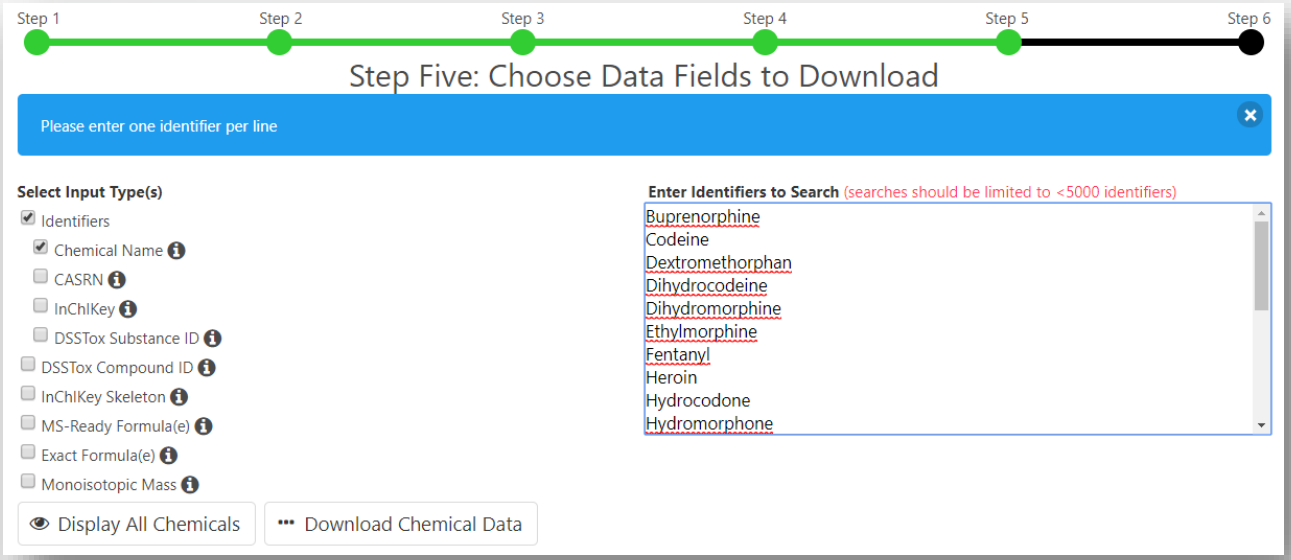
Show more

https://doi.org/10.1016/j.teac.2018.e00059

Get rights and content

Batch Search Names

Buprenorphine
Codeine
Dextromethorphan
Dihydrocodeine
Dihydromorphine
Ethylmorphine
Fentanyl
Heroin
Hydrocodone
Hydromorphone
Ketamine
Meperidine
Methadone
Morphine
Morphinone
Naloxone
Naltriben
Oxycodone
Oxymorphone
Propoxyphene
Sufentanil
Tramadol



Excel Download

INPUT	FOUND_BY	DTXSID
Buprenorphine	Approved Name	DTXSID2022705
Codeine	Approved Name	DTXSID2020341
Dextromethorphan	Approved Name	DTXSID3022908
Dihydrocodeine	Approved Name	DTXSID5022936
Dihydromorphine	Approved Name	DTXSID7048908
Ethylmorphine	Approved Name	DTXSID1046760
Fentanyl	Approved Name	DTXSID9023049
Heroin	Synonym	DTXSID6046761
Hydrocodone	Approved Name	DTXSID8023131
Hydromorphone	Approved Name	DTXSID8023133
Ketamine	Approved Name	DTXSID8023187
Meperidine	Approved Name	DTXSID9023253
Methadone	Approved Name	DTXSID7023273
Morphine	Approved Name	DTXSID9023336

Include Other Data of Interest

Chemical Identifiers

- [x] DTXSID
- [x] Chemical Name
- [ ] DTXCID
- [x] CAS-RN
- [x] InChIKey
- [ ] IUPAC Name

Structures

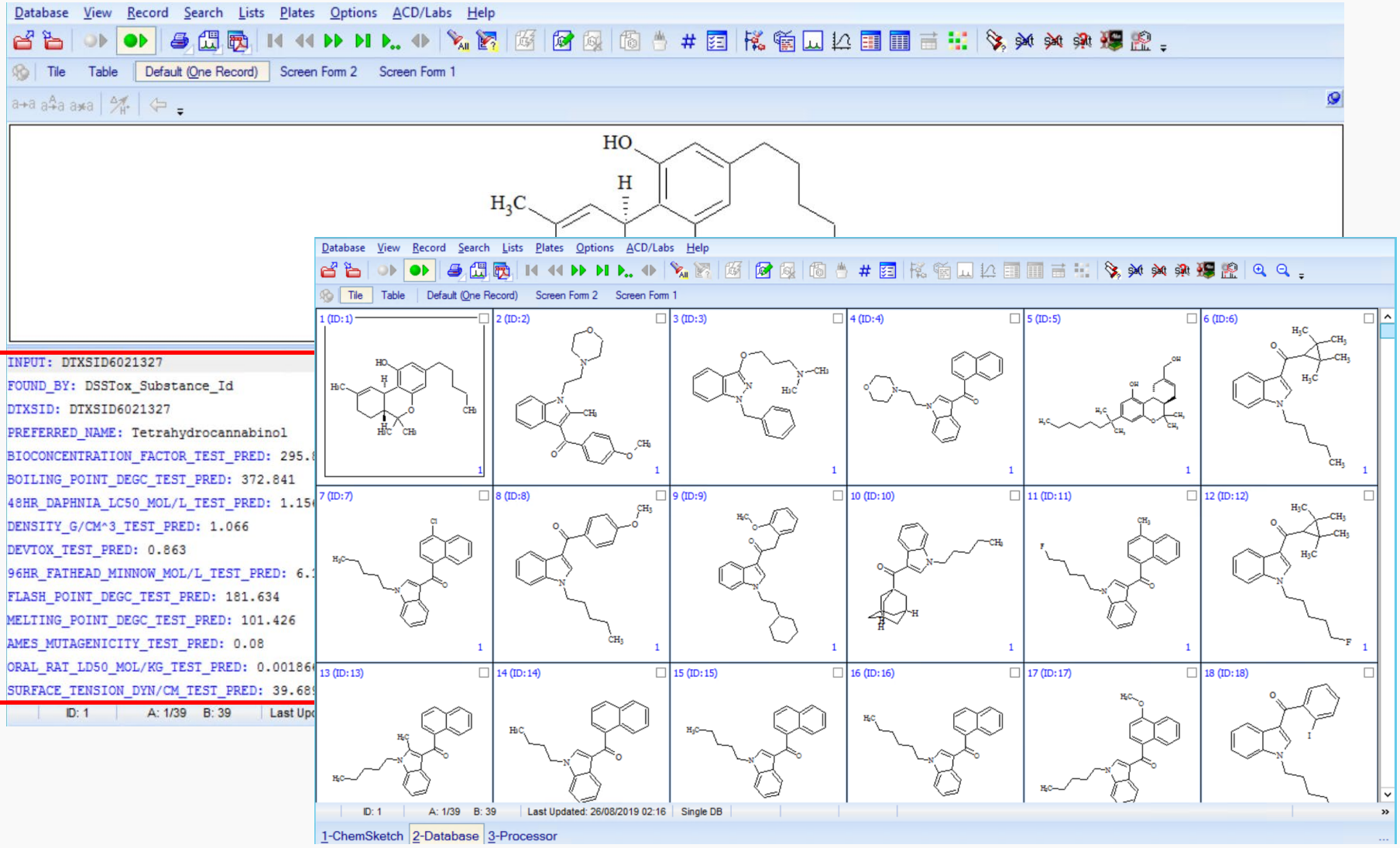
- [ ] Mol File
- [ ] SMILES
- [ ] InChI String
- [x] MS-Ready SMILES
- [ ] QSAR-Ready SMILES

Intrinsic And Predicted Properties

- [x] Molecular Formula
- [ ] Average Mass
- [x] Monoisotopic Mass
- [ ] TEST Model Predictions
- [ ] OPERA Model Predictions

INPUT	DTXSID	CASRN	MOLECULAR_F	MONOISOTOPIC	MS_READY_SMI
Buprenorph	DTXSID202	52485-79-7	C29H41NO4	467.3035588	[H]C12CC3=C4C(
Codeine	DTXSID202	76-57-3	C18H21NO3	299.1521435	[H]C12CC3=C4C(
Dextrometh	DTXSID302	125-71-3	C18H25NO	271.1936144	[H]C12CC3=C(C=
Dihydrocod	DTXSID502	125-28-0	C18H23NO3	301.1677936	[H]C12CC3=C4C(
Dihydromor	DTXSID704	509-60-4	C17H21NO3	287.1521435	[H]C12CC3=C4C(
Ethylmorph	DTXSID104	76-58-4	C19H23NO3	313.1677936	[H]C12CC3=C4C(
Fentanyl	DTXSID902	437-38-7	C22H28N2O	336.2201635	CCC(=O)N(C1CC
Heroin	DTXSID604	561-27-3	C21H23NO5	369.1576228	[H]C12CC3=C4C(
Hydrocodor	DTXSID802	125-29-1	C18H21NO3	299.1521435	[H]C12CC3=C4C(
Hydromorph	DTXSID802	466-99-9	C17H19NO3	285.1364935	[H]C12CC3=C4C(
Ketamine	DTXSID802	6740-88-1	C13H16ClNO	237.0920418	CNC1(CCCCC1=(
Meperidine	DTXSID902	57-42-1	C15H21NO2	247.1572289	CCOC(=O)C1(CC
Methadone	DTXSID702	76-99-3	C21H27NO	309.2092645	CCC(=O)C(CC(C)
Morphine	DTXSID902	57-27-2	C17H19NO3	285.1364935	[H]C12CC3=C4C(
Morphinone	DTXSID501	467-02-7	C17H17NO3	283.1208434	[H]C12CC3=C4C(
Naloxone	DTXSID802	465-65-6	C19H21NO4	327.1470582	[H]C12CC3=C4C(
Naltriben	-	-	-	-	-
Oxycodone	DTXSID502	76-42-6	C18H21NO4	315.1470582	[H]C12CC3=C4C(
Oxymorpho	DTXSID502	76-41-5	C17H19NO4	301.1314081	[H]C12CC3=C4C(
Propoxyphe	DTXSID102	469-62-5	C22H29NO2	339.2198292	CCC(=O)OC(CC1
Sufentanil	DTXSID602	56030-54-7	C22H30N2O2S	386.2027994	CCC(=O)N(C1=C
Tramadol	DTXSID908	27203-92-5	C16H25NO2	263.188529	COC1=CC=CC(=

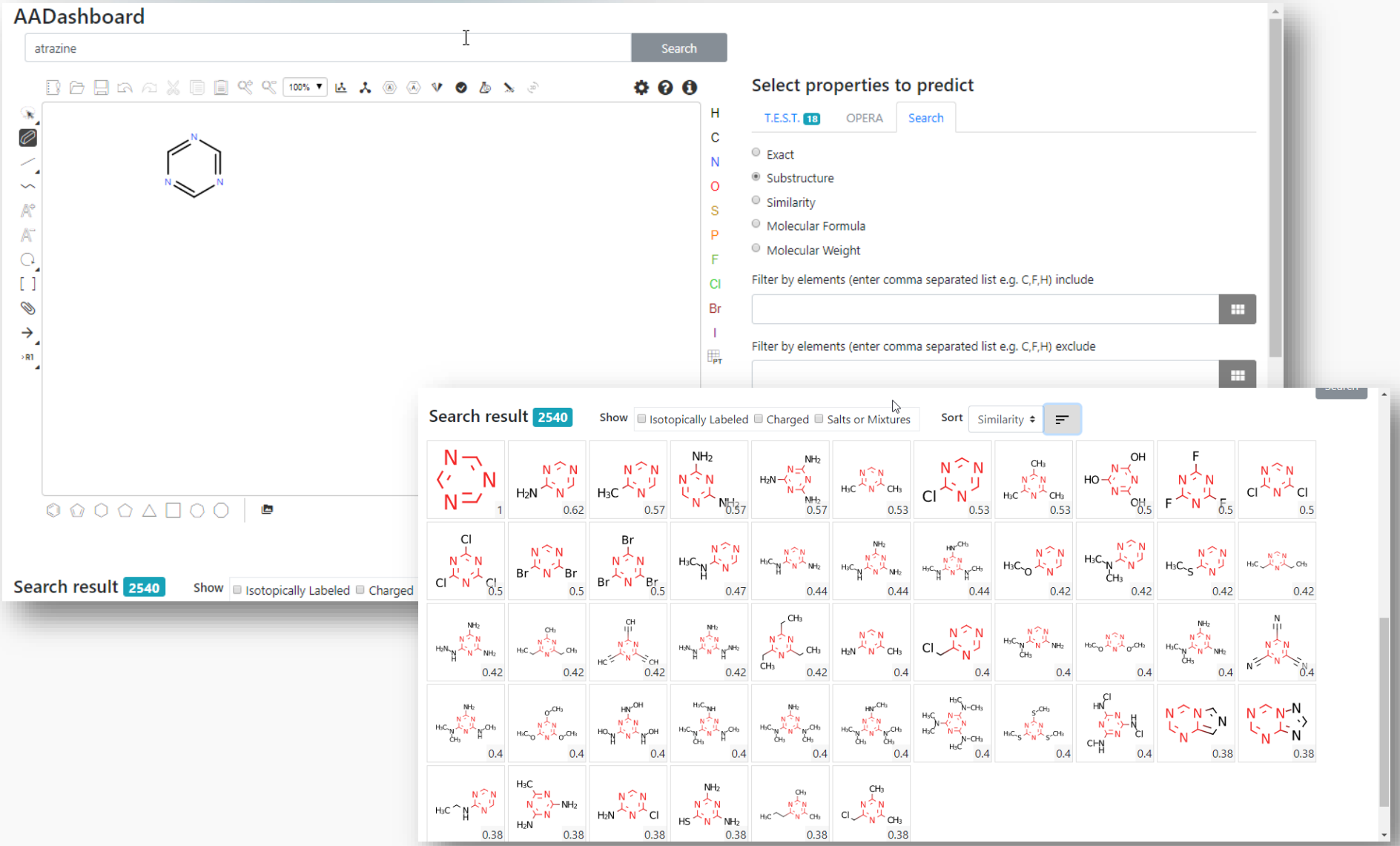
Batch search results - SDF



EPA United States Environmental Protection Agency

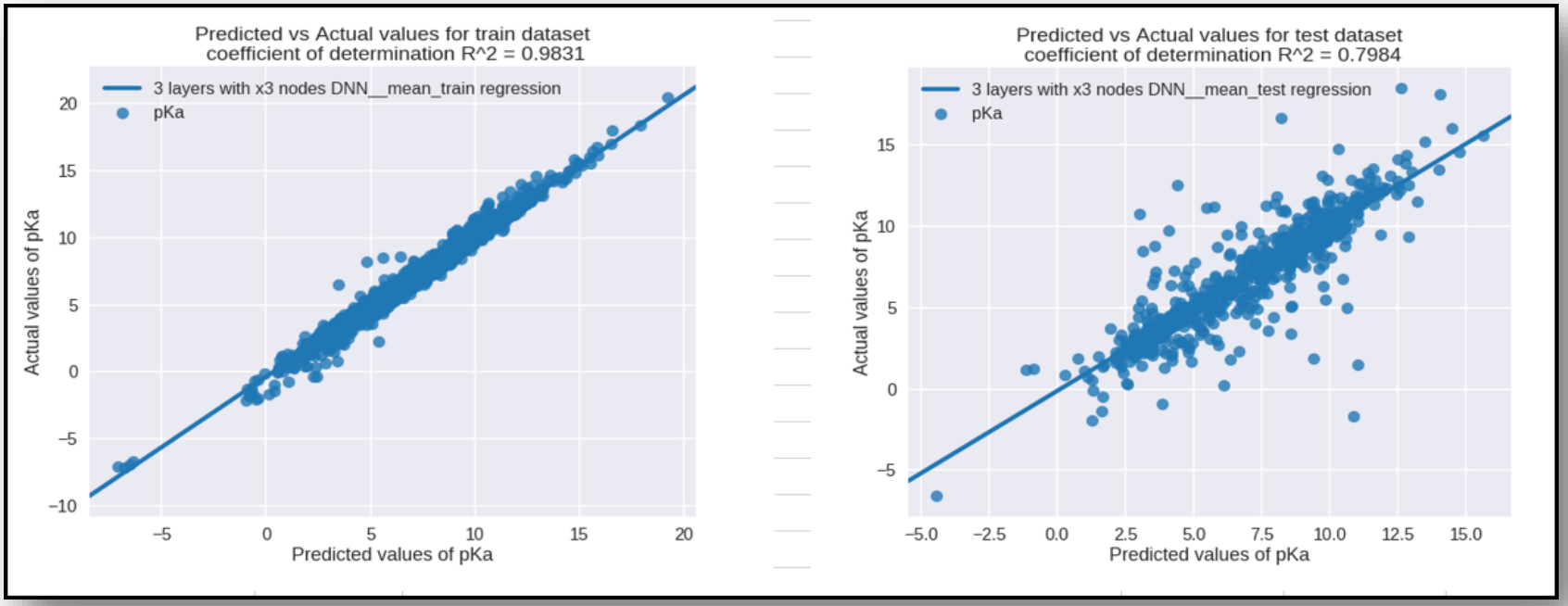
Work in Progress

Prototype Development
Structure/substructure search



OPERA pKa Prediction Model

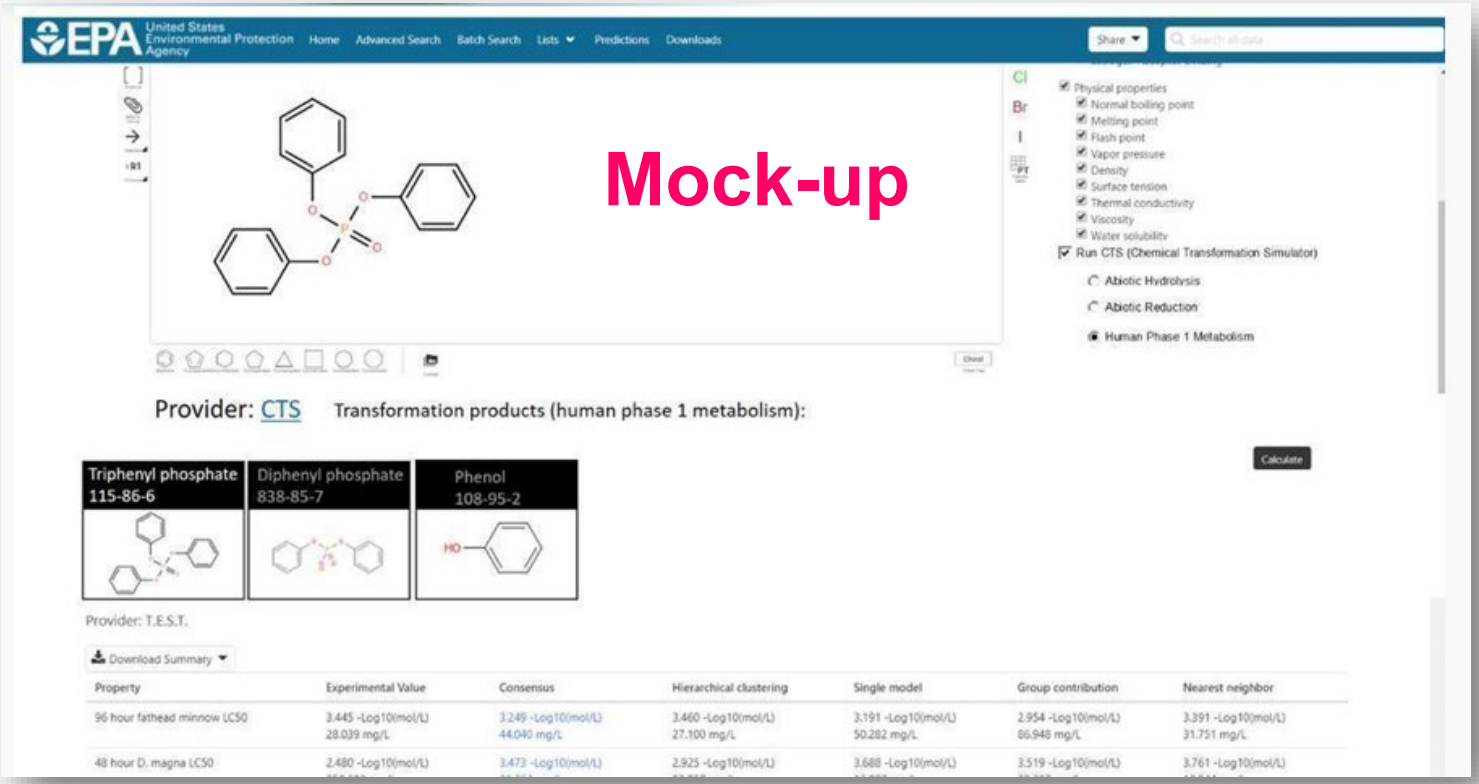
- pKa prediction models based on Open Data Set of 8000 chemicals – acidic, basic and amphoteric chemicals



- *Accepted for publication to Journal of Cheminformatics*

“Chemical Transformation Simulator”

- Chemical Transformation Simulator has public web services already available
  - (1) Abiotic Hydrolysis
  - (2) Abiotic Reduction
  - (3) Phase 1 Metabolism

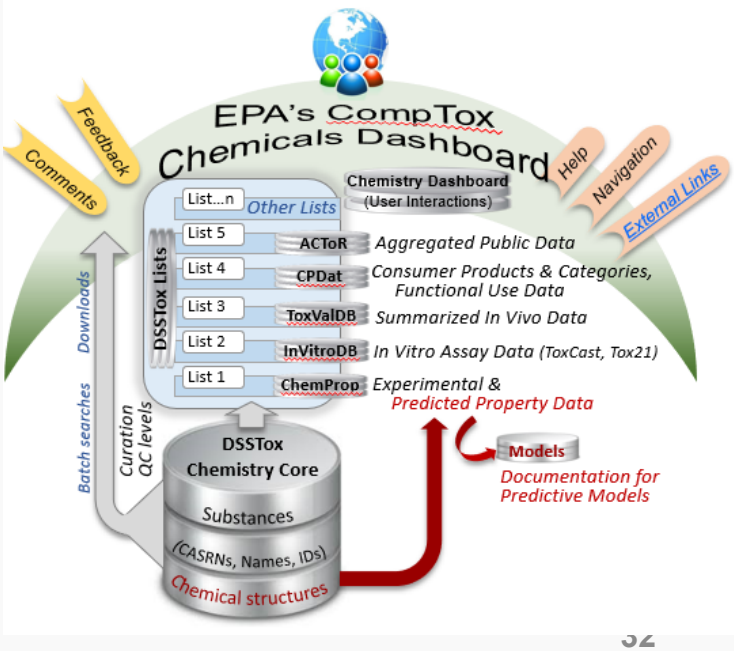


Ongoing Extraction of Data

- Data are extracted from literature based on agency priorities
- Specific data sets of interest at present include those available for PFAS chemicals
- Data releases are every 6 months at present

Conclusion

- Dashboard access to data for ~875,000 chemicals
- Ongoing aggregation of physicochemical property and environmental fate and transport data
- Retraining and rebuilding of models will occur as new data are assembled
- Web services already available for TEST predictions with services for OPERA next
- Future developments include integration of chemical transformation simulator



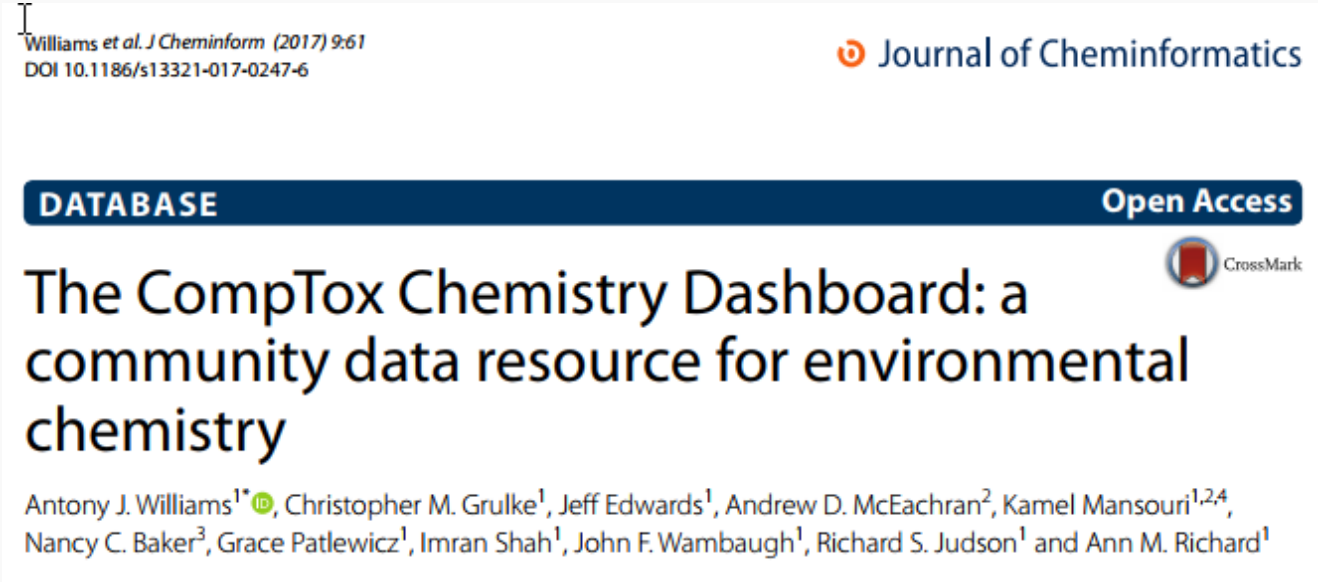
Acknowledgements

EPA-RTP

- *An enormous team of contributors from NCCT, especially the IT software development team and data curation team*
- *US-EPA ECOTOX for sharing data from their database*
- *Valery Tkachenko for development of TEST web services*

Contact

Antony Williams
NCCT, US EPA Office of Research and Development,
Williams.Antony@epa.gov
ORCID: https://orcid.org/0000-0002-2668-4821



Williams et al. J Cheminform (2017) 9:61
DOI 10.1186/s13321-017-0247-6

Journal of Cheminformatics

DATABASE Open Access

The CompTox Chemistry Dashboard: a community data resource for environmental chemistry

Antony J. Williams[1*], Christopher M. Grulke[1], Jeff Edwards[1], Andrew D. McEachran[2], Kamel Mansouri[1,2,4], Nancy C. Baker[3], Grace Patlewicz[1], Imran Shah[1], John F. Wambaugh[1], Richard S. Judson[1] and Ann M. Richard[1]

https://doi.org/10.1186/s13321-017-0247-6