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

Web-based access to experimental and predicted data for environmental fate, transport and toxicity data

Antony Williams[1], Todd Martin[2], Valery Tkachenko[3], Kamel Mansouri[4] and Chris Grulke[1]

1) National Center for Computational Toxicology, U.S. Environmental Protection Agency, RTP, NC
2) National Risk Management Research Laboratory, U.S. Environmental Protection Agency, Cincinnati, OH
3) Science Data Software, LLC, Rockville, MD 20850
4) Integrated Laboratory Systems, Research Triangle Park, North Carolina, United States

August 2018
ACS Fall Meeting, Boston

Abstract

- The EPA CompTox Chemistry Dashboard provides access to data associated with ~760,000 chemical substances. The available data includes experimental and predicted physicochemical properties, environmental fate and transport data, *in vivo* and *in* silico toxicity data, *in vitro* bioassay data, exposure data and a variety of other types of information. The data are under continuous expansion and curation and the experimental data have been used to develop QSAR and QSPR models. A number of these models are available via a web interface so that users can submit a chemical structure and predict properties in real time. The dashboard also provides access to pre-compiled chemical lists and categories, including pesticides, and chemicals detected in the environment via non-targeted mass spectrometry analysis. The data are searchable using chemical identifiers (systematic names, trade names, CAS Registry Numbers), by structure, mass and formula. Batch searches allow for data associated with thousands of chemicals to be obtained in a few seconds, with just a few button clicks, and downloaded to the desktop. This presentation will provide an overview of the Dashboard and its applications to accessing source data associated with agriculturally related chemicals.

Who is NCCT?

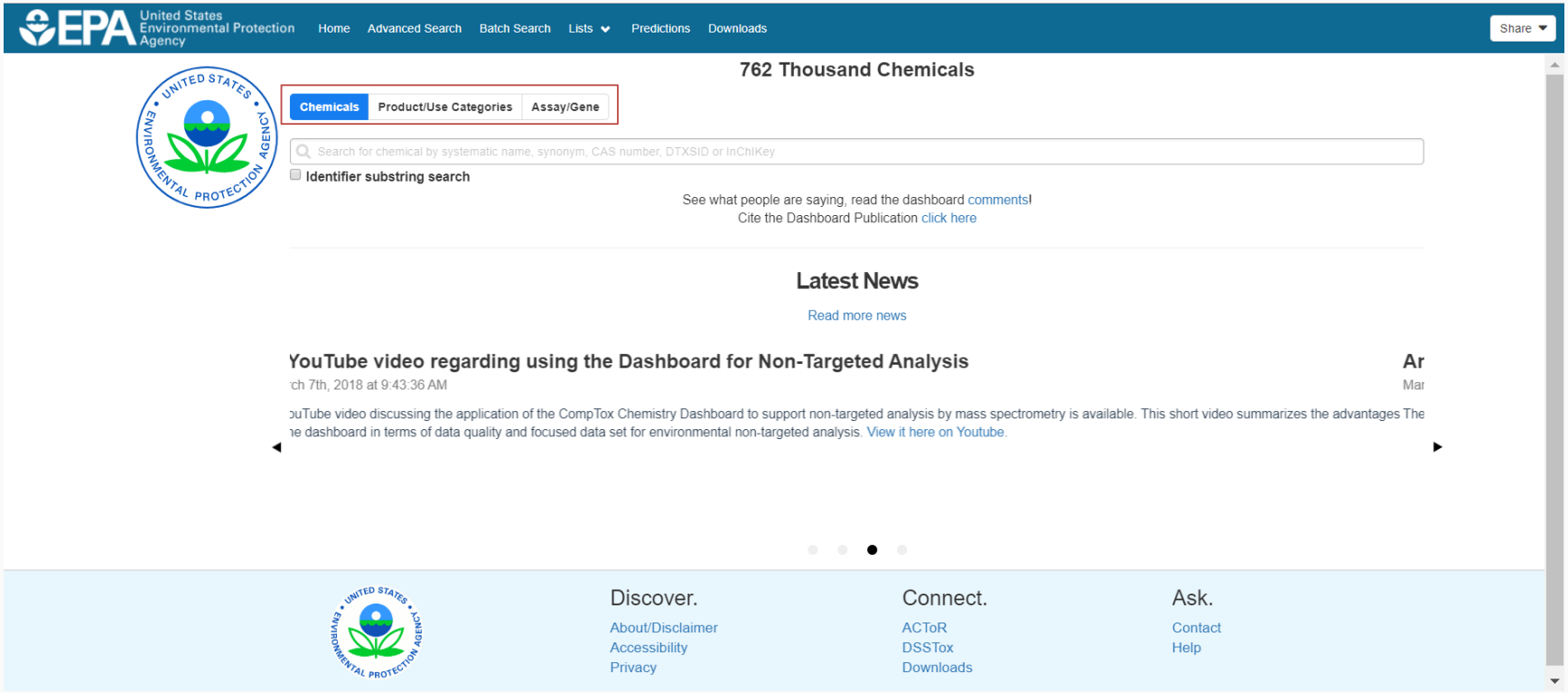
- National Center for Computational Toxicology – part of EPA's Office of Research and Development
- Research driven by EPA's *Chemical Safety for Sustainability Research Program*
  - Develop new approaches to **evaluate** the **safety** of chemicals
  - Integrate advances in biology, biotechnology, chemistry, exposure science and computer science
- Goal - To identify **chemical exposures** that may disrupt biological processes and cause adverse outcomes.
- Prediction models and predicted data are some of our major outputs

The CompTox ~~Chemistry~~ Dashboard

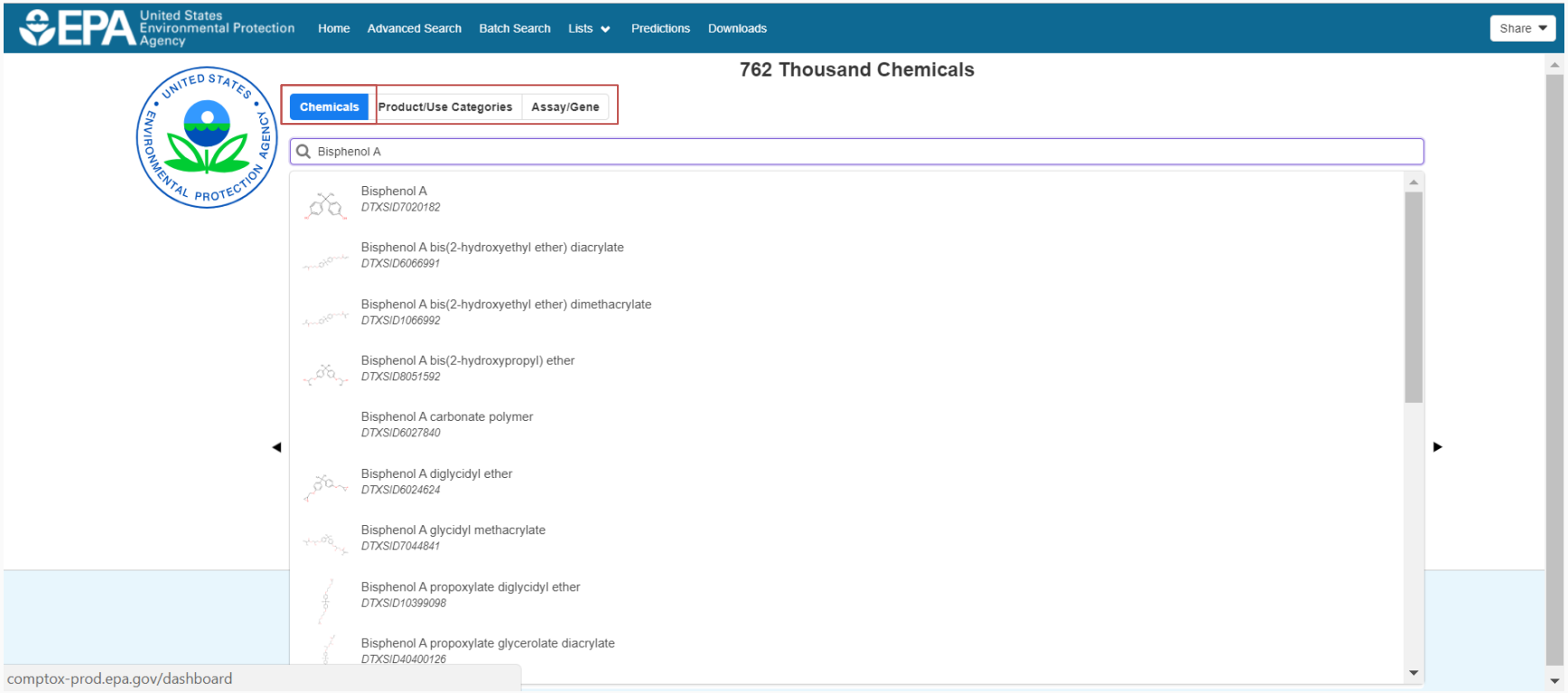
- A publicly accessible website delivering access:
  - ~762,000 chemicals with related property data
  - Searchable by chemical, product use, gene and assay
  - Experimental and **predicted** physicochemical property data
  - “Bioactivity data” for the ToxCast/Tox21 project – plus derived **models**
  - Generalized **Read-Across** (GenRA) module
  - “Batch searching” of **predicted** data for thousands of chemicals

CompTox Dashboard

https://comptox.epa.gov/dashboard



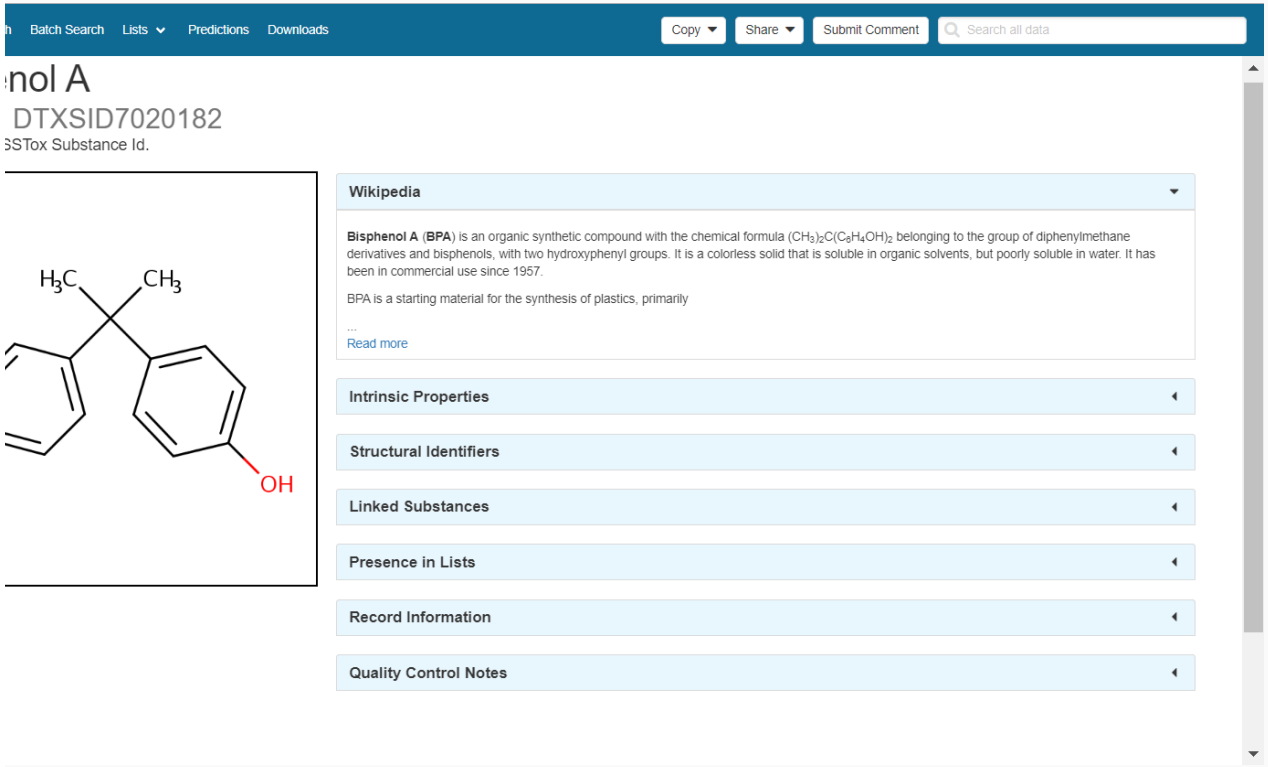
CompTox Dashboard Chemicals



Detailed Chemical Pages

EPA United States Environmental Protection Agency

- DETAILS
- EXECUTIVE SUMMARY
- PROPERTIES
- ENV. FATE/TRANSPORT
- HAZARD
- ADME
- EXPOSURE
- BIOACTIVITY
- SIMILAR COMPOUNDS
- GENRA (BETA)
- RELATED SUBSTANCES
- SYNONYMS
- LITERATURE
- LINKS
- COMMENTS



Physicochemical properties

EPA United States Environmental Protection Agency

Property

Summary

- Summary
- LogP: Octanol-Water
- Melting Point
- Boiling Point
- Water Solubility
- Vapor Pressure
- Flash Point
- Surface Tension
- Index of Refraction
- Molar Refractivity
- Polarizability
- Density
- Molar Volume
- Thermal Conductivity
- Viscosity
- Henry's Law
- LogKoa: Octanol-Air

Summary

Search query

d average	Experimental median	Predicted median	Experimental range	Predicted range	Unit
		3.43	3.32	2.40 to 3.64	
	156	138	153 to 156	125 to 157	°C
		360	200	343 to 401	°C
		1.00e-3	5.26e-4	5.44e-4 to 1.31e-3	mol/L
		3.43e-7	-	6.83e-8 to 2.59e-6	mmHg
		190	-	188 to 192	°C
			-	46.0	dyn/cm
			-	1.60	
			-	68.2	cm^3
			-	27.0	Å^3
		1.17	-	1.14 to 1.20	g/cm^3
			-	200	cm^3
			-	150	mW/(m*K)

OPERA Predicted Properties

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 automate
datasets use
DOI: 10.108(

To link to th

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]

Curation to QSAR Ready Files

Property	Initial file	Curated Data	Curated QSAR ready
AOP	818	818	745
BCF	685	618	608
BioHC	175	151	150
Biowin	1265	1196	1171
BP	5890	5591	5436
HL	1829	1758	1711
KM	631	548	541
KOA	308	277	270
LogP	15809	14544	14041
MP	10051	9120	8656
PC	788	750	735
VP	3037	2840	2716
WF	5764	5076	4836
WS	2348	2046	2010

Detailed OPERA Prediction Reports

EPA United States Environmental Protection Agency

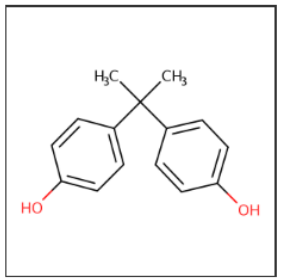
Source	Result	Calculation Details
Experimental Values		
PhysPropNCCT		
Predicted Values		
EPISUITE		
NICEATM		
ACD/Labs Conse		
ACD/Labs		
OPERA		

OPERA Models: LogP: Octanol-Water

Save PDF

Bisphenol A

80-05-7 | DTXSID7020182



Model Results

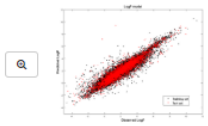
Predicted value: 3.35

Global applicability domain: Inside

Local applicability domain index: 0.88

Confidence level: 0.75

Model Performance

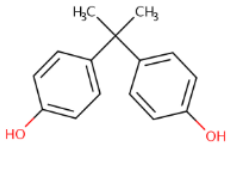


Weighted KNN model

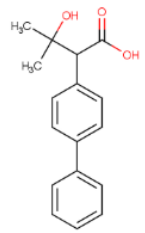
QMRF

5-fold CV (75%)		Training (75%)		Test (25%)	
Q2	RMSE	R2	RMSE	R2	RMSE
0.85	0.69	0.86	0.67	0.86	0.78

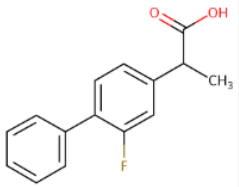
Nearest Neighbors from the Training Set



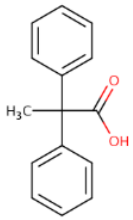
Bisphenol A
Measured: 3.32
Predicted: 3.35



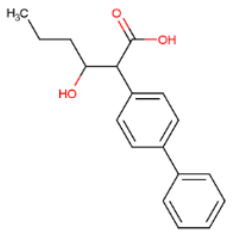
BUTANOIC ACID,2-(4-BIPHENYLYL)-3-HYDROX...
Measured: 3.25
Predicted: 3.45



Flurbiprofen
Measured: 4.16
Predicted: 3.83



2,2-Diphenylpropionic acid
Measured: 2.69
Predicted: 2.93



3-OH-2-(4-BIPHENYLYL)HEXANOIC ACID
Measured: 3.75
Predicted: 3.68

Prediction Details and QMRF Report

EPA United States Environmental Protection Agency

Model Results

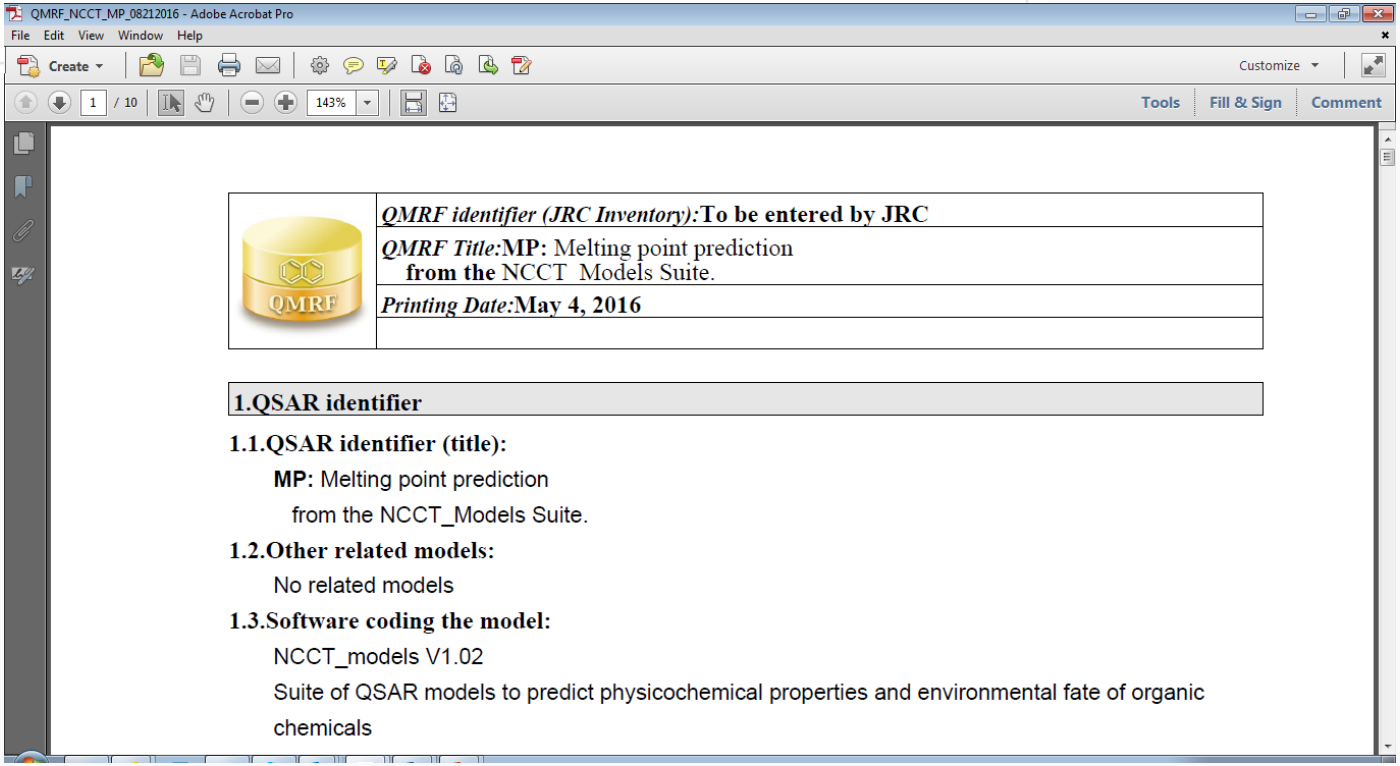
Predicted value: 144 °C

Global applicability domain: Inside

Applicability domain using the leverage approach. All training set space considered. More details in QMRF.

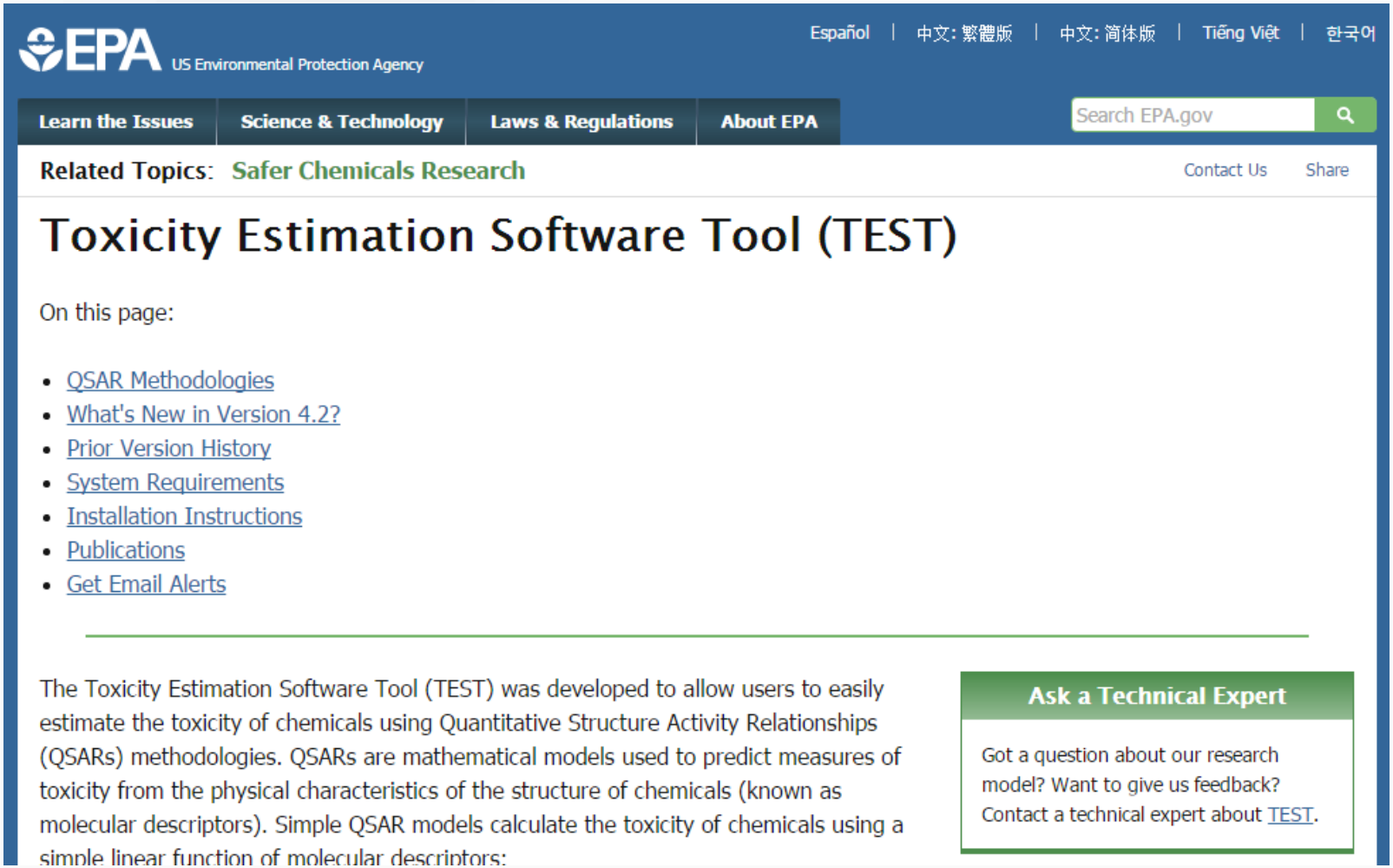
Local applicability domain index: 0.91

Confidence level: 0.65



EPA T.E.S.T

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



Physical properties in TEST

Endpoint	Definition
Viscosity	A measure of the resistance of a fluid to flow (cP) defined as the proportionality constant between shear rate and shear stress
Surface tension	A property of the surface of a liquid (dyn/cm) that allows it to resist an external force
Water solubility	The amount of a chemical (mg/L) that will dissolve in liquid water to form a homogeneous solution

Other Dashboard Predictions

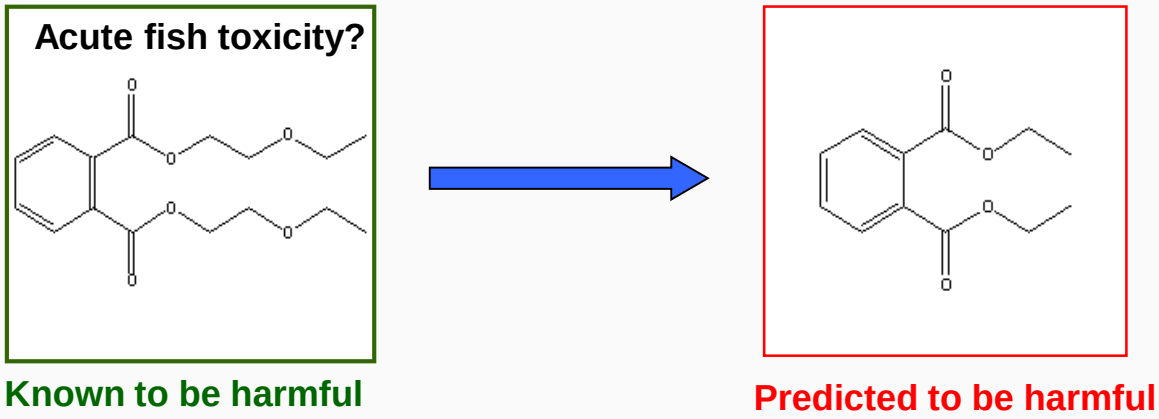
- Predictions and models expand outside of simply physicochemical and environmental fate and transport
- Examples
  - Read-across for Toxicity Endpoints
  - Quantitative Structure–Use Relationship (QSUR) models
  - High-Throughput ToxicoKinetics (HTTK)
  - Models based on high throughput bioactivity data

Definitions: Read-Across

- Known information on the property of a substance (source) is used to make a prediction of the same property for another substance (target) that is considered “similar”

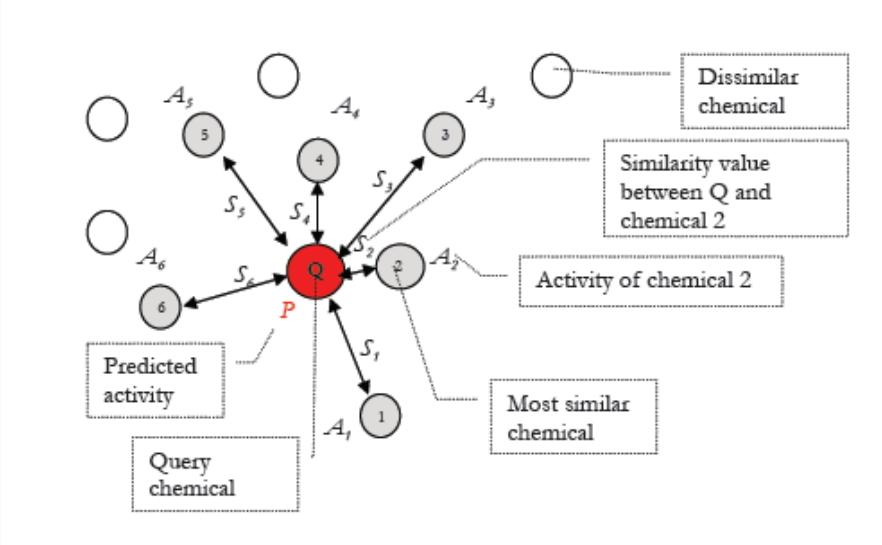
	Source chemical	Target chemical
Property	● →	○

● Reliable data
○ Missing data



GenRA (Generalised Read-Across)

- Predicting toxicity as a similarity-weighted activity of nearest neighbors based on chemistry and/or bioactivity descriptors
- Goal: to systematically evaluate read-across performance and uncertainty using available data
- The approach enabled a performance baseline for read-across predictions of toxicity effects within specific study outcomes to be established



GenRA (Generalised Read-Across)

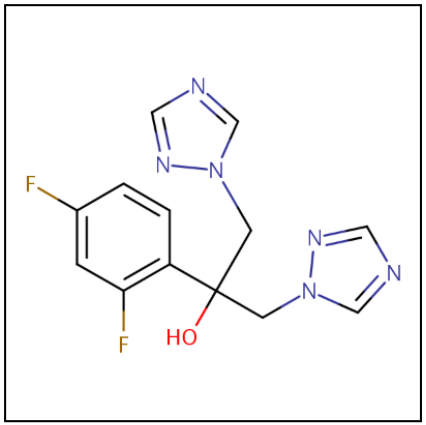
EPA United States Environmental Protection Agency

Fluconazole

86386-73-4 | DTXSID3020627

Searched by DSSTox Substance Id.

- DETAILS
- EXECUTIVE SUMMARY
- PROPERTIES
- ENV. FATE/TRANSPORT
- HAZARD
- ADME
- EXPOSURE
- BIOACTIVITY
- SIMILAR COMPOUNDS
- GENRA
- RELATED SUBSTANCES
- SYNONYMS
- LITERATURE
- LINKS
- COMMENTS



Wikipedia

Fluconazole is an antifungal medication used for a number of fungal infections. This includes candidiasis, blastomycosis, coccidiodomycosis, cryptococcosis, histoplasmosis, dermatophytosis, and pityriasis versicolor. It is also used to prevent candidiasis in those who are at high risk such as following organ transplantation, low birth weight babies, and those with low blood neutrophil counts. It is given either by mouth or by injection into a vein.

Common side effects include vomiting

...

Read more

Intrinsic Properties

Molecular Formula: $C_{13}H_{12}F_2N_6O$ Mol File

Find All Chemicals

Average Mass: 306.277 g/mol Isotope Mass Distribution

Monoisotopic Mass: 306.104065 g/mol

Structural Identifiers

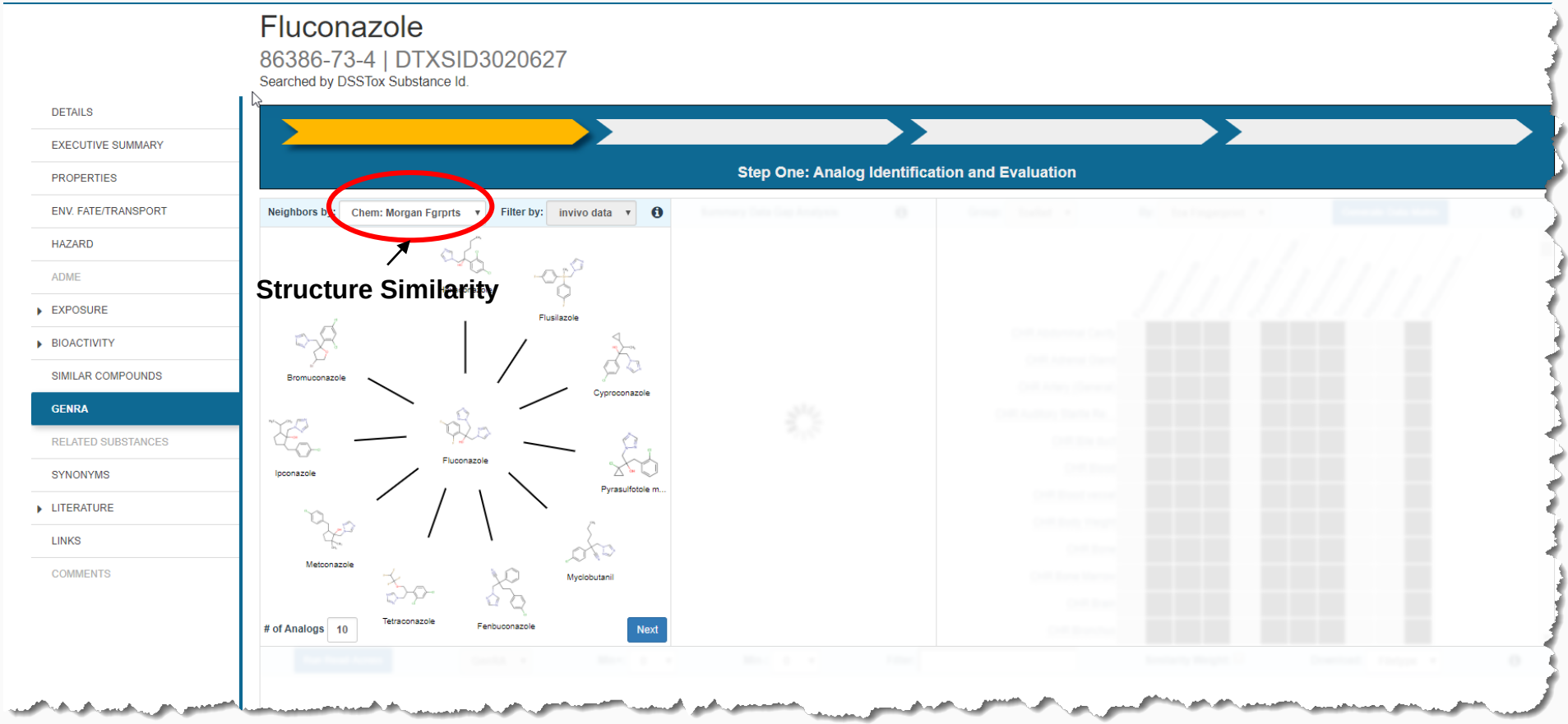
Linked Substances

Presence in Lists

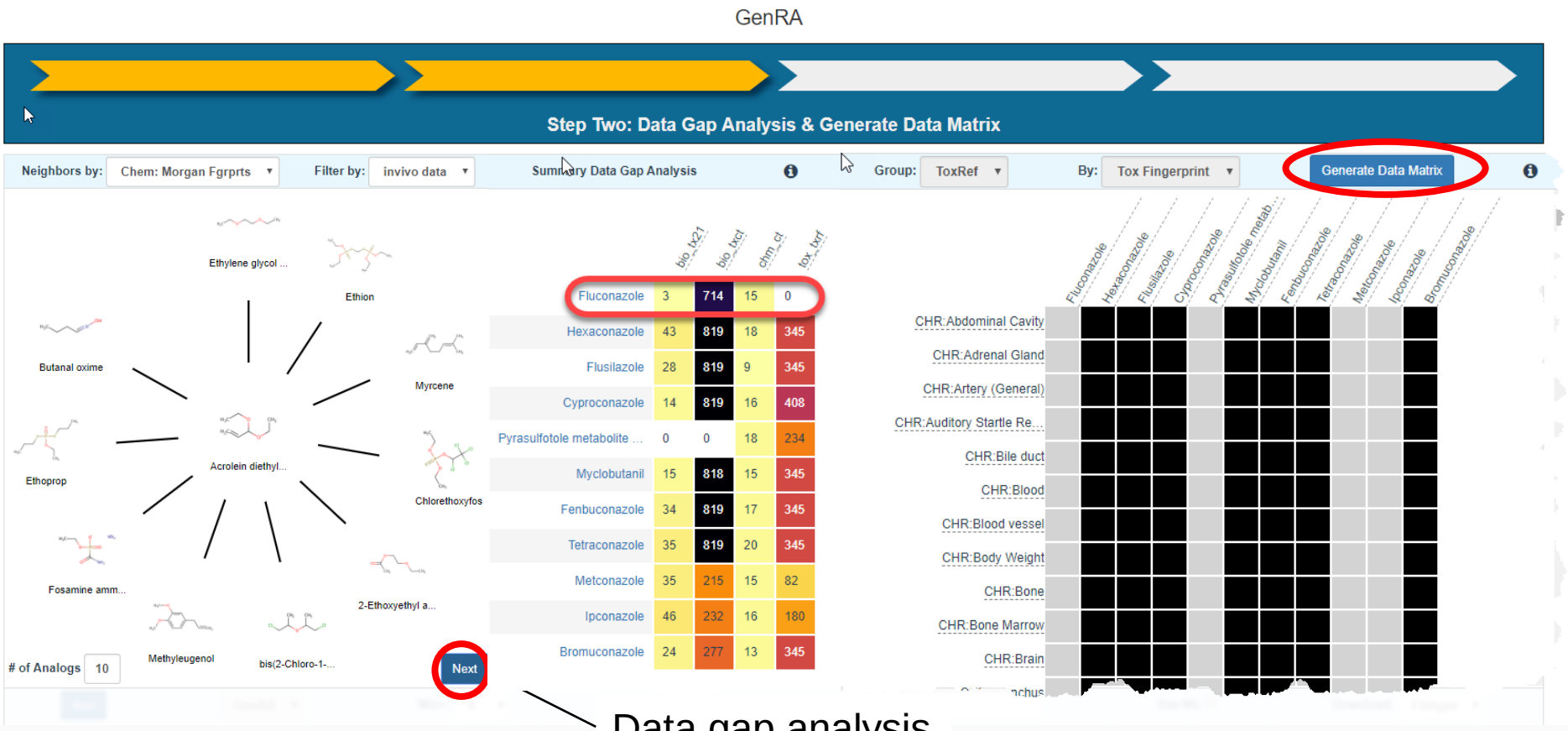
Record Information

Quality Control Notes

GenRA (Generalised Read-Across)

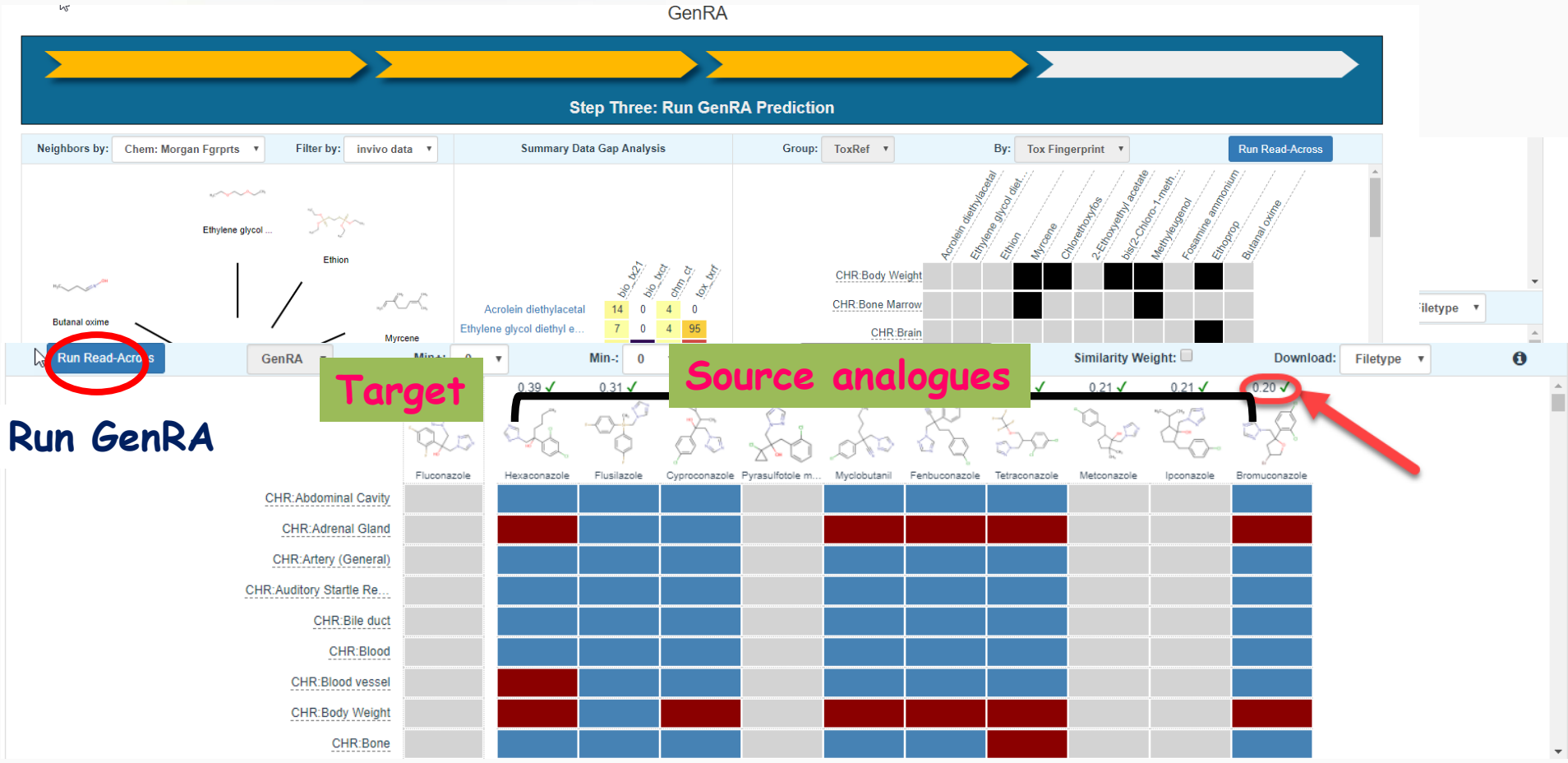


GenRA (Generalised Read-Across)



Data gap analysis

GenRA (Generalised Read-Across)

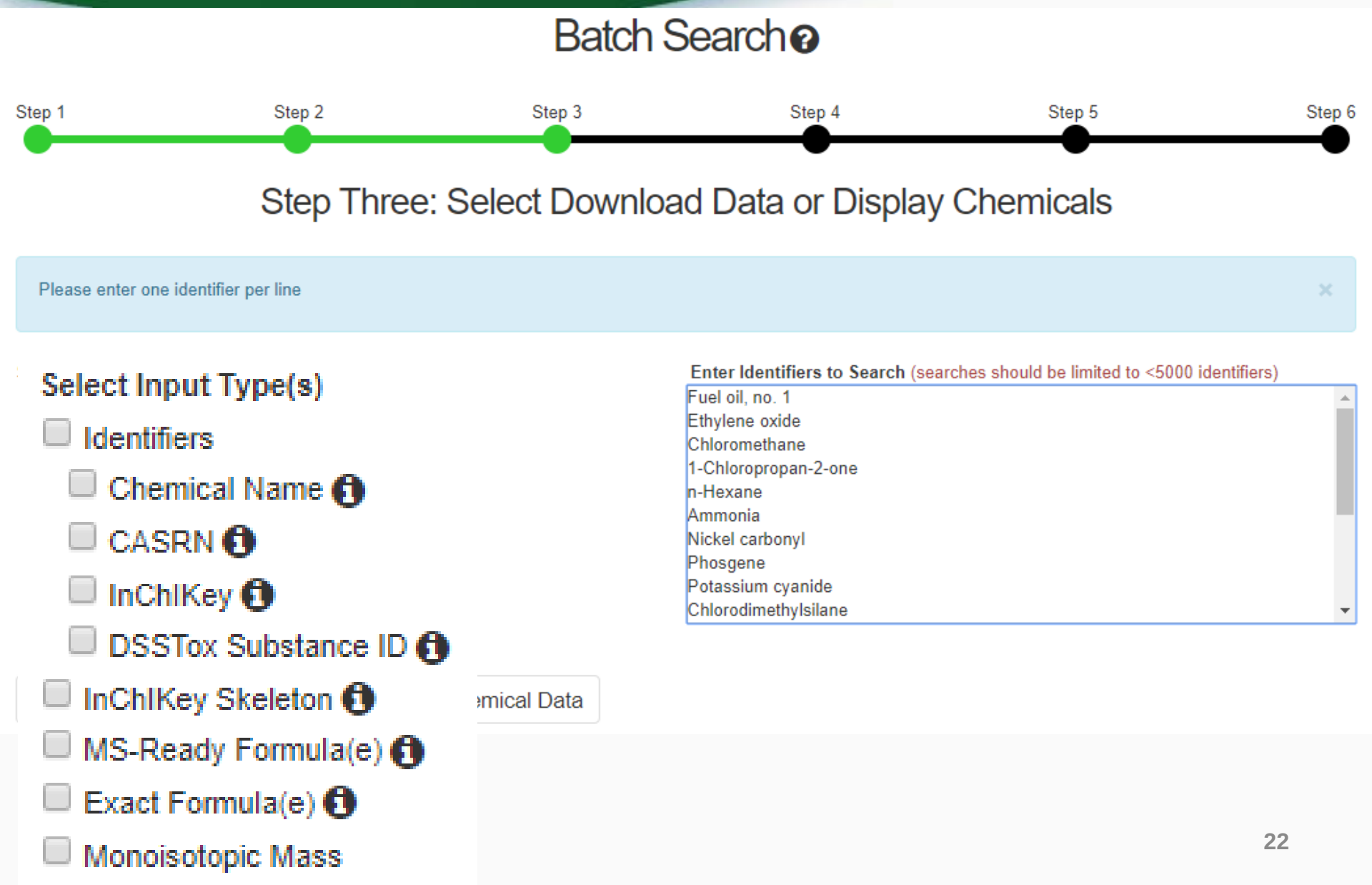


Batch Searching

- Singleton searches are useful but we work with thousands of chemicals!

- Typical questions
  - What are the SMILES strings for a list of 1000 chemicals?
  - Do any of this list of chemicals have XXX type of data?
  - What are the predicted logP values for a list of chemicals?
  - Can I get lists of predicted properties in Excel files? In SDF files?

Batch Searching



Batch Searching

Select Output Format:

Excel

Download

Customize Results

- Select All
- Select All in Lists

Chemical Identifiers

- DTXSID
- Chemical Name
- CAS-RN
- InChIKey
- IUPAC Name

Structures

- Mol File
- SMILES
- InChI String
- MS-Ready SMILES
- QSAR-Ready SMILES

Intrinsic And Predicted Properties

- Molecular Formula
- Average Mass
- TEST Model Predictions
- OPERA Model Predictions

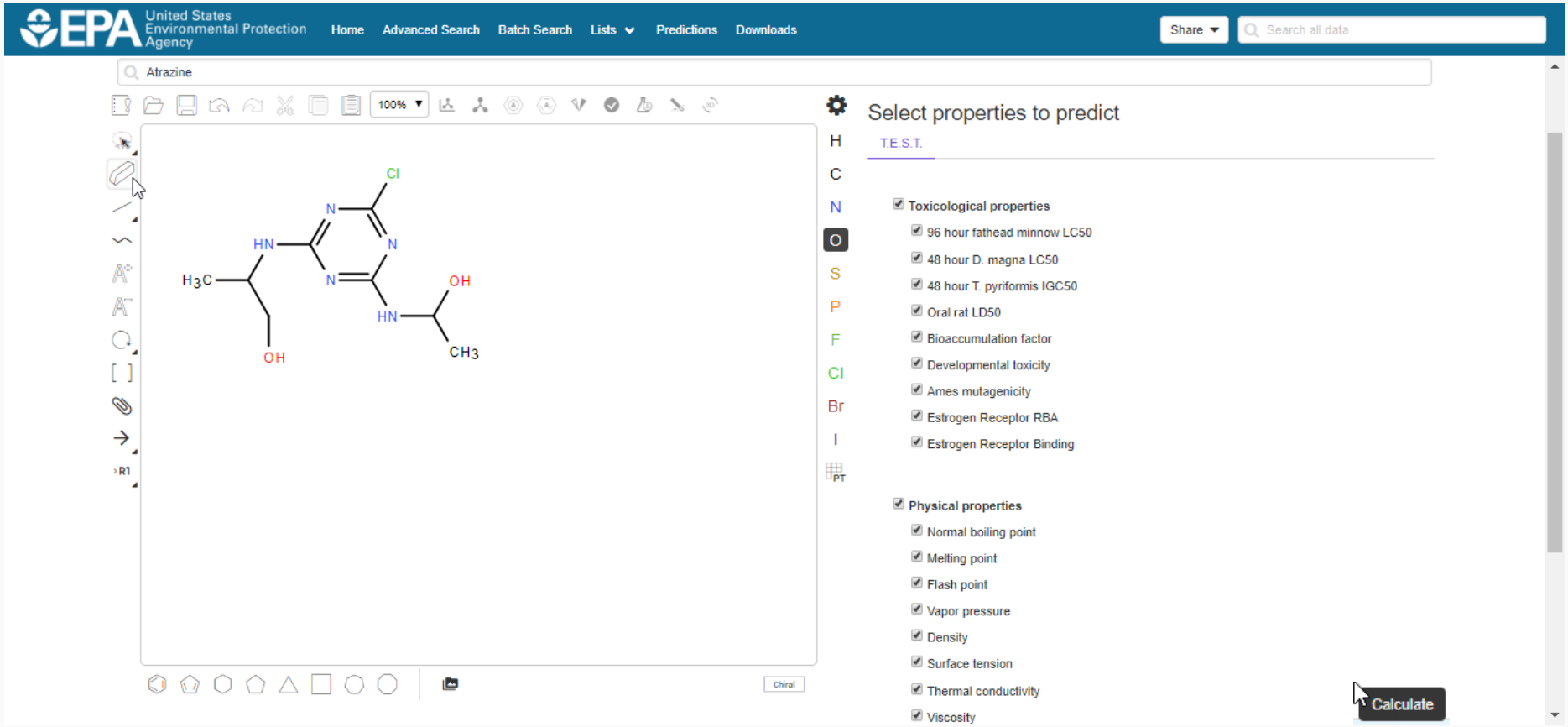
Presence in Lists:

- ICCVAM test method evaluation report: in vitro ocular toxicity test methods
- 40CFR355
- A list of all PBDEs (Polybrominated diphenyl ethers)
- A list of all PCBs (Polychlorinated biphenyls)
- A list of polycyclic aromatic hydrocarbons
- Acute exposure guideline levels
- Algal Toxins
- Androgen Receptor Chemicals
- APCRA Chemicals for Prospective Analysis
- APCRA Chemicals for Retrospective Analysis
- APCRA Chemicals for Retrospective Analysis_App_List_448_Chemicals
- ATSDR Minimal Risk Levels (MRLs) for Hazardous Substances
- ATSDR Toxic Substances Portal Chemical List
- Bisphenol Compounds
- California Office of Environmental Health Hazard Assessment
- Chemicals with interesting names
- CMAP
- DNT Screening Library
- Drinking Water Suspects, KWR Water, Netherlands
- EDSP Universe
- EPA Chemicals associated with hydraulic fracturing

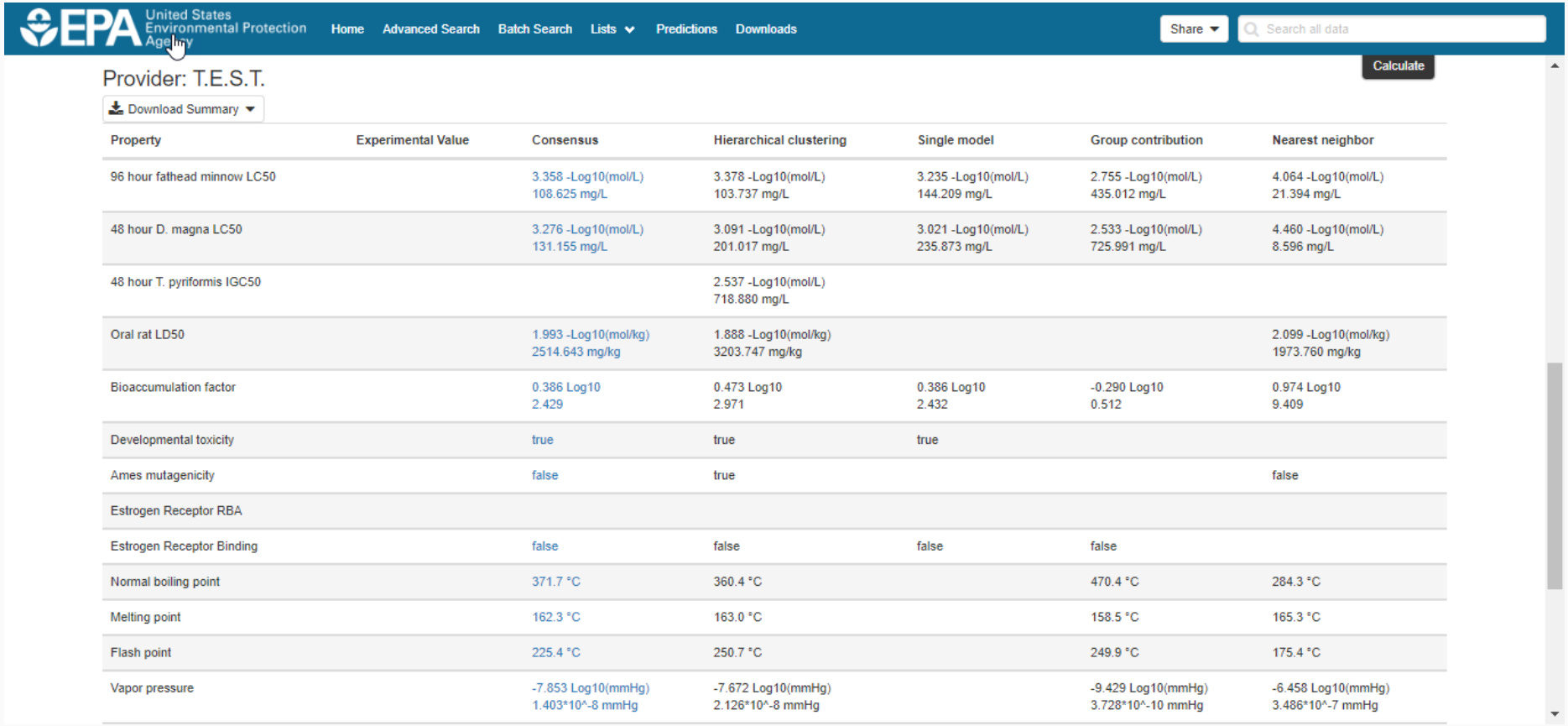
Excel Output

	A	B	C	D	E	F
1	DTXSID	PREFERRED_NAME	VAPOR_PRESSURE_	WATER_SOLUBILITY	ATMOSPHERIC_HYDRO)	BIOCONCENTRATIOI
2	DTXSID0020523	2,4-Dinitrophenol	0.000058479	0.0055847	8.57649E-13	5.31642
3	DTXSID0021256	Sulfasalazine	6.85488E-14	1.07895E-06	1.05999E-11	13.1817
4	DTXSID0021333	Tetramethylthiuram mon	-	0.0076913	2.89566E-11	5.1961
5	DTXSID0021337	Thiabendazole	7.06318E-07	0.000156675	4.3347E-11	7.89835
6	DTXSID0022436	Diphenolic acid	3.40408E-08	0.00125893	1.70869E-11	2.94716
7	DTXSID0023745	D-Xylose	7.60326E-08	3.52371	3.57198E-11	1.66292
8	DTXSID0023901	Bentazone	-	5.40754E-05	1.67809E-11	1.4173
9	DTXSID0023951	Carboxin	1.06905E-07	0.000304789	1.79673E-11	10.2339
10	DTXSID0024216	Nitrapyrin	0.0107895	0.000162181	2.28302E-11	661.452
11	DTXSID0024345	Triasulfuron	1.47231E-11	0.00022856	1.3866E-11	11.1975
12	DTXSID0027983	Dipropylene glycol monc	-	-	-	-
13	DTXSID0032493	Triadimenol	1.58489E-07	0.000267301	1.65682E-11	111.018
14	DTXSID0032520	Azoxystrobin	9.4189E-11	1.53815E-06	1.6811E-11	12.6717
15	DTXSID0032601	Cyproconazole	5.24807E-08	0.000289734	1.66415E-11	43.1754
16	DTXSID0032655	Triticonazole	9.88553E-09	0.000014388	2.98831E-11	102.292
17	DTXSID0039223	Aldicarb	0.000138676	0.0254683	1.5544E-11	5.90478
18	DTXSID0040707	4-Pentylaniline	0.00272898	0.000567545	2.73005E-11	94.115
19	DTXSID0041270	3,5,5-Trimethylhexyl ace	0.21727	0.000521195	1.32332E-11	43.3424
20	DTXSID0044361	Hexanedihydrazide	-	0.341193	5.54119E-12	1.78013
21	DTXSID0044494	2-Phenylethyl phenylace	0.000101859	0.000628058	1.43965E-11	342.896

Real-Time Predictions



Real-Time Predictions



Property	Experimental Value	Consensus	Hierarchical clustering	Single model	Group contribution	Nearest neighbor
96 hour fathead minnow LC50		3.358 -Log10(mol/L) 108.625 mg/L	3.378 -Log10(mol/L) 103.737 mg/L	3.235 -Log10(mol/L) 144.209 mg/L	2.755 -Log10(mol/L) 435.012 mg/L	4.064 -Log10(mol/L) 21.394 mg/L
48 hour D. magna LC50		3.276 -Log10(mol/L) 131.155 mg/L	3.091 -Log10(mol/L) 201.017 mg/L	3.021 -Log10(mol/L) 235.873 mg/L	2.533 -Log10(mol/L) 725.991 mg/L	4.460 -Log10(mol/L) 8.596 mg/L
48 hour T. pyriformis IGC50			2.537 -Log10(mol/L) 718.880 mg/L			
Oral rat LD50		1.993 -Log10(mol/kg) 2514.643 mg/kg	1.888 -Log10(mol/kg) 3203.747 mg/kg			2.099 -Log10(mol/kg) 1973.760 mg/kg
Bioaccumulation factor		0.386 Log10 2.429	0.473 Log10 2.971	0.386 Log10 2.432	-0.290 Log10 0.512	0.974 Log10 9.409
Developmental toxicity		true	true	true		
Ames mutagenicity		false	true			false
Estrogen Receptor RBA						
Estrogen Receptor Binding		false	false	false	false	
Normal boiling point		371.7 °C	360.4 °C		470.4 °C	284.3 °C
Melting point		162.3 °C	163.0 °C		158.5 °C	165.3 °C
Flash point		225.4 °C	250.7 °C		249.9 °C	175.4 °C
Vapor pressure		-7.853 Log10(mmHg) 1.403*10^-8 mmHg	-7.672 Log10(mmHg) 2.126*10^-8 mmHg		-9.429 Log10(mmHg) 3.728*10^-10 mmHg	-6.458 Log10(mmHg) 3.486*10^-7 mmHg

Real-Time Predictions

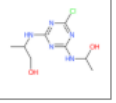
Predicted Oral rat LD50 for ClC=1N=C(N=C(N1)NC(C)CO)NC(O)C from Consensus method

Prediction results

Endpoint	Experimental value	Predicted value
Oral rat LD_{50} -Log10(mol/kg)	N/A	1.99
Oral rat LD_{50} mg/kg	N/A	2514.64

Individual Predictions

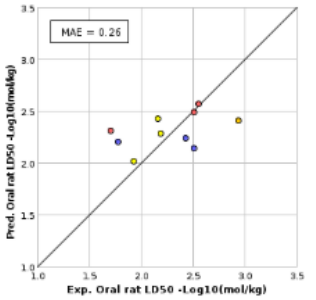
Method	Predicted value -Log10(mol/kg)
Hierarchical clustering	1.89
Nearest neighbor	2.10



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

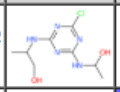
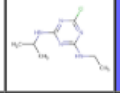
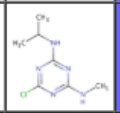
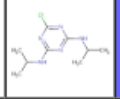
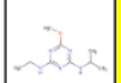
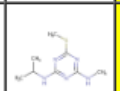
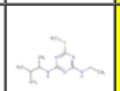
If the predicted value matches the experimental values for similar chemicals in the test set (and the similar chemicals were predicted well), one has greater confidence in the predicted value.

Prediction results (colors defined in table below)



Chemicals	MAE*
Entire set	0.43
Similarity coefficient ≥ 0.5	0.26

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

CAS	Structure	Similarity Coefficient	Experimental value -Log10(mol/kg)	Predicted value -Log10(mol/kg)
ClC=1N=C(N=C(N1)NC(C)CO)NC(O)C (test chemical)			N/A	1.99
1912-24-9		0.84	2.51	2.15
3004-71-5		0.83	2.43	2.24
139-40-2		0.82	1.78	2.20
1610-17-9		0.76	2.16	2.43
1014-69-3		0.74	2.19	2.28
22936-75-0		0.71	1.93	2.02

API in development
Prototype services available

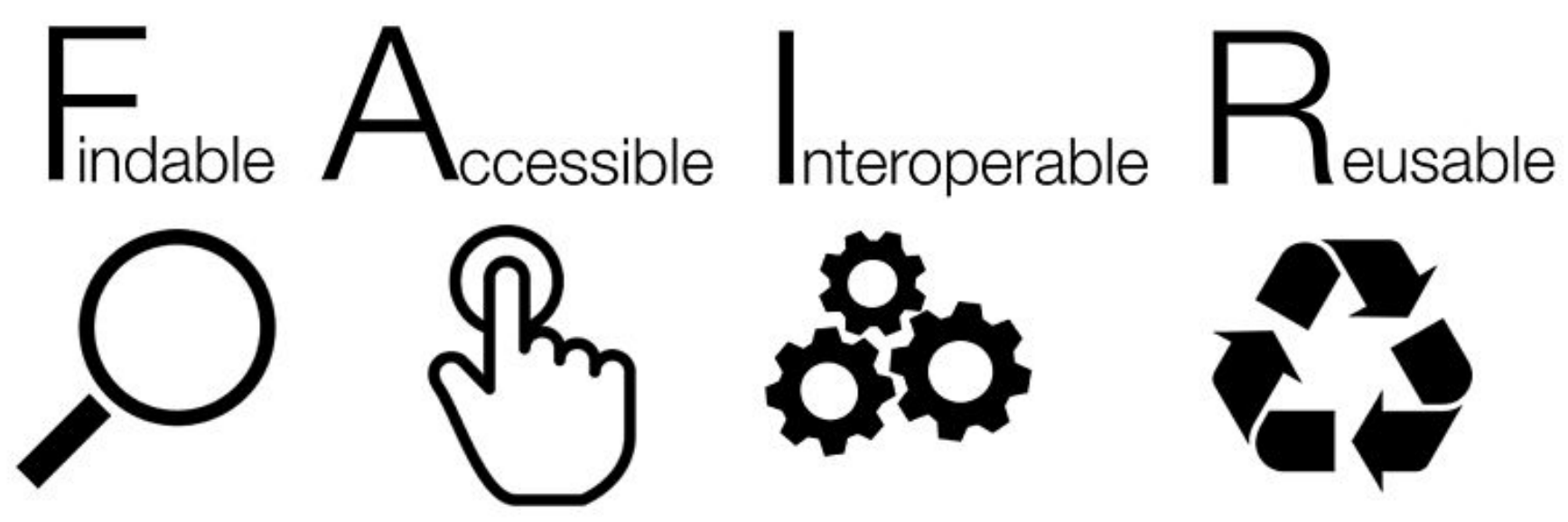
https://comptox.epa.gov/dashboard/web-test/WS?smiles=CCO&method=hc

JSON Raw Data Headers

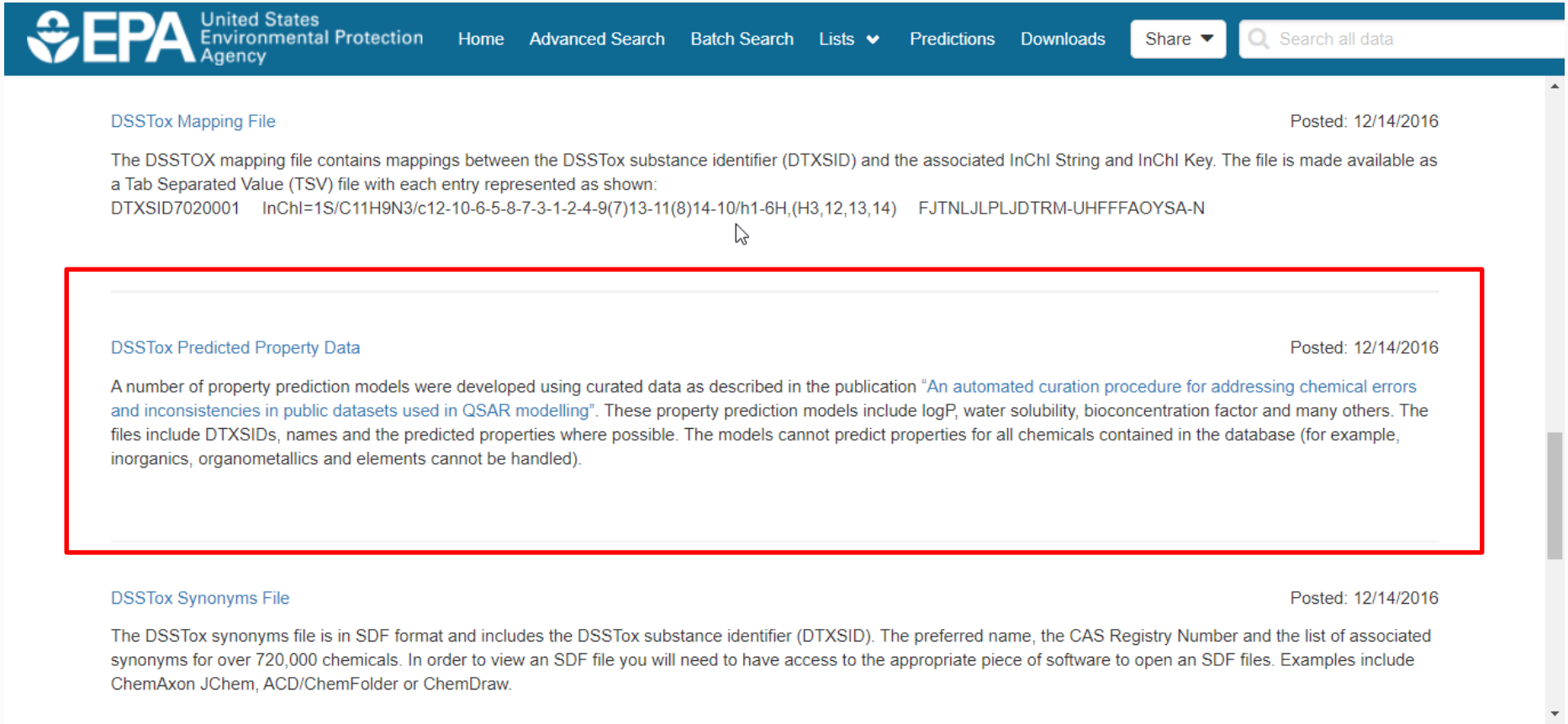
Save Copy


```
uuid:                "55547f4f-f966-48e8-b831-a0d217998064"
predictionTime:      1520539090089
software:            "T.E.S.T (Toxicity Estimation Software Tool)"
softwareVersion:     "5.01"
condition:           "25°C"
endpoint:            "Water solubility at 25°C"
method:              "Hierarchical clustering"
predictions:
  0:
    id:              "C_1520539090089"
    smiles:          "OCC"
    expValMolarLog:  "-1.337"
    expValMass:      "1001180.703"
    predValMolarLog: "-1.338"
    predValMass:     "1002625.241"
    molarLogUnits:   "-Log10(mol/L)"
    massUnits:       "mg/L"
```

Our support for FAIR Data



Downloadable Data

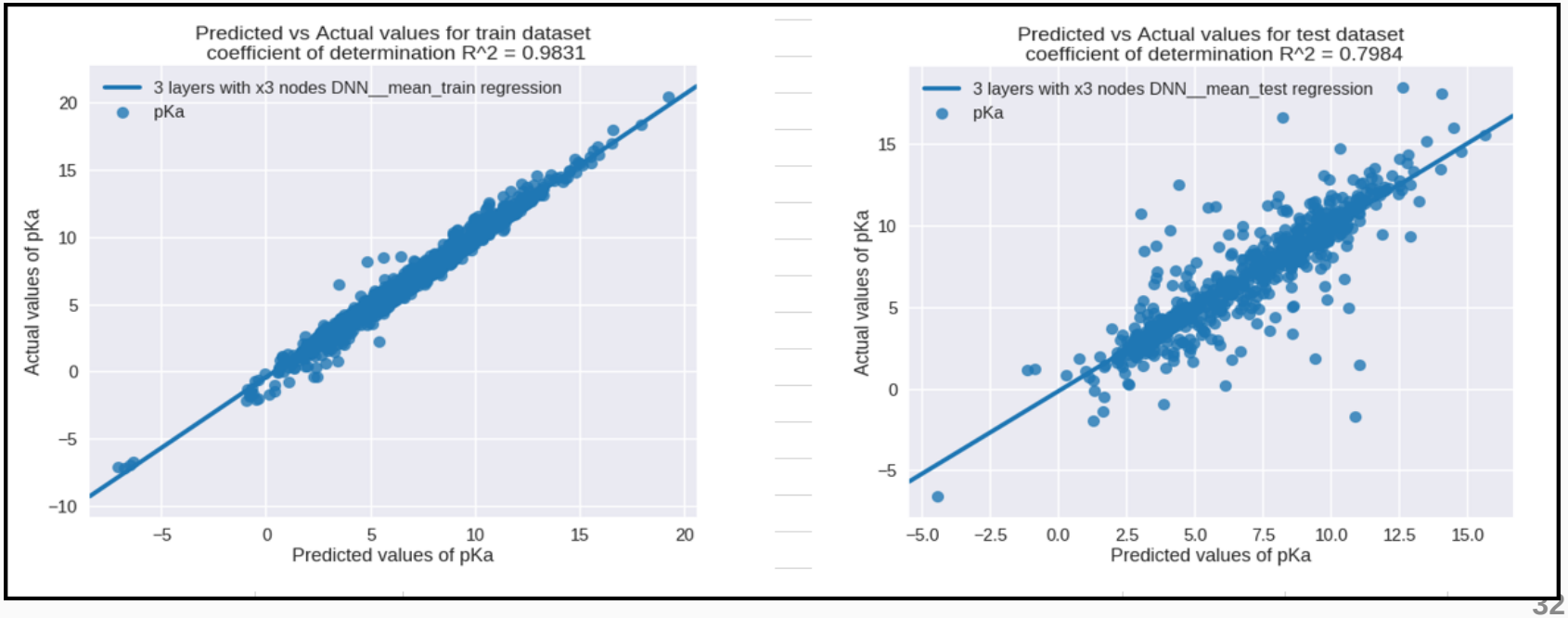


Work in Progress

- Present work in development
  - Development of OPERA model web services
  - Development of pKa and logD prediction models
  - Display of TEST Toxicity endpoints predicted data
  - Analytical Data support
    - Spectral searching against predicted Mass Spectra

pKa Prediction Model

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



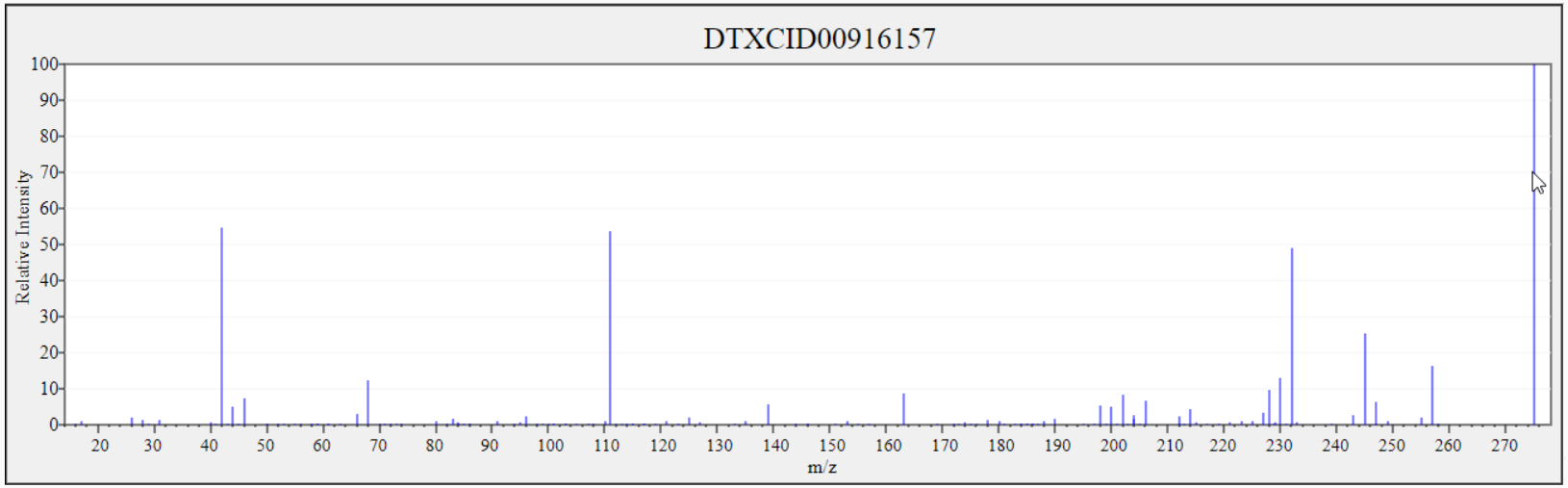
Toxicity Endpoints in TEST

Endpoint	Definition
96 hour fathead minnow LC_{50}	Concentration in mg/L that causes 50% of fathead minnow to die after 96 hours
48 hour *Daphnia magna* LC_{50}	Concentration in mg/L that causes 50% of *Daphnia magna* to die after 48 hours
48 hour *T. pyriformis* IGC_{50}	Concentration in mg/L that causes 50% growth inhibition to *T. pyriformis* after 48 hours
Oral rat LD_{50}	Amount of chemical in mg/kg body weight that causes 50% of rats to die after oral ingestion
Developmental toxicity	Whether or not a chemical causes developmental toxicity effects to humans or animals
Ames mutagenicity	A compound is positive for mutagenicity if it induces revertant colony growth in any strain of Salmonella typhimurium

Predicted Mass Spectra

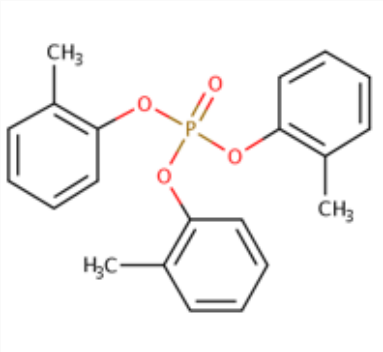
http://cfmid.wishartlab.com/

- MS/MS spectra prediction for ESI+, ESI-, and EI
- Predictions generated and stored for >700,000 structures, to be accessible via Dashboard

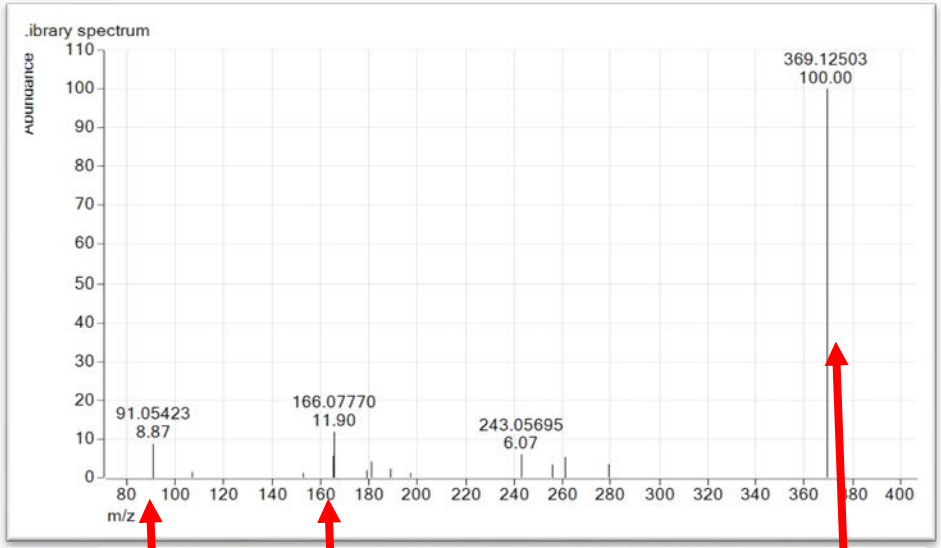


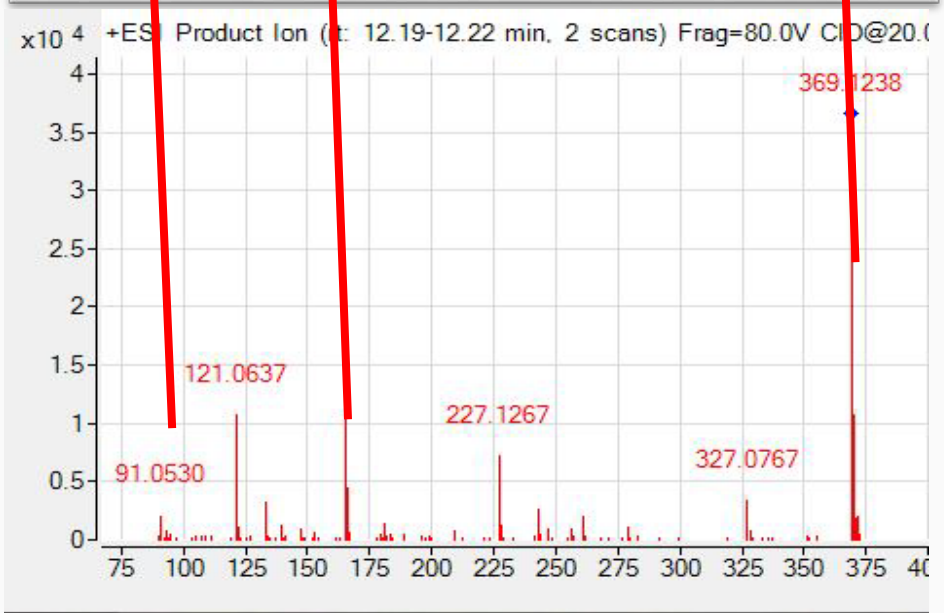
Predicted Mass Spectra

Library Fragmentation Spectra (20eV)



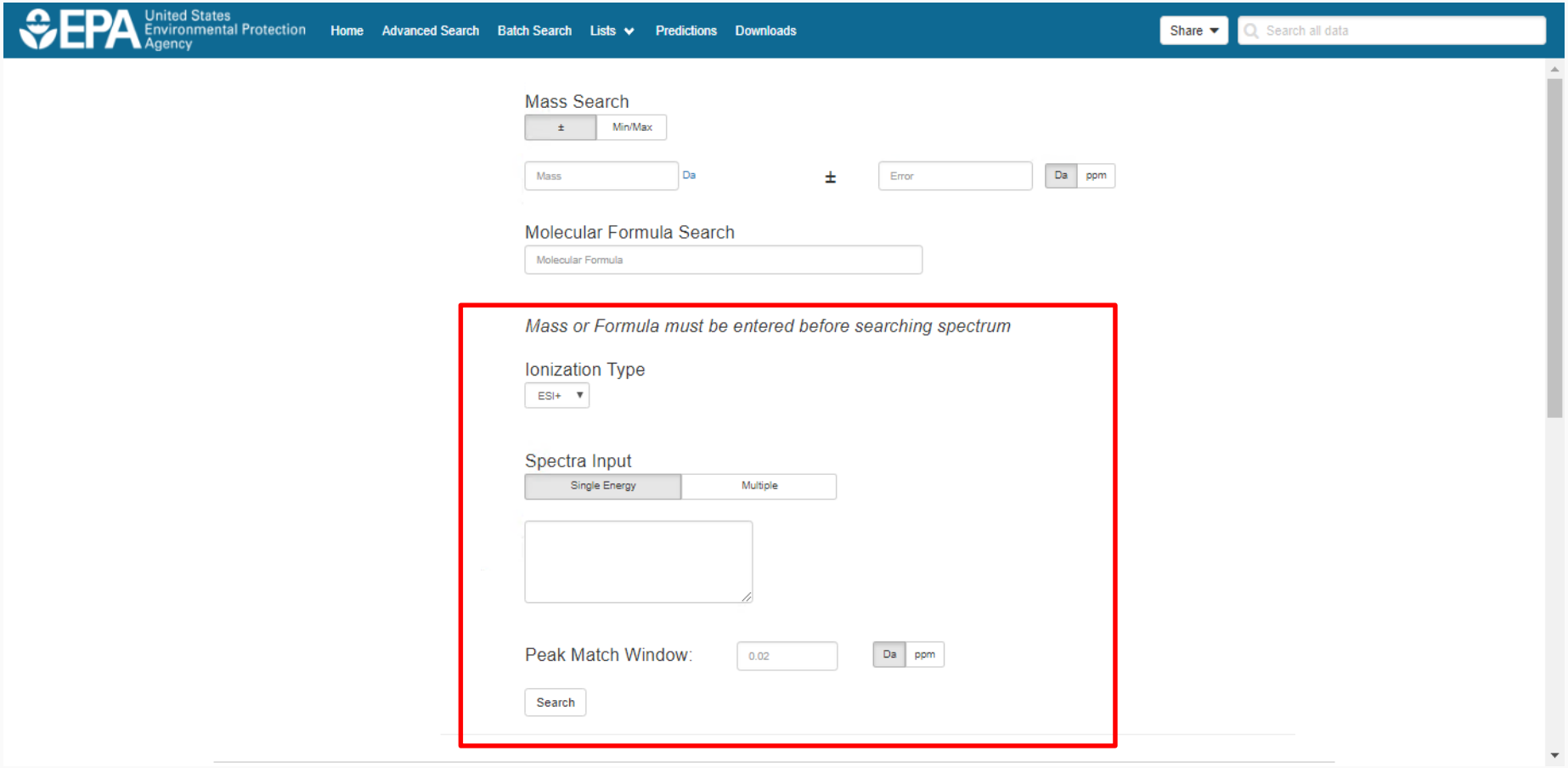
Observed Fragmentation Spectra (20eV)





Match Score

Search Expt. vs. Predicted Spectra



Conclusion

- The CompTox Dashboard provides access to data for ~762,000 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 delivered
- Real time prediction models expanding in number
- Web services available for some physchem and toxicity endpoints

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