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

New developments in delivering public access to data from the National Center for Computational Toxicology at the EPA

Antony Williams[1], Chris Grulke[1], Andrew McEachran[2], Grace Patlewicz[1], Imran Shah[1], John Wambaugh[1], Ann Richard[1], Richard Judson[1] and Jeff Edwards[1]

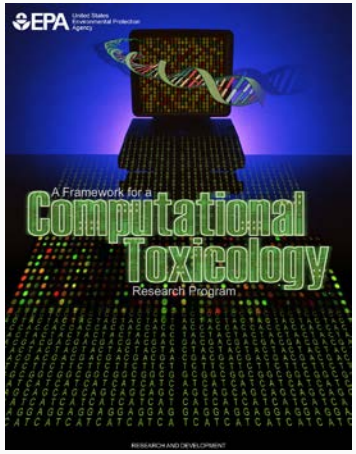
1) National Center for Computational Toxicology, U.S. Environmental Protection Agency, RTP, NC

2) Oak Ridge Institute of Science and Education (ORISE) Research Participant, Research Triangle Park, NC

March 2018

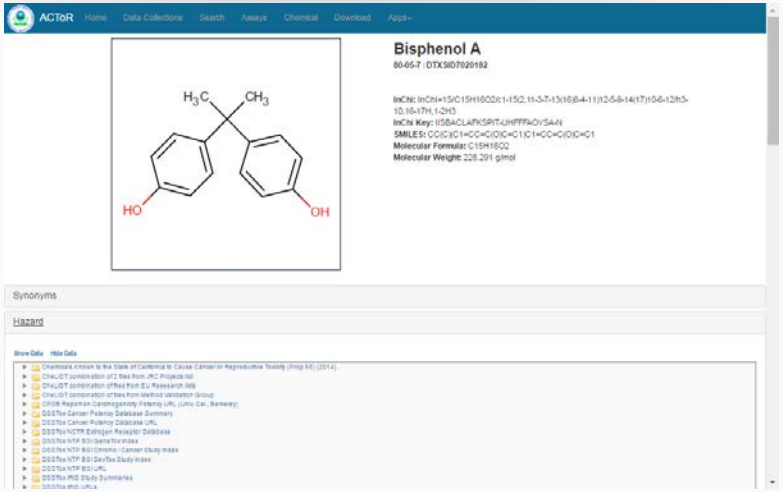
ACS Spring Meeting, New Orleans

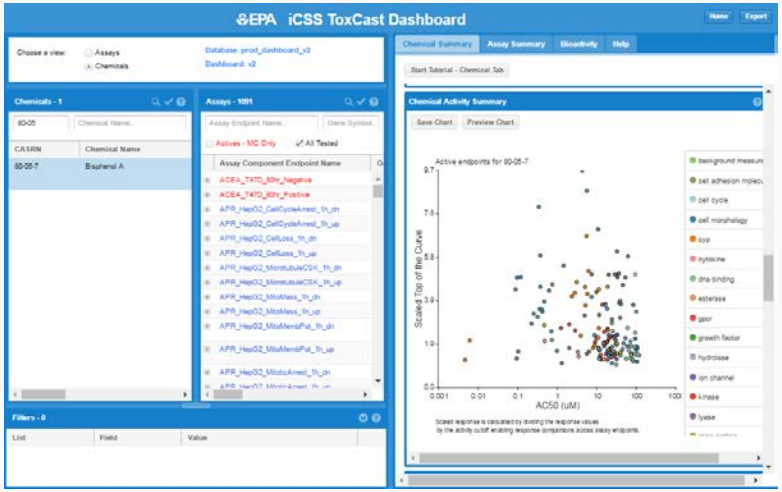
National Center for Computational Toxicology

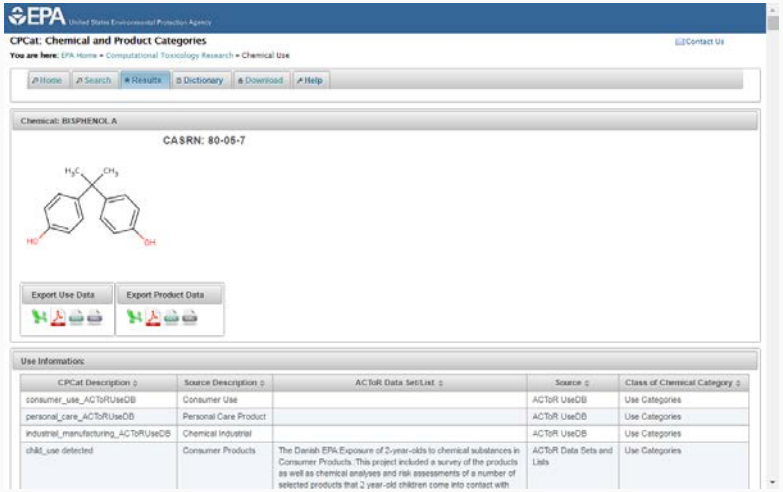


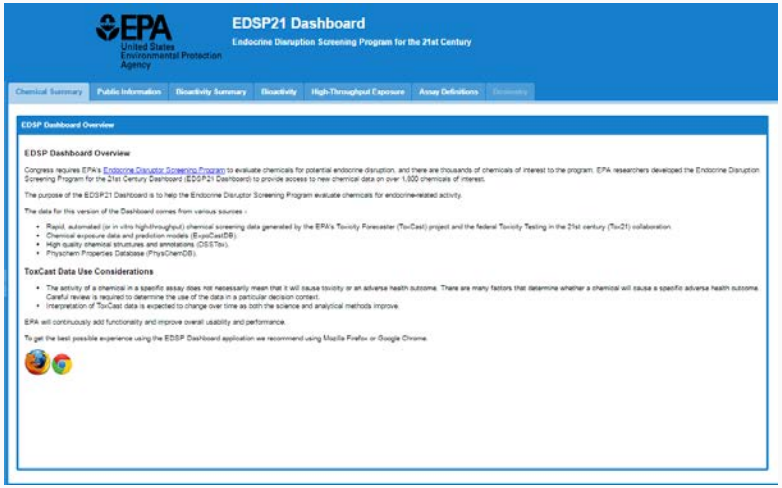
- National Center for Computational Toxicology established in 2005 to integrate:
  - High-throughput and high-content technologies
  - Modern molecular biology
  - Data mining and statistical modeling
  - Computational biology and chemistry
- Outputs: a lot of data, models, algorithms and software applications
- Open Data – we want scientists to interrogate it, learn from it, develop understanding

Early Dashboard Applications









Earlier Dashboards

- Chemistry data mashed together based on CAS Number/Names
- Chemistry data quality issues
- Multiple applications requiring maintenance

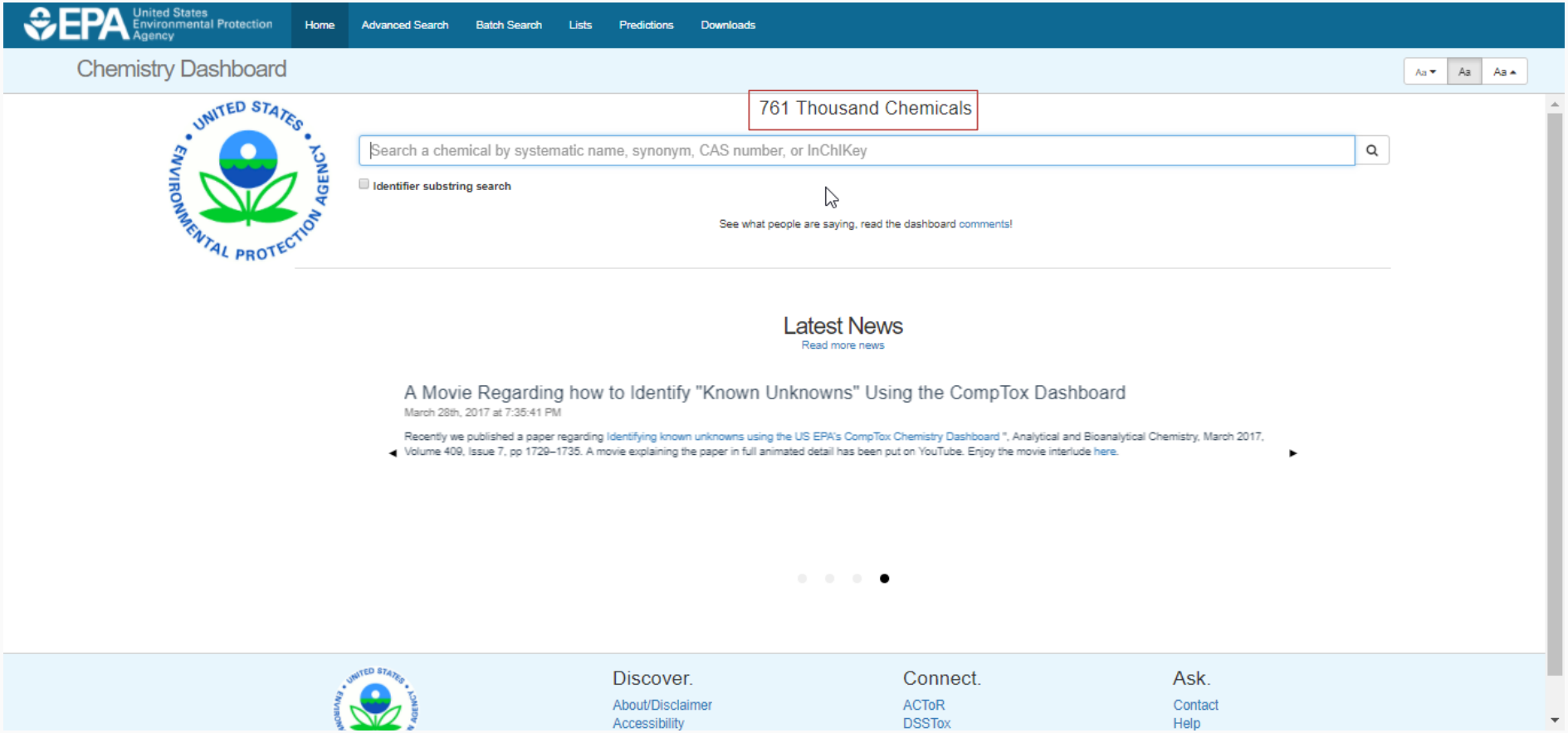
- April 2016 – beta release of the CompTox Chemistry Dashboard as an integration hub

The CompTox Chemistry Dashboard

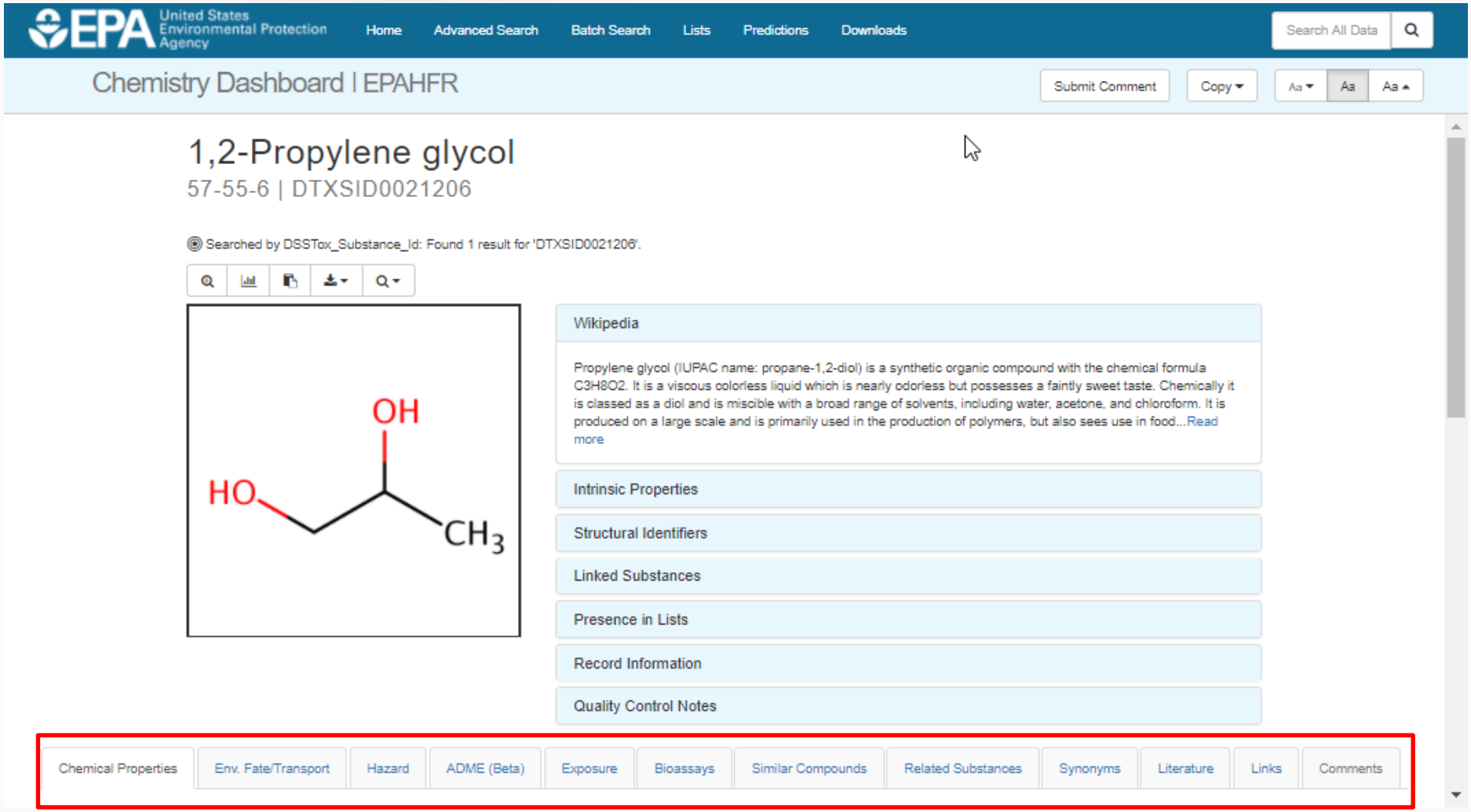
- A **publicly accessible website** delivering access:
  - ~760,000 chemicals with related property data
  - Experimental and predicted physicochemical property data
  - Integration to “biological assay data” for 1000s of chemicals
  - Information regarding consumer products containing chemicals
  - Links to other agency websites and public data resources
  - “Literature” searches for chemicals using public resources
  - “Batch searching” for thousands of chemicals
  - **DOWNLOADABLE Open Data for reuse and repurposing**

CompTox Chemistry Dashboard

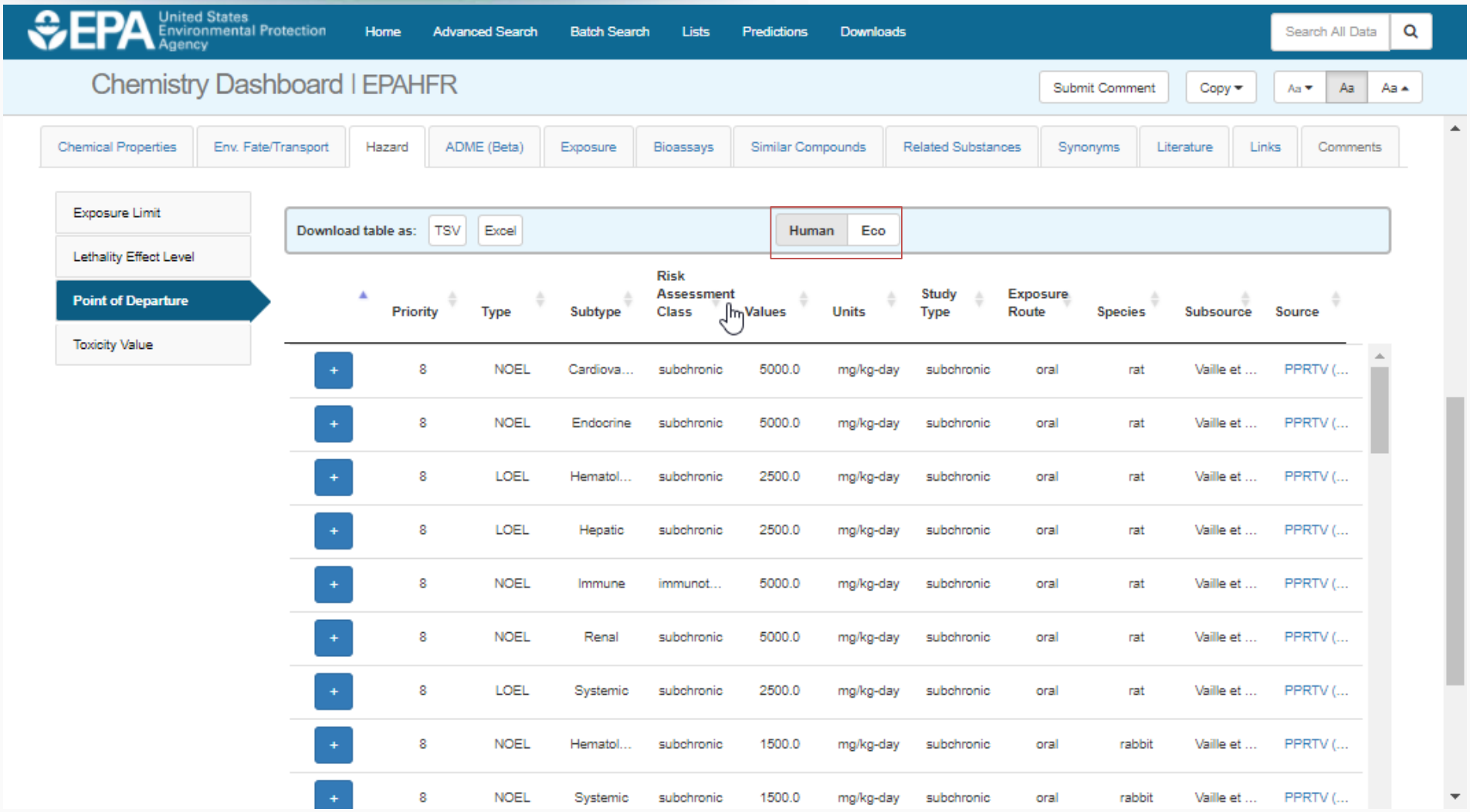
https://comptox.epa.gov/dashboard



Detailed Chemical Pages



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

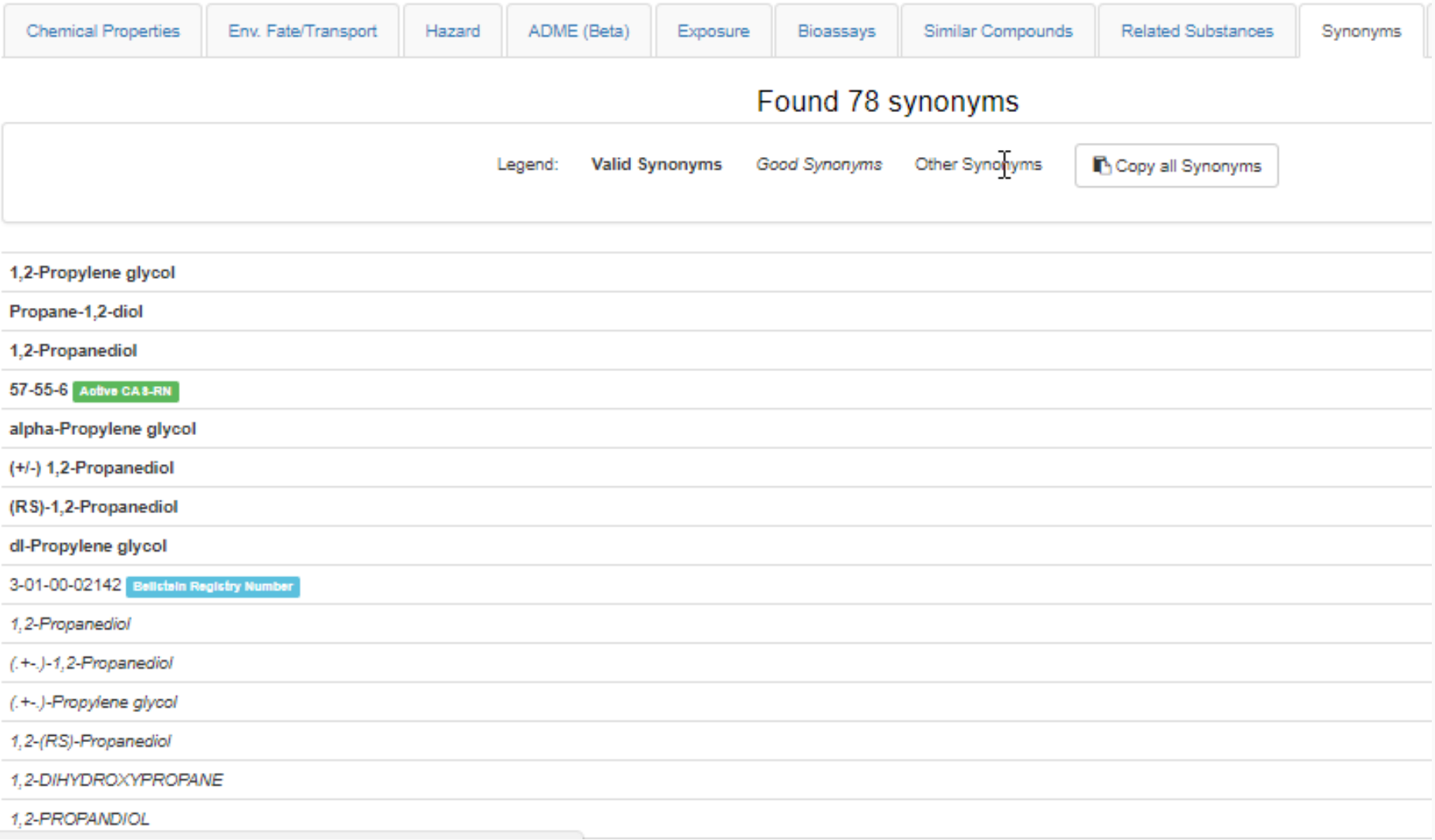
osure | Bioassays | Similar Compounds | Related Substances | Synonyms | Literature | Links | Comments

Product & Use Categories (PUCs)

Categorization type	Number of Unique Products
PUC	288
PUC	206
PUC	117
PUC	107
PUC	107
PUC	101
PUC	101
PUC	90
PUC	89

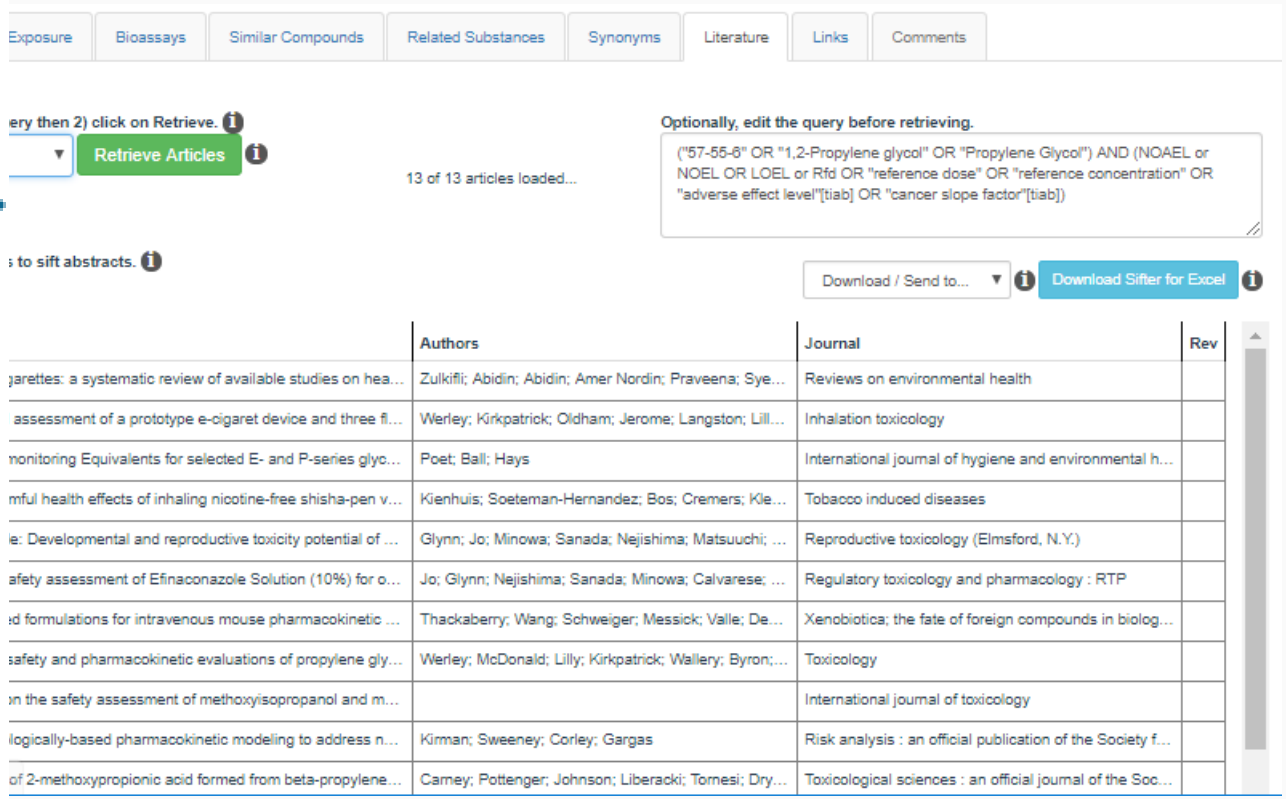
Identifiers to Support Searches

EPA United States Environmental Protection Agency

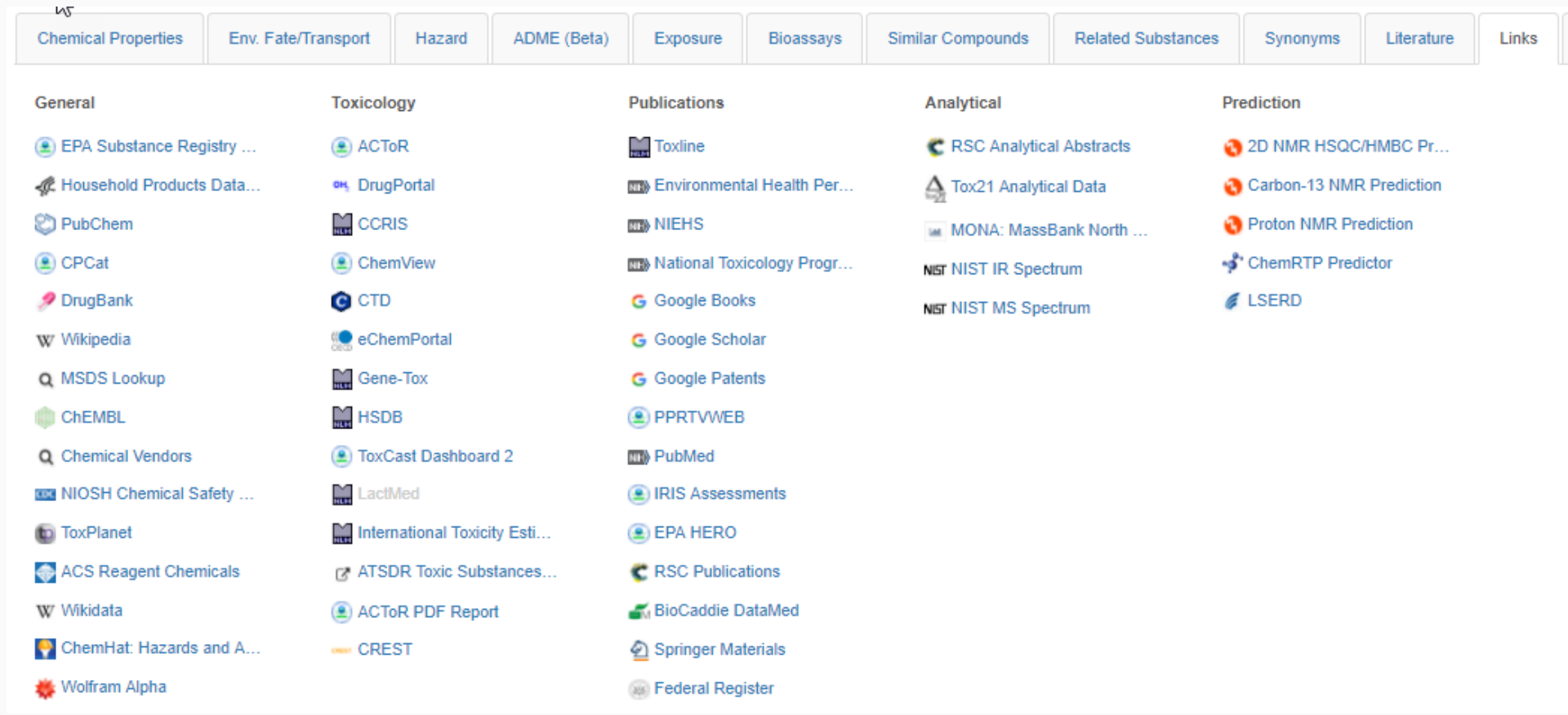


Literature Searches and Links

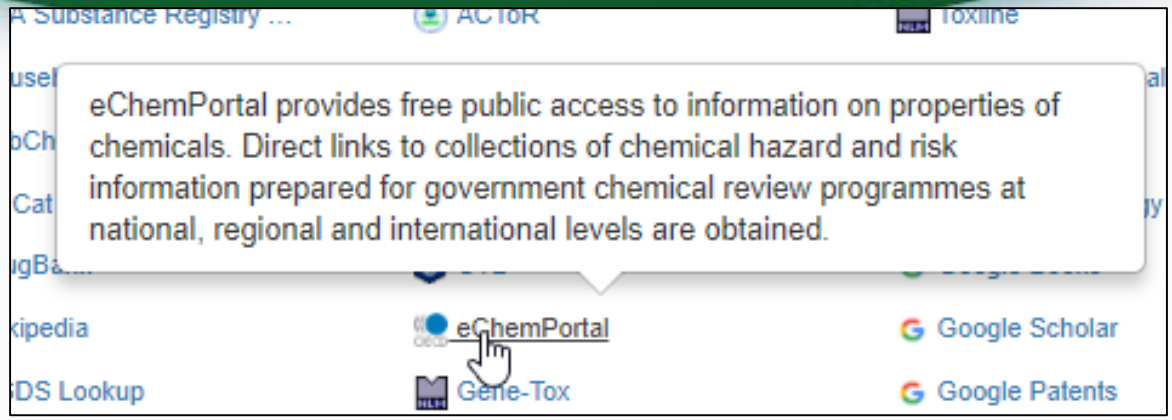
- Google Scholar
- **PubMed Abstract Sifter**
- PubChem Articles
- PubChem Patents
- PPRTV
- IRIS

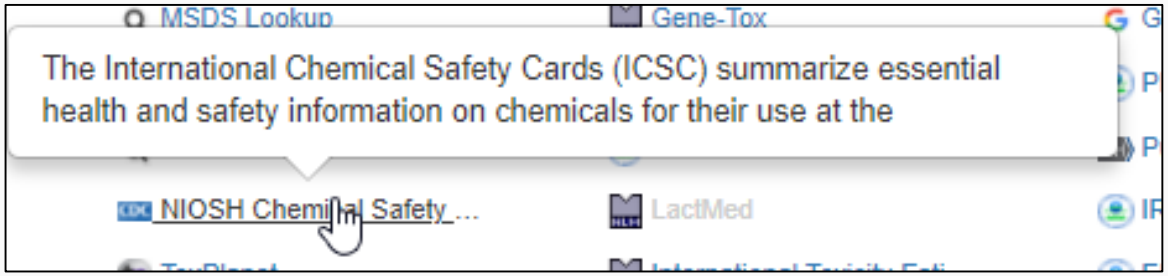


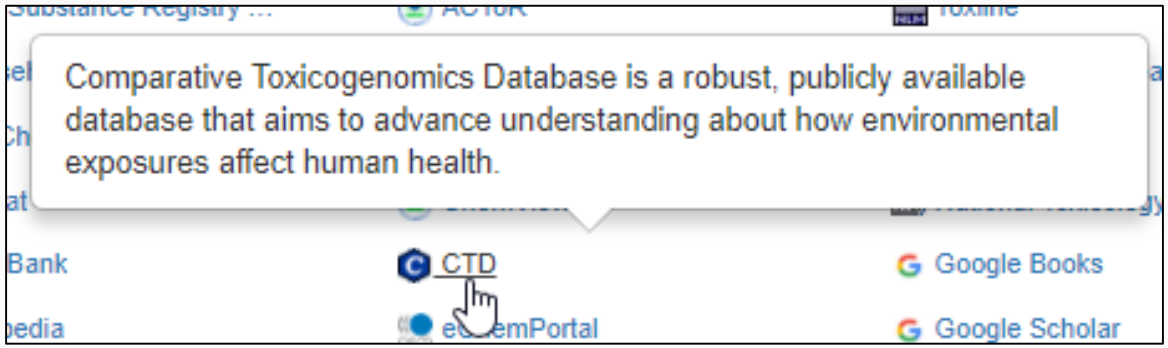
External Links to Data and Services



Integrated Linkouts



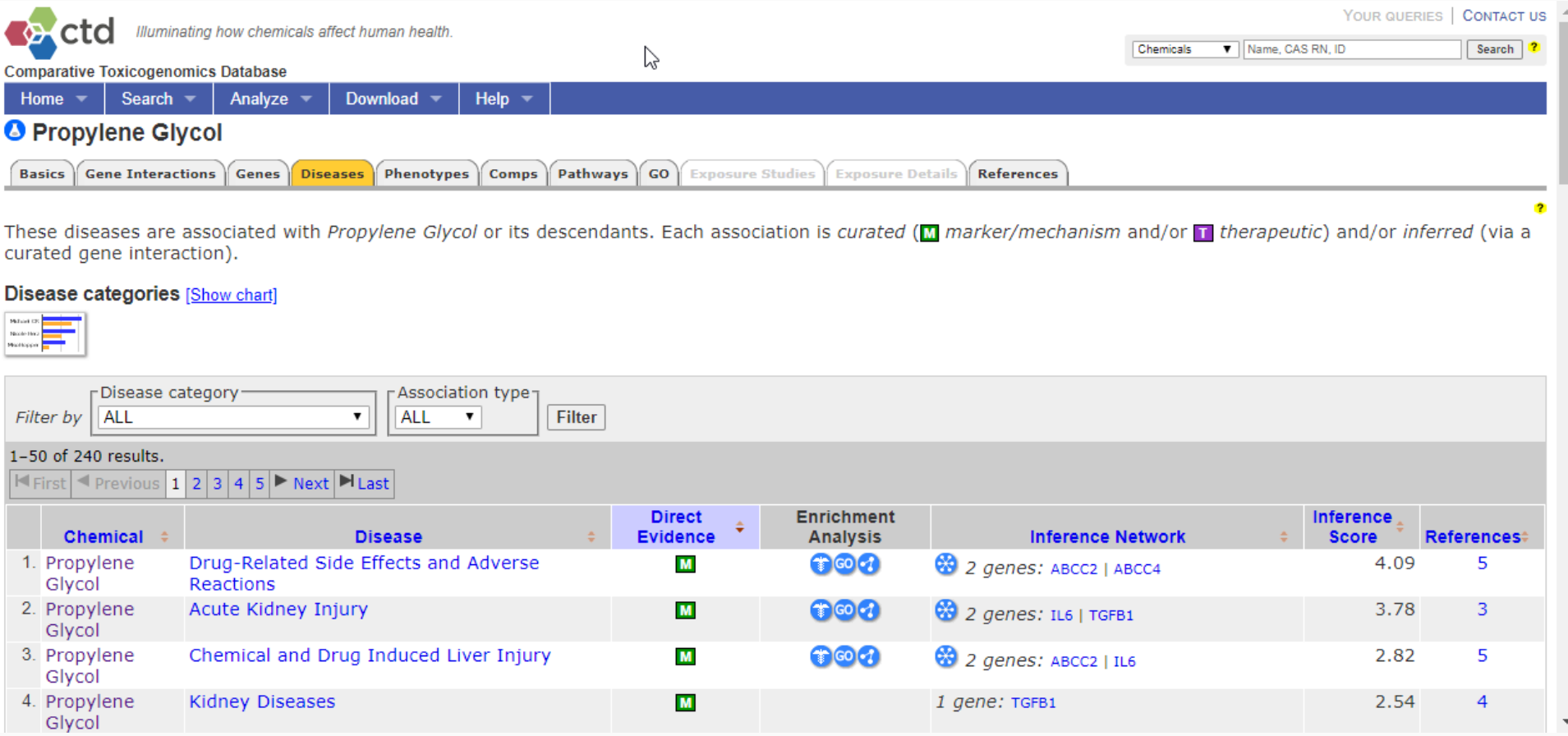




Integrated Linkouts

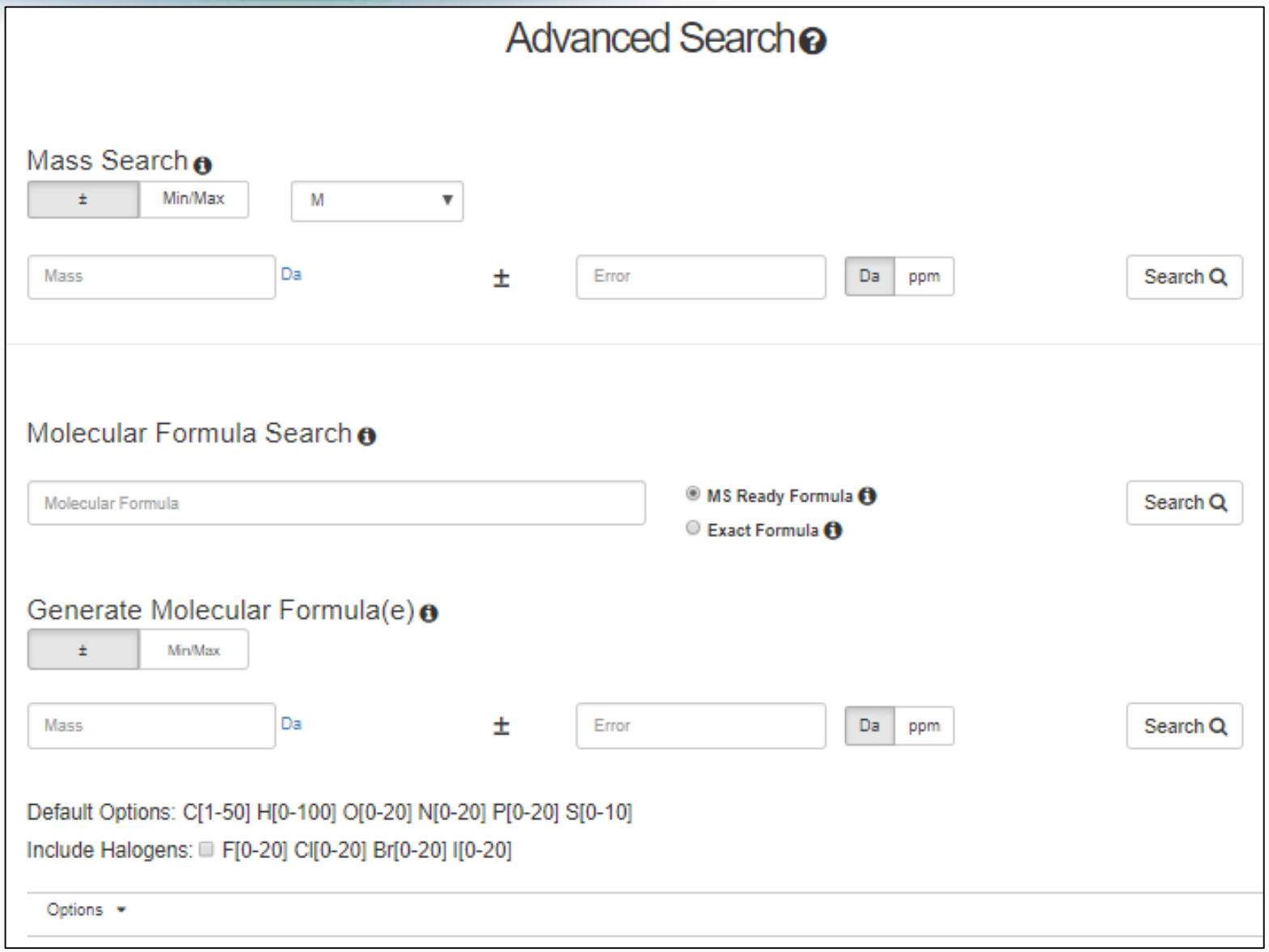
Comparative Toxicogenomics DB

EPA United States Environmental Protection Agency



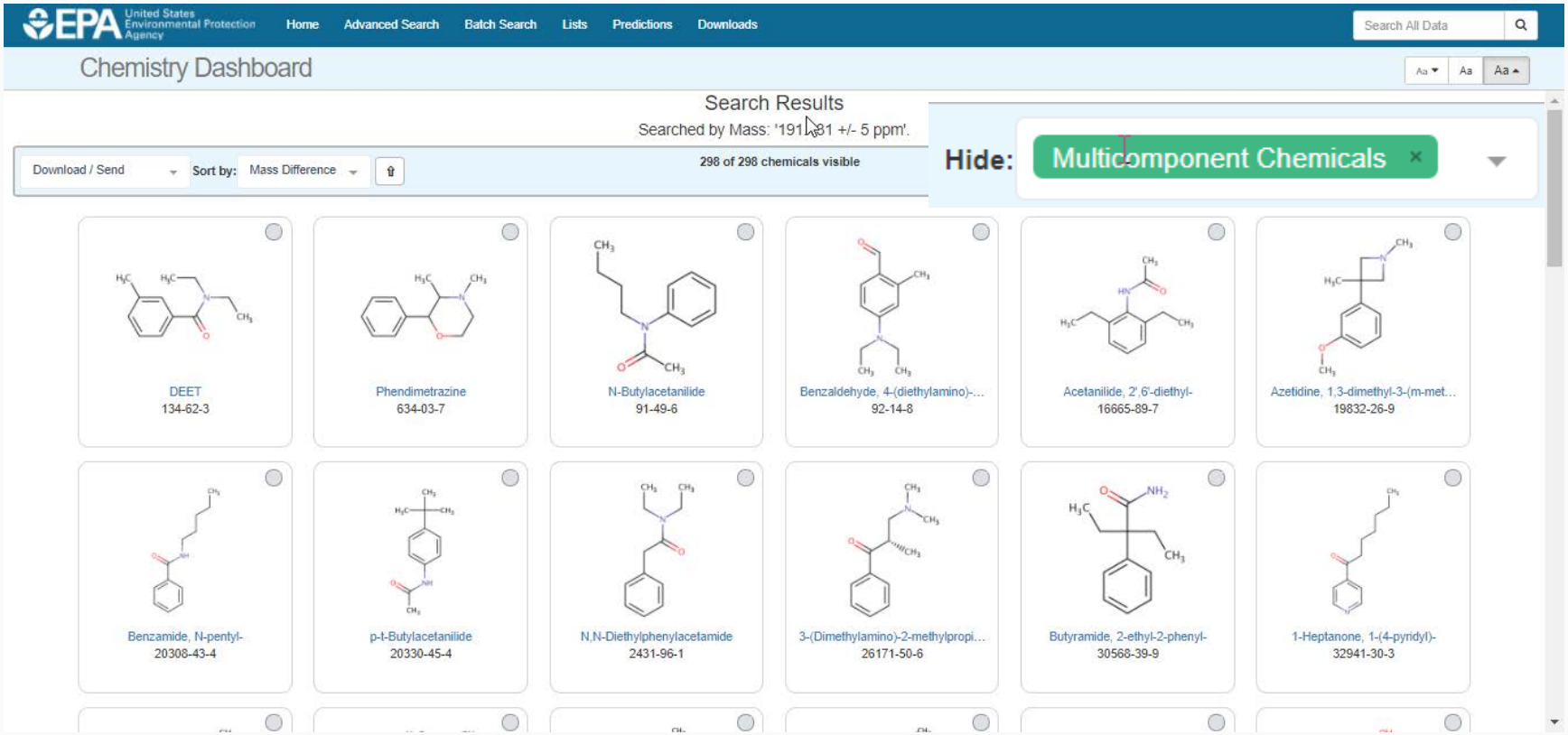
	Chemical	Disease	Direct Evidence	Enrichment Analysis	Inference Network	Inference Score	References
1.	Propylene Glycol	Drug-Related Side Effects and Adverse Reactions	M		2 genes: ABCC2 | ABCC4	4.09	5
2.	Propylene Glycol	Acute Kidney Injury	M		2 genes: IL6 | TGFB1	3.78	3
3.	Propylene Glycol	Chemical and Drug Induced Liver Injury	M		2 genes: ABCC2 | IL6	2.82	5
4.	Propylene Glycol	Kidney Diseases	M		1 gene: TGFB1	2.54	4

Advanced Searches

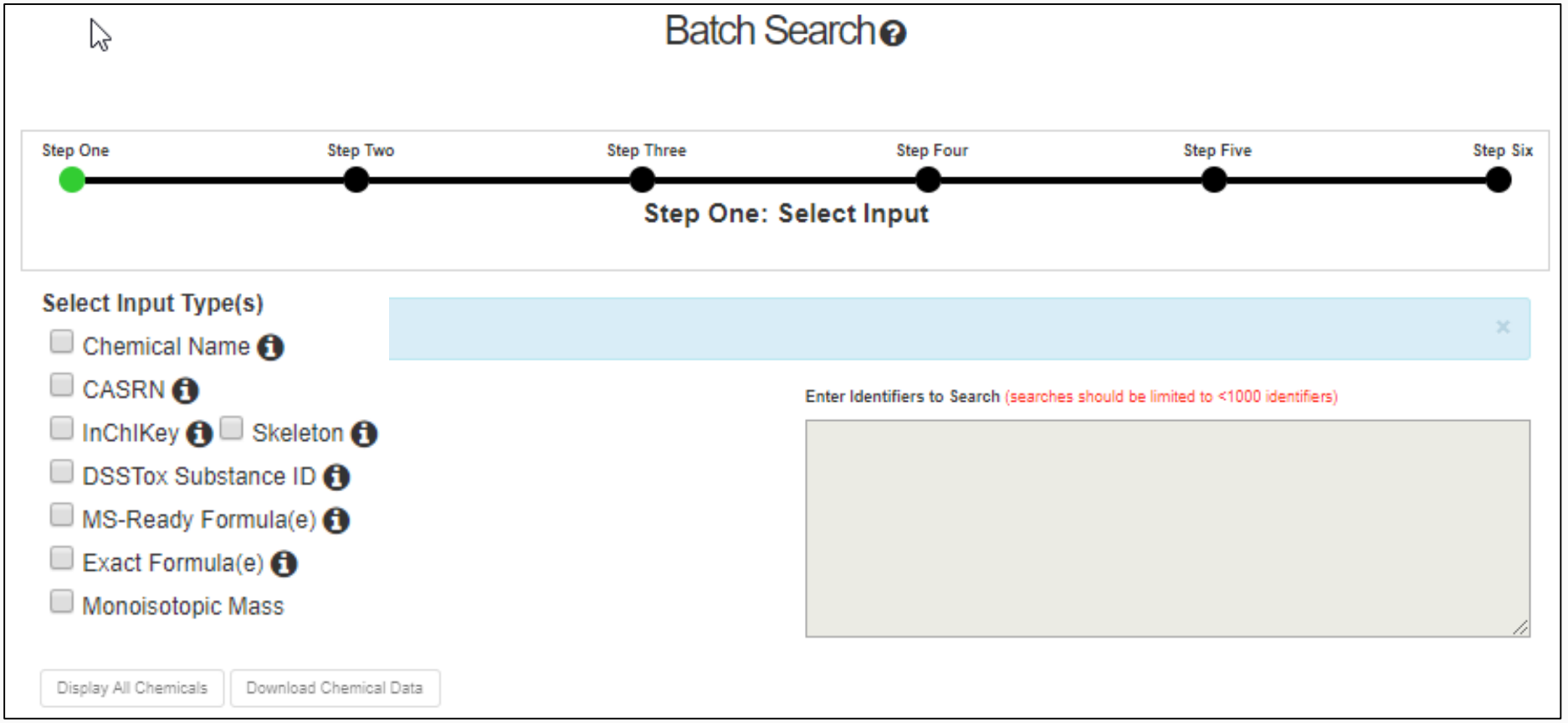


Advanced Searches
Mass Based Search

Advanced Searches



Batch Searches



Batch Search

Select Input Type(s)

- [ ] Chemical Name
- [ ] CASRN
- [ ] InChIKey [ ] Skeleton
- [ ] DSSTox Substance ID
- [x] MS-Ready Formula(e)
- [ ] Exact Formula(e)
- [ ] Monoisotopic Mass

Enter Identifiers to Search (searches should be limited to <1000 identifiers)

C6H12O3
C7H7N3
C8H11NO
C7H5NOS
C9H15NO
C11H12O
C9H8O3
C6H12O5
C9H15NO2

Metadata

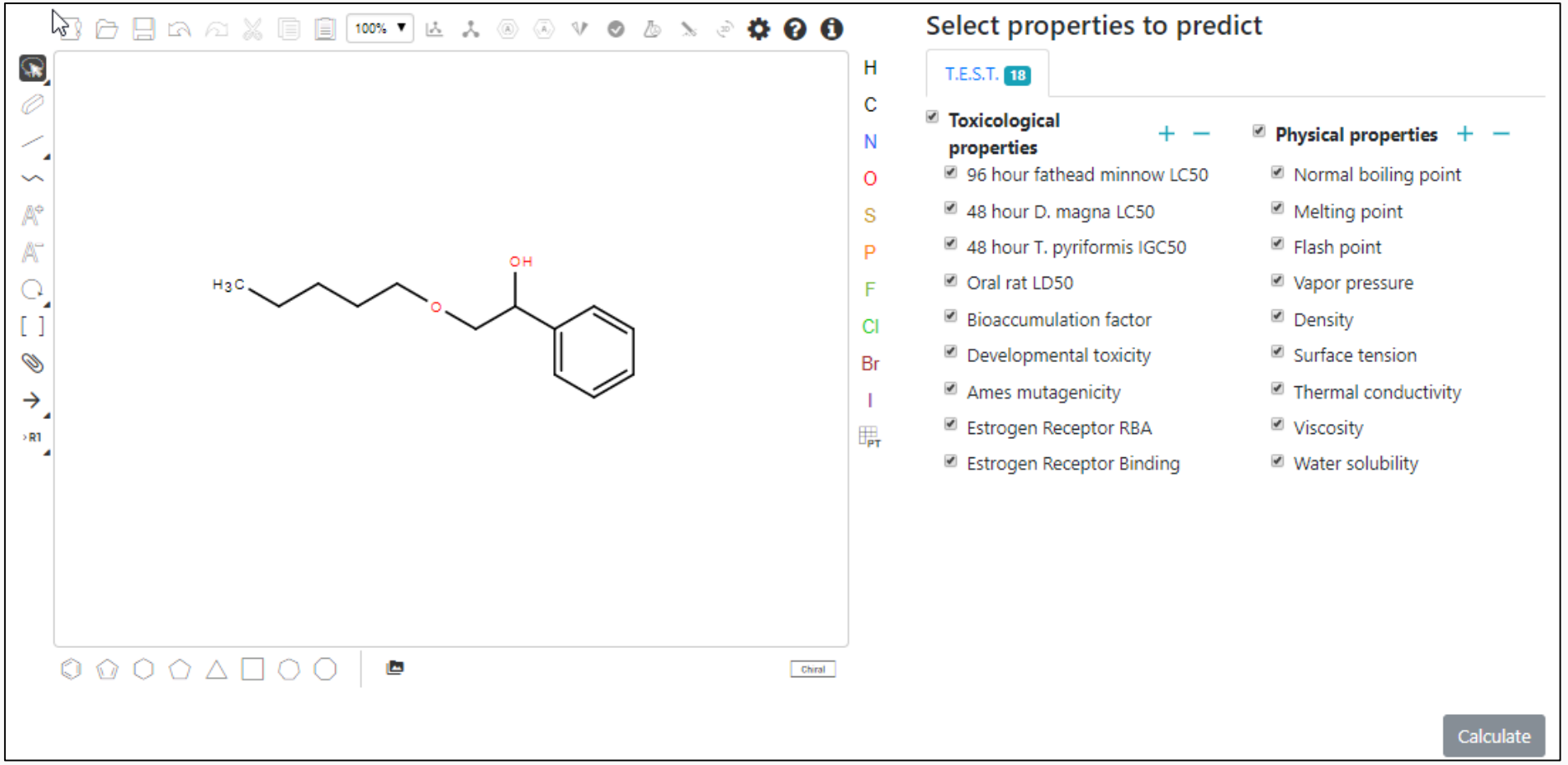
- [ ] Curation Level Details
- [x] Data Sources
- [x] Assay Hit Count
- [ ] Include links to ACToR reports - SLOW! (BETA)
- [x] NHANES/Predicted Exposure
- [x] Include ToxVal Data Availability
- [x] Number of PubMed Articles
- [ ] Abstract Sifter Input File (Beta)
- [ ] MetFrag Input File(Beta)
- [x] IRIS
- [x] PPRTV
- [x] PubChem Data Sources
- [ ] ToxPrint fingerprints

- [ ] NIOSH IDLH Values
- [ ] NIOSH International Chemical Safety Cards
- [ ] NIOSH Pocket Guide to Chemical Hazards
- [ ] NIOSH Skin Notation Profiles
- [ ] NORMAN Collaborative Trial 2015 Targets and Suspects
- [ ] Norman Network PFAS (KEMI Report)
- [ ] NORMAN Network Priority List
- [ ] NormaNEWS: Norman Early Warning System
- [ ] PFAS list provided by X.Trier et al
- [ ] Pharmaceutical List with EU, Swiss and US Consumption Data
- [ ] Provisional Peer Reviewed Toxicity Values
- [ ] Stockholm Convention on Organic Pollutants
- [x] STOFF-IDENT Database of Water-Relevant Substances
- [ ] Superfund Chemical Data Matrix
- [x] Surfactant List Screened in Swiss Wastewater (2014)

Excel Output

INPUT	FOUND_BY	DTXCID_IN	DATA_SOU	TOXVAL_D	TOXCAST_	TOXCAST_	NUMBER_C	PUBCHEM_	STO
C6H12O3	MS Ready I	DTXCID701	91	Y	0.36	2/562	24	83	Y
C6H12O3	MS Ready I	DTXCID003	67	Y	0.36	1/276	376	80	Y
C6H12O3	MS Ready I	DTXCID106	65	Y	4.42	5/113	6	77	Y
C6H12O3	MS Ready I	DTXCID105	45	Y	0.0	0/163	3	94	-
C6H12O3	MS Ready I	DTXCID901	38	Y	-	-	14	110	Y
C6H12O3	MS Ready I	DTXCID402	34	Y	0.0	0/113	-	53	Y
C6H12O3	MS Ready I	DTXCID202	31	Y	-	-	-	36	Y
C6H12O3	MS Ready I	DTXCID202	30	-	2.54	7/276	-	54	-
C6H12O3	MS Ready I	DTXCID109	26	Y	-	-	-	46	-
C6H12O3	MS Ready I	DTXCID202	24	Y	0.0	0/113	-	47	-
C6H12O3	MS Ready I	DTXCID303	22	Y	-	-	-	89	-
C6H12O3	MS Ready I	DTXCID302	20	Y	-	-	2	25	Y
C6H12O3	MS Ready I	DTXCID407	19	Y	-	-	12	62	-
C6H12O3	MS Ready I	DTXCID704	17	Y	-	-	-	64	-
C6H12O3	MS Ready I	DTXCID704	16	Y	-	-	3	49	-

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

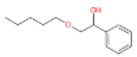
Predicted Water solubility at 25°C for OC(C=1C=CC=CC1)COCCCCC from Consensus method

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

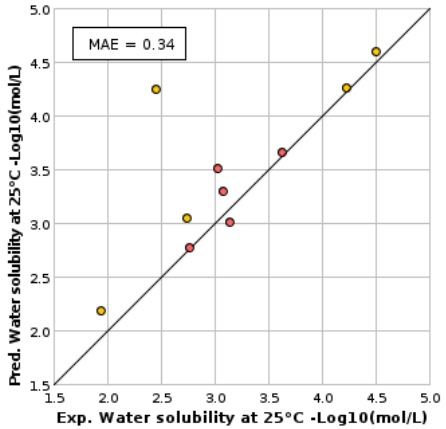
Individual Predictions

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



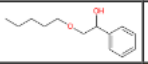
Predictions for the test chemical and for the most similar chemicals

Prediction results (colors defined in table below)

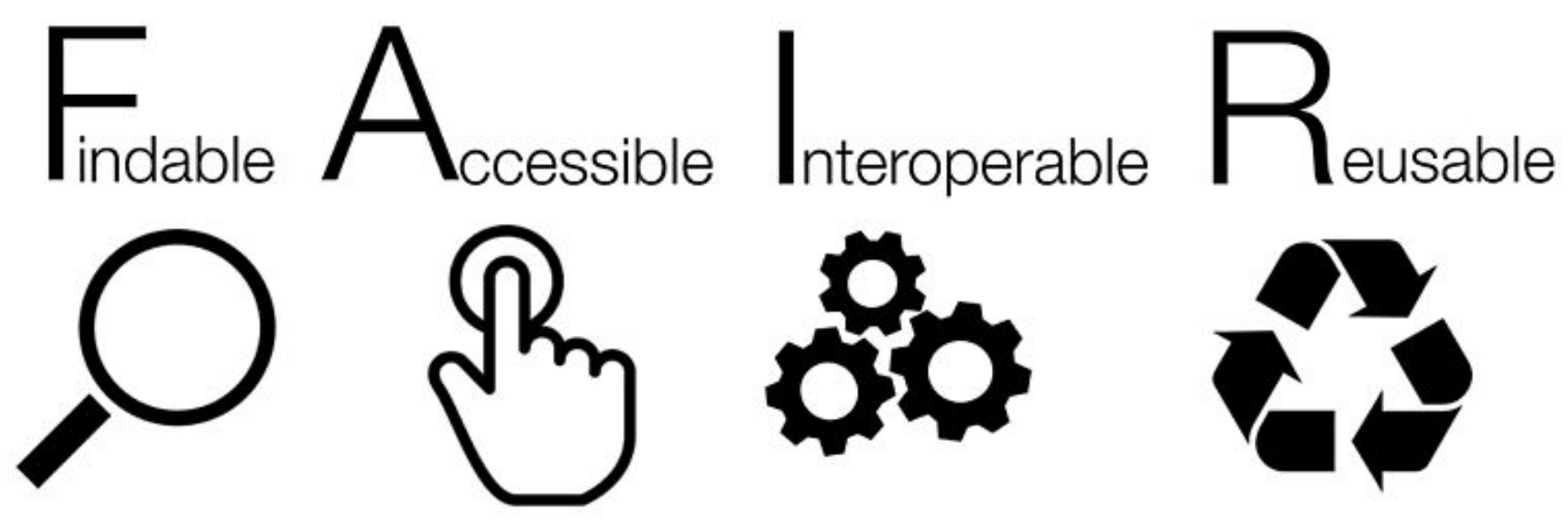


Chemicals	MAE*
Entire set	0.58
Similarity coefficient ≥ 0.5	0.34

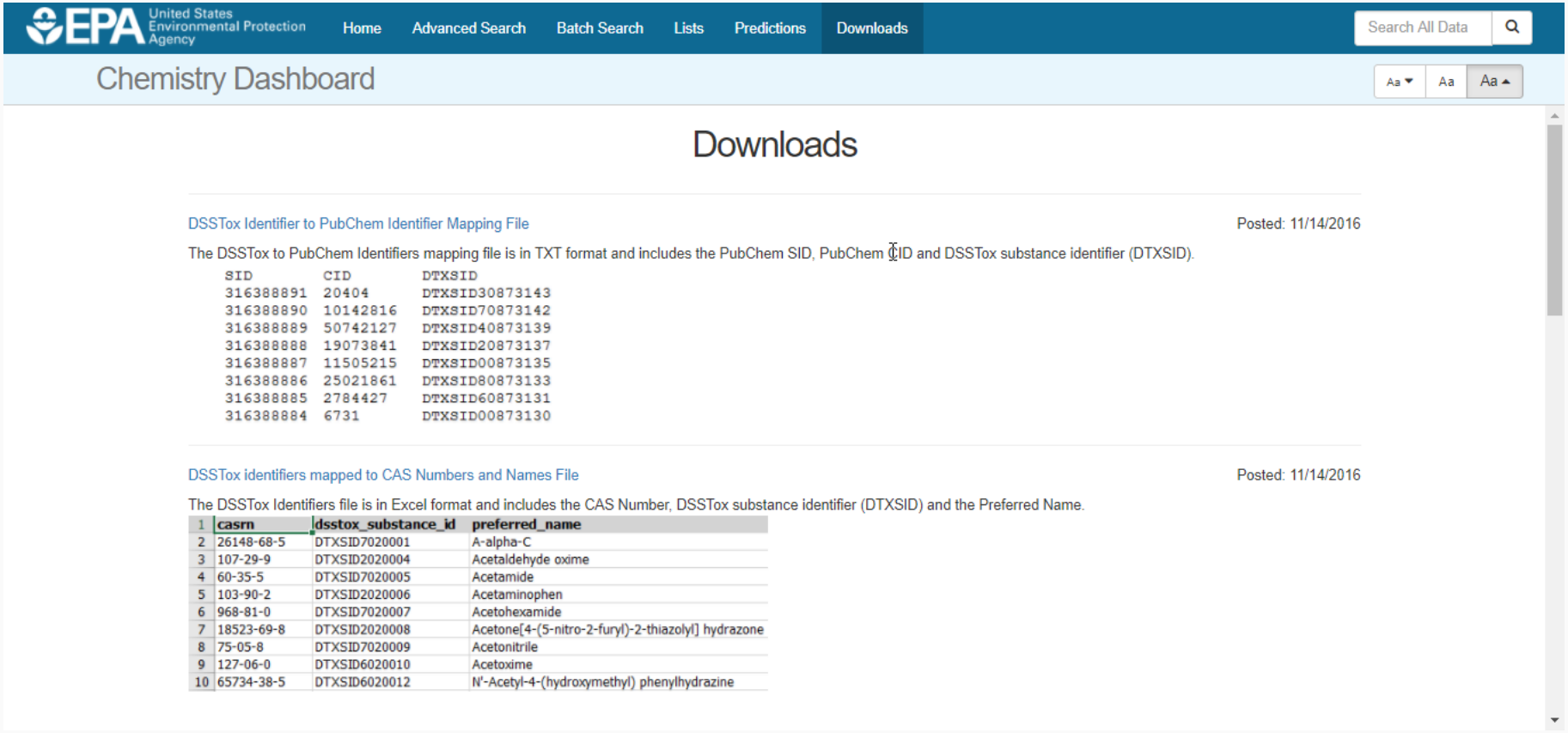
*Mean absolute error in -Log10(mol/L)

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

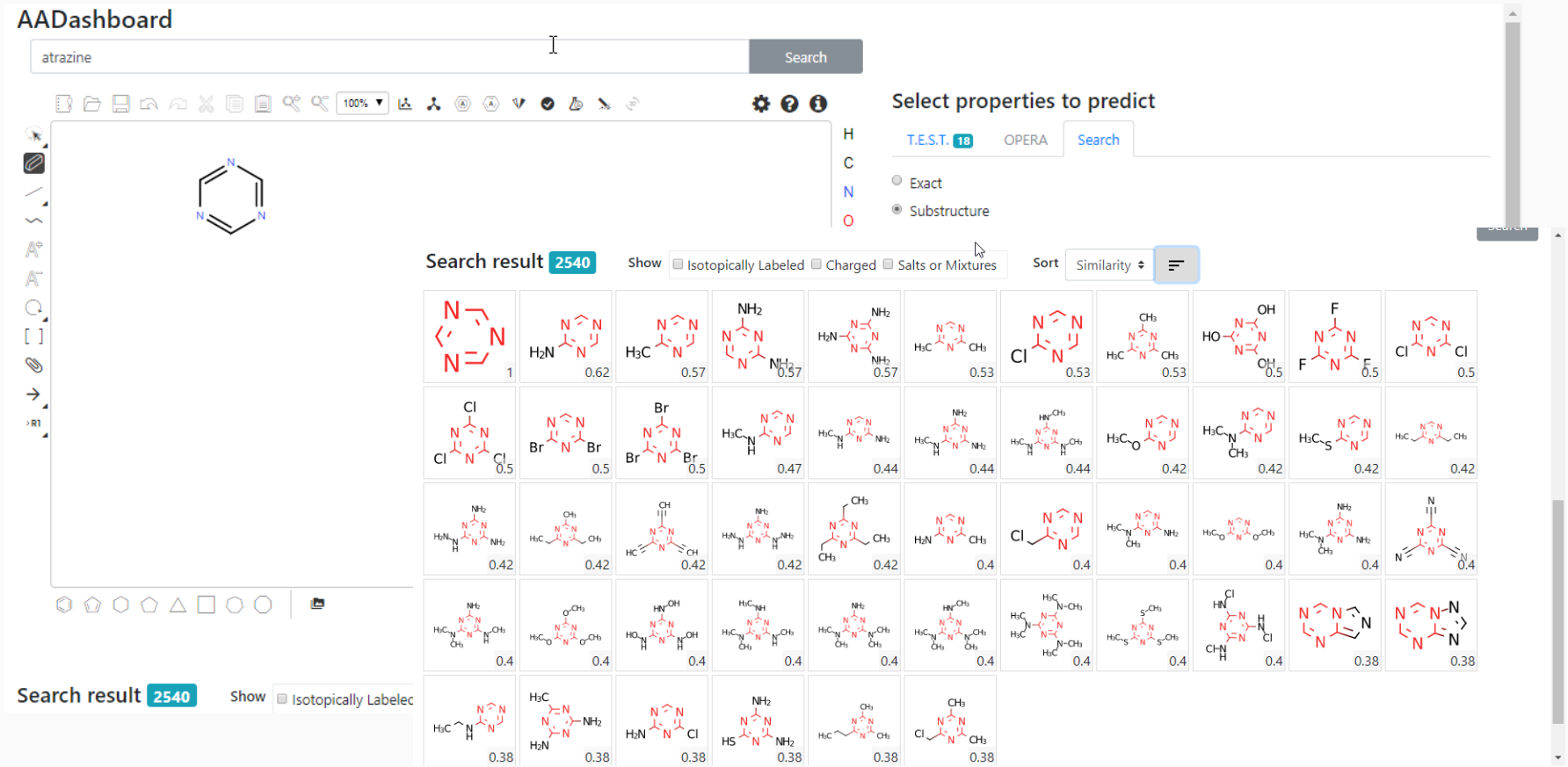
Our support for FAIR Data



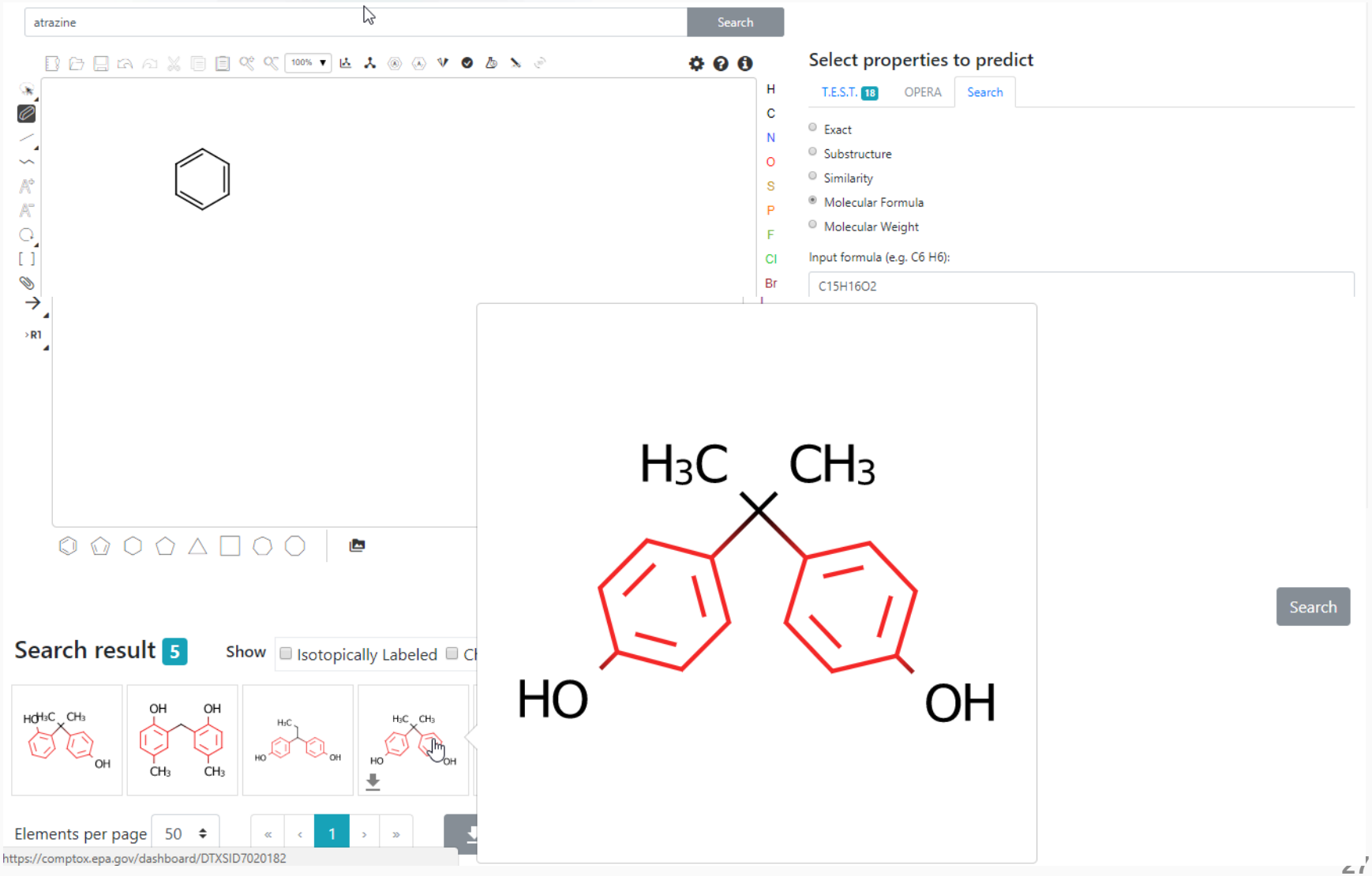
Downloadable Data



Prototype Development



Prototype Development



Future Development
Real Time OPERA Predictions

EPA United States Environmental Protection Agency

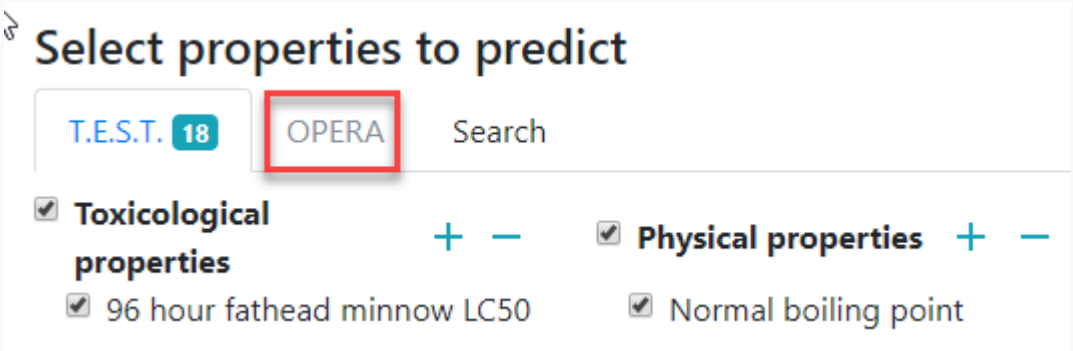
Mansouri *et al. J Cheminform (2018) 10:10*
https://doi.org/10.1186/s13321-018-0263-1

Journal of Cheminformatics

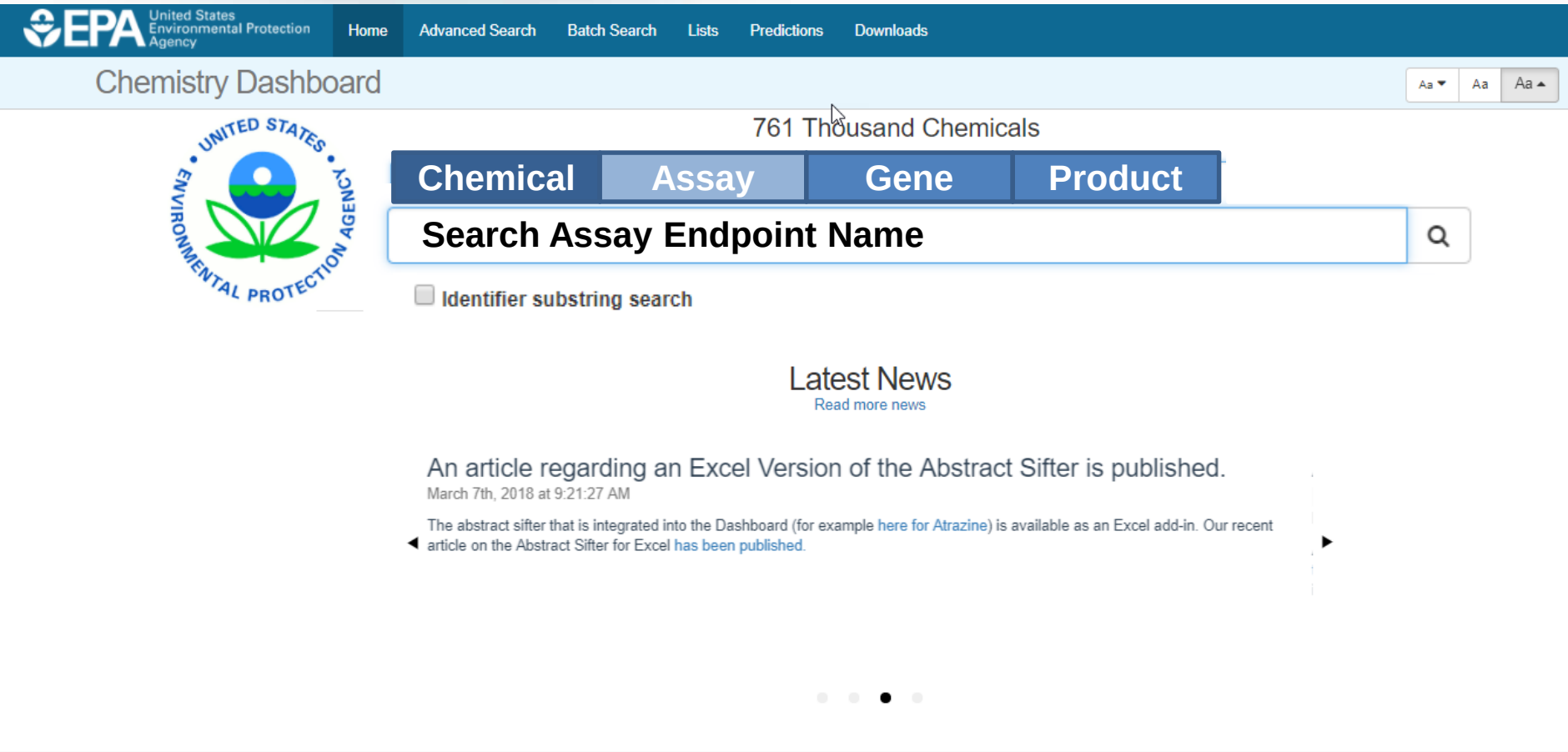
RESEARCH ARTICLE Open Access

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]



Future Search Possibilities



Conclusion

- The CompTox Chemistry Dashboard provides access to data for ~760,000 chemicals
- High quality data from ongoing curation efforts
- An integration hub for multiple “modules”
  - Experimental and predicted properties
  - Human and Ecological Hazard data
  - Exposure data – products, data in the environment
  - *In vitro* bioassay data – ToxCast/Tox21
  - Literature searching – Google Scholar and PubMed
  - Specialized searches – mass/formula for analytical support
  - Batch searching and Real Time Predictions
- The primary architecture for NCCT data

Acknowledgments

- The NCCT CompTox Chemistry Dashboard Development Team
- NERL scientists (Jon Sobus, Elin Ulrich) – Mass Spectrometry
- Kamel Mansouri – OPERA models
- Todd Martin – TEST predictions
- Nancy Baker – Abstract Sifter

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