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

Real-time prediction of physicochemical & toxicological endpoints using the web-based CompTox Chemistry Dashboard

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

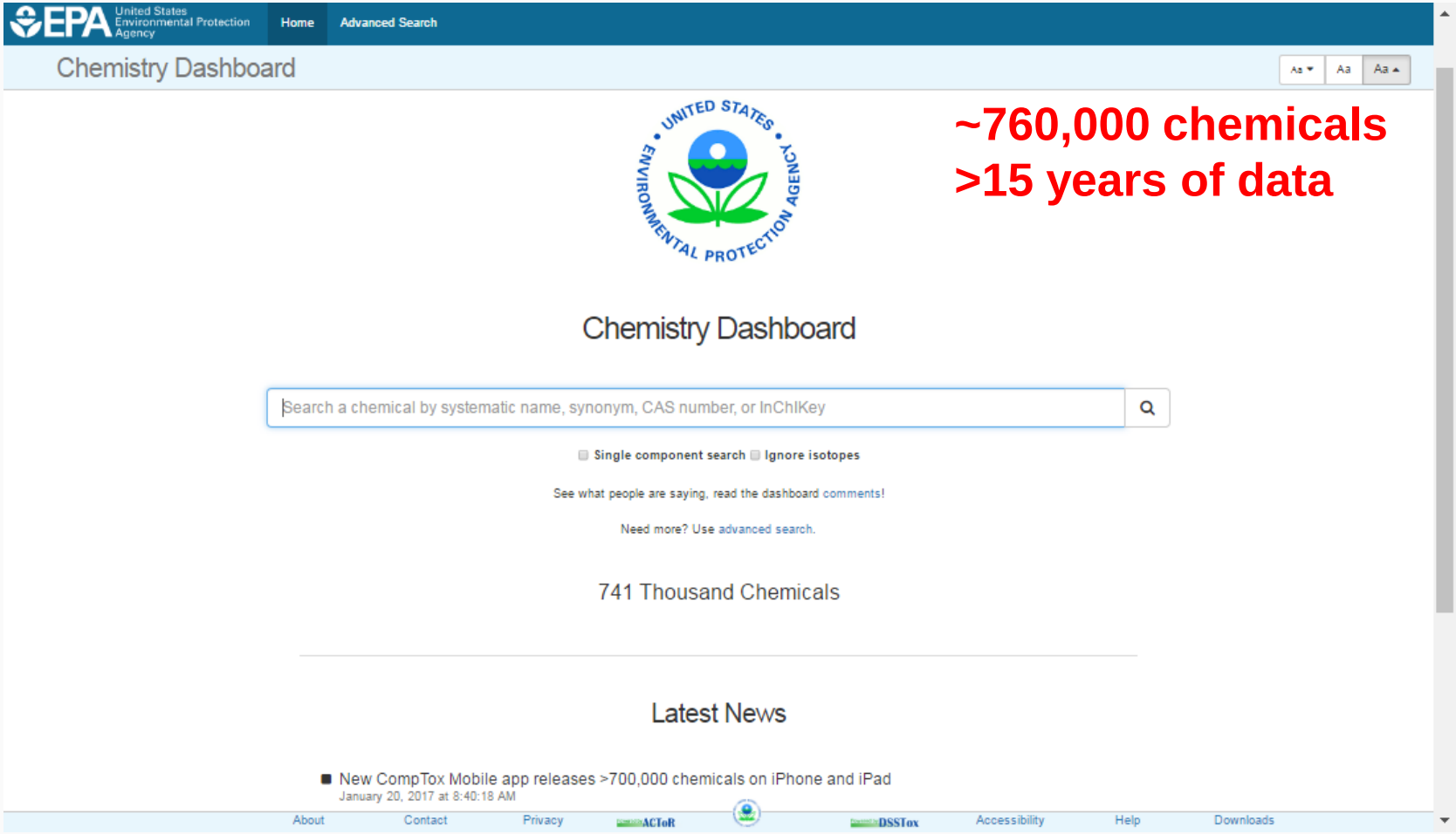
1. *National Center for Computational Toxicology, U.S. Environmental Protection Agency, RTP, NC*
2. *Oak Ridge Institute of Science and Education (ORISE) Research Participant, Research Triangle Park, NC*
3. *Science Data Experts, MD, USA*
4. *Scitovation, RTP-NC*

August 2017
ACS Fall Meeting, Washington, DC

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
  - **In progress : public web services & real time prediction**

Comptox Chemistry Dashboard
https://comptox.epa.gov

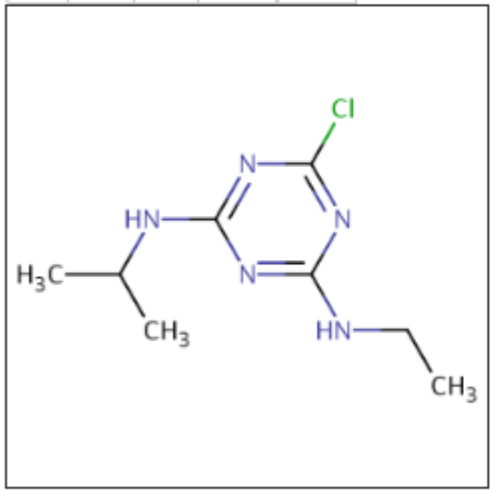


Chemical Page

Atrazine

1912-24-9 | DTXSID9020112

Searched by Approved Name: Found 1 result for 'atrazine'.



Wikipedia

Intrinsic Properties

Molecular Formula: C8H14ClN5 — Find All Chemicals

Average Mass: 215.69 g/mol

Monoisotopic Mass: 215.093773 g/mol

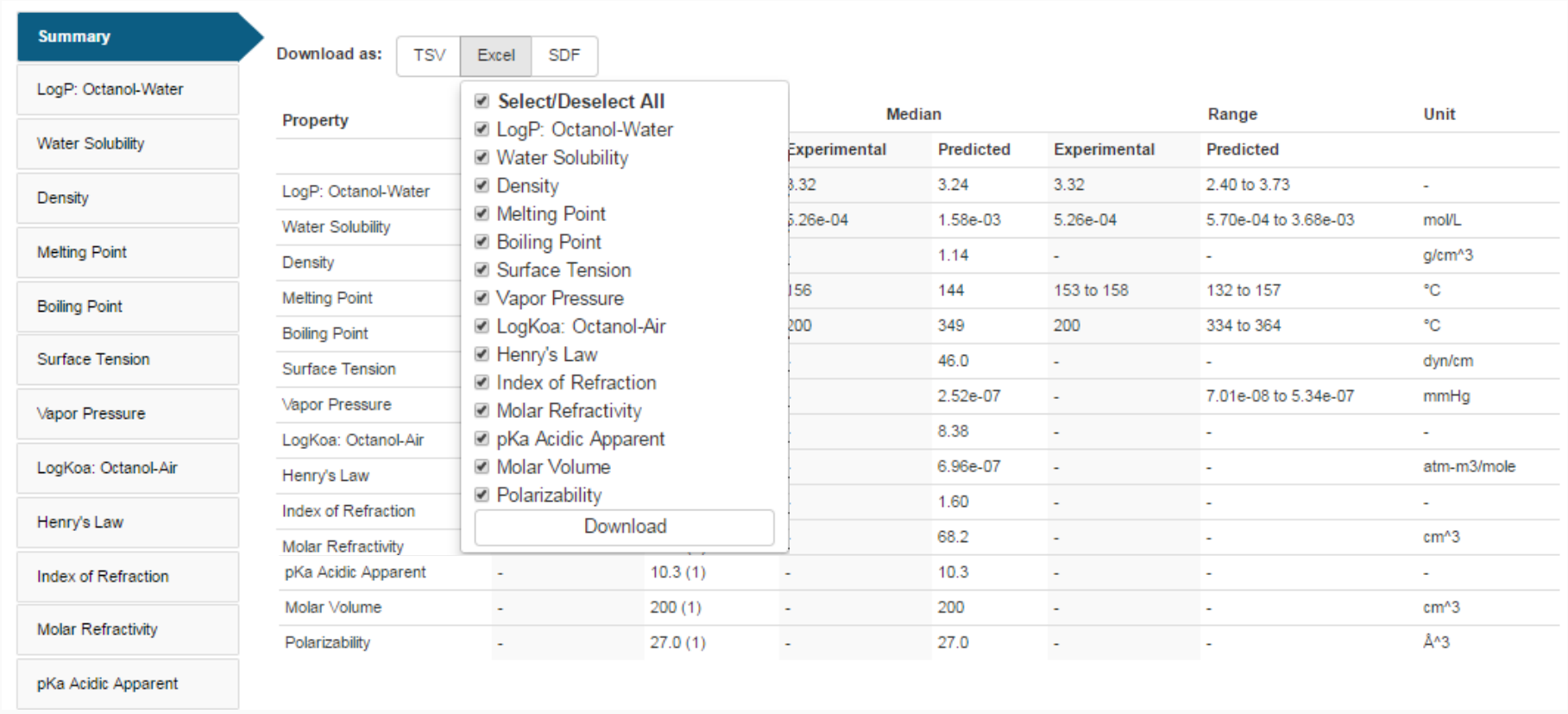
Structural Identifiers

Record Information

Chemical Properties | Synonyms | External Links | Product Composition | Bioassays | Exposure | Analytical | Literature | Comments

Prediction Algorithms

- Multiple prediction algorithms used – OPERA, TEST, NICEATM, EPI Suite



Data Distribution

	A	B	C	D	E	F	G	H
1	**Property**	**Average**		**Median**		**Range**		**Unit**
2		**Experimental**	**Predicted**	**Experimental**	**Predicted**	**Experimental**	**Predicted**	
3	LogP: Octanol-Water	3.32 (1)	3.24 (4)	3.32	3.24	3.32	2.40 to 3.73	-
4	Water Solubility	5.26e-04 (1)	1.58e-03 (4)	0.000526	0.00158	0.000526	5.70e-04 to 3.68e-03	mol/L
5	Density	-	1.14 (1)	-	1.14	-	-	g/cm^3
6	Melting Point	155 (7)	144 (3)	156	144	153 to 158	132 to 157	°C
7	Boiling Point	200 (1)	349 (3)	200	349	200	334 to 364	°C
8	Surface Tension	-	46.0 (1)	-	46	-	-	dyn/cm
9	Vapor Pressure	-	2.52e-07 (3)	-	0.000000252	-	7.01e-08 to 5.34e-07	mmHg
10	LogKoa: Octanol-Air	-	8.38 (1)	-	8.38	-	-	-
11	Henry's Law	-	6.96e-07 (1)	-	0.000000696	-	-	atm-m3/mole
12	Index of Refraction	-	1.60 (1)	-	1.6	-	-	-
13	Molar Refractivity	-	68.2 (1)	-	68.2	-	-	cm^3
14	pKa Acidic Apparent	-	10.3 (1)	-	10.3	-	-	-
15	Molar Volume	-	200 (1)	-	200	-	-	cm^3
16	Polarizability	-	27.0 (1)	-	27	-	-	Å^3
17								

Developing “OPERA Models”

- Our approach to modeling:
  - Obtain **high quality** training sets
  - Apply appropriate modeling approaches
  - **Validate** performance of models
  - Define the applicability domain and model limitations
  - Use models to predict properties across our full datasets
  - Release as **Open Data and Open Models**

PHYSPROP Data Files

http://esc.syrres.com/interkow/EpiSuiteData_ISIS_SDF.htm

EPA United States Environmental Protection Agency

EPI Suite Data - ISIS/Base & SDF

The downloaded files are provided in "zip" format ... the downloaded file must be "un-zipped" with common utility programs such as WinZip.

... ***Updated September 15, 2010***

Basic Instructions:

(1) Download the zip file
(2) Un-Zip the file

NOTE ... zipped files extract to **Folders** containing the individual data files ... Folders named **EPI_ISIS_Data** and **EPI_SDF_Data**

Substructure Searching Files:

ISIS™/Base & SD Files of the EPI Suite Program Experimental Data Files are now available ... The ISIS™/Base files require the commercial program for use ... The SD Files can be imported into other commercial chemical structure programs (such as ChemFinder).

... **Click here to download EPI_ISIS_Data.zip** ... (about 11 MB)

... **Click here to download EPI_SDF_Data.zip** ... (about 10 MB)

NOTE ... **EPI Suite Data Files (some in Excel, Text, Word format) available at:**

http://esc.syrres.com/interkow/EpiSuiteData.htm

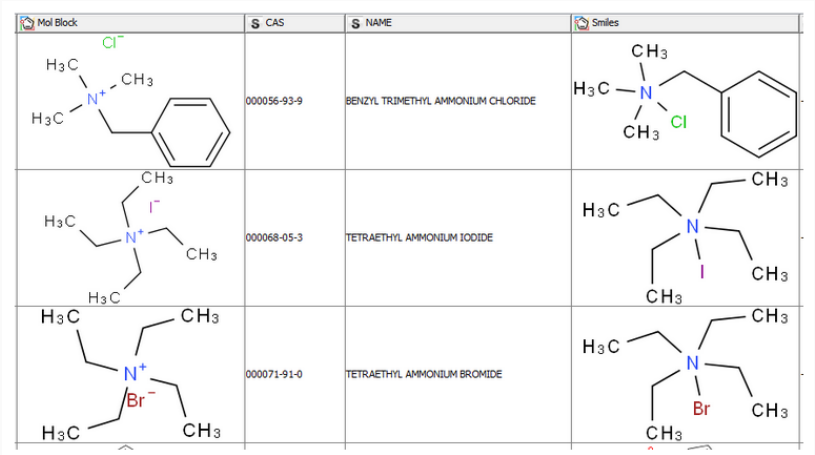
Available Properties

- Solubility
- Melting Point
- Boiling Point
- LogP (Octanol-water partition coefficient)
- Atmospheric Hydroxylation Rate
- LogBCF (Bioconcentration Factor)
- Biodegradation Half-life
- Henry's Law Constant
- Fish Biotransformation Half-life
- LogKOA (Octanol/Air Partition Coefficient)
- LogKOC (Soil Adsorption Coefficient)
- Vapor Pressure

Curating Public Data

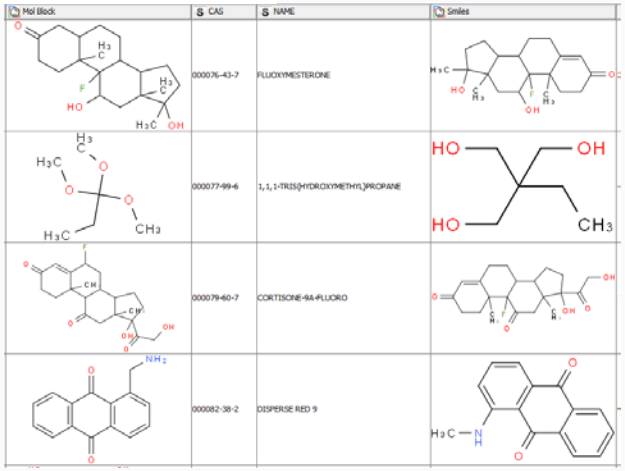
EPA United States Environmental Protection Agency

Public data should be curated prior to modeling



Covalent Halogens

Identical Chemicals



Mismatches

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%)

Workflow Details and Data

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

OPERA Model Predictions

EPA United States Environmental Protection Agency

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

Melting Point			
	Average	Median	Range
Experimental	155 (7)	156	153 to 158
Predicted	139 (4)	139	125 to 157

Download as: TSV | Excel | SDF

Experimental

Source	Result
Jean-Claude Bradley Open Melting Point Dataset	153 °C
Jean-Claude Bradley Open Melting Point Dataset	156 °C
TCI	156 °C
Merck Millipore	156 °C
Alfa Aesar	156 °C
Alfa Aesar	156 °C
PhysPropNCCT	153 °C

Predicted

Source	Result	Calculation Details	QMRF
EPISUITE	132 °C	Not Available	Not Available
NICEATM	157 °C	Not Available	Available
TEST	125 °C	Not Available	Not Available
OPERA	144 °C	OPERA Model Report	Available

GREEN – Inside Applicability Domain

OPERA Model Report

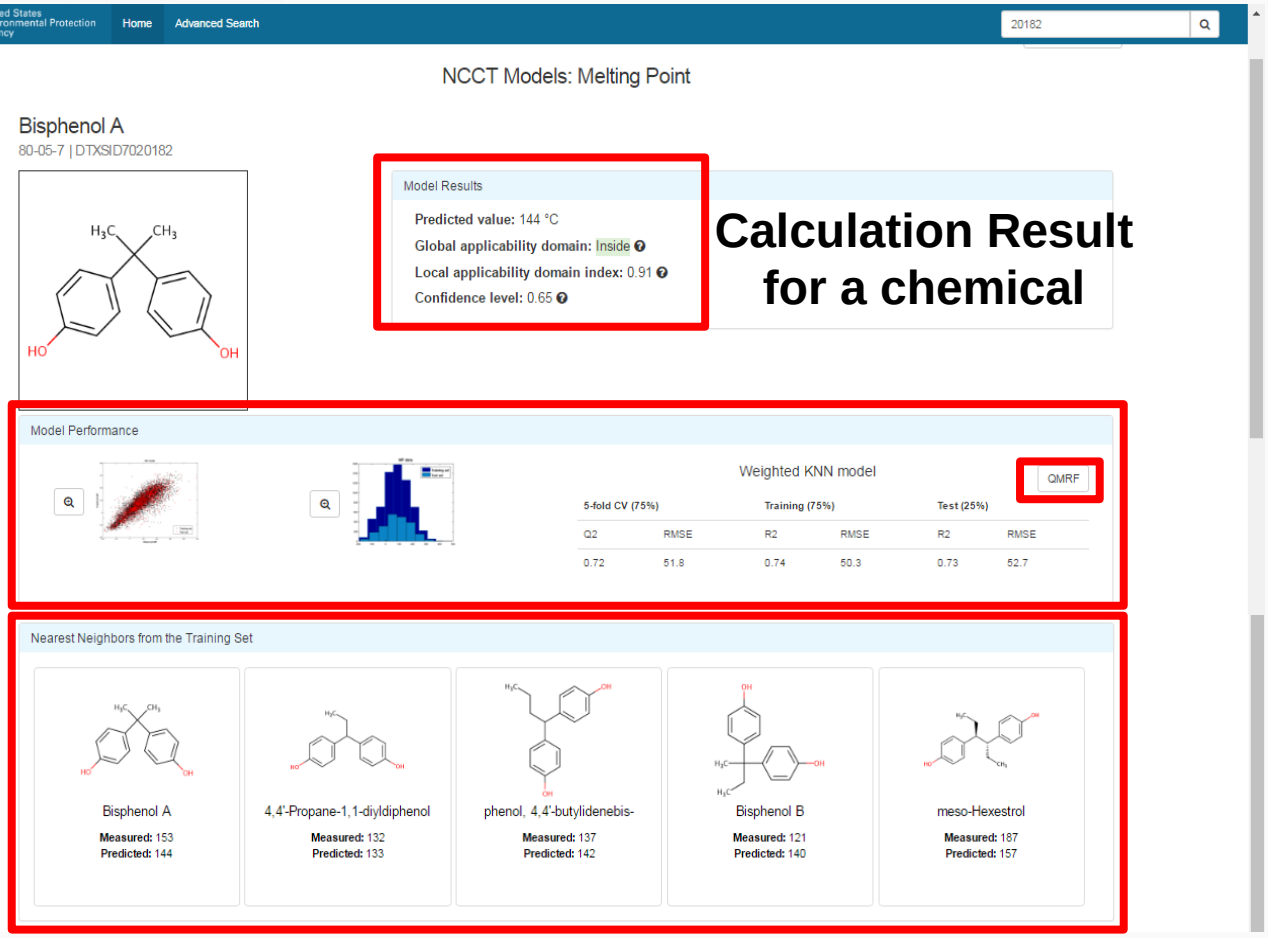
OPERA Model Report

Chemical Properties | Env. Fate/Transport | Synonyms | External Links | Toxicity Values (Beta) | Exposure | Bioassays | Similar Molecules (Beta) | Literature | Comments

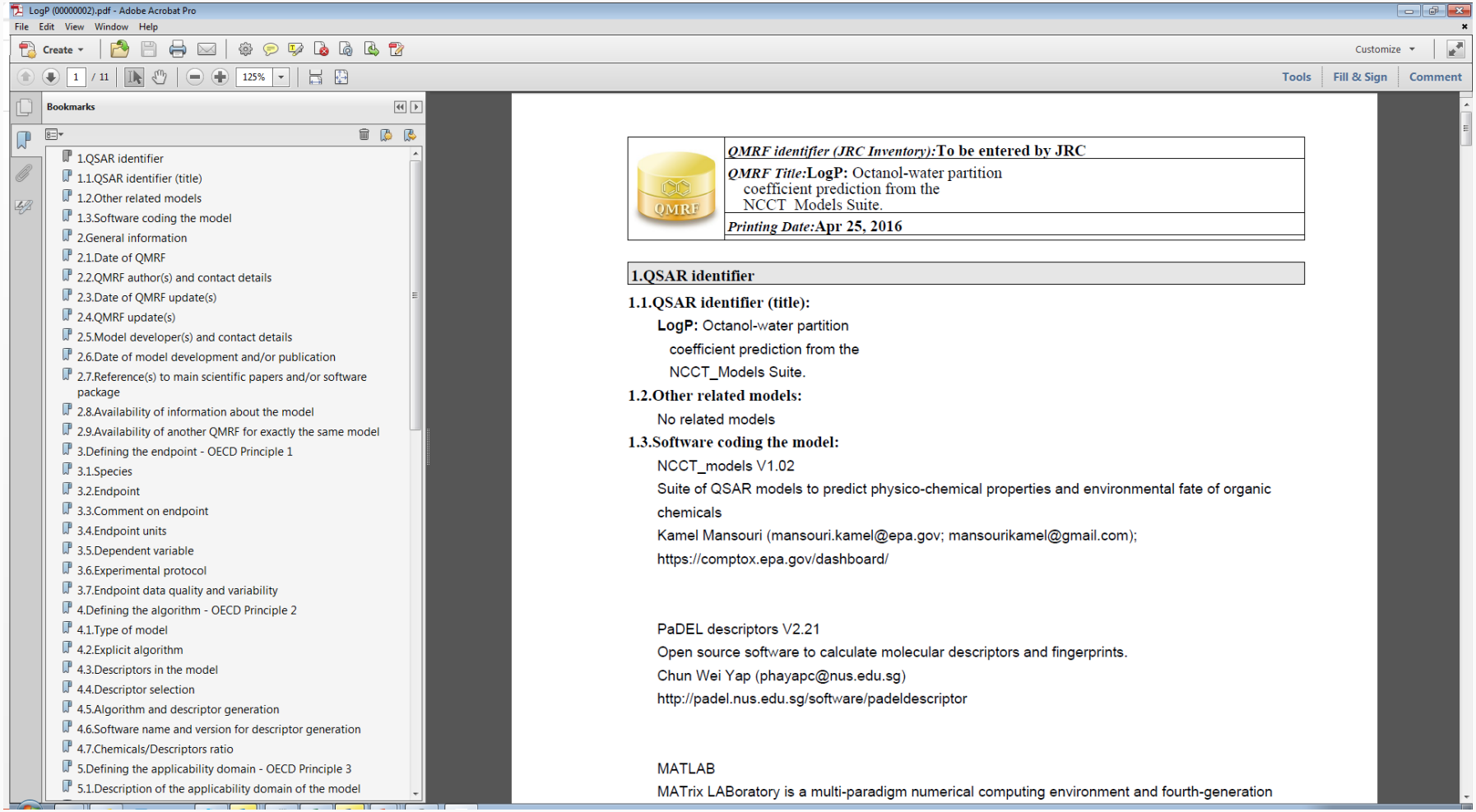
Modeling Details

Model Performance with full QMRF

Nearest Neighbors from Training Set



QSAR Modeling Reporting Format



Downloadable Data File

https://comptox.epa.gov/dashboard/downloads

Downloads

DSSTox Identifier to PubChem Identifier Mapping File — Posted: 11/14/2016

The DSSTox to PubChem Identifiers mapping file is in TXT format and includes the PubChem SID, PubChem CID and DSSTox substance identifier (DTXSID).
image

DSSTox identifiers mapped to CAS Numbers and Names File — Posted: 11/14/2016

The DSSTox Identifiers file is in Excel format and includes the CAS Number, DSSTox substance identifier (DTXSID) and the Preferred Name.
image

DSSTox Mapping File — Posted: 12/14/2016

The DSSTOX mapping file contains mappings between the DSSTox substance identifier (DTXSID) and the associated InChI String and InChI Key. The file is made available as a Tab Separated Value (TSV) file with each entry represented as shown:
DTXSID7020001 InChI=1S/C11H9N3/c12-10-6-5-8-7-3-1-2-4-9(7)13-11(8)14-10/h1-6H,(H3,12,13,14) FJTNLJLPLJDTRM-UHFFFAOYSA-N

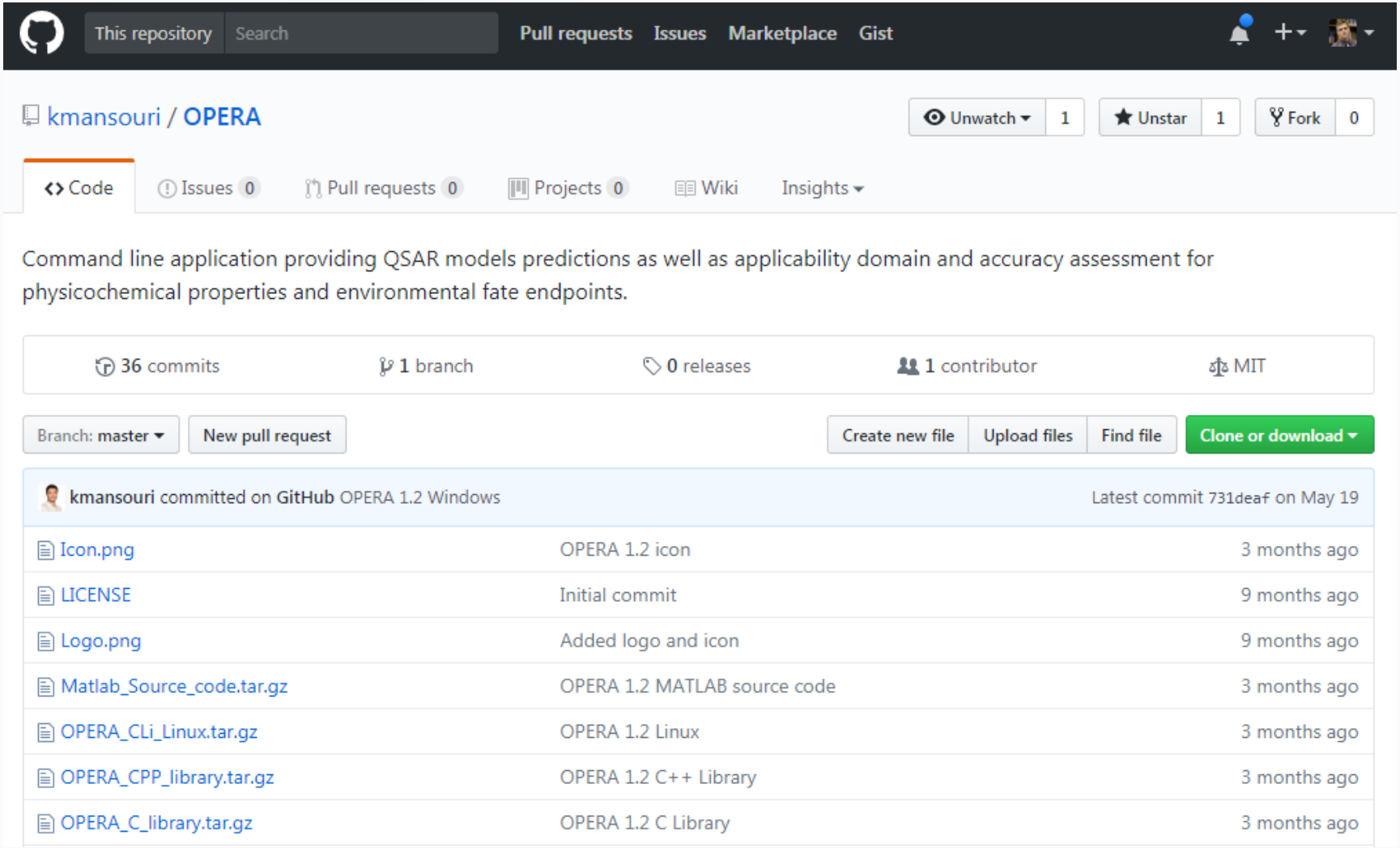
DSSTox Predicted Property Data — Posted: 12/14/2016

A number of property prediction models were developed using curated data as described in the publication "An automated curation procedure for addressing chemical errors and inconsistencies in public datasets used in QSAR modelling". These property prediction models include logP, water solubility, bioconcentration factor and many others. The files include DTXSIDs, names and the predicted properties where possible. The models cannot predict properties for all chemicals contained in the database (for example, inorganics, organometallics and elements cannot be handled).

DSSTox Synonyms File — Posted: 12/14/2016

The DSSTox synonyms file is in SDF format and includes the DSSTox substance identifier (DTXSID). The preferred name, the CAS Registry Number and the list of associated synonyms for over 720,000 chemicals. In order to view an SDF file you will need to have access to the appropriate piece of software to open an SDF files. Examples include ChemAxon JChem, ACD/ChemFolder or ChemDraw.

OPERA on GitHub



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

OPERA Standalone application:

Input:

- MATLAB .mat file, an ASCII file with only a matrix of variables
- SDF file or SMILES strings of QSAR-ready structures. In this case the program will calculate PaDEL 2D descriptors and make the predictions.
- The program will extract the molecules names from the input csv or SDF (or assign arbitrary names if not) as IDs for the predictions.

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

OPERA Standalone application:

Output:

- Depending on the extension, the can be text file or csv with
  - A list of molecules IDs and predictions
  - Applicability domain
  - Accuracy of the prediction
  - Similarity index to the 5 nearest neighbors
  - The 5 nearest neighbors from the training set: Exp. value, Prediction, InChI key

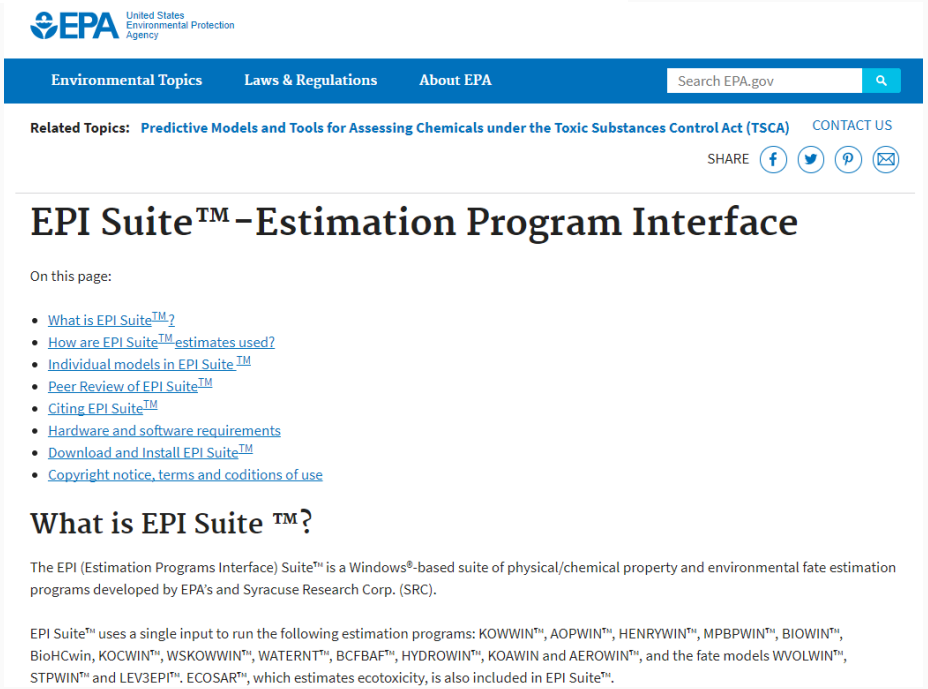
https://github.com/kmansouri/OPERA.git

EPI Suite Models

https://www.epa.gov/tsca-screening-tools/epi-suitetm-estimation-program-interface

EPA United States Environmental Protection Agency

Predicted			
Source	Result	Calculation Details	QMRF
EPISUITE	364 °C	Not Available	Not Available
NICEATM	350 °C	Not Available	Available
ACD/Labs	401 °C	Not Available	Not Available
TEST	360 °C	Not Available	Not Available
OPERA	343 °C	OPERA Model Report	Available



EPA United States Environmental Protection Agency

Environmental Topics | Laws & Regulations | About EPA | Search EPA.gov

Related Topics: Predictive Models and Tools for Assessing Chemicals under the Toxic Substances Control Act (TSCA) CONTACT US

SHARE

EPI Suite™–Estimation Program Interface

On this page:

- What is EPI Suite™?
- How are EPI Suite™ estimates used?
- Individual models in EPI Suite™
- Peer Review of EPI Suite™
- Citing EPI Suite™
- Hardware and software requirements
- Download and Install EPI Suite™
- Copyright notice, terms and coditions of use

What is EPI Suite ™?

The EPI (Estimation Programs Interface) Suite™ is a Windows®-based suite of physical/chemical property and environmental fate estimation programs developed by EPA's and Syracuse Research Corp. (SRC).

EPI Suite™ uses a single input to run the following estimation programs: KOWWIN™, AOPWIN™, HENRYWIN™, MPBPWIN™, BIOWIN™, BioHCwin, KOCWIN™, WSKOWWIN™, WATERNT™, BCFBAF™, HYDROWIN™, KOAWIN and AEROWIN™, and the fate models WVOLWIN™, STPWIN™ and LEV3EPI™. ECOSAR™, which estimates ecotoxicity, is also included in EPI Suite™.

Now Testing EPI Suite Services

Enter SMILES string to test EPI Web Services.

CCNC1=NC(NC(C)C)=NC(C | Get Kow | Get Koc | Get BCF/BAF | Get Melting Point | Get Boiling Point | Get Vapor Pressure | Get Henry's Law Constant | Get Water Solubility | Get AOP | Get BIOWIN | Get Fugacity

XML response:


```
<?xml version='1.0' encoding='UTF-8'?>
<S:Envelope
        xmlns:S="http://schemas.xmlsoap.org/soap/envelope/">
        <S:Body>
                <ns2:getKowResultResponse
                        xmlns:ns2="http://epiwebservices.srcinc.com/">
                        <return>
                                <LogKow>2.82</LogKow>
                                <LogKowExperimental>2.609999895095825</LogKowExperimental>
                                <ExperimentalReferences>HANSCH,C ET AL.</ExperimentalReferences>
                                <ExperimentalReferences
                                        xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:nil="true"/>
                                        <ExperimentalReferences
                                                xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:nil="true"/>
                                                <ExperimentalReferences
                                                        xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:nil="true"/>
                                                        <ExperimentalReferences
                                                                xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:nil="true"/>
                                                                <MolecularWeight>215.6873016357422</MolecularWeight>
                                                                <MolecularFormula>C8 H14 Cl1 N5 </MolecularFormula>
                                                                <Factors>
                                                                        <ValueType>Fragment</ValueType>
                                                                        <NumberOfFragments>3</NumberOfFragments>
                                                                        <FragmentDescription>-CH3    [aliphatic carbon]               </FragmentDescription>
                                                                        <Coefficient>0.5472999811172485</Coefficient>
                                                                        <TotalCoefficient>1.6418999433517456</TotalCoefficient>
                                                                </Factors>
                                                                <Factors>
                                                                        <ValueType>Fragment</ValueType>
                                                                        <NumberOfFragments>1</NumberOfFragments>
                                                                        <FragmentDescription>-CH2-   [aliphatic carbon]               </FragmentDescription>
                                                                        <Coefficient>0.491100013256073</Coefficient>
                                                                        <TotalCoefficient>0.491100013256073</TotalCoefficient>
```

NICEATM Models
logP, WS, BP, MP, VP, BCF

EPA United States Environmental Protection Agency

Predicted			
Source	Result	Calculation Details	QMRF
EPISUITE	364 °C	Not Available	Not Available
NICEATM	350 °C	Not Available	Available
ACD/Labs	401 °C	Not Available	Not Available
TEST	360 °C	Not Available	Not Available
OPERA	343 °C	OPERA Model Report	Available

JOURNAL OF CHEMICAL INFORMATION AND MODELING

Search | Ci

Enter search text

J. Chem. Inf. Mc

Browse the Journal | Articles ASAP | Current Issue | Submission & Review | Open Access | About the Jou

Article

Previo

In Silico Prediction of Physicochemical Properties of Environmental Chemicals Using Molecular Fingerprints and Machine Learning

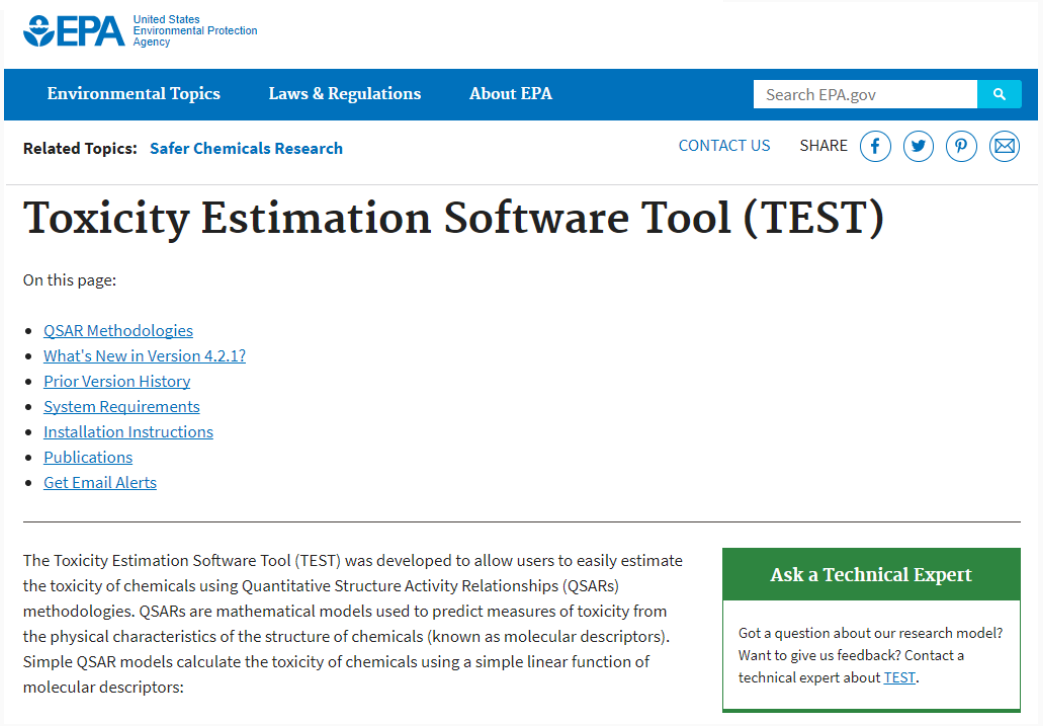
Qingda Zang†, Kamel Mansouri‡, Antony J. Williams‡, Richard S. Judson‡, David G. Allen†, Warren M. Casey∥, and Nicole C. Kleinstreuer*∥

T.E.S.T Models
logP, WS, BP, MP, VP, BCF

EPA United States Environmental Protection Agency

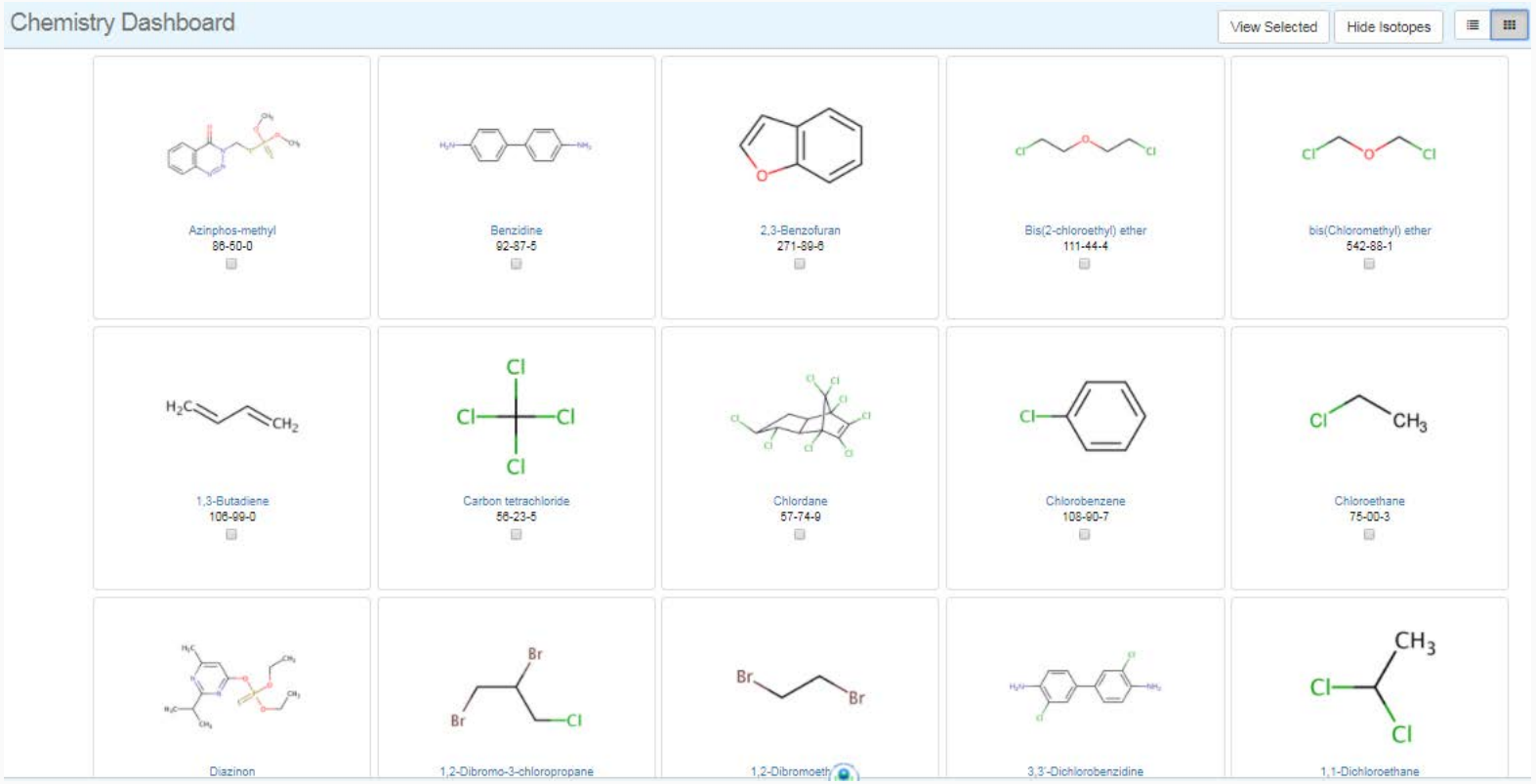
Predicted

Source	Result	Calculation Details	QMRF
EPISUITE	364 °C	Not Available	Not Available
NICEATM	350 °C	Not Available	Available
ACD/Labs	401 °C	Not Available	Not Available
TEST	360 °C	Not Available	Not Available
OPERA	343 °C	OPERA Model Report	Available

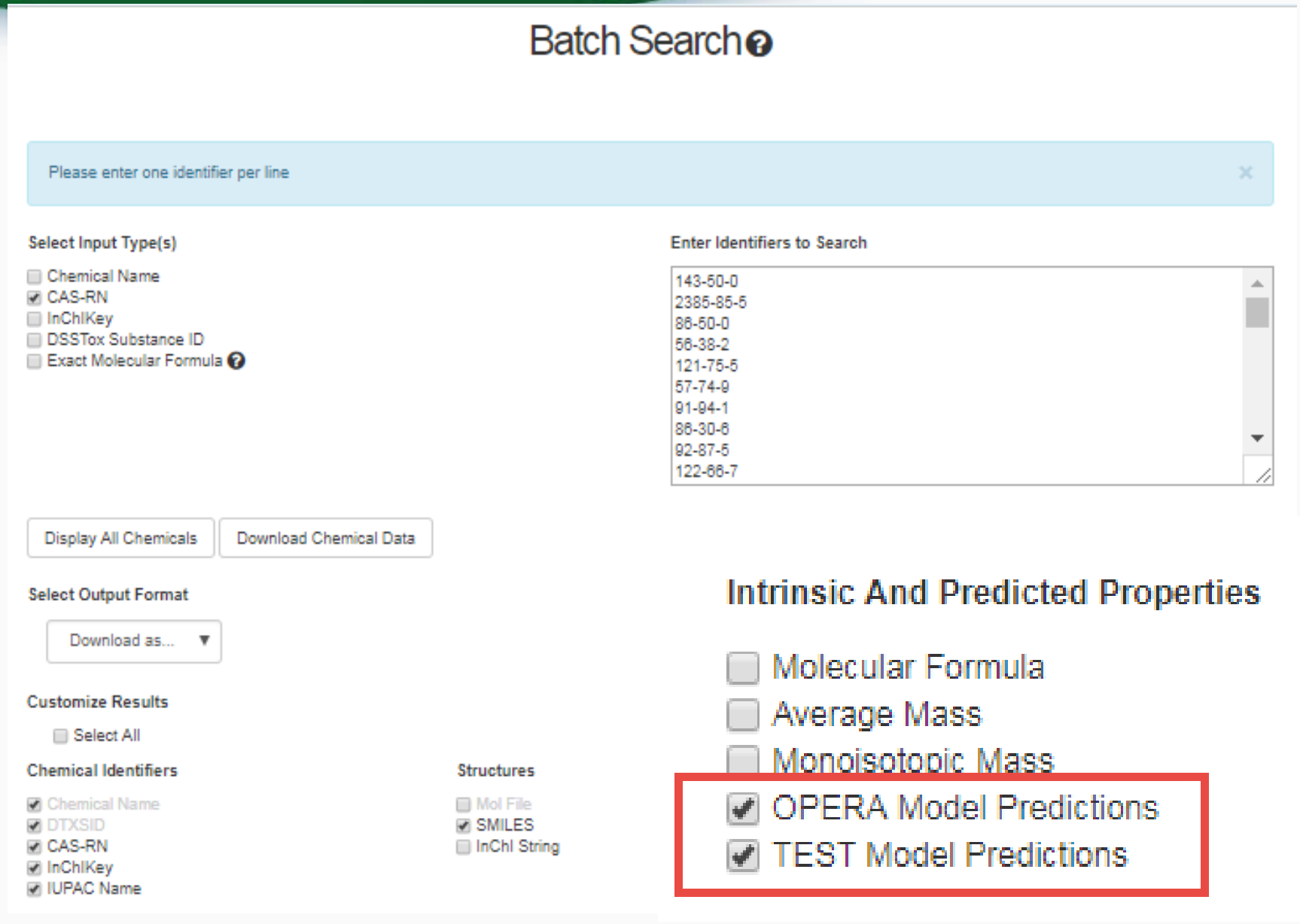


Batch Predictions for Chemicals

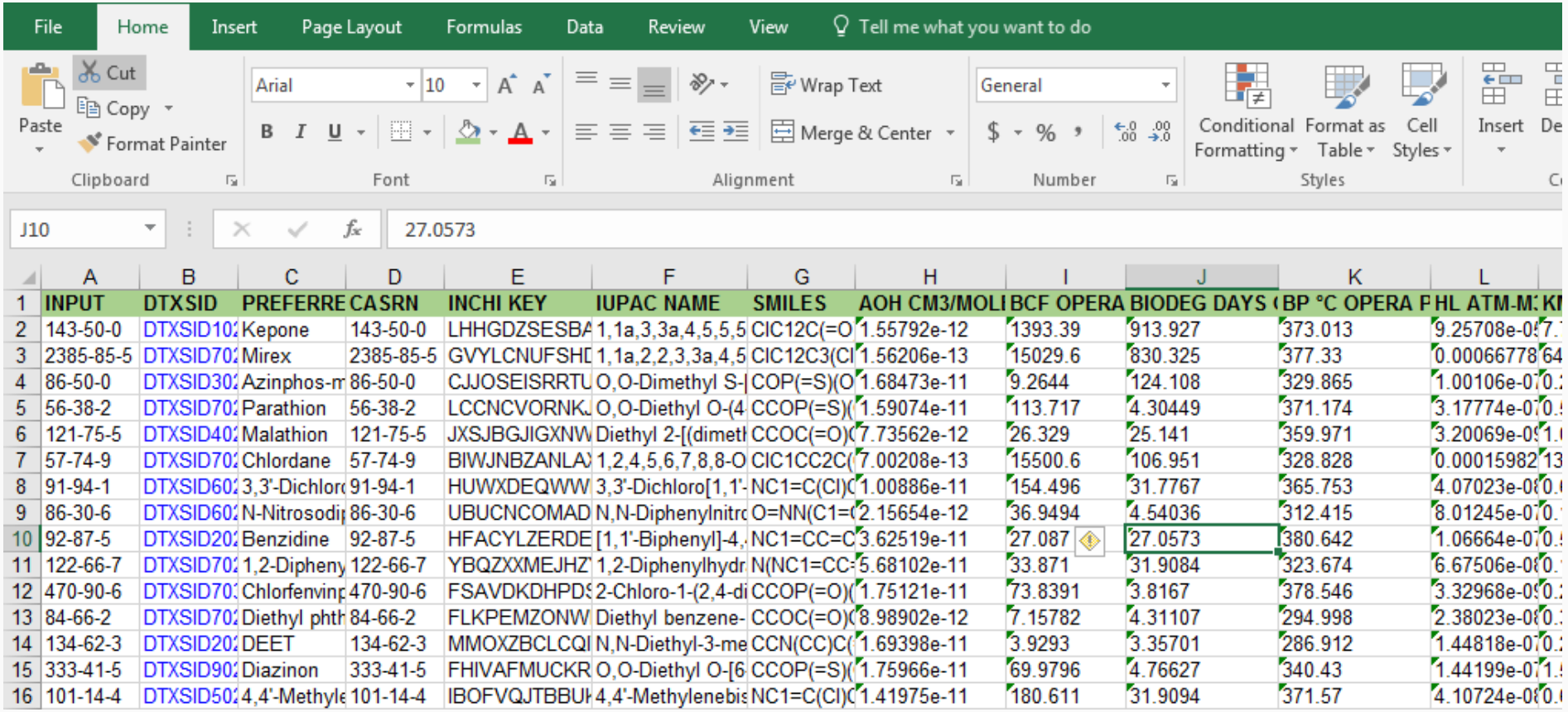
- I want OPERA predictions for a set of chemicals – I have their CASRNs (or names)



Batch Predictions for Chemicals



Excel Download



File Home Insert Page Layout Formulas Data Review View Tell me what you want to do

Cut Copy Format Painter Paste — Clipboard

Arial 10 — Font

Wrap Text Merge & Center — Alignment

General — Number

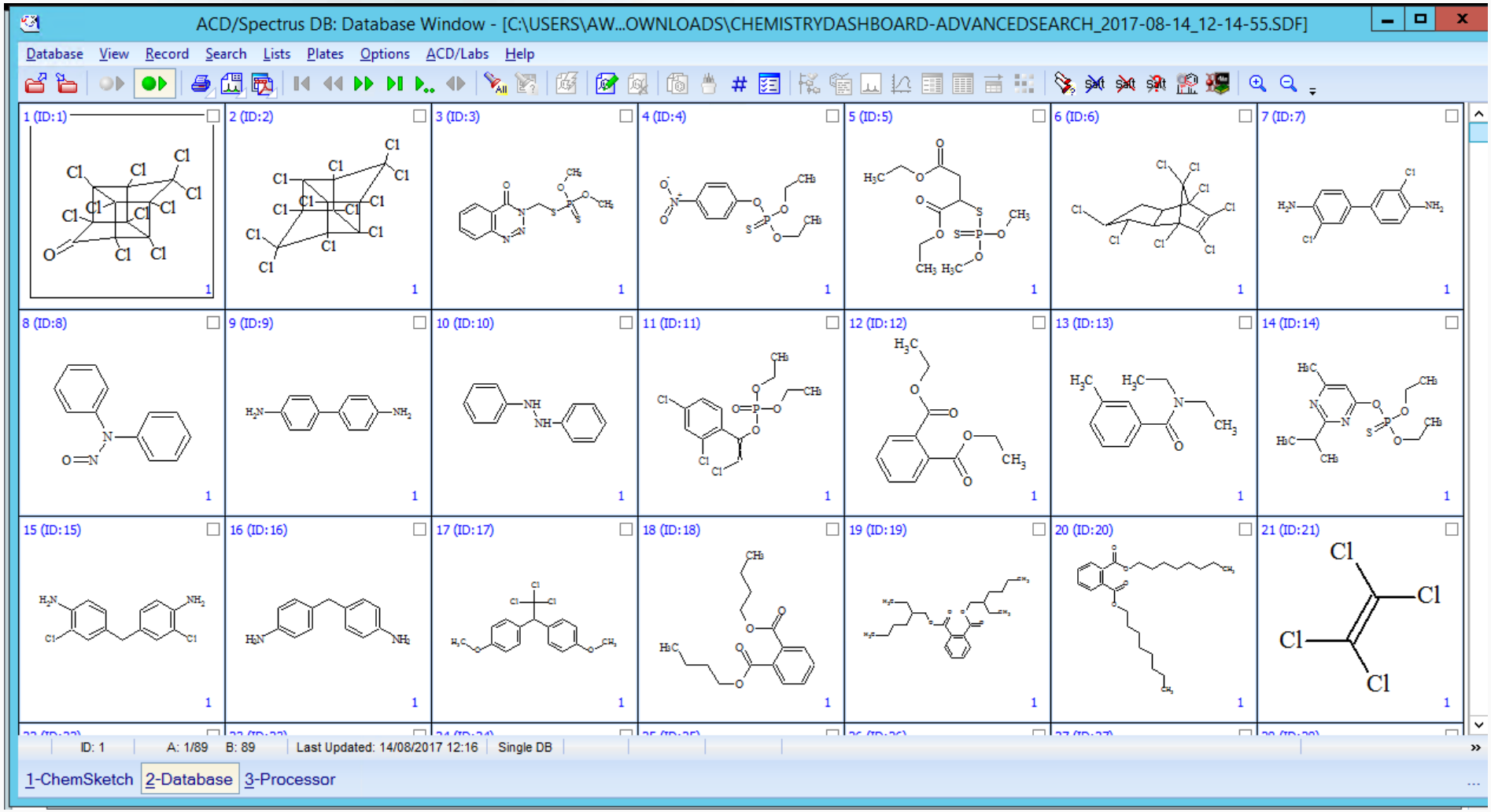
Conditional Formatting Format as Table Cell Styles — Styles

Insert

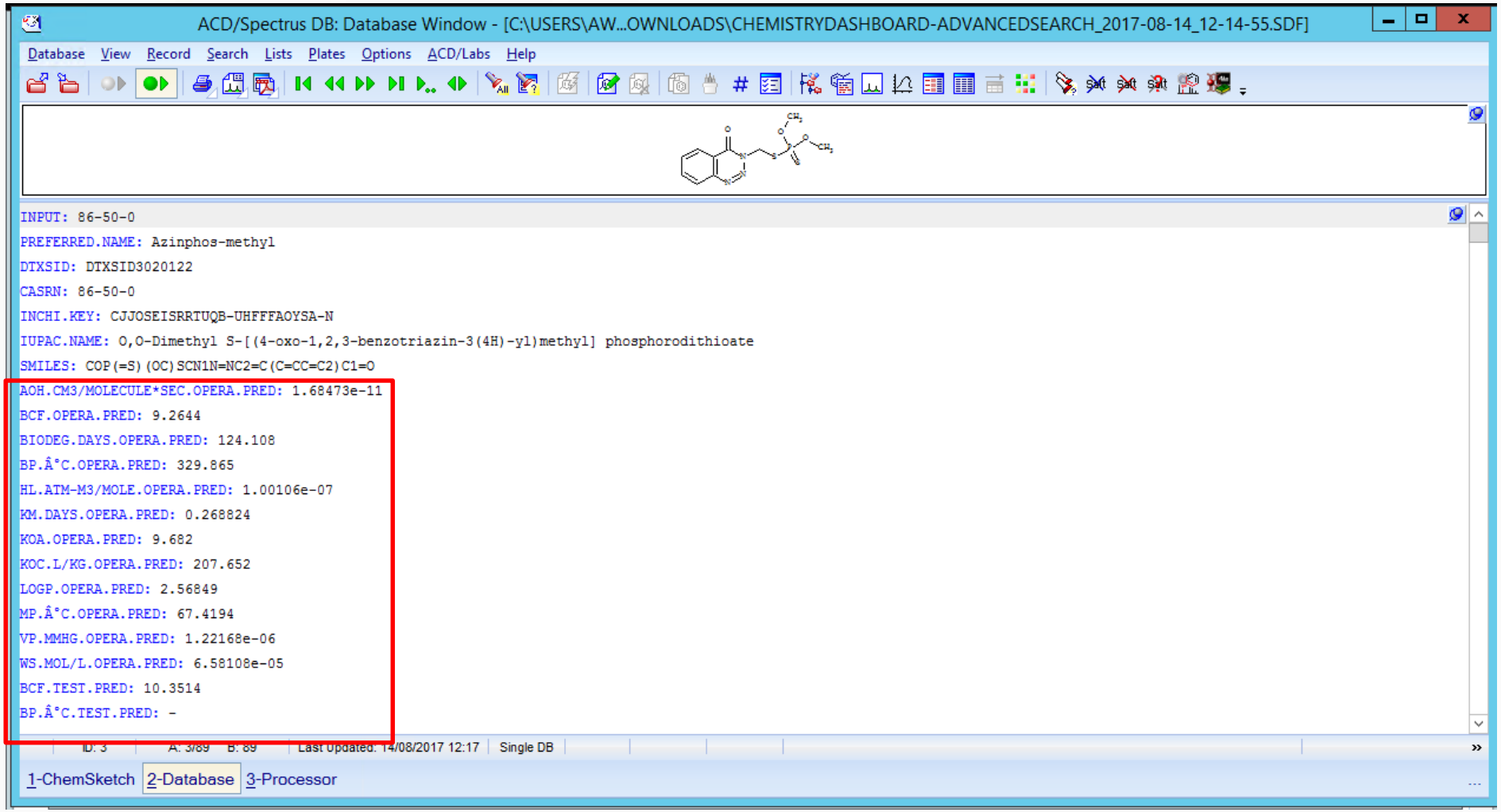
J10 27.0573

	A	B	C	D	E	F	G	H	I	J	K	L
1	INPUT	DTXSID	PREFERRE	CASRN	INCHI KEY	IUPAC NAME	SMILES	AOH CM3/MOLI	BCF OPERA	BIODEG DAYS	BP °C OPERA	HL ATM-M
2	143-50-0	DTXSID10	Kepone	143-50-0	LHHGDZSESBA	1,1a,3,3a,4,5,5,5	ClC12C(=O	1.55792e-12	1393.39	913.927	373.013	9.25708e-0
3	2385-85-5	DTXSID70	Mirex	2385-85-5	GVYLCNUFSHI	1,1a,2,2,3,3a,4,5	ClC12C3(Cl	1.56206e-13	15029.6	830.325	377.33	0.00066778
4	86-50-0	DTXSID30	Azinphos-m	86-50-0	CJJOSEISRRTU	O,O-Dimethyl S-[	COP(=S)(O	1.68473e-11	9.2644	124.108	329.865	1.00106e-0
5	56-38-2	DTXSID70	Parathion	56-38-2	LCCNCVORNK	O,O-Diethyl O-(4	CCOP(=S)(	1.59074e-11	113.717	4.30449	371.174	3.17774e-0
6	121-75-5	DTXSID40	Malathion	121-75-5	JXSJBGJIGXNW	Diethyl 2-[(dimeth	CCOC(=O)C	7.73562e-12	26.329	25.141	359.971	3.20069e-0
7	57-74-9	DTXSID70	Chlordane	57-74-9	BIWJNBZANLA	1,2,4,5,6,7,8,8-O	ClC1CC2C(	7.00208e-13	15500.6	106.951	328.828	0.00015982
8	91-94-1	DTXSID60	3,3'-Dichlor	91-94-1	HUWXDEQWW	3,3'-Dichloro[1,1'-	NC1=C(Cl)C	1.00886e-11	154.496	31.7767	365.753	4.07023e-0
9	86-30-6	DTXSID60	N-Nitrosodi	86-30-6	UBUCNCOMAD	N,N-Diphenylnitro	O=NN(C1=C	2.15654e-12	36.9494	4.54036	312.415	8.01245e-0
10	92-87-5	DTXSID20	Benzidine	92-87-5	HFACYLZERDE	[1,1'-Biphenyl]-4,4	NC1=CC=C	3.62519e-11	27.087	27.0573	380.642	1.06664e-0
11	122-66-7	DTXSID70	1,2-Dipheny	122-66-7	YBQZXXMEJHZ	1,2-Diphenylhydr	N(NC1=CC=	5.68102e-11	33.871	31.9084	323.674	6.67506e-0
12	470-90-6	DTXSID70	Chlorfenvinp	470-90-6	FSAVDKDHPDS	2-Chloro-1-(2,4-di	CCOP(=O)(	1.75121e-11	73.8391	3.8167	378.546	3.32968e-0
13	84-66-2	DTXSID70	Diethyl phth	84-66-2	FLKPEMZONW	Diethyl benzene-	CCOC(=O)C	8.98902e-12	7.15782	4.31107	294.998	2.38023e-0
14	134-62-3	DTXSID20	DEET	134-62-3	MMOXZBCLCQI	N,N-Diethyl-3-me	CCN(CC)C(	1.69398e-11	3.9293	3.35701	286.912	1.44818e-0
15	333-41-5	DTXSID90	Diazinon	333-41-5	FHIVAFMUCKR	O,O-Diethyl O-[6	CCOP(=S)(	1.75966e-11	69.9796	4.76627	340.43	1.44199e-0
16	101-14-4	DTXSID50	4,4'-Methyle	101-14-4	IBOFVQJTBBUI	4,4'-Methylenebis	NC1=C(Cl)C	1.41975e-11	180.611	31.9094	371.57	4.10724e-0

SDF Download



SDF Download



Real-Time Property Prediction

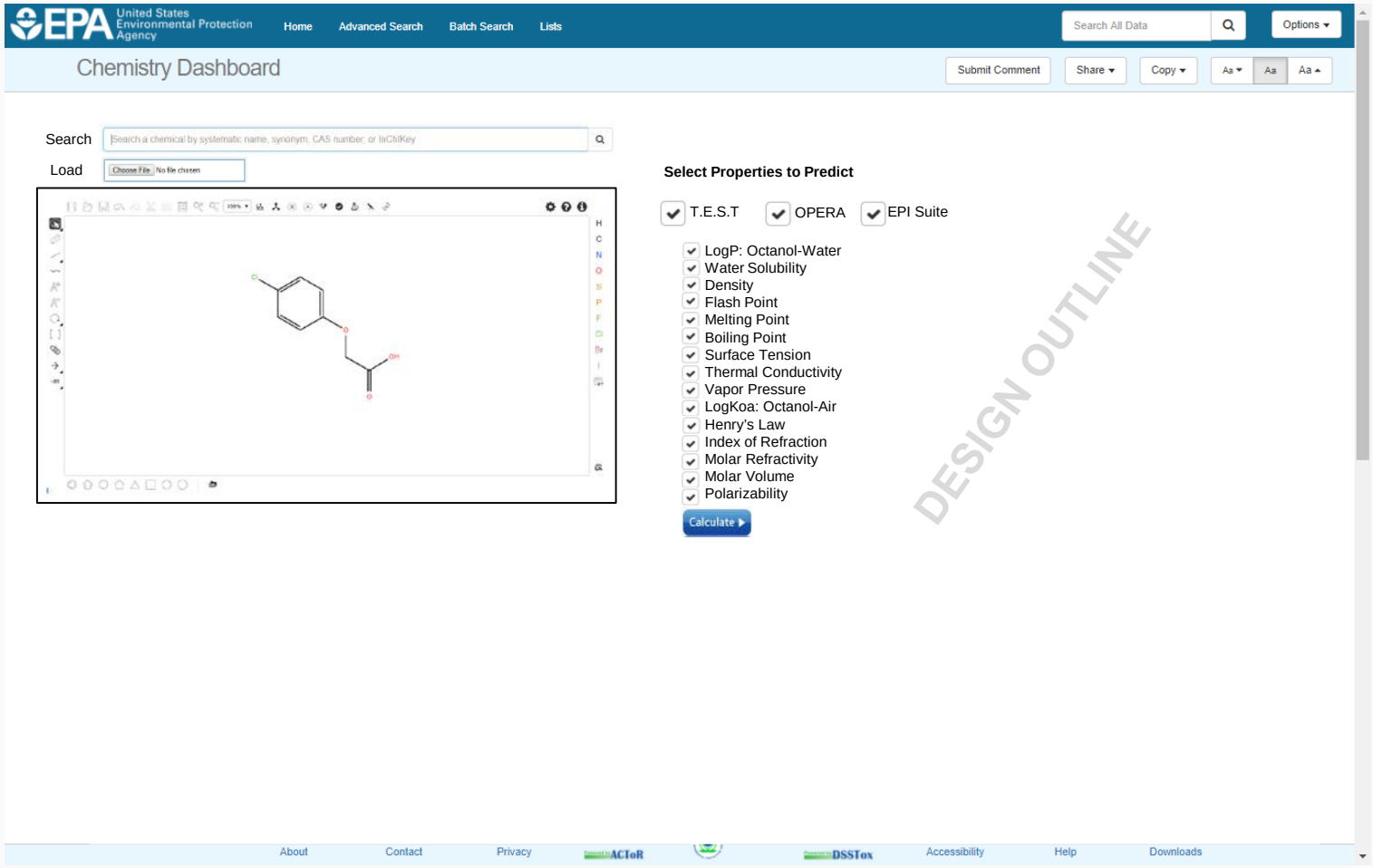
- We are registering chemicals everyday
- At present, prior to release, we do batch predictions of all properties. INEFFICIENT
- At registration we want to calculate all properties based on OPERA, T.E.S.T and EPI Suite models
- The best way to do this is using web service-based calculations

Now working on T.E.S.T services

- 96hr fathead minnow 50% lethal concentration (LC50)
- 48hrr daphnia magna 50% lethal concentration (LC50)
- Tetrahymena pyriformis 50% growth inhibition conc. (IGC50)
- Oral rat 50% lethal dose (LD50)
- Bioconcentration Factor (BCF)
- Developmental Toxicity (DevTox)
- Ames Mutagenicity (Mutagenicity)
- Normal boiling point, Flash point, Melting point
- Surface tension, Viscosity, Water Solubility
- Thermal Conductivity, Vapor Pressure, Density

- EXAMPLE: https://comptox.epa.gov/dashboard/web-test/WS?smiles=ClC(Cl)(Cl)Cl

Services to Support Real-Time Property Prediction



Future Work

- Provide “programmatic access” to all data – rollout API and web services

- Load test existing WebTEST public services
- Load test existing EPI Suite public services
- Create OPERA web services and make public

- Develop web input page for real time predictions

Conclusion

- The CompTox Chemistry Dashboard provides access to data for ~760,000 chemicals
- An **Integration Hub** to predicted data from different models (e.g. TEST, EPI Suite, OPERA)
- Data downloads allows for reuse in other systems and integration of resources to support research
- Real-time prediction based on developing web services is on the horizon

Acknowledgments

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