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

US EPA CompTox Chemistry Dashboard as a source of data to fill data gaps for chemical sources of risk

Antony Williams[1], Chris Grulke[1], Kamel Mansouri[4], Kathie Dionisio[2], Katherine Phillips[2], Grace Patlewicz[1], Imran Shah[1], Kristin Isaacs[2], Todd Martin[3], John Wambaugh[1], Ann Richard[1] and Richard Judson[1]

1) National Center for Computational Toxicology, U.S. Environmental Protection Agency, RTP, NC
2) National Exposure Research Laboratory, U.S. Environmental Protection Agency, RTP, NC
3) National Risk Management Research Laboratory, U.S. Environmental Protection Agency, Cincinnati, OH
4) Scitovation, RTP, NC

March 2018
ACS Spring Meeting, New Orleans

Two Years of Development for the CompTox Chemistry Dashboard

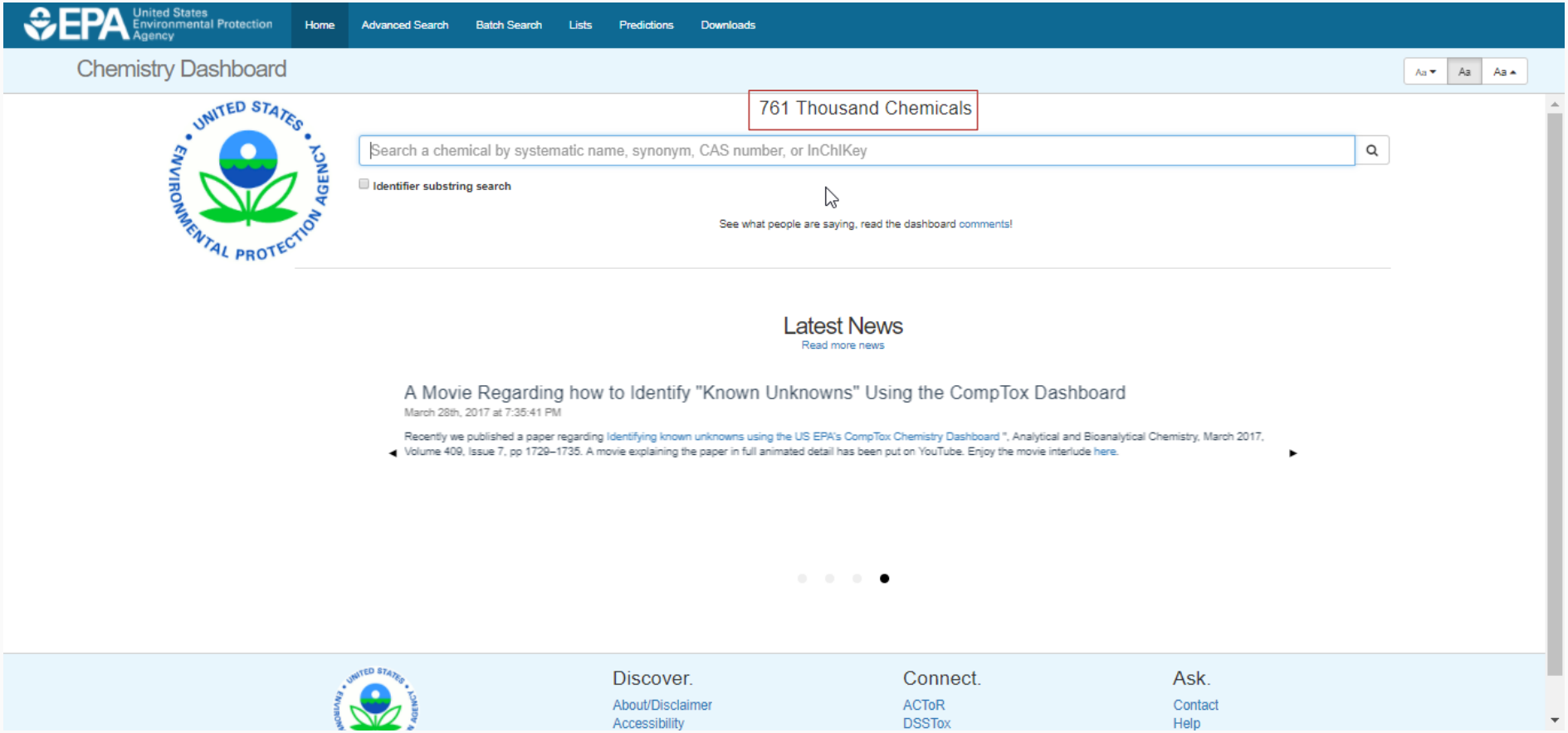
- The Chemistry Dashboard went online on April 1st 2016
- Initial concept was as an integration hub for NCCT chemistry data
- Two years later, with 10k users a month, it is fulfilling the promise with an underlying architecture for integrating CompTox data
- Hazard and Exposure data to fill data gaps

The CompTox Chemistry Dashboard

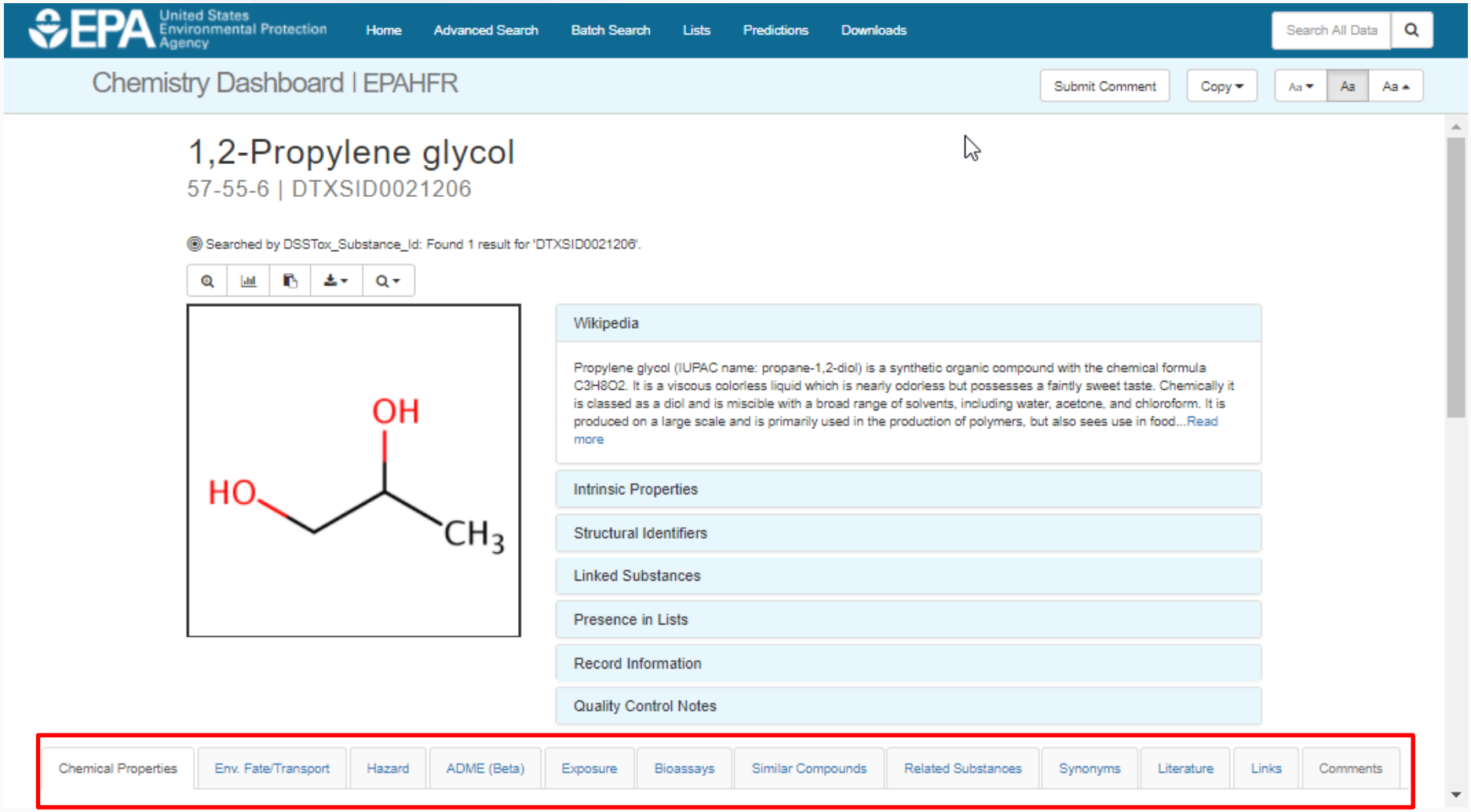
- A **publicly accessible website** delivering access:
  - ~760,000 chemicals with related property data
  - Experimental and predicted physicochemical property data
  - Experimental Human and Ecological hazard data
  - Integration to “biological assay data” for 1000s of chemicals
  - Information regarding consumer products containing chemicals
  - Links to other agency websites and public data resources
  - “Literature” searches for chemicals using public resources
  - “Batch searching” for thousands of chemicals
  - Real time prediction of physchem and toxicity endpoints

CompTox Chemistry Dashboard

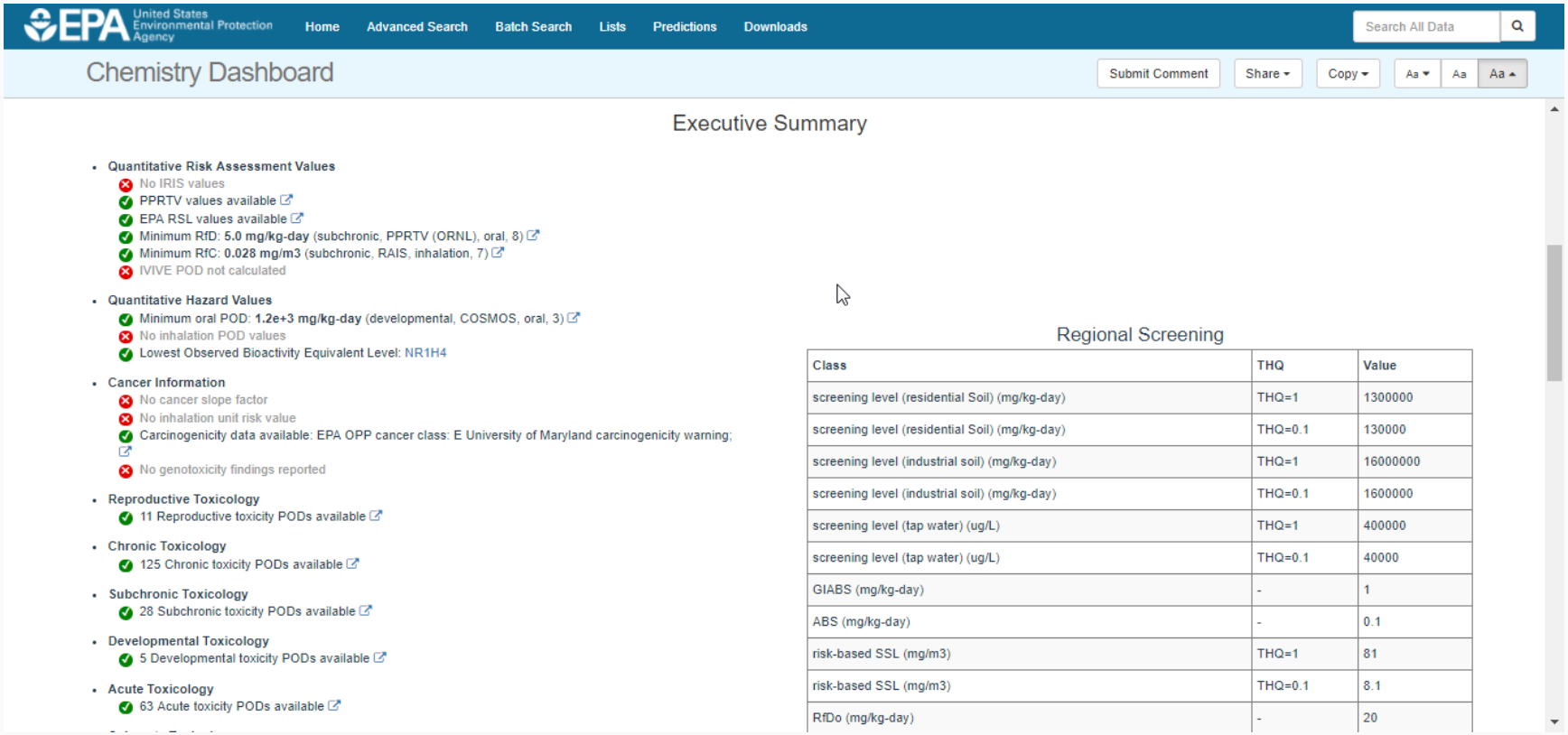
https://comptox.epa.gov/dashboard



Detailed Chemical Pages

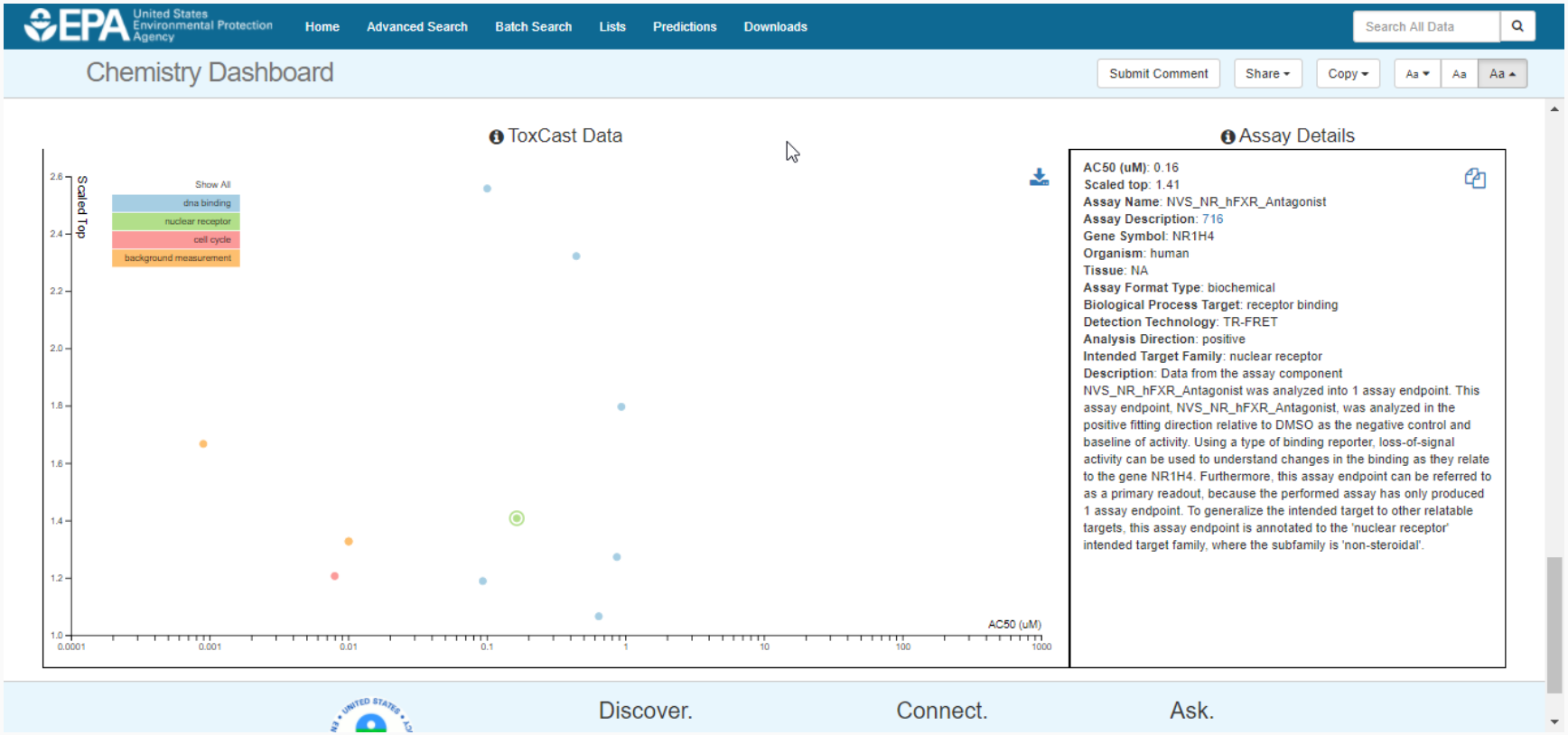


The Executive Summary (NEW)

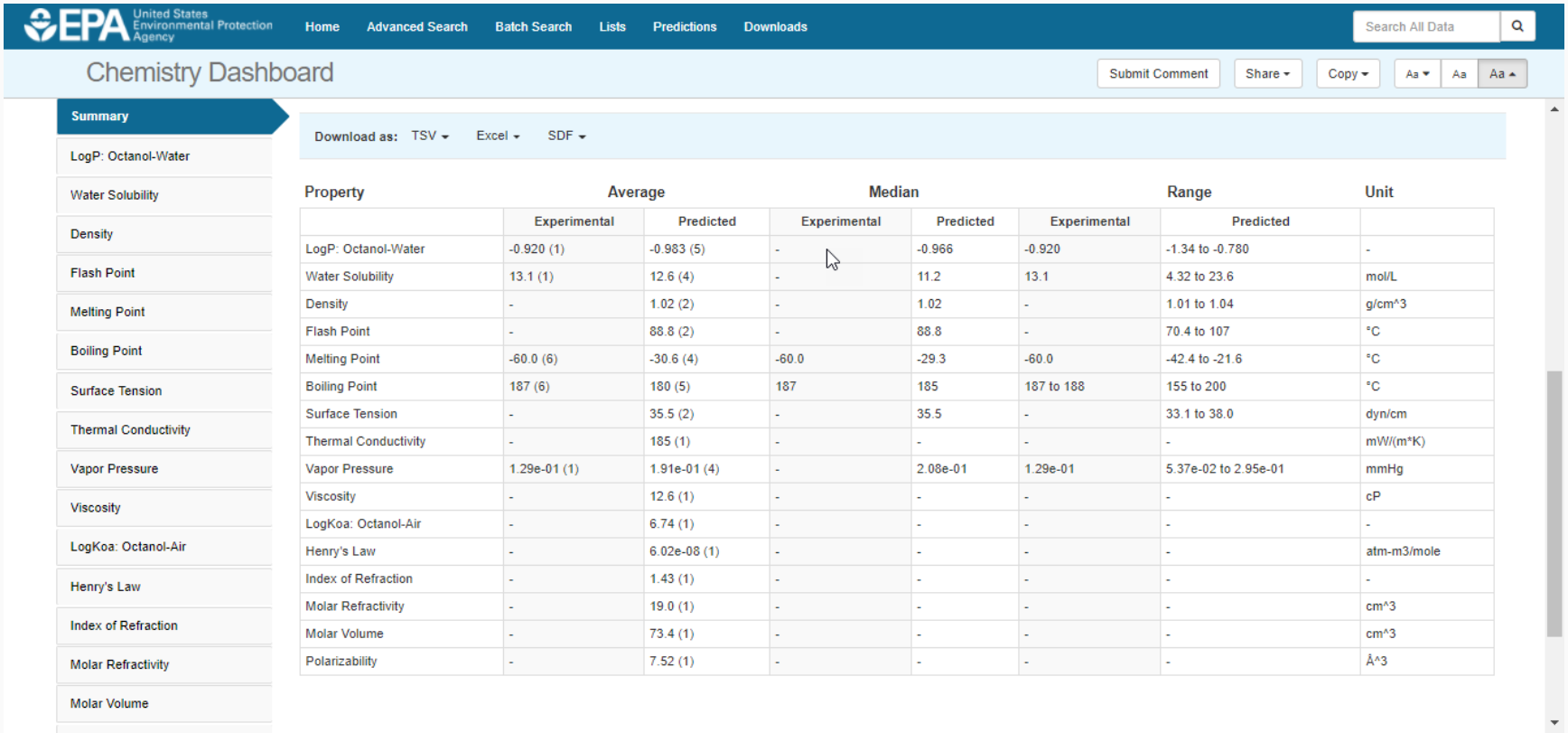


The Executive Summary (NEW)

The Executive Summary (NEW)



Properties, Fate and Transport



Property	Average		Median		Range		Unit
	Experimental	Predicted	Experimental	Predicted	Experimental	Predicted	
LogP: Octanol-Water	-0.920 (1)	-0.983 (5)	-	-0.966	-0.920	-1.34 to -0.780	-
Water Solubility	13.1 (1)	12.6 (4)	-	11.2	13.1	4.32 to 23.6	mol/L
Density	-	1.02 (2)	-	1.02	-	1.01 to 1.04	g/cm^3
Flash Point	-	88.8 (2)	-	88.8	-	70.4 to 107	°C
Melting Point	-60.0 (6)	-30.6 (4)	-60.0	-29.3	-60.0	-42.4 to -21.6	°C
Boiling Point	187 (6)	180 (5)	187	185	187 to 188	155 to 200	°C
Surface Tension	-	35.5 (2)	-	35.5	-	33.1 to 38.0	dyn/cm
Thermal Conductivity	-	185 (1)	-	-	-	-	mW/(m*K)
Vapor Pressure	1.29e-01 (1)	1.91e-01 (4)	-	2.08e-01	1.29e-01	5.37e-02 to 2.95e-01	mmHg
Viscosity	-	12.6 (1)	-	-	-	-	cP
LogKoa: Octanol-Air	-	6.74 (1)	-	-	-	-	-
Henry's Law	-	6.02e-08 (1)	-	-	-	-	atm-m3/mole
Index of Refraction	-	1.43 (1)	-	-	-	-	-
Molar Refractivity	-	19.0 (1)	-	-	-	-	cm^3
Molar Volume	-	73.4 (1)	-	-	-	-	cm^3
Polarizability	-	7.52 (1)	-	-	-	-	Å^3

Properties, Fate and Transport

- When we don't have experimental data we predict it…

Property	Average	
	Experimental	Predicted
Soil Adsorp. Coeff.	2.29 (1)	1.94 (2)
Atmos. Hydroxylation Rate	1.20e-11 (1)	1.30e-11 (1)
Biodeg. Half-Life	-	4.28 (1)
Fish Biotrans. Half-Life (Km)	-	8.30e-02 (1)
Bioaccumulation Factor	-	8.96e-01 (1)
Bioconcentration Factor	-	2.13 (5)

Model Performance Details

EPA United States Environmental Protection Agency

EPA United States Environmental Protection Agency | Home | Advanced Search | Batch Search | Lists | Predictions | Downloads | 21206

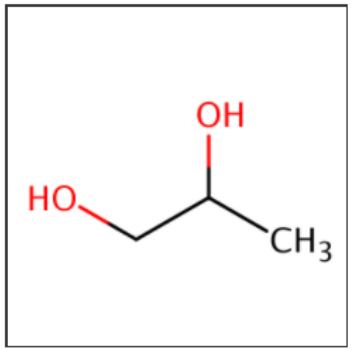
Chemistry Dashboard

OPERA Models: Fish Biotrans. Half-Life (Km)

1,2-Propylene glycol

57-55-6 | DTXSID0021206

Save PDF



Model Results

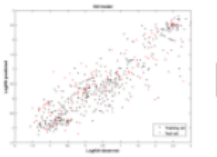
Predicted value: 8.30e-02 days

Global applicability domain: Inside

Local applicability domain index: 0.28

Confidence level: 0.71

Model Performance

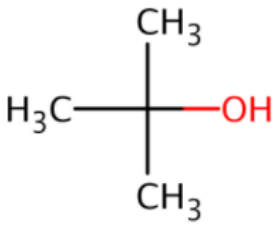


Weighted KNN model

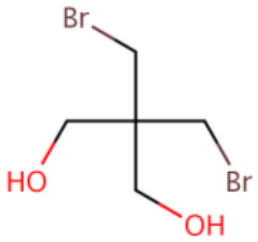
QMRF

5-fold CV (75%)		Training (75%)		Test (25%)	
Q2	RMSE	R2	RMSE	R2	RMSE
0.83	0.49	0.82	0.5	0.73	0.62

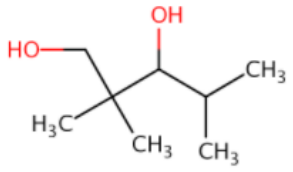
Nearest Neighbors from the Training Set



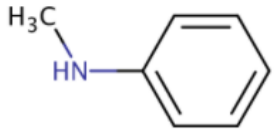
tert-Butyl alcohol
Measured: 7.41e-02
Predicted: 7.66e-02



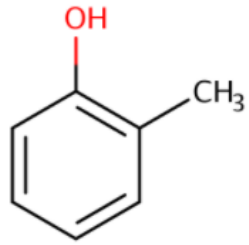
Pentaerythritol dibromide
Measured: 3.98e-02
Predicted: 5.44e-02



2,2,4-Trimethyl-1,3-pentanediol
Measured: 4.57e-02
Predicted: 5.13e-02



N-Methylaniline
Measured: 1.32e-01
Predicted: 1.50e-01



o-Cresol
Measured: 3.02e-01
Predicted: 2.42e-01

OPERA: OPEN Data and OPEN Models

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

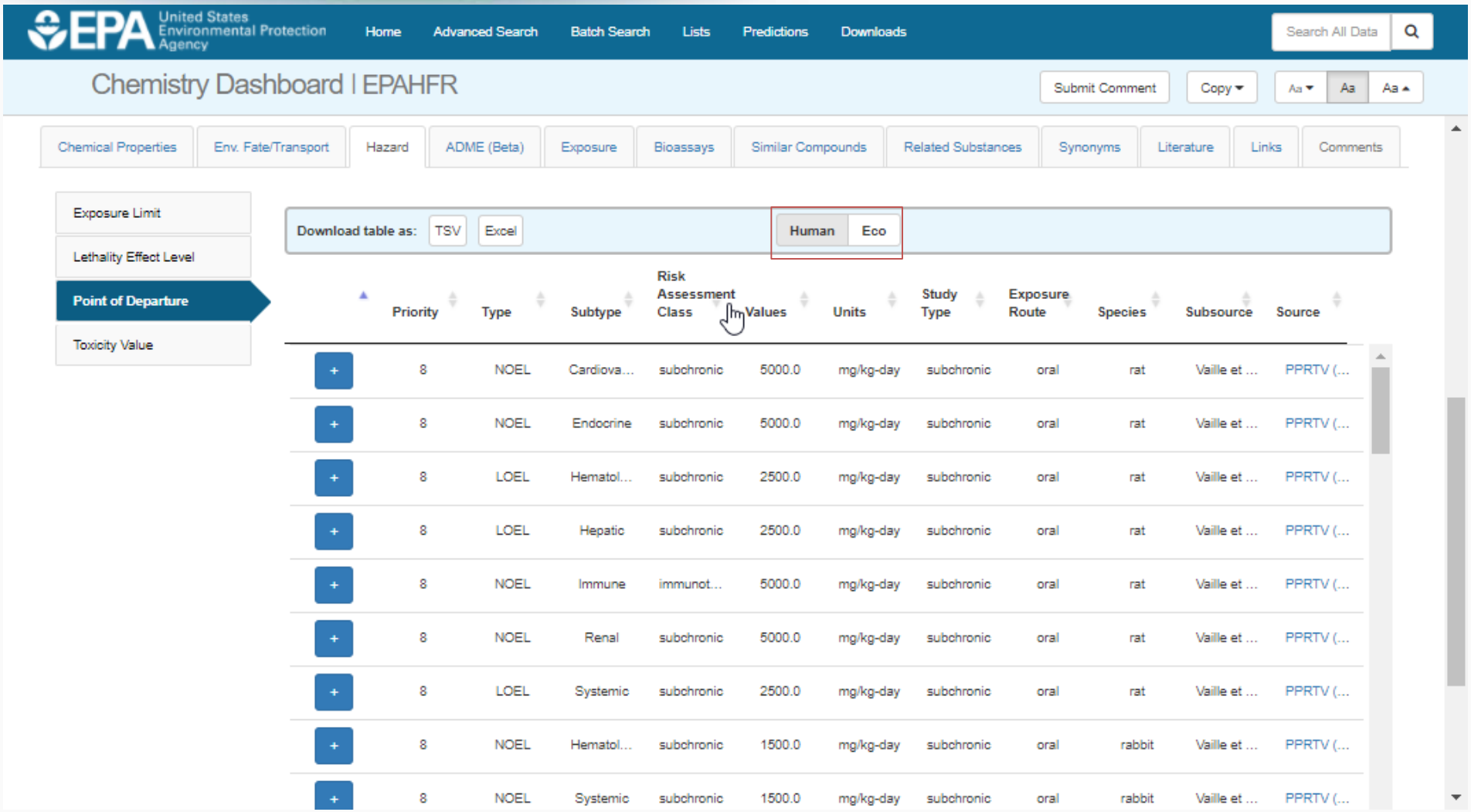
Journal of Cheminformatics

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]

Access to Chemical Hazard Data



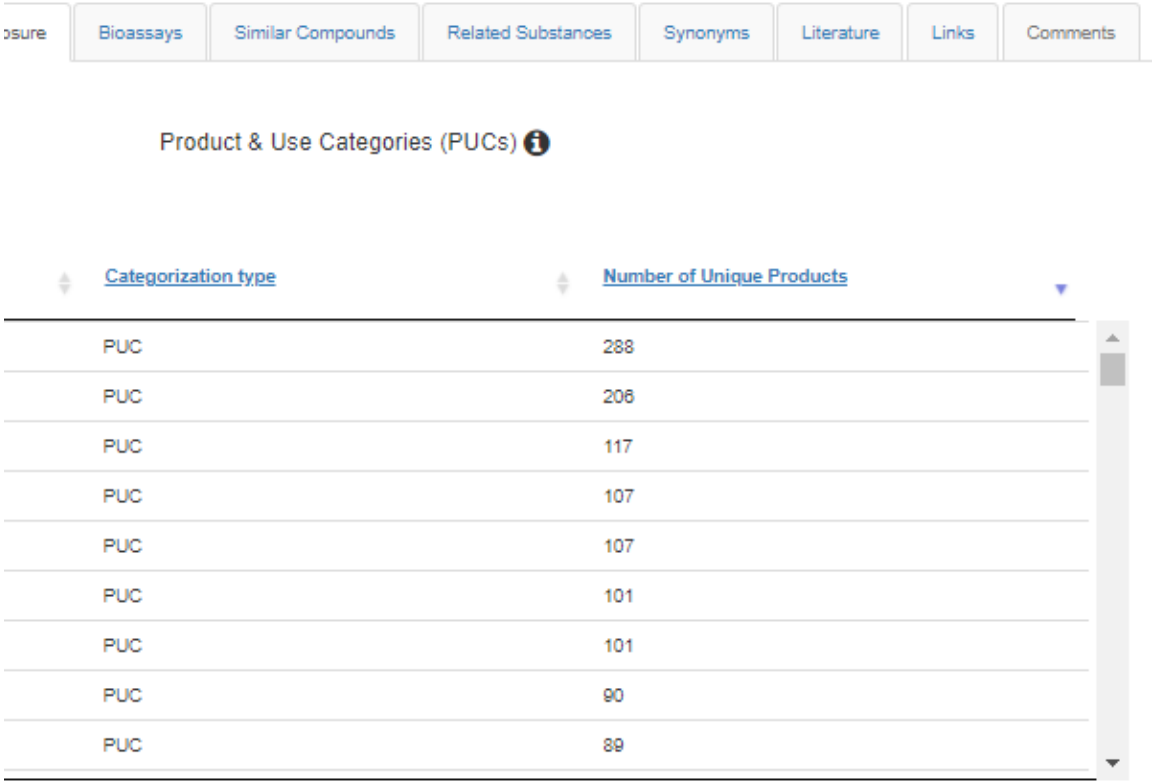
	Priority	Type	Subtype	Risk Assessment Class	Values	Units	Study Type	Exposure Route	Species	Subsource	Source
+	8	NOEL	Cardiova...	subchronic	5000.0	mg/kg-day	subchronic	oral	rat	Vaille et ...	PPRTV (...
+	8	NOEL	Endocrine	subchronic	5000.0	mg/kg-day	subchronic	oral	rat	Vaille et ...	PPRTV (...
+	8	LOEL	Hematol...	subchronic	2500.0	mg/kg-day	subchronic	oral	rat	Vaille et ...	PPRTV (...
+	8	LOEL	Hepatic	subchronic	2500.0	mg/kg-day	subchronic	oral	rat	Vaille et ...	PPRTV (...
+	8	NOEL	Immune	immunot...	5000.0	mg/kg-day	subchronic	oral	rat	Vaille et ...	PPRTV (...
+	8	NOEL	Renal	subchronic	5000.0	mg/kg-day	subchronic	oral	rat	Vaille et ...	PPRTV (...
+	8	LOEL	Systemic	subchronic	2500.0	mg/kg-day	subchronic	oral	rat	Vaille et ...	PPRTV (...
+	8	NOEL	Hematol...	subchronic	1500.0	mg/kg-day	subchronic	oral	rabbit	Vaille et ...	PPRTV (...
+	8	NOEL	Systemic	subchronic	1500.0	mg/kg-day	subchronic	oral	rabbit	Vaille et ...	PPRTV (...

In Vitro Bioassay Screening

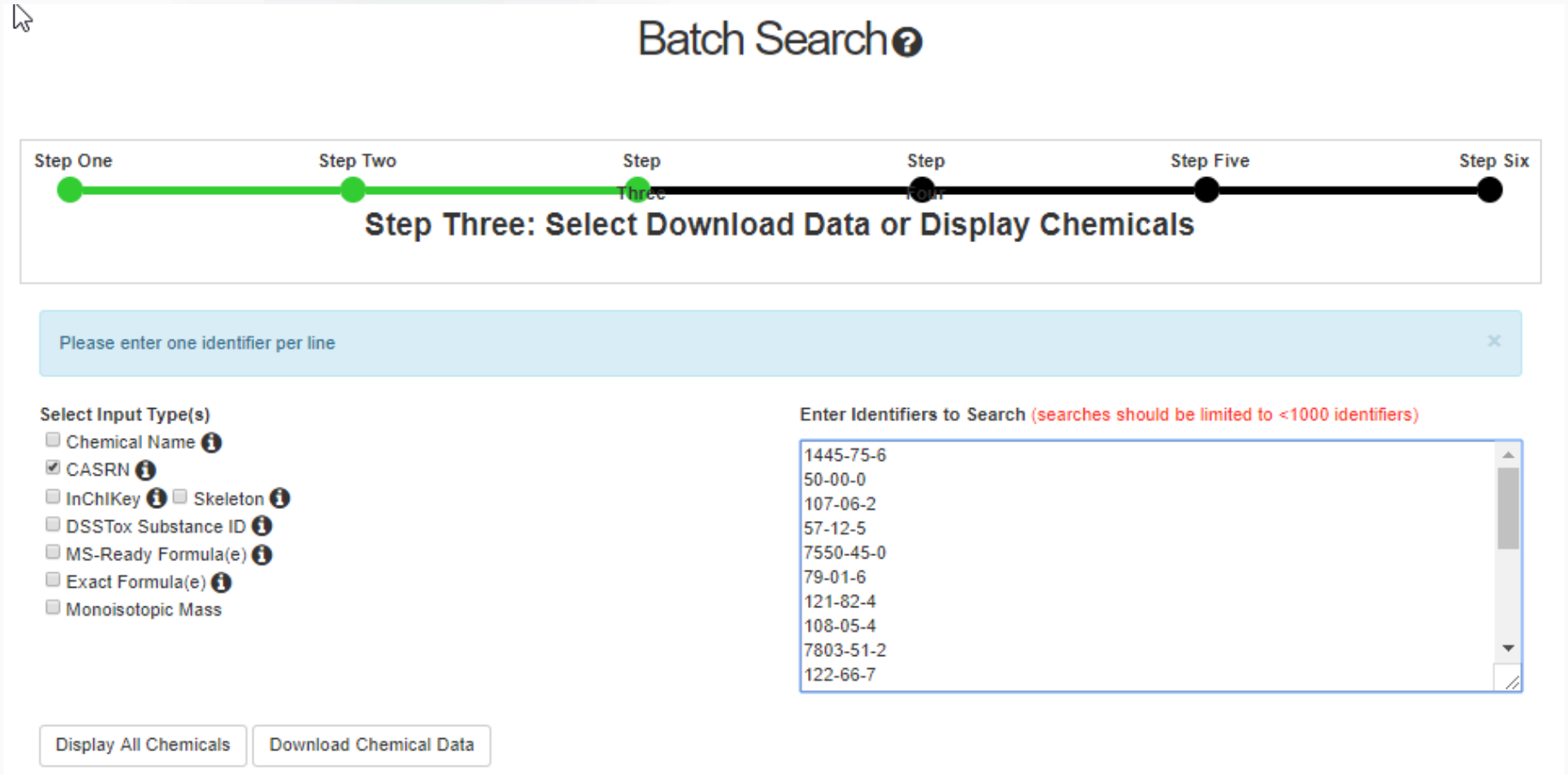
ToxCast and Tox21

Sources of Exposure to Chemicals

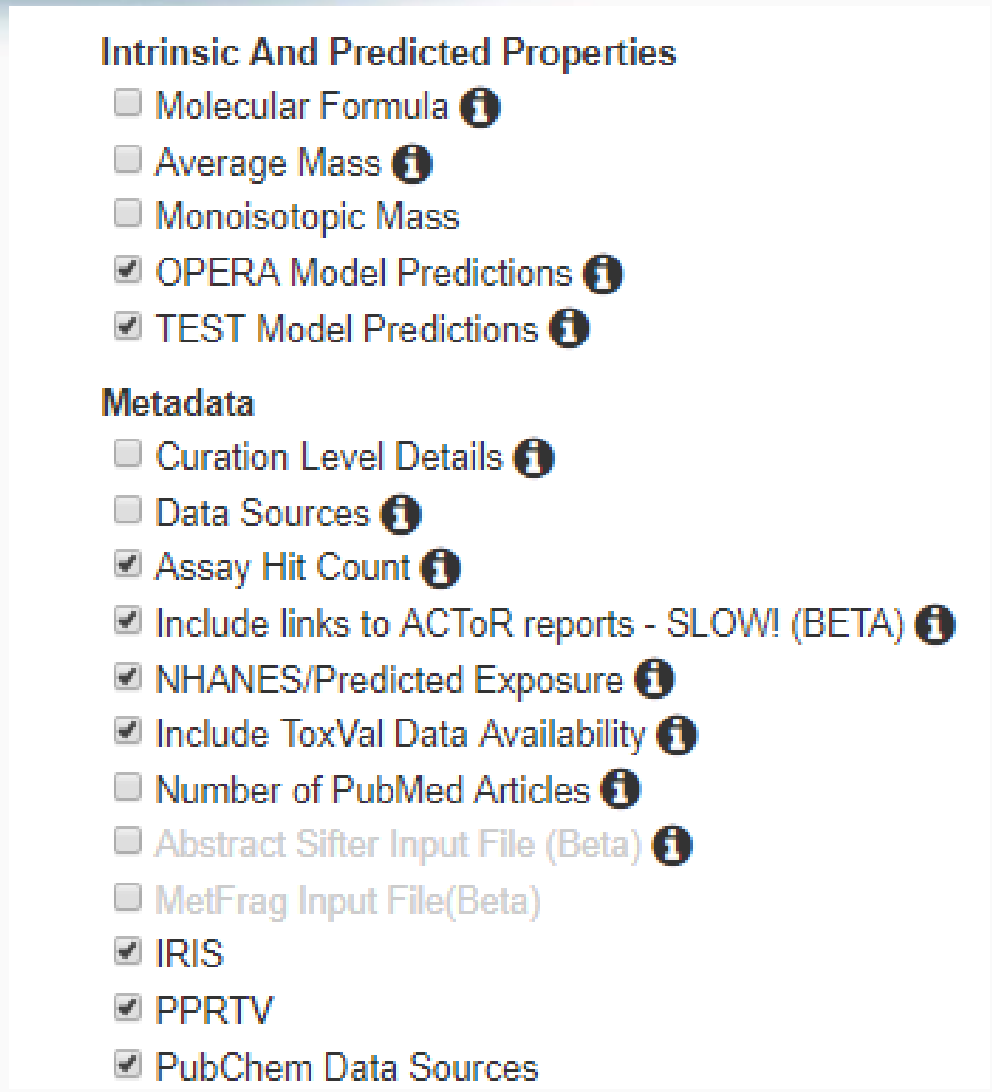
- **Product & Use Categories**
- Chemical Weight Fraction
- Chemical Functional Use
- Monitoring Data
- Exposure Predictions
- Production Volume



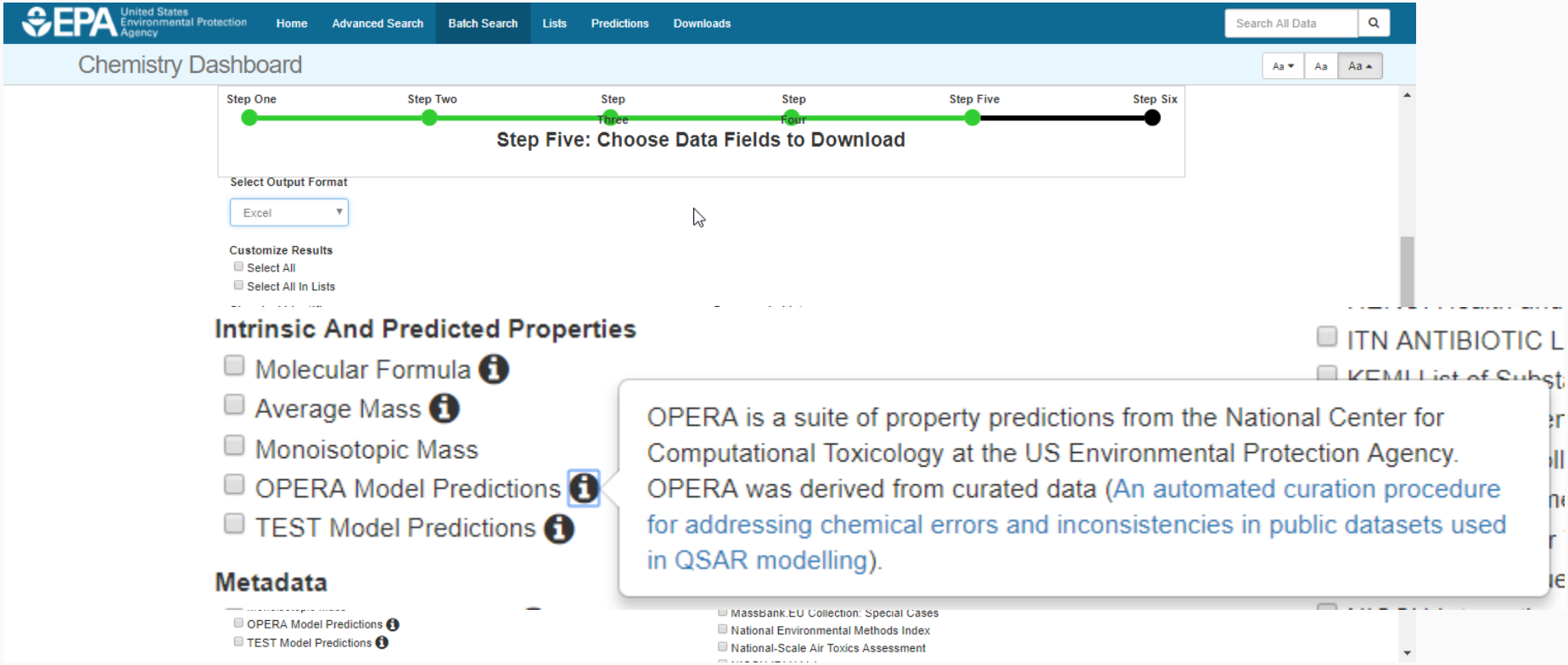
Batch Searches



Batch Search



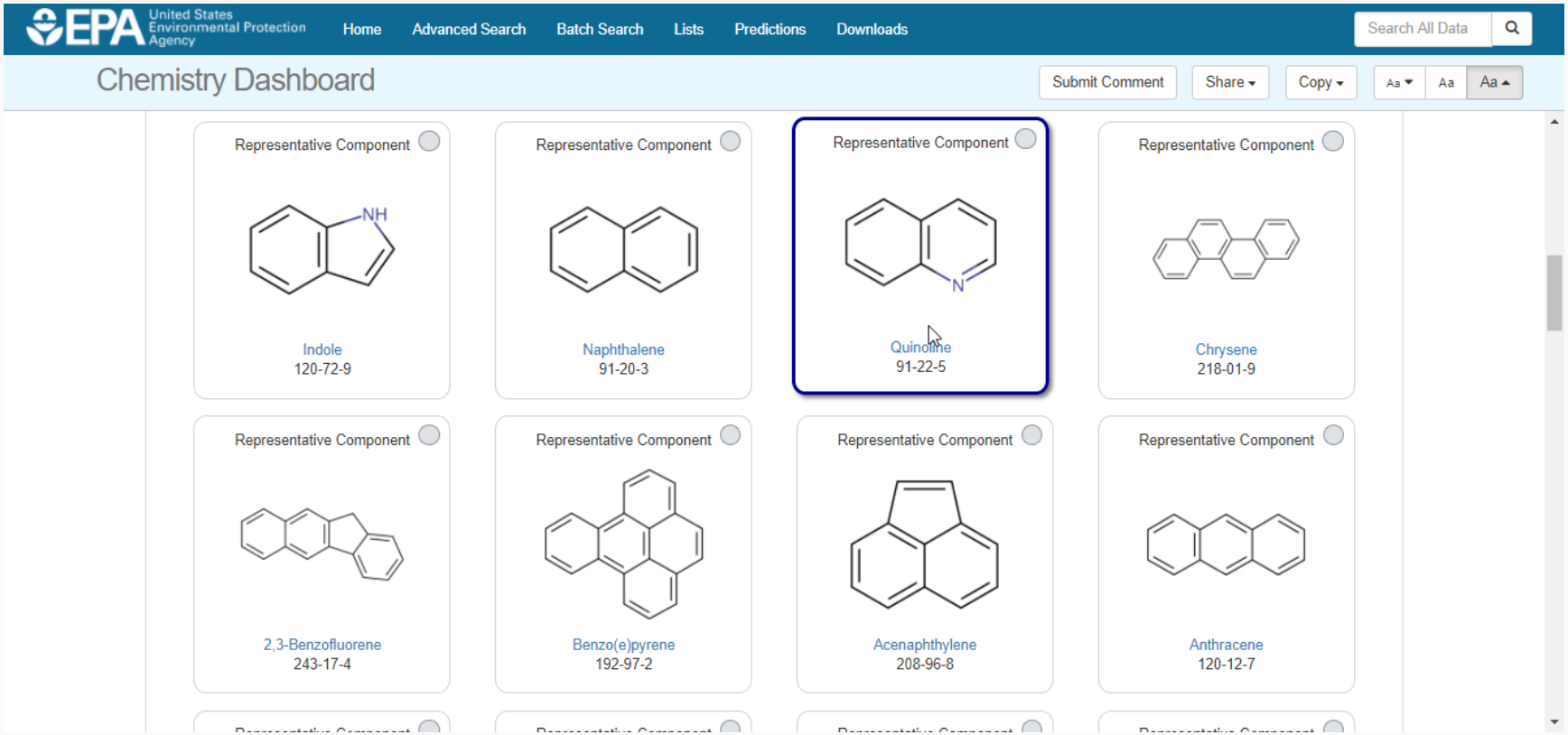
OPERA and TEST in Batch



Excel Output

	A	B	C	D	E	F	G	H
1	INPUT	FOUND_BY	DTXSID	PREFERRED_NAME	EXPOCAST	EXPOCAST	NHANES	TOXVAL_D
2	1445-75-6	CAS-RN	DTXSID5024051	Diisopropyl methylpho	2.09e-08	Y	-	Y
3	50-00-0	CAS-RN	DTXSID7020637	Formaldehyde	1.32e-06	Y	-	Y
4	107-06-2	CAS-RN	DTXSID6020438	1,2-Dichloroethane	4.9e-06	Y	-	Y
5	57-12-5	CAS-RN	DTXSID6023991	Cyanide	-	-	-	Y
6	7550-45-0	CAS-RN	DTXSID8042476	Titanium tetrachloride	-	-	-	Y
7	79-01-6	CAS-RN	DTXSID0021383	Trichloroethylene	7.27e-06	Y	-	Y
8	121-82-4	CAS-RN	DTXSID9024142	Cyclonite	6.72e-08	Y	-	Y
9	108-05-4	CAS-RN	DTXSID3021431	Vinyl acetate	8.3e-05	Y	-	Y
10	7803-51-2	CAS-RN	DTXSID2021157	Phosphine	-	-	-	Y
11	122-66-7	CAS-RN	DTXSID7020710	1,2-Diphenylhydrazine	1.49e-07	Y	-	Y
12	101-77-9	CAS-RN	DTXSID6022422	4,4'-Methylenedianiline	6.08e-06	Y	-	Y
13	14017-34-6	CAS-RN	DTXSID90161250	Selenium difluoride	-	-	-	-
14	75-44-5	CAS-RN	DTXSID0024260	Phosgene	-	-	-	Y
15	621-64-7	CAS-RN	DTXSID6021032	N-Nitrosodipropylamine	4.55e-07	Y	-	Y
16	75-09-2	CAS-RN	DTXSID0020868	Dichloromethane	2.02e-06	Y	-	Y
17	100-41-4	CAS-RN	DTXSID3020596	Ethylbenzene	8.32e-05	Y	-	Y
18	7440-28-0	CAS-RN	DTXSID2036035	Thallium	-	-	-	Y
19	108-88-3	CAS-RN	DTXSID7021360	Toluene	8.61e-05	Y	-	Y
20	111-44-4	CAS-RN	DTXSID9020168	Bis(2-chloroethyl) ethe	2.82e-07	Y	-	Y
21	7440-42-8	CAS-RN	DTXSID3023922	Boron	-	-	-	Y
22	7440-29-1	CAS-RN	DTXSID6049800	Thorium	-	-	-	Y

Families of chemicals
Polyaromatic Hydrocarbons

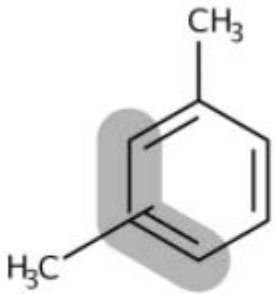


Not all chemicals are “structures”

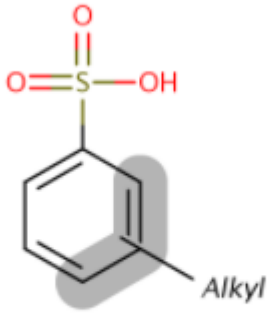
- UVCBs are chemical substances of unknown or variable composition, complex reaction products and biological materials
  - Surfactants (C11-14 linear alkyl sulfonates)
  - Reaction mass of p-t-butylphenyldiphenyl phosphate and bis(p-t-butylphenyl)phenyl phosphate and triphenyl phosphate
  - Almond Oil

“Markush Structures”

https://en.wikipedia.org/wiki/Markush_structure

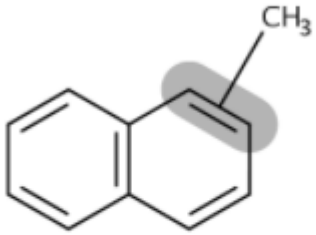


Xylenes
1330-20-7

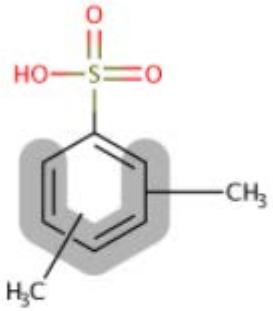


(C10-C16) Alkylbenzenesulfonic acid
68584-22-5

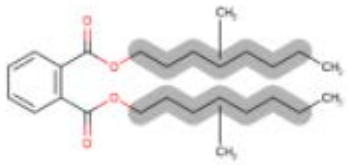
n-Nonylphenol
25154-52-3



Methylnaphthalene
1321-94-4



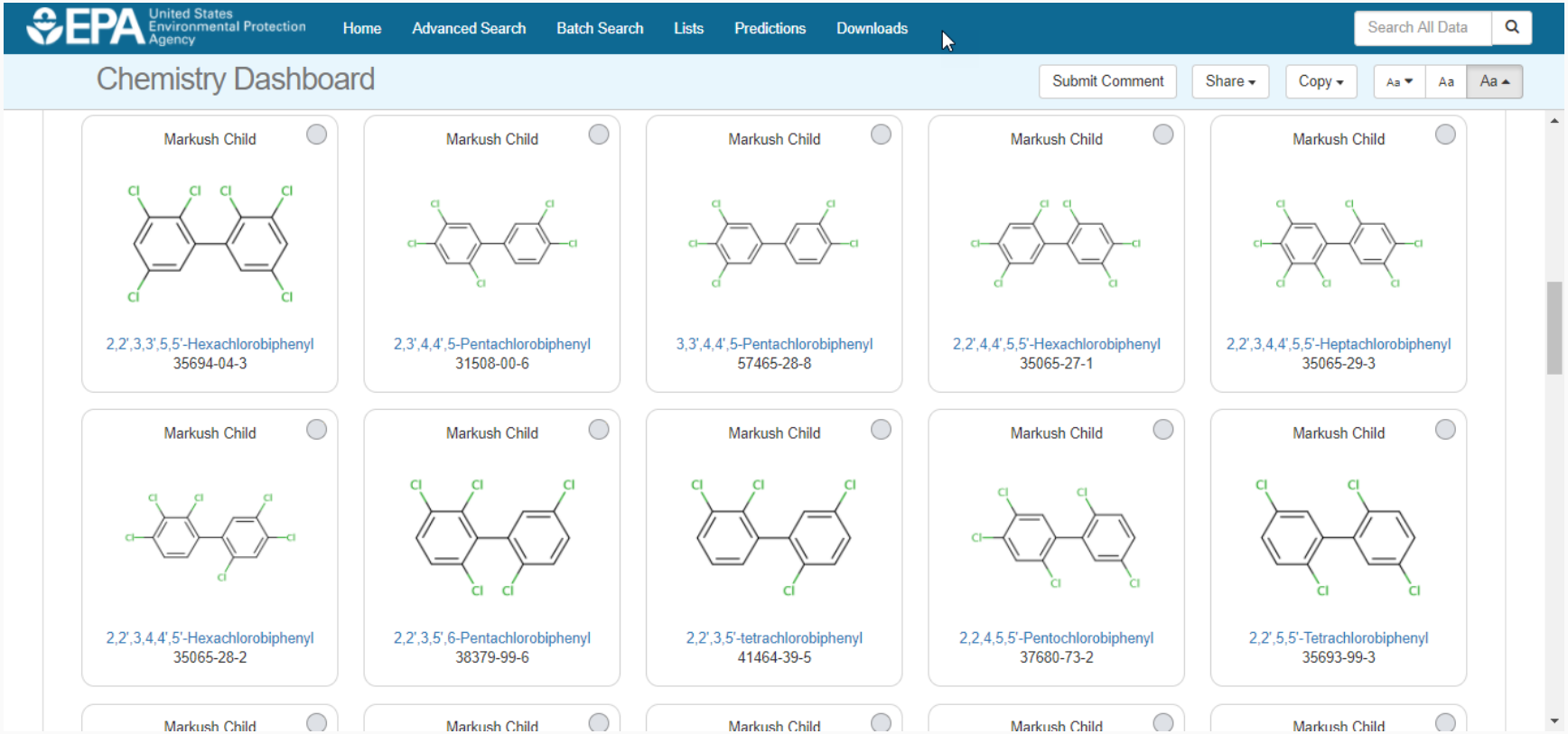
Sodium xylenesulfonate
1300-72-7



Diisononyl phthalate
28553-12-0

Enumeration of Markush

- Markush structures can be enumerated into chemical families



From Dashboard to Prioritization

- Can we bring data together to prioritize risk?
  - Integrated dashboard data can be used to rank order and prioritize risk

- Internal application in development to use scoring to prioritize risk – uses combination of available experimental data and new assessment methods (NAMs)

Potential Data Sources

***In Vivo* Human Hazard:**

- Mammalian toxicity studies – guideline-like, use POD
- System-specific *in vivo* data (Cancer, developmental)
- Models (QSAR) to predict POD and organ-specific effects
- Genotoxicity
- *In vitro*-derived endocrine disruption and neurotoxicity models

***In Vivo* Eco Hazard**

- Aquatic *in vivo* studies – POD
- Models (QSAR) of POD

Potential Data Sources

EPA United States Environmental Protection Agency

Human Exposure

- Data on production volume and releases
- Quantitative biomonitoring data
- Predictions of oral and inhalation exposure

Eco Exposure

- Biomonitoring data
- Predictions of water concentrations

Physchem

- Persistence and Bioaccumulation models (OPERA)


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

Journal of Cheminformatics

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]

Scoring approaches

- For each chemical, each domain receives a score of 1 (Low), 2 (Moderate), or 3 (High) concern
- **Hazard score** = maximum of human and eco hazard scores
- **Exposure score** = maximum of human and eco exposure scores
- **Total score** = hazard score + exposure score + physchem score
- If no data is available for a domain, it is given the "missing data score", currently 1 (Low)
- Scoring can include or exclude NAM

Implemented Scoring Methods

EPA United States Environmental Protection Agency

Method 1

- Maximum score from human and eco hazard: 1 – 3
- Maximum score from human and eco exposure: 1 – 3
- Maximum score from persistence/ bioaccumulation (P/B): 1 – 3
- No NAM
- Add hazard, exposure, and P/B
- Categorical bins
  - High: 7-9
  - Moderate: 5-6
  - Low: 3-4

Method 2: NAM Equal

- Same as Method 1 except NAM is incorporated with equal weighting in all domains
- Add hazard, exposure, and P/B
- Categorical bins
  - High: 7-9
  - Moderate: 5-6
  - Low: 3-4

Method 3: NAM Deferential

- Same as Method 1 except human hazard NAM is incorporated in the absence of traditional *in vivo* studies
- In other domains, NAM is given equal weight.
- Add hazard, exposure, and P/B
- Categorical bins
  - High: 7-9
  - Moderate: 5-6
  - Low: 3-4

Method 4: Sum of Scores

- Sum all components (incl. NAM) from human and eco hazard
- Sum all components (incl. NAM) from human and eco exposure
- Sum all components (incl. NAM) from persistence/ bioaccumulation
- Add hazard, exposure, and P/B
- Categorical bins
  - High: >30
  - Medium: 10-30
  - Low: ≤10

Method 5: H/BER*

- Ratio of the minimum effect level from *in vivo* toxicity studies or the quantitative human hazard NAM data divided by the maximum oral exposure
- Categorical bins
  - High: $\leq 10^4$
  - Medium: $10^4 – 10^6$
  - Low: $\geq 10^6$

*Hazard/Bioactivity Exposure Ratio

Overall Scoring Page

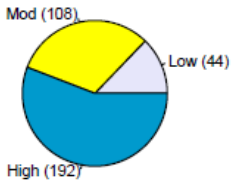
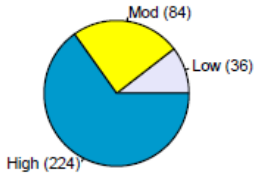
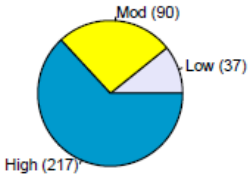
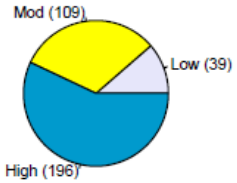
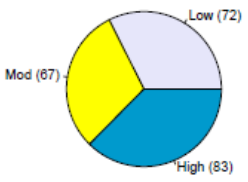
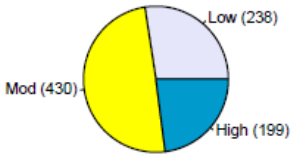
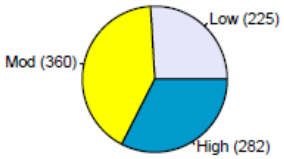
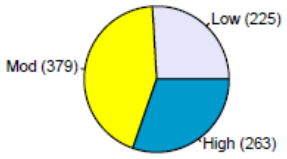
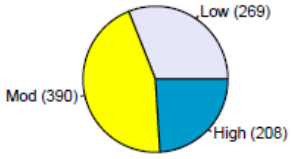
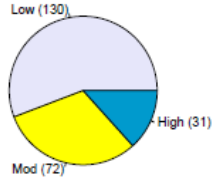
Cutoffs | Domain Help | Human Hazard Prioritization | Eco Hazard Prioritization | Human Exposure Prioritization | Eco Exposure Prioritization

Physchem Prioritization | Overall Prioritization | Data Coverage Summary | Distribution Graphs

Show 25 entries Search:

DSSTox_ID	CASRN	Name	Method 1 score	Method 1 bin	Method 2 score	Method 2 bin	Method 3 score	Method 3 bin	Method 4 score	Method 4 bin
DTXSID124356	1234-56-7	Name1	5	Moderate (ExeHzeHzh)	6	Moderate (ExeHzeHzh)	6	Moderate (ExeHzeHzh)	25	Moderate (ExeHzeHzh)
DTXSID124357	1234-56-8	Name2	5	Moderate (PbExeHzeHzh)	7	High (PbExeHzeHzh)	7	High (PbExeHzeHzh)	24	Moderate (PbExeHzeHzh)
DTXSID124358	1234-56-9	Name3	5	Moderate (ExeHzeHzh)	6	Moderate (ExeHzeHzh)	6	Moderate (ExeHzeHzh)	41	High (ExeHzeHzh)
DTXSID124359	1234-56-10	Name4	5	Moderate (ExeHzeHzh)	5	Moderate (ExeHzeHzh)	5	Moderate (ExeHzeHzh)	13	Moderate (ExeHzeHzh)
DTXSID124360	1234-56-11	Name5	5	Moderate (PbExeHzeHzh)	5	Moderate (PbExeHzeHzh)	5	Moderate (PbExeHzeHzh)	20	Moderate (PbExeHzeHzh)
DTXSID124361	1234-56-12	Name6	5	Moderate (ExeHzeHzh)	5	Moderate (ExeHzeHzh)	5	Moderate (ExeHzeHzh)	35	High (ExeHzeHzh)
DTXSID124362	1234-56-13	Name7	5	Moderate (PbExeHze)	5	Moderate (PbExeHze)	5	Moderate (PbExeHze)	21	Moderate (PbExeHze)
DTXSID124363	1234-56-14	Name8	6	Moderate (PbExeHze)	6	Moderate (PbExeHze)	6	Moderate (PbExeHze)	16	Moderate (PbExeHze)
DTXSID124364	1234-56-15	Name9	5	Moderate (PbExeHze)	5	Moderate (PbExeHze)	5	Moderate (PbExeHze)	26	Moderate (PbExeHze)
DTXSID124365	1234-56-16	Name10	6	Moderate (PbExeHze)	6	Moderate (PbExeHze)	6	Moderate (PbExeHze)	23	Moderate (PbExeHze)
DTXSID124366	1234-56-17	Name11	5	Moderate	5	Moderate	5	Moderate	17	Moderate

Fraction of chemicals in each bin

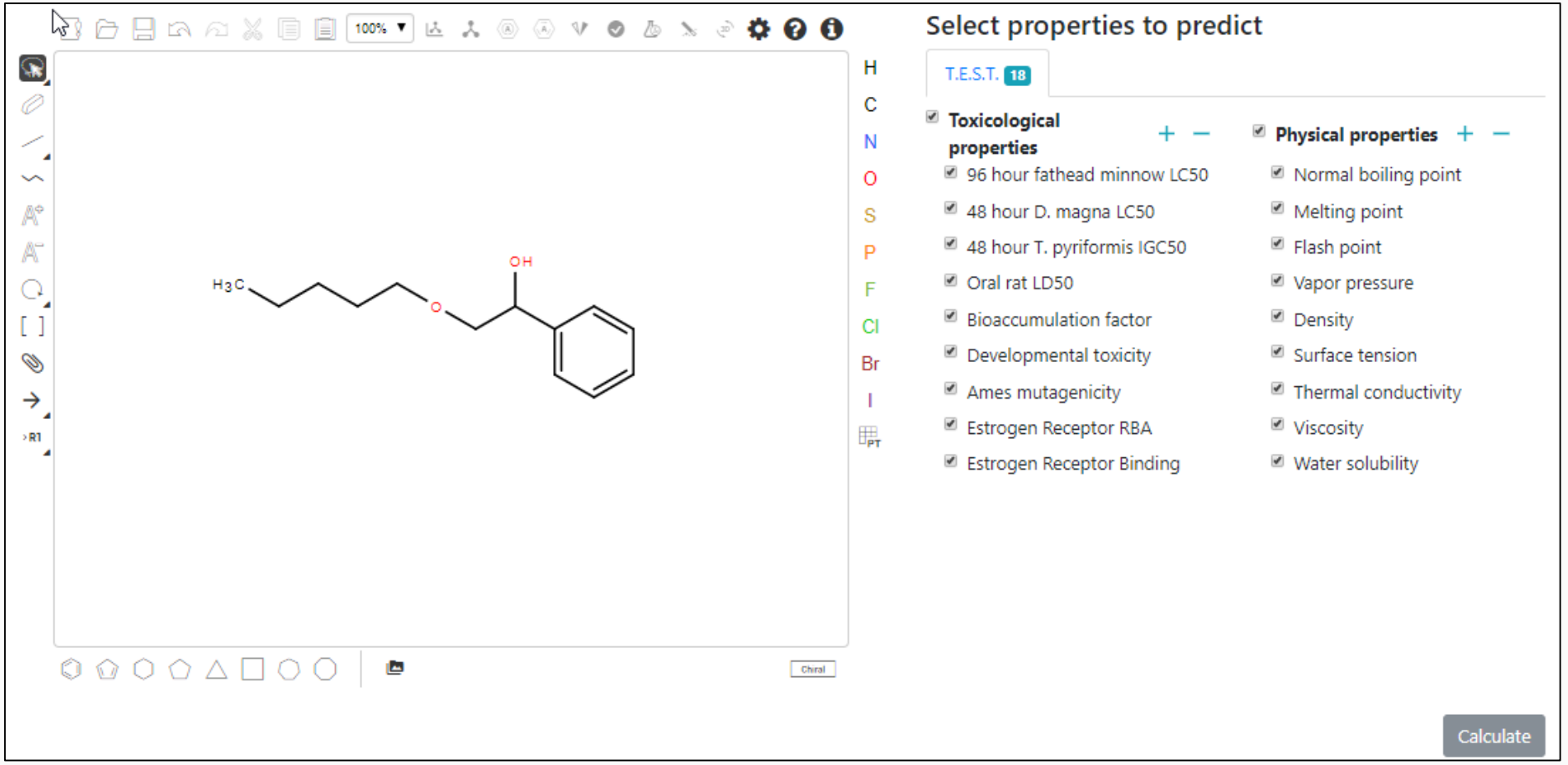
	General	NAM Equal	NAM Differential	Sum of Scores	H/BER
Set 1 (344)					 (222/344)
Set 2 (867)					 (233/867)

High Pre-Priority Moderate Pre-Priority Low Pre-Priority

Work in Progress

- Present work in development
  - Real time OPERA predictions – TEST predictions done

Real-Time Predictions



Real-Time Predictions

Property	Experimental Value	Prediction				
		Consensus	Hierarchical clustering	Single model	Group contribution	Nearest neighbor
96 hour fathead minnow LC50		4.477 -Log10(mol/L) 6.954 mg/L	4.195 -Log10(mol/L) 13.288 mg/L	3.994 -Log10(mol/L) 21.110 mg/L	3.478 -Log10(mol/L) 69.224 mg/L	6.238 -Log10(mol/L) 0.120 mg/L
48 hour D. magna LC50		4.398 -Log10(mol/L) 8.328 mg/L	3.877 -Log10(mol/L) 27.677 mg/L	4.039 -Log10(mol/L) 19.026 mg/L	4.084 -Log10(mol/L) 17.173 mg/L	5.593 -Log10(mol/L) 0.532 mg/L
48 hour T. pyriformis IGC50		4.063 -Log10(mol/L) 18.039 mg/L	3.731 -Log10(mol/L) 38.668 mg/L		3.386 -Log10(mol/L) 85.610 mg/L	5.070 -Log10(mol/L) 1.773 mg/L
Oral rat LD50		1.758 -Log10(mol/kg) 3640.950 mg/kg	1.982 -Log10(mol/kg) 2172.756 mg/kg			1.533 -Log10(mol/kg) 6101.245 mg/kg
Bioaccumulation factor		1.797 Log10 62.700	2.202 Log10 159.310	1.287 Log10 19.346	1.181 Log10 15.157	2.520 Log10 330.834
Developmental toxicity		false	false	false		true
Ames mutagenicity		false	false			false
Estrogen Receptor RBA		-3.075 Log10 $8.418*10^{-4}$	-3.078 Log10 $8.356*10^{-4}$	-3.720 Log10 $1.907*10^{-4}$		-2.427 Log10 0.004
Estrogen Receptor Binding		true	true	true	false	true

Real-Time Predictions

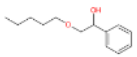
Predicted Water solubility at 25°C for OC(C=1C=CC=CC1)COCCCCC from Consensus method

Prediction results

Endpoint	Experimental value	Predicted value
Water solubility at 25°C -Log10(mol/L)	N/A	2.46
Water solubility at 25°C mg/L	N/A	723.26

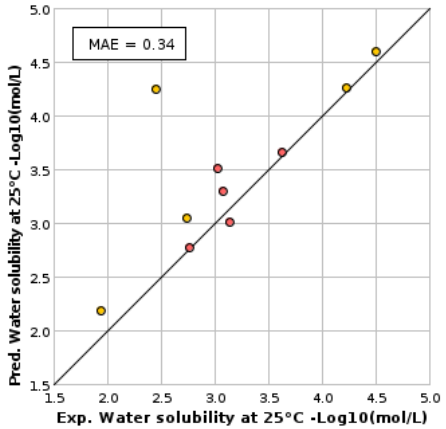
Individual Predictions

Method	Predicted value -Log10(mol/L)
Hierarchical clustering	2.42
Group contribution	2.32
Nearest neighbor	2.64



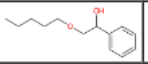
Predictions for the test chemical and for the most similar chemicals

Prediction results (colors defined in table below)



Chemicals	MAE*
Entire set	0.58
Similarity coefficient ≥ 0.5	0.34

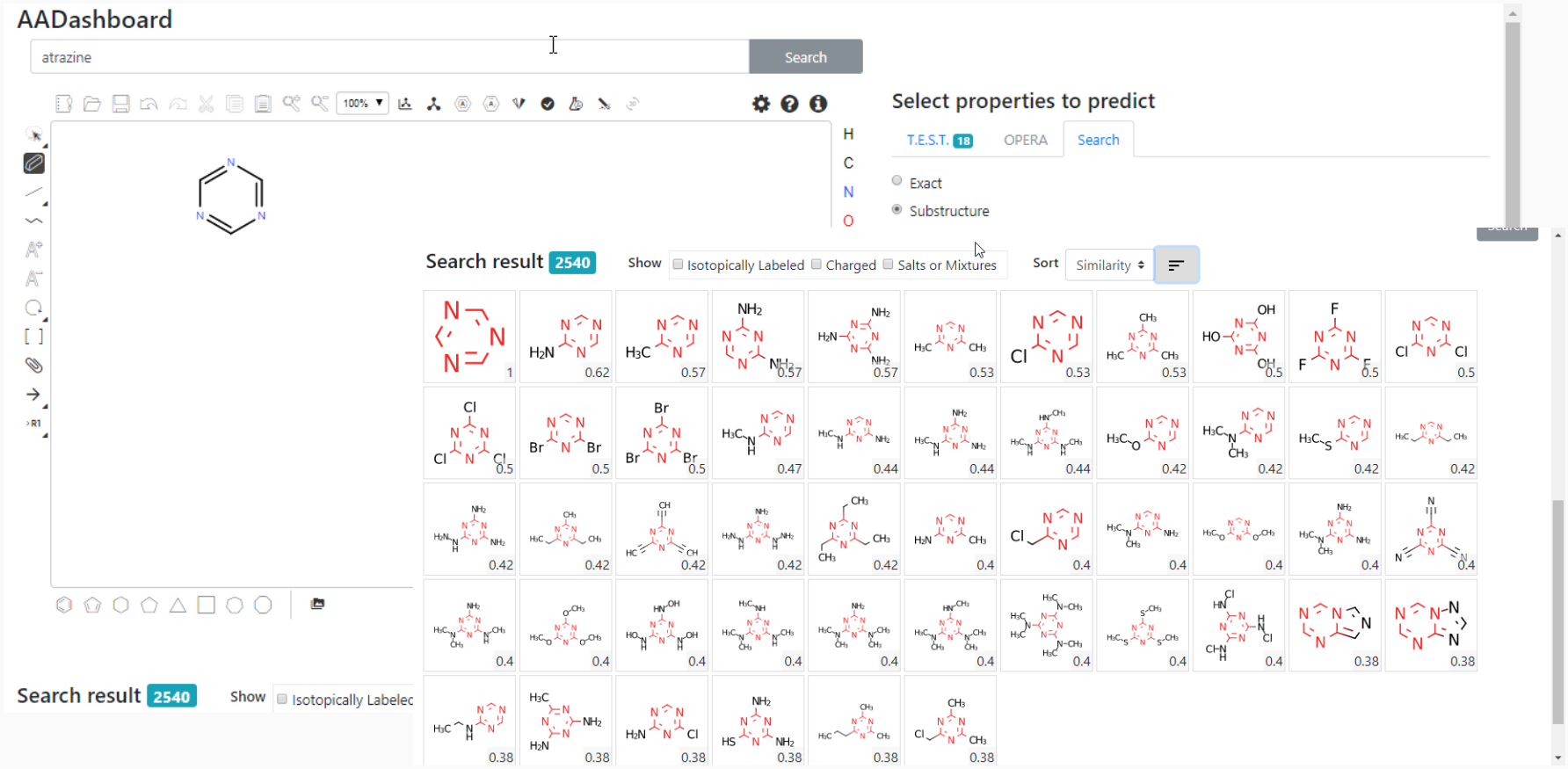
*Mean absolute error in -Log10(mol/L)

CAS	Structure	Similarity Coefficient	Experimental value -Log10(mol/L)	Predicted value -Log10(mol/L)
OC(C=1C=CC=CC1)COCCCCC (test chemical)			N/A	2.46
104-40-5		0.68	4.50	4.60
1219-38-1		0.67	4.22	4.26

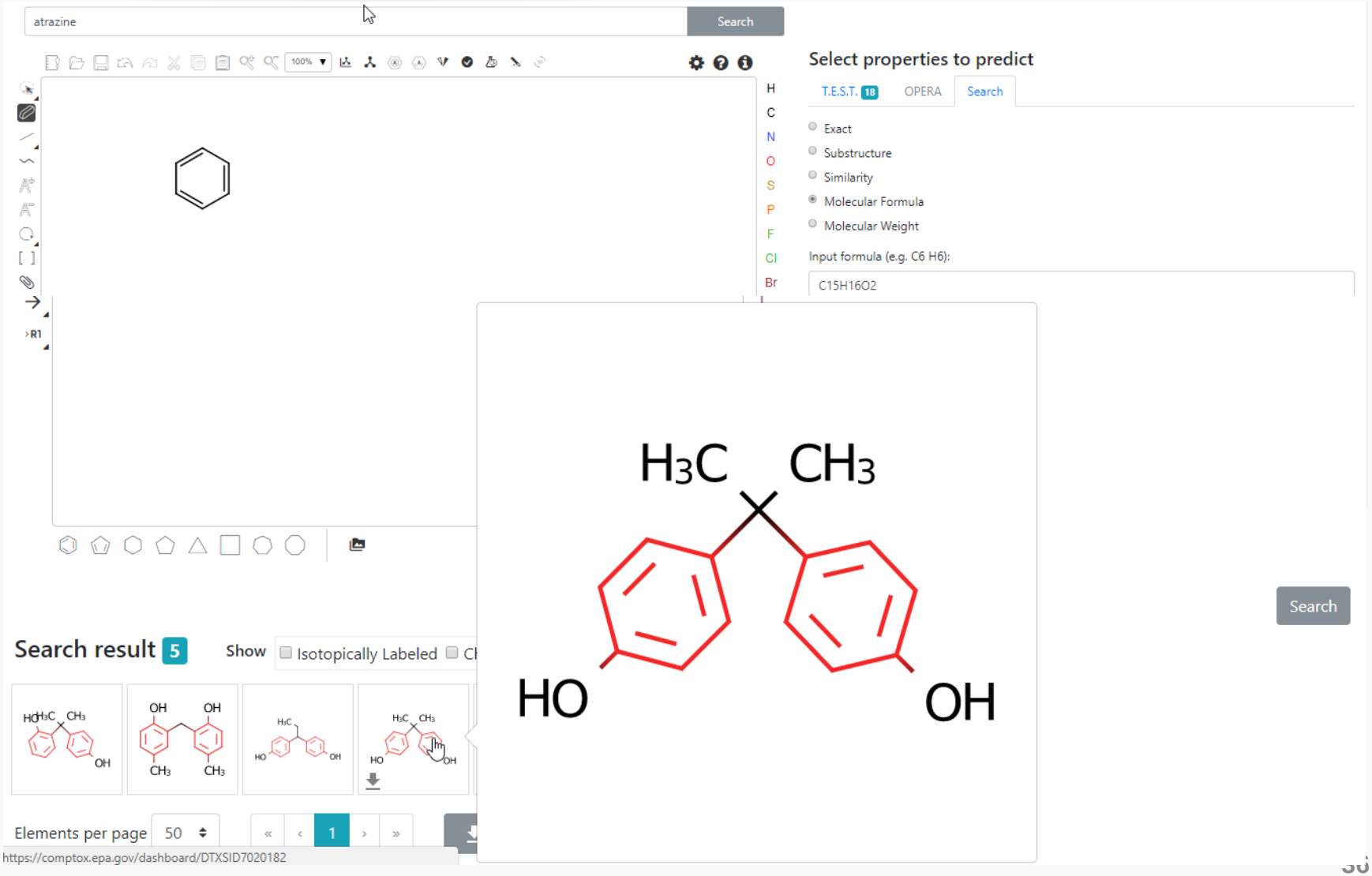
Work in Progress

- Present work in development
  - Real time OPERA prediction
  - Structure/substructure/similarity search

Prototype Development



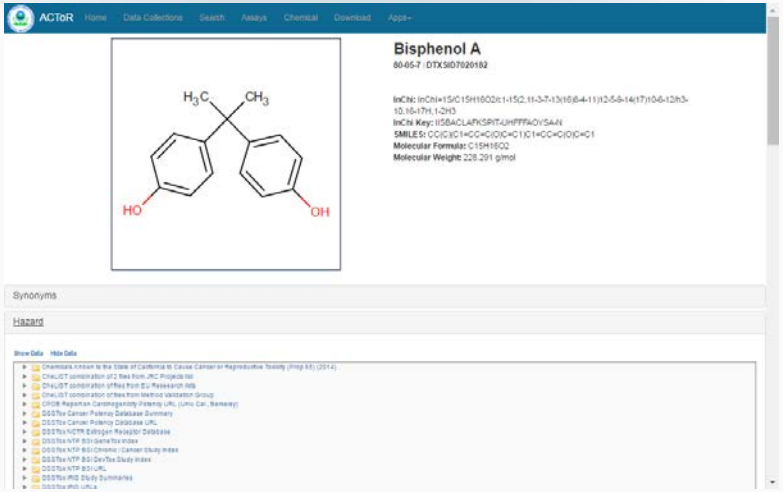
Prototype Development

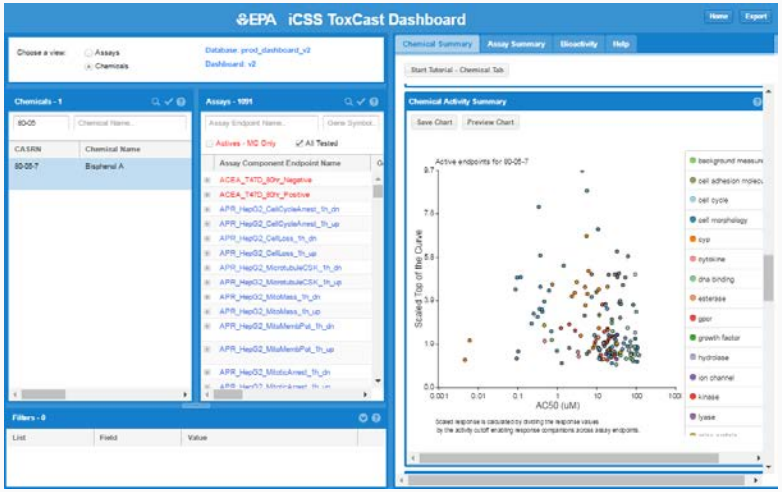


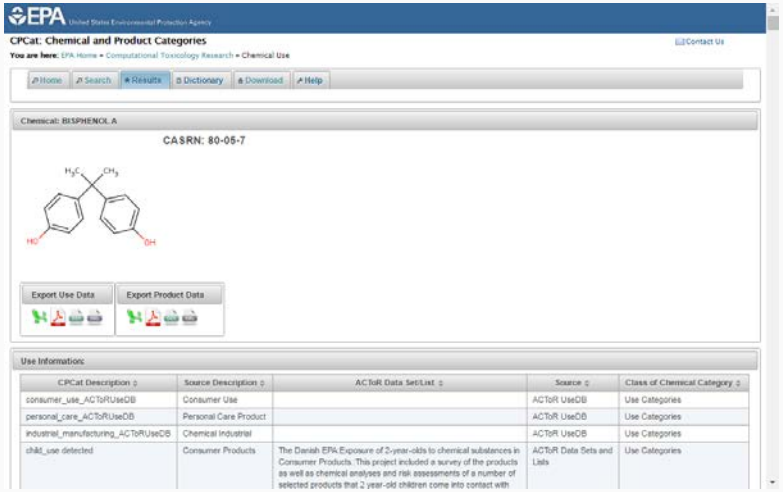
Work in Progress

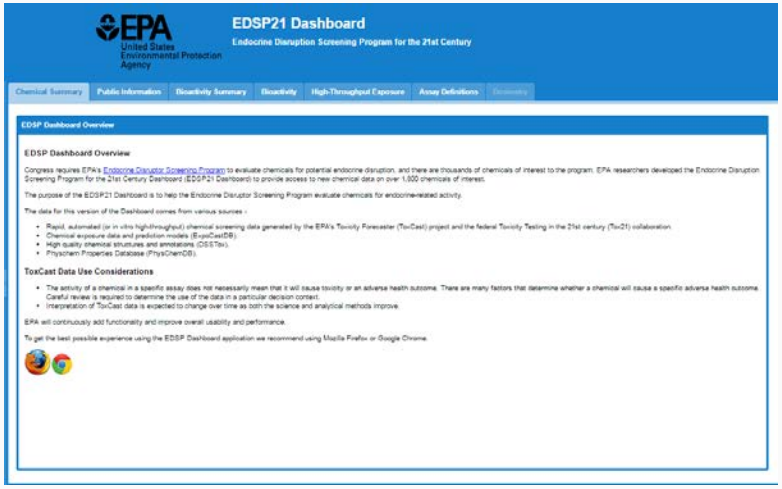
- Present work in development
  - Real time OPERA prediction
  - Structure/substructure/similarity search
  - Merging in other NCCT dashboard capabilities

Earlier Dashboard Applications









Work in Progress

- Present work in development
  - Real time OPERA prediction
  - Structure/substructure/similarity search
  - Merging in other NCCT dashboard capabilities
  - Web-services for community consumption
    - TEST predictions already available
    - Chemical Resolver service
    - Embeddable widgets

Conclusion

- The CompTox Chemistry Dashboard provides access to data for ~760,000 chemicals
- Multiple prediction models available for data gap filling
  - OPERA models and TEST models – PhysChem and Tox endpoints
  - Models based on *in vitro* data – classification models
  - Generalized Read-Across development in progress
- Real time prediction models rollout has started
- Web services available for some physchem and toxicity endpoints
- 2 years development as a CompTox Integration Hub

Contact

Antony Williams
US EPA Office of Research and Development
National Center for Computational Toxicology (NCCT)
Williams.Antony@epa.gov
ORCID: https://orcid.org/0000-0002-2668-4821