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

# Benefits of making data from the EPA National Center for Computational Toxicology available for reuse

*Antony Williams[1], Kamel Mansouri[2], Valery Tkachenko[3], Kirill Blinov[4] and Chris Grulke[1]*

*1. National Center for Computational Toxicology, U.S. Environmental Protection Agency, RTP, NC*
*2. Scitovation, RTP, NC; 3. Science Data Software, LLC, MD, USA; 4. Molecule Apps LLC, 97330, Corvallis, OR*

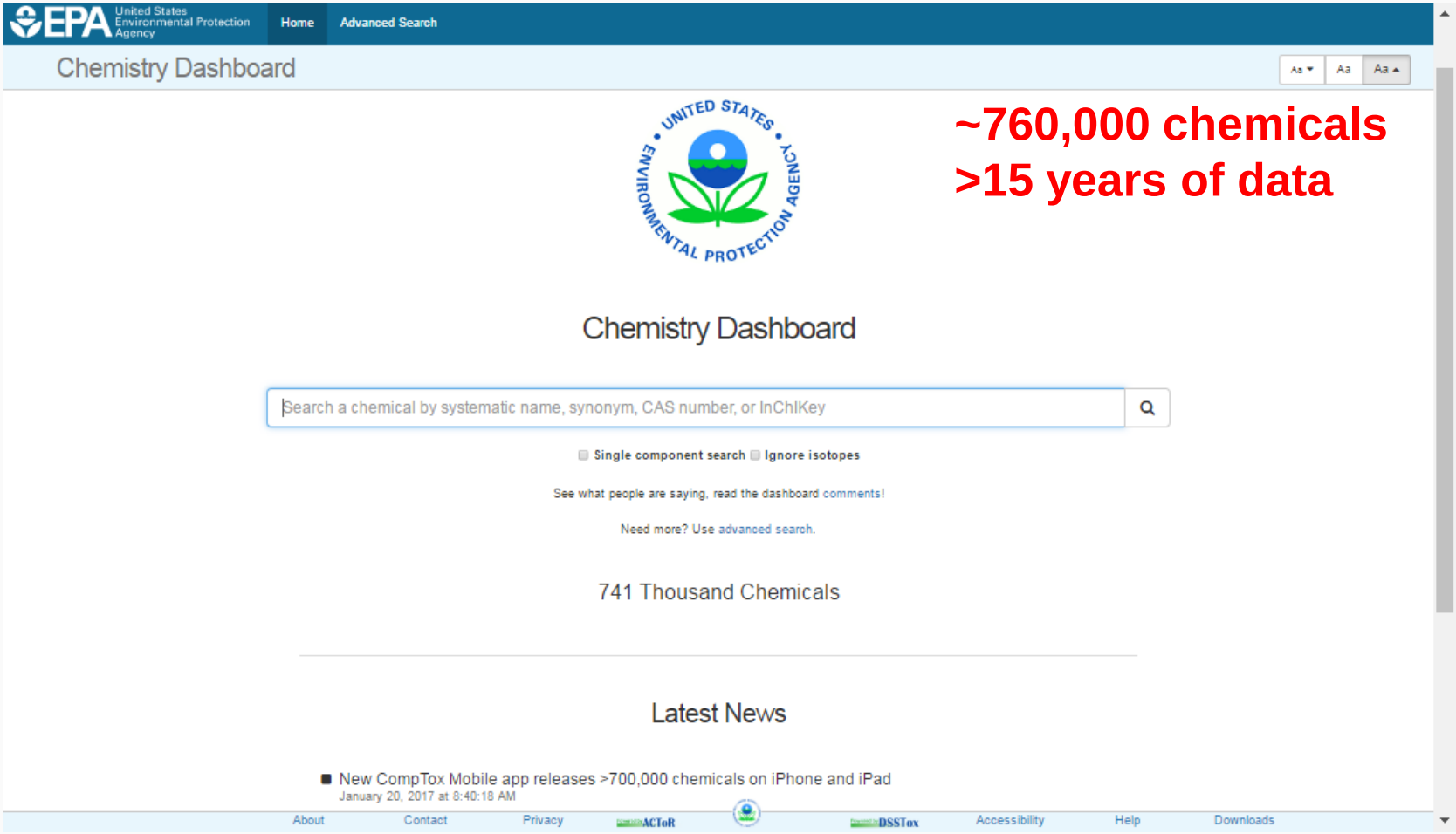


*August 2017*
*ACS Fall Meeting, Washington, DC*

# 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**

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

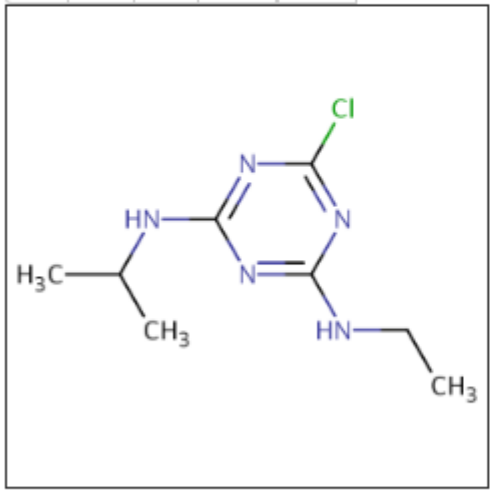


# Chemical Page

## 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

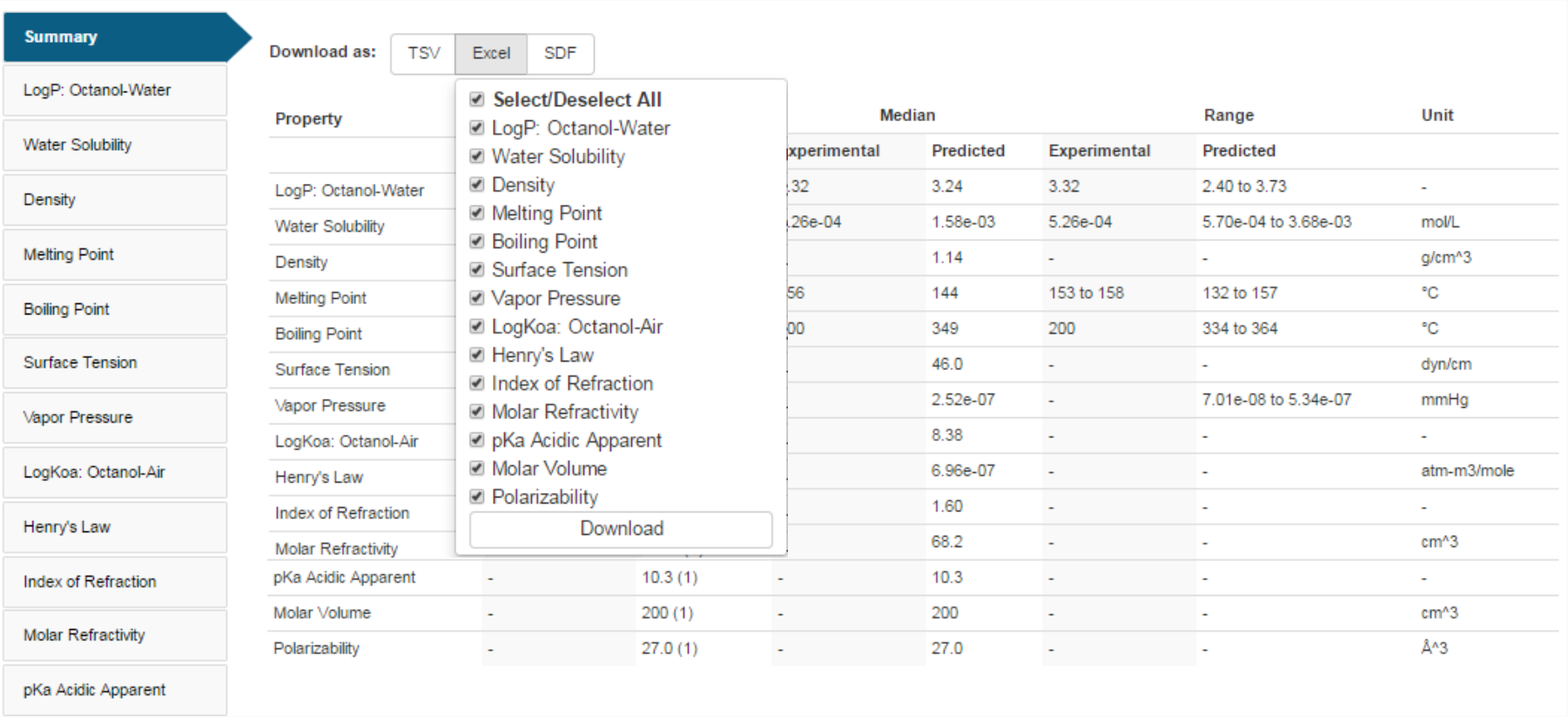
Structural Identifiers

Record Information

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

# Data Distribution

- Consuming and producing open data



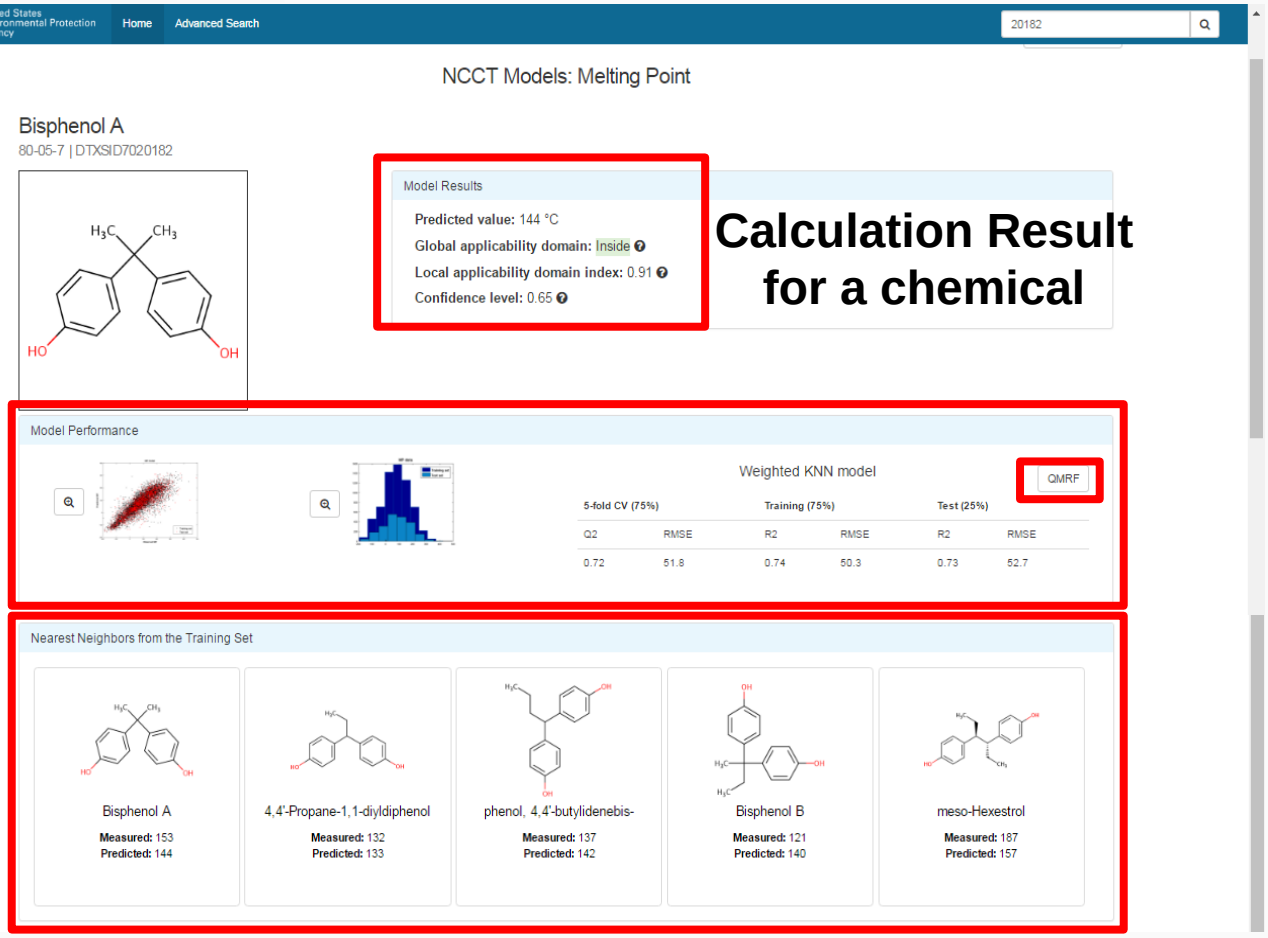
# Data Distribution

| | 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 | | | | | | | | |

# Modeling Details

Model Performance with full QMRF

Nearest Neighbors from Training Set



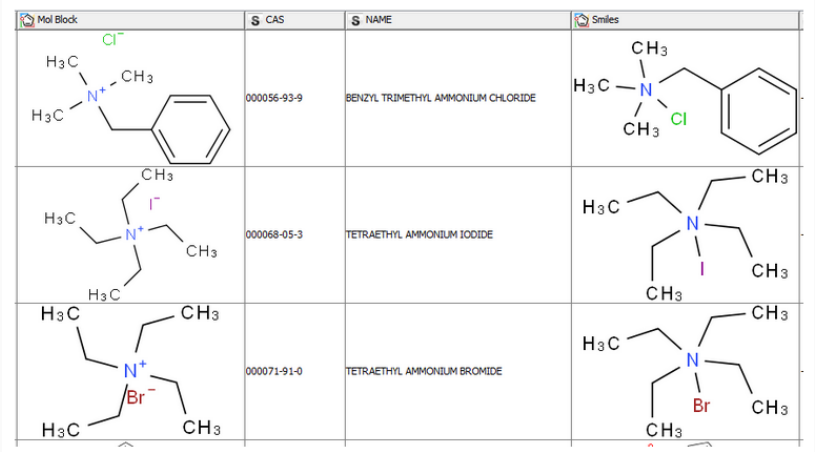
# Developing “NCCT Models”

- 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**

# Consuming and Curating Public Data

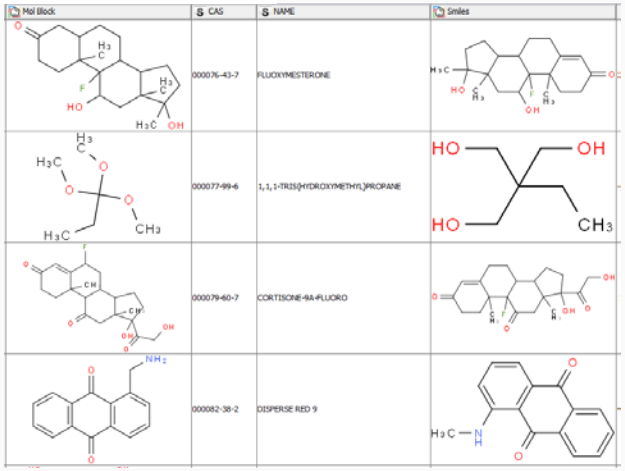
EPA United States Environmental Protection Agency

## 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

Journal
SAR and QSAR in Environmental Research ›
Volume 27, 2016 - Issue 11: 17th International Conference on QSAR in Environmental and Health Sciences (QSAR 2016) - Part II. Guest Editors: C.G. Barber and G.J. Myatt

Enter keywords, authors, DOI etc.

258
Views
4
CrossRef citations
16
Altmetric

Articles

**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

Pages 911-937 | Received 03 Sep 2016, Accepted 24 Oct 2016, Published online: 25 Nov 2016

Download citation http://dx.doi.org/10.1080/1062936X.2016.1253611 Check for updates

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

# 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

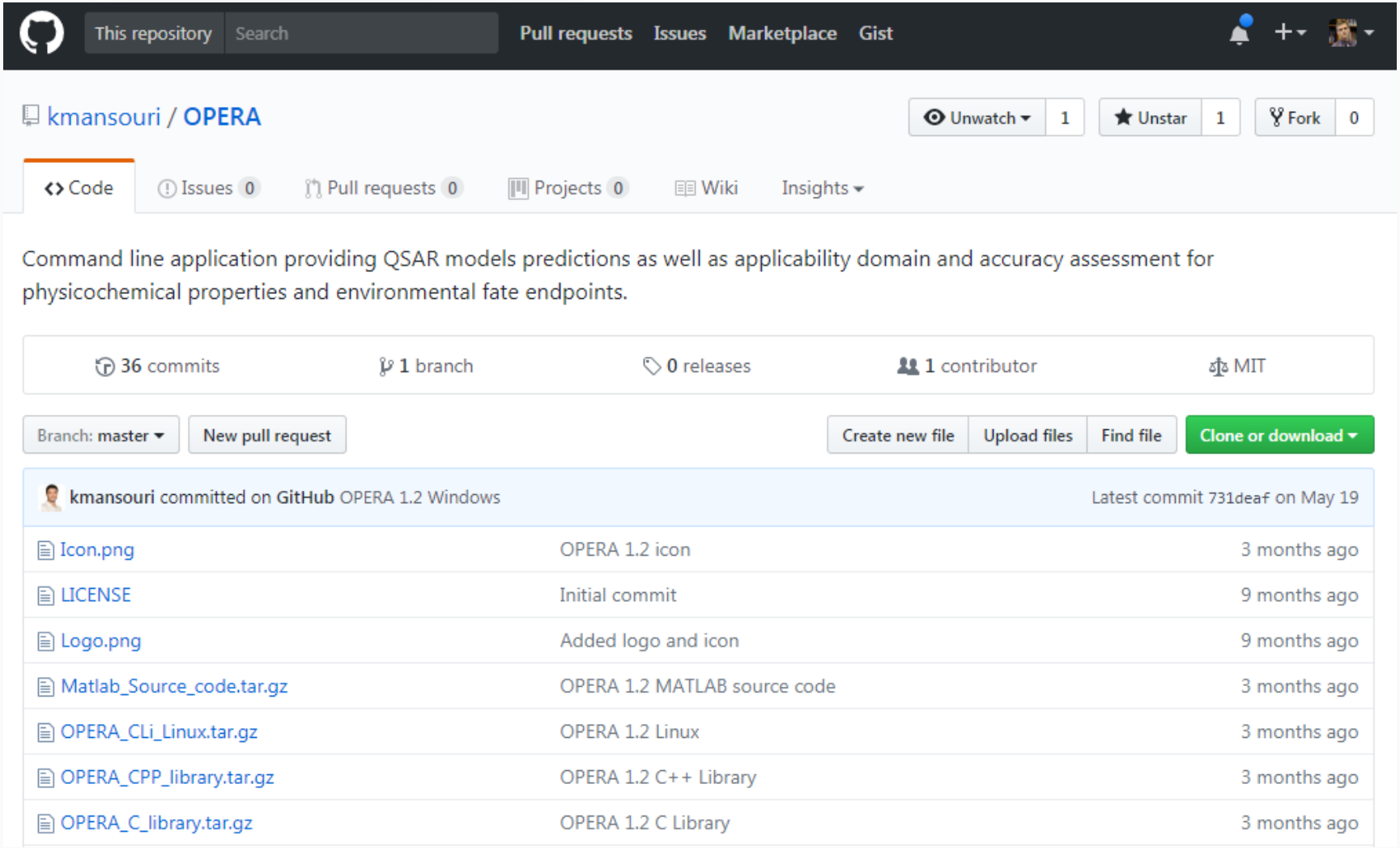
DSSTox Predicted Property Data — Posted: 12/14/2016

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:

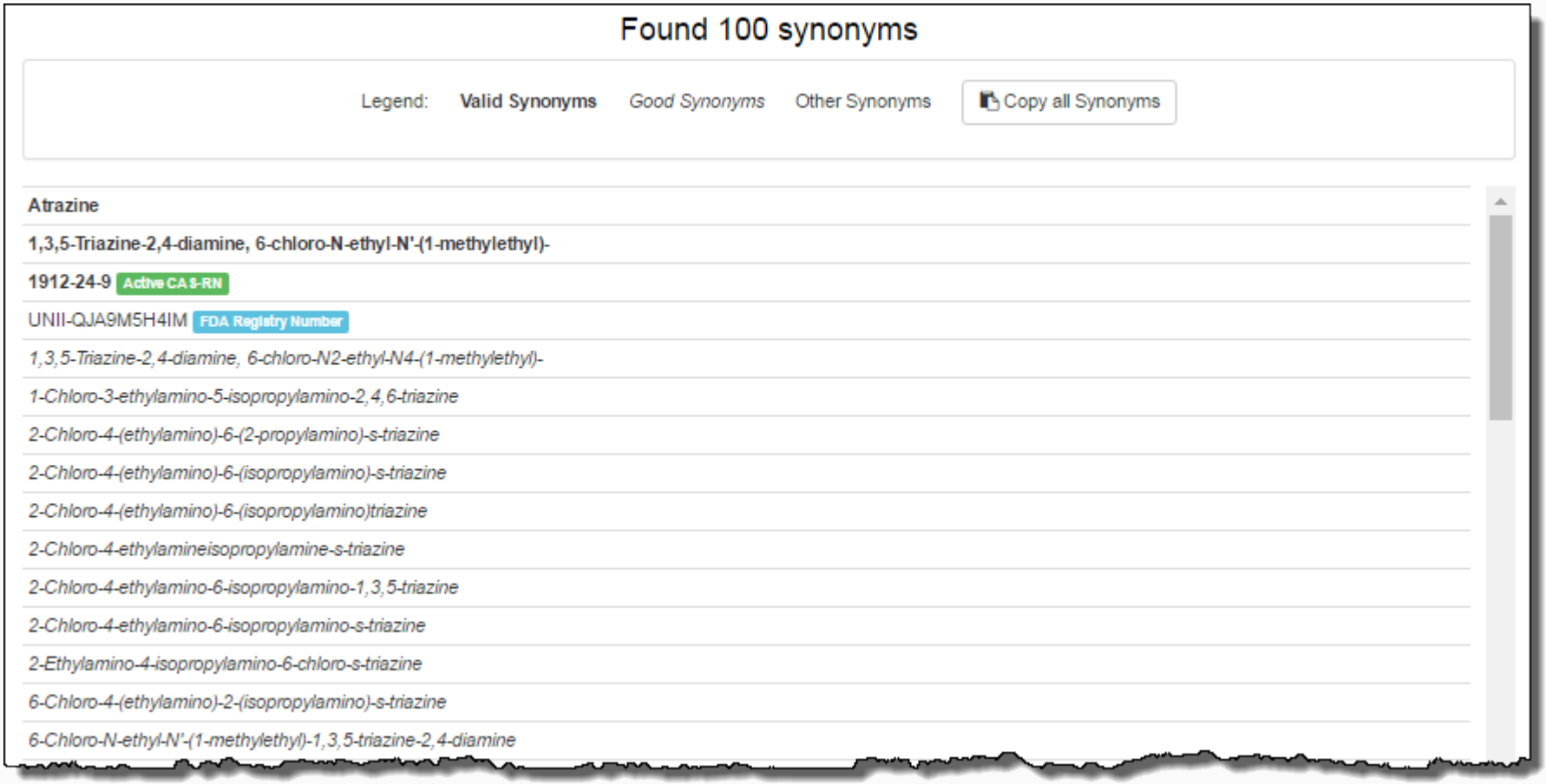
**Output**:

- 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

# Names and Identifiers

EPA United States Environmental Protection Agency



## Found 100 synonyms

Legend: **Valid Synonyms** *Good Synonyms* Other Synonyms Copy all Synonyms

**Atrazine**

**1,3,5-Triazine-2,4-diamine, 6-chloro-N-ethyl-N'-(1-methylethyl)-**

**1912-24-9** Active CAS-RN

UNII-QJA9M5H4IM FDA Registry Number

*1,3,5-Triazine-2,4-diamine, 6-chloro-N2-ethyl-N4-(1-methylethyl)-*

*1-Chloro-3-ethylamino-5-isopropylamino-2,4,6-triazine*

*2-Chloro-4-(ethylamino)-6-(2-propylamino)-s-triazine*

*2-Chloro-4-(ethylamino)-6-(isopropylamino)-s-triazine*

*2-Chloro-4-(ethylamino)-6-(isopropylamino)triazine*

*2-Chloro-4-ethylamineisopropylamine-s-triazine*

*2-Chloro-4-ethylamino-6-isopropylamino-1,3,5-triazine*

*2-Chloro-4-ethylamino-6-isopropylamino-s-triazine*

*2-Ethylamino-4-isopropylamino-6-chloro-s-triazine*

*6-Chloro-4-(ethylamino)-2-(isopropylamino)-s-triazine*

*6-Chloro-N-ethyl-N'-(1-methylethyl)-1,3,5-triazine-2,4-diamine*

# Downloadable Data File

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

EPA United States Environmental Protection Agency

## 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

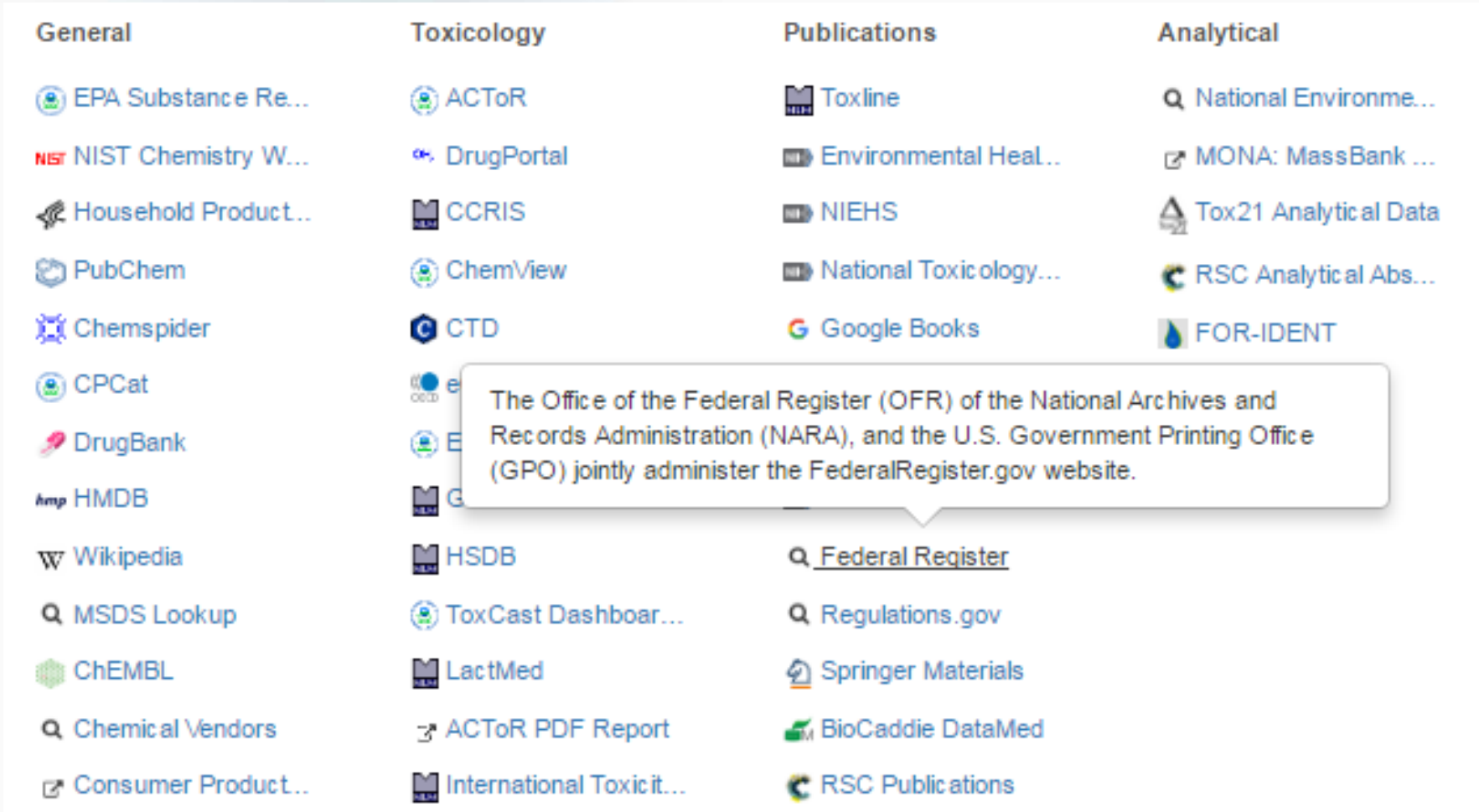
DSSTox MS Ready Mapping File — Posted: 11/14/2016

The CompTox Chemistry Dashboard can be used by mass spectrometrists for the purpose of structure identification. A normal formula search would search the exact formula associated with any chemical, whether it include solvents of hydration, salts or multiple components. However, mass spectrometry detects ionized chemical structures and molecular formulae searches should be based on desalted, and desolvated structures with stereochemistry removed. We refer to these as "MS ready structures" and the MS-ready mappings are delivered as Excel Spreadsheets containing the Preferred Name, CAS-RN. DTXSID, Formula, Formula of the MS-ready structure and associated masses, SMILES and InChI Strings/Keys.
image

DSSTox SDF File — Posted: 12/14/2016

This zip file contains the entire chemical structure collection of over 700,000 chemicals from the DSSTox database contained in one large SDF file. The file contains the structure, The DSSTox Structure Identifier (DTXCID), The DSSTOX Substance Identifier (DTXSID listed as PubChem External Data Source), the associated Dashboard URL, associated synonyms and Quality Control Level details. 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.
image

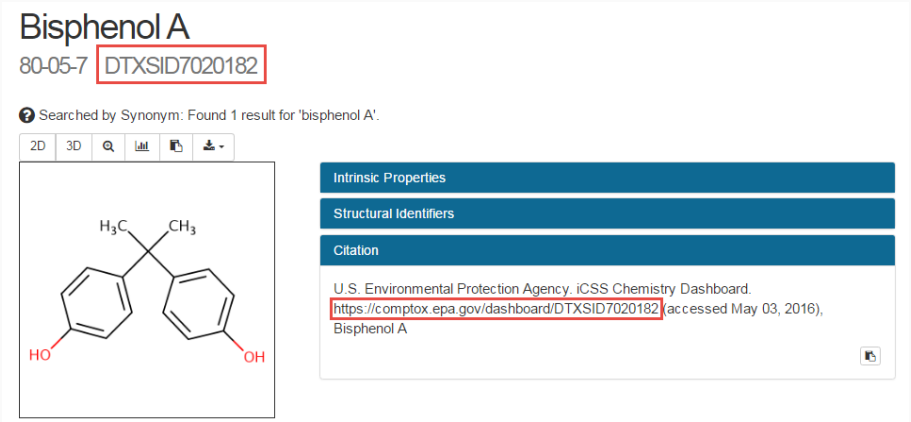
# Links to Other Resources




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

# In-Links into the Dashboard

- Linkages into the Dashboard are simple: using the associated identifiers




- For integration we can supply files of structures and identifiers mapped to DTXSIDs. Contact us…
- PubChem, EBI's UNICHEM, ChemSpider, etc.

# In-linking Resources

EPA United States Environmental Protection Agency

NIH NLM NCBI National Center for Biotechnology Information

PubChem Open Chemistry Database

HOME SERVICES SUBMIT ABOUT CONTACT

## EPA DSSTox

PUBCHEM > DATA SOURCES > EPA DSSTOX

| | |
|---|---|
| **Organization:** | National Center for Computational Toxicology (NCCT) Office of Research and Development US Environmen |
| **Category:** | Research and Development, Governmental Organizations, Curation Efforts |
| **URL:** | https://www.epa.gov/chemical-research/distributed-structure-searchable-toxicity-dsstox-database |

ChemSpider
Search and share chemistry

Simple Structure Advanced History

## Data source details

* **Name** EPA DSSTox
**Primary user** None
**Contact name** Antony Williams
**Secondary users**

None

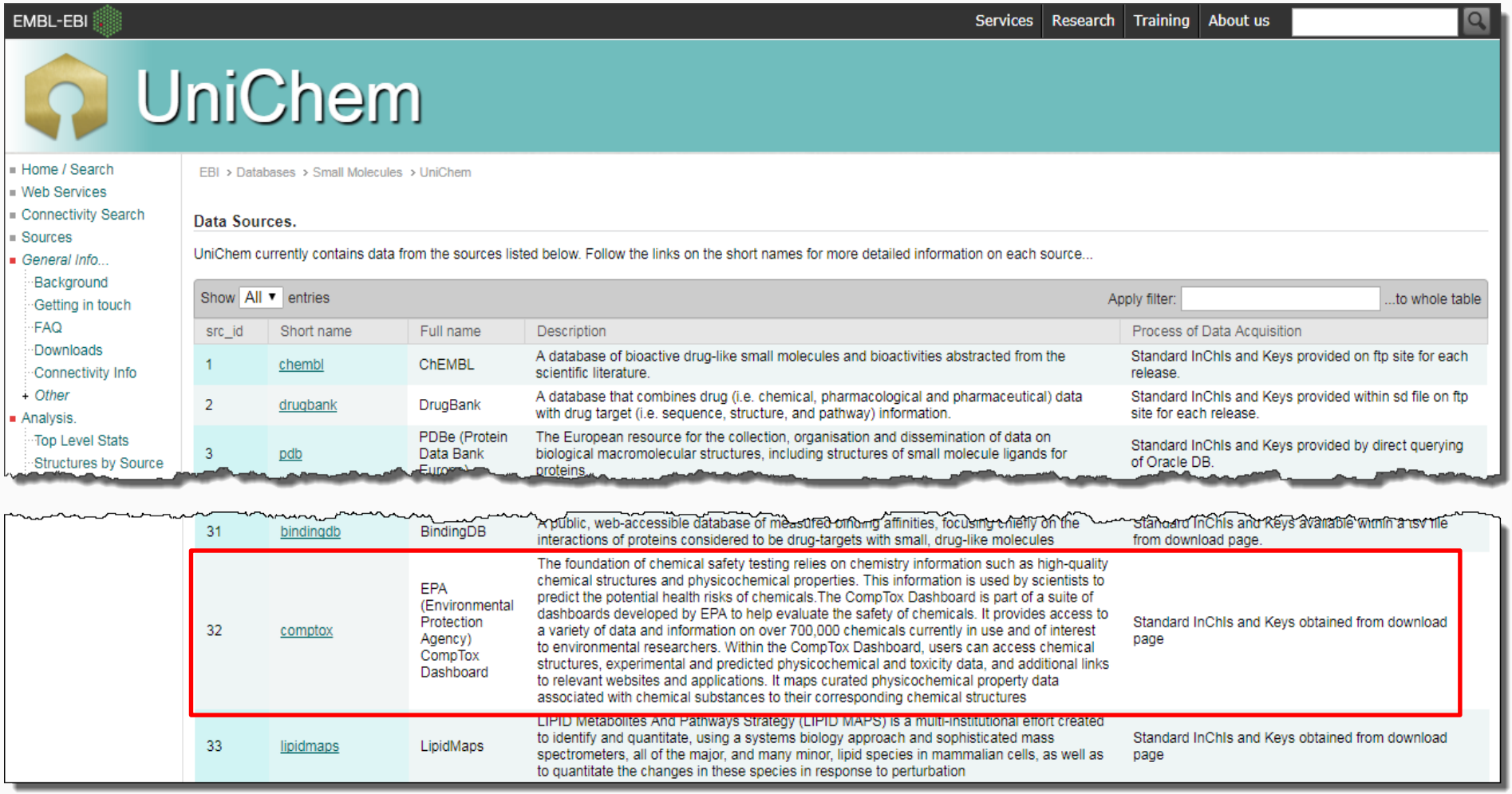
**Description**
The foundation of chemical safety testing relies on chemistry information such as high-quality chemical structures and physicochemical properties. This information is used by scientists to predict the potential health risks of chemicals.The CompTox Dashboard is part of a suite of dashboards developed by EPA to help evaluate the safety of chemicals. It provides access to a variety of data and information on over 700,000 chemicals currently in use and of interest to environmental researchers. Within the CompTox Dashboard, users can access chemical structures, experimental and predicted physicochemical and toxicity data, and additional links to relevant websites and applications. It maps curated physicochemical property data associated with chemical substances to their corresponding chemical structures.

**Data collections** None

**Company name** National Center for Computational Toxicology (NCCT)

# In-linking Resources

EPA United States Environmental Protection Agency



EMBL-EBI | Services | Research | Training | About us

UniChem

- Home / Search
- Web Services
- Connectivity Search
- Sources
- *General Info...*
  - Background
  - Getting in touch
  - FAQ
  - Downloads
  - Connectivity Info
  - *Other*
- Analysis.
  - Top Level Stats
  - Structures by Source

EBI > Databases > Small Molecules > UniChem

**Data Sources.**

UniChem currently contains data from the sources listed below. Follow the links on the short names for more detailed information on each source...

Show All entries | Apply filter: ...to whole table

| src_id | Short name | Full name | Description | Process of Data Acquisition |
|---|---|---|---|---|
| 1 | chembl | ChEMBL | A database of bioactive drug-like small molecules and bioactivities abstracted from the scientific literature. | Standard InChIs and Keys provided on ftp site for each release. |
| 2 | drugbank | DrugBank | A database that combines drug (i.e. chemical, pharmacological and pharmaceutical) data with drug target (i.e. sequence, structure, and pathway) information. | Standard InChIs and Keys provided within sd file on ftp site for each release. |
| 3 | pdb | PDBe (Protein Data Bank Europe) | The European resource for the collection, organisation and dissemination of data on biological macromolecular structures, including structures of small molecule ligands for proteins. | Standard InChIs and Keys provided by direct querying of Oracle DB. |
| 31 | bindingdb | BindingDB | A public, web-accessible database of measured binding affinities, focusing chiefly on the interactions of proteins considered to be drug-targets with small, drug-like molecules | Standard InChIs and Keys available within a tsv file from download page. |
| 32 | comptox | EPA (Environmental Protection Agency) CompTox Dashboard | The foundation of chemical safety testing relies on chemistry information such as high-quality chemical structures and physicochemical properties. This information is used by scientists to predict the potential health risks of chemicals.The CompTox Dashboard is part of a suite of dashboards developed by EPA to help evaluate the safety of chemicals. It provides access to a variety of data and information on over 700,000 chemicals currently in use and of interest to environmental researchers. Within the CompTox Dashboard, users can access chemical structures, experimental and predicted physicochemical and toxicity data, and additional links to relevant websites and applications. It maps curated physicochemical property data associated with chemical substances to their corresponding chemical structures | Standard InChIs and Keys obtained from download page |
| 33 | lipidmaps | LipidMaps | LIPID Metabolites And Pathways Strategy (LIPID MAPS) is a multi-institutional effort created to identify and quantitate, using a systems biology approach and sophisticated mass spectrometers, all of the major, and many minor, lipid species in mammalian cells, as well as to quantitate the changes in these species in response to perturbation | Standard InChIs and Keys obtained from download page |

# DTXSIDs on WikiData

English Not logged in Talk Contributions Create account Log in

WIKIDATA

Project page Discussion Read Edit View history Search Wikidata

## Wikidata:Property proposal/DTXSID

< Wikidata:Property proposal

Main page
Community portal
Project chat
Create a new item
Recent changes
Random item
Query Service
Nearby
Help
Donate

Print/export
Create a book
Download as PDF
Printable version

Tools
What links here
Related changes
Special pages
Permanent link
Page information

### DSSTOX substance identifier [ edit ]

Originally proposed at Wikidata:Property proposal/Natural science

Done. DSSTOX substance identifier (P3117) (Talk and documentation)

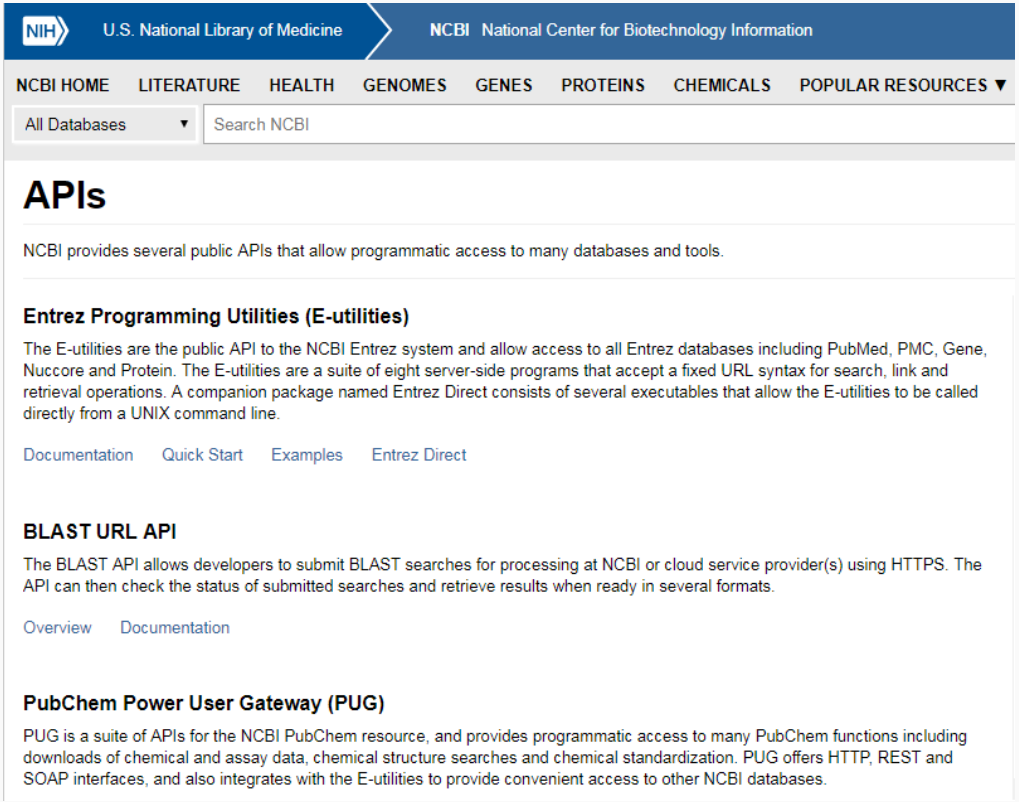
| | |
|---|---|
| **Description** | Ths DTXSID is the DSSTox substance identifier used in the Environmental Protection Agency CompTox Dashboard. |
| **Data type** | External identifier |
| **Domain** | chemical substance (Q79529) |
| **Example** | benzene (Q2270) → DTXSID3039242 |
| **Planned use** | Adding DTXSIDs to @wikidata using a bot or possible Mix&Match, based on InChIKey matches and this CCZero data on Figshare: https://figshare.com/articles/Mapping_file_of_InChIStrings_InChIKeys_and_DTXSIDs_for_the_EPA_CompTox_Dashboard/3578313/1 . |
| **Formatter URL** | https://comptox.epa.gov/dashboard/**$1** |

**Motivation**

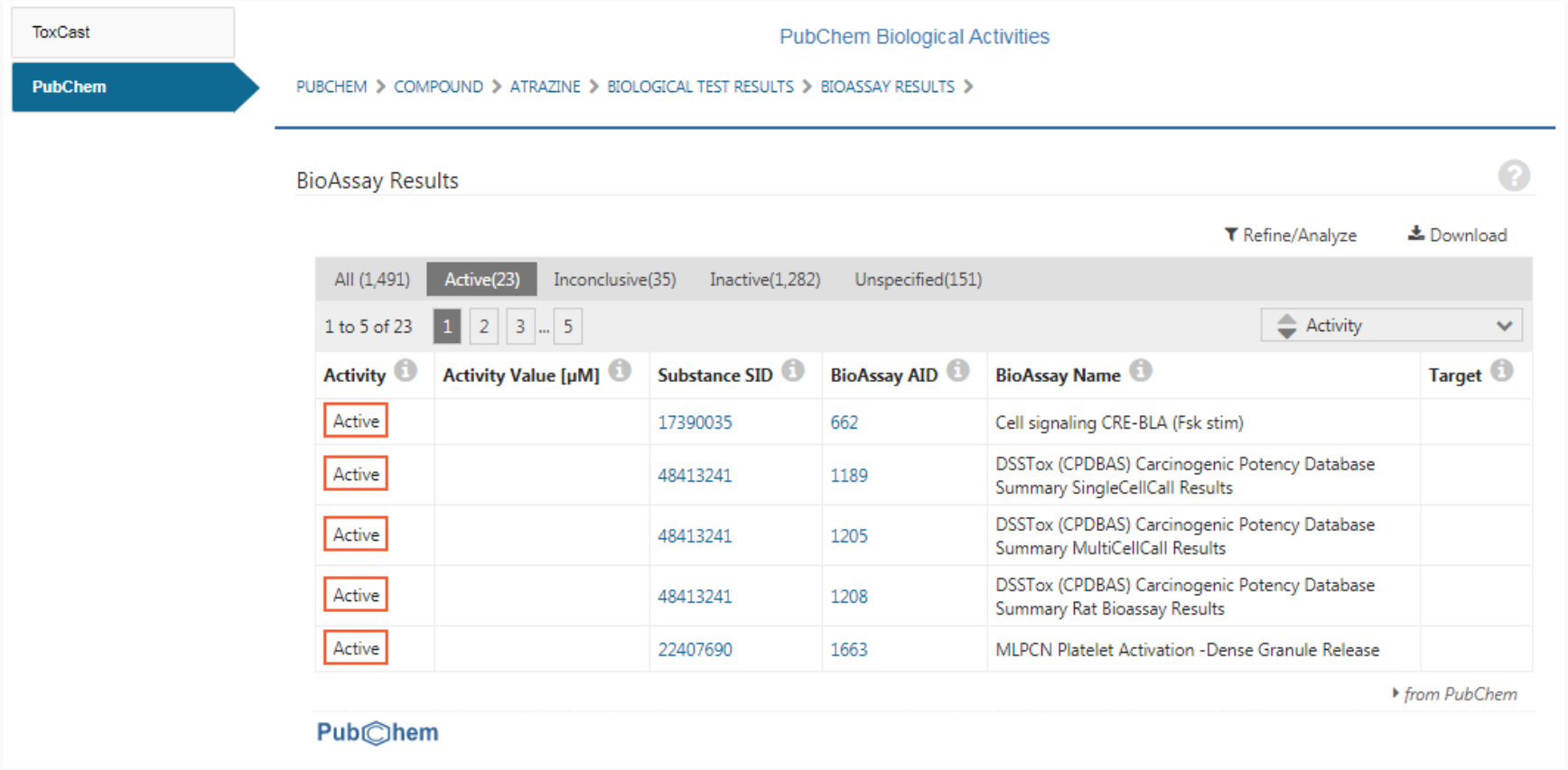
The Environmental Protection Agency (Q16919823)'s CompTox Dashboard is a fairly new website, but the DSSTox project exists for much longer (Distributed structure-searchable toxicity (DSSTox) public database network: a proposal (Q26701395)). The Dashboard aggregates over 720 thousand chemical substances and is open data. The CompTox dashboard also provides access to synonyms, experimental and predicted property data, product

# Open APIS are of Value!

- Open APIS provide valuable integrations
  - PubChem API and widgets
  - NCBI Literature

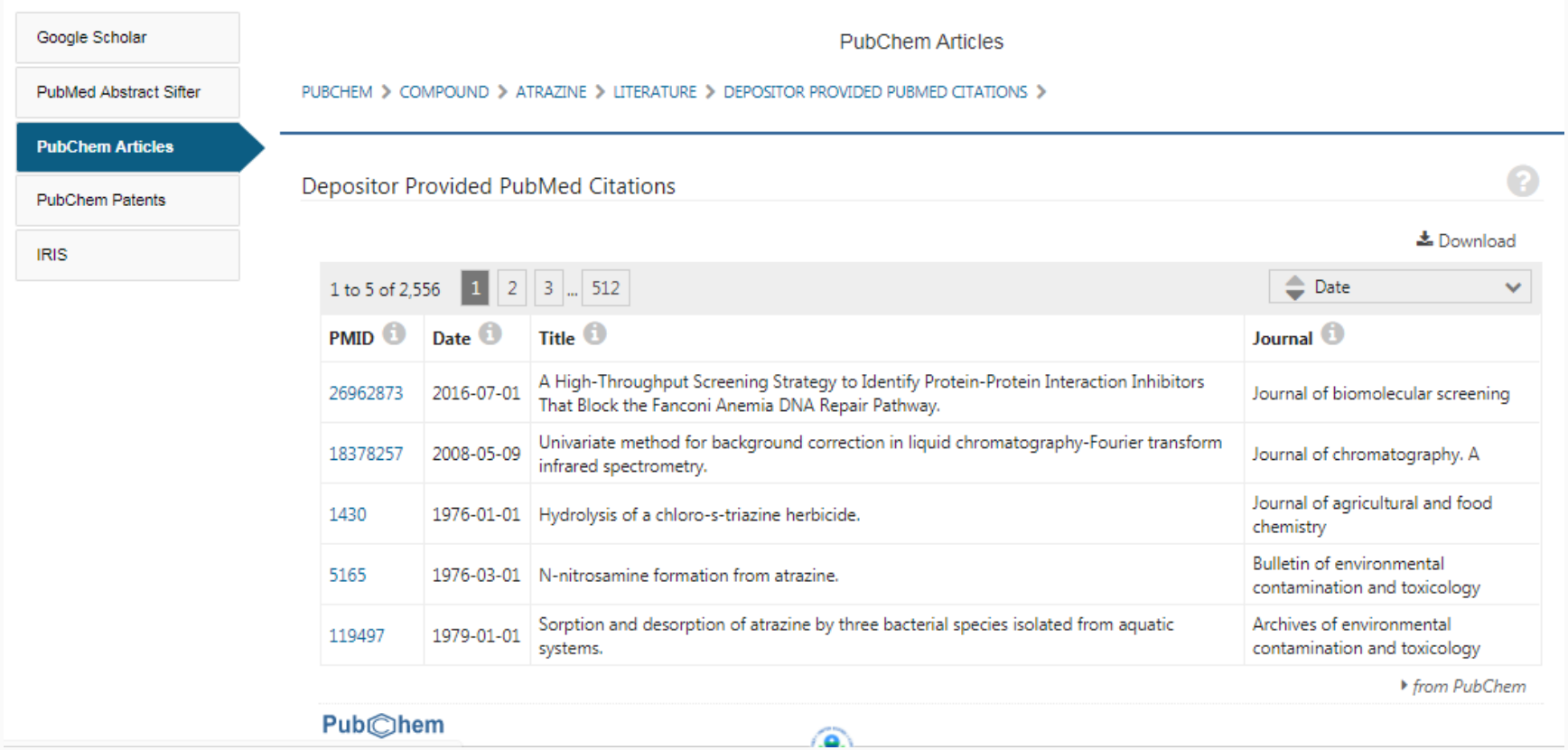


# PubChem Bioassay Data Widget

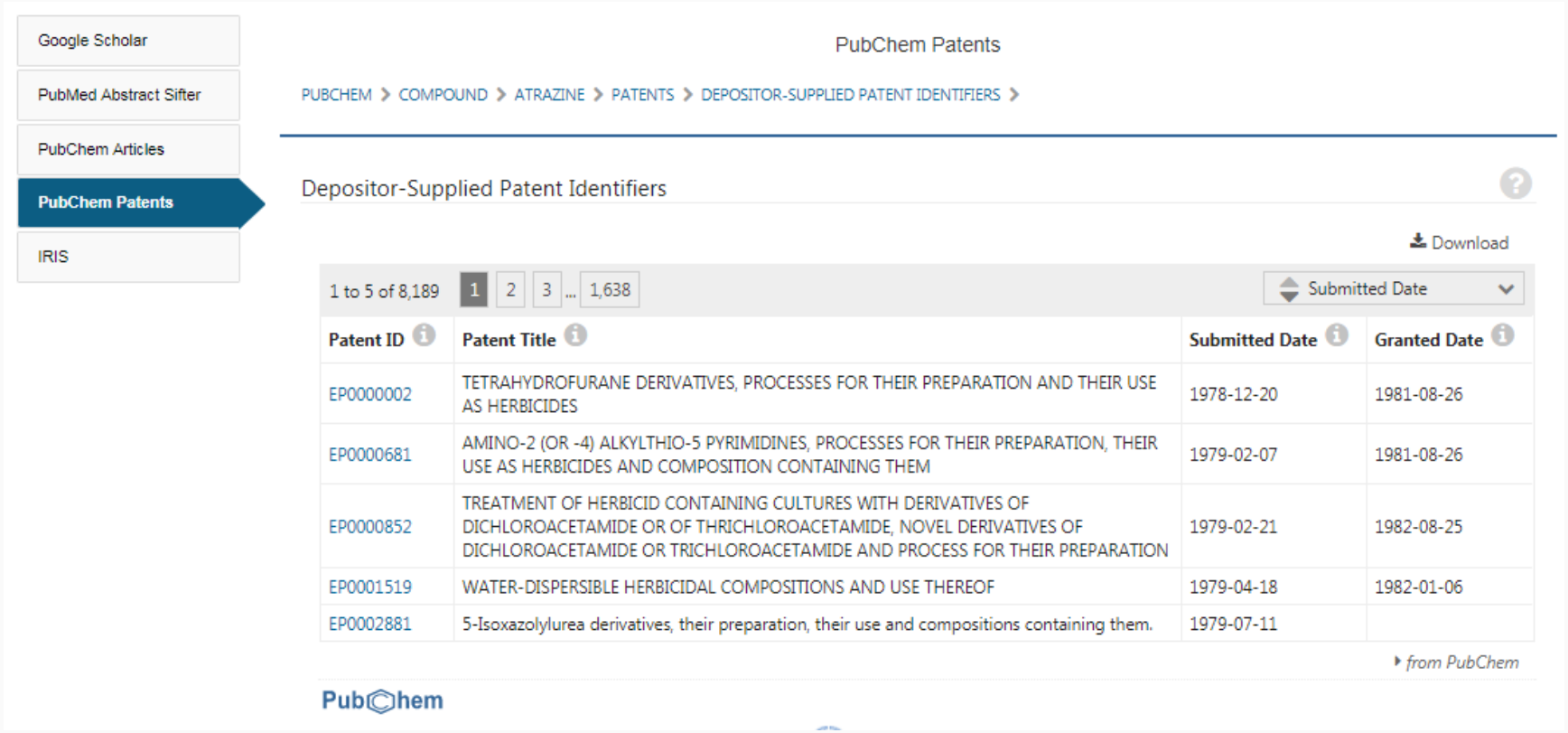



Chemical Properties | Synonyms | External Links | Product Composition | **Bioassays** | Exposure | Analytical | Literature | Comments

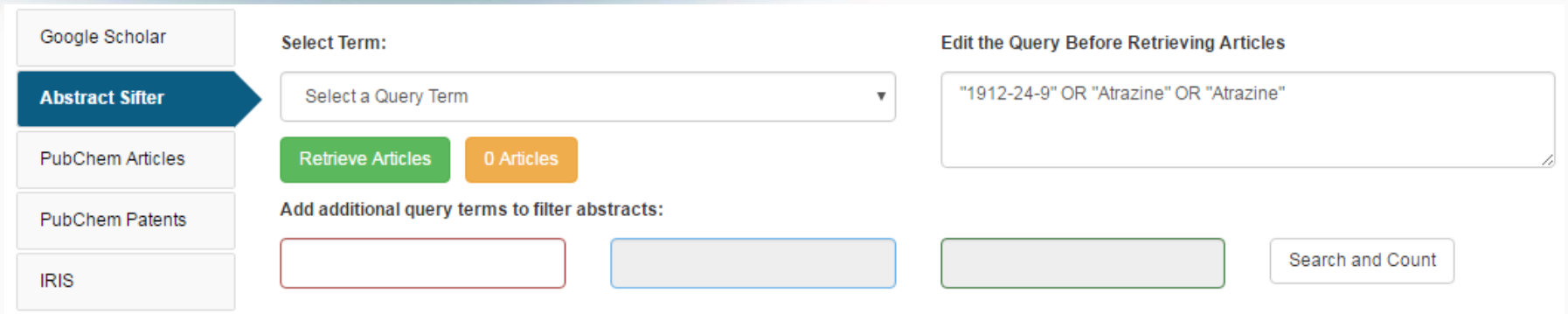
# PubChem Articles Widget

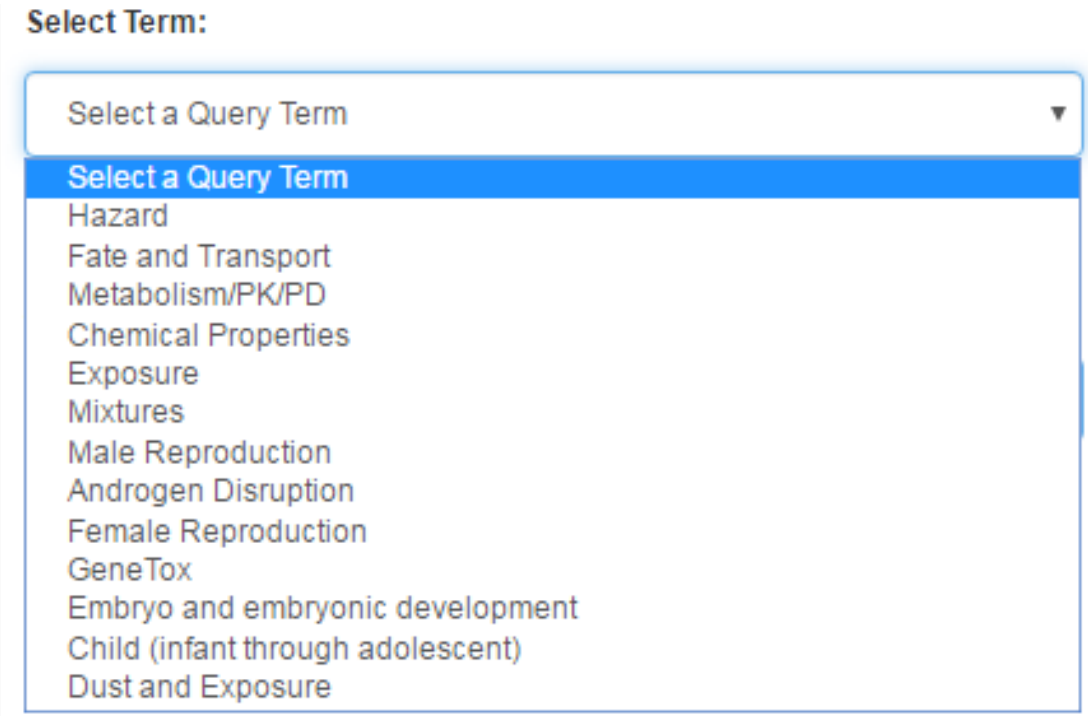


# PubChem Patents Widget



# PubMed Literature Searching




# Integrated Literature Searching

EPA United States Environmental Protection Agency

**Select Term:**

Hazard

Retrieve Articles | 28 Articles (out of 28)

**Edit the Query Before Retrieving Articles**

("1912-24-9" OR "Atrazine" OR "Atrazine") AND (NOAEL or NOEL OR LOEL or Rfd OR "reference dose" OR "reference concentration" OR "adverse effect level" [tiab] OR "cancer slope factor"[tiab])

**Add additional query terms to filter abstracts:**

estrogen | endocrine | | Search and Count

| est... | end... | Ter... | Total | PMID | Pub.. | Title |
|---|---|---|---|---|---|---|
| 3 | 1 | 0 | 4 | 10613397 | 1999 | Estrous cycle patterns of Sprague-Dawley rats during acute and chronic atrazine administration. |
| 2 | 1 | 0 | 3 | 15655806 | 2005 | A risk assessment of atrazine use in California: human health and ecological aspects. |
| 1 | 1 | 0 | 2 | 15590006 | 2004 | Impact of endocrine toxicants on survival, development, and reproduction of the estuarine copepod Eurytemora affinis... |
| 0 | 1 | 0 | 1 | 8829248 | 1996 | Effect of atrazine on ovarian function in the rat. |
| 0 | 0 | 0 | 0 | 28089211 | 2017 | Toxicity effects of an environmental realistic herbicide mixture on the seagrass Zostera noltei. |

Record: 1 of 28

**Title:** Effect of atrazine on ovarian function in the rat.

**Abstract:** The effect of the chlorotriazine herbicide, atrazine, on ovarian function was studied in Long-Evans hooded (LE-hooded) and Sprague-Dawley (SD) rats. Atrazine was administered by gavage for 21 d to females displaying regular 4-d estrous cycles. In both strains, 75 mg/kg/d disrupted the 4-d ovarian cycle; however, no distinct alteration (i.e., irregular cycles but not persistent estrus or diestrus) was apparent at this dose. At 150 mg/kg/d, atrazine induced repetitive pseudopregnancies in females of both strains. The highest dose tested (300 mg/kg/d) also induced repetitive pseudopregnancies in the SD females, while the ovaries of the LE-hooded female appeared regressed and the smear cytology was indicative of the anestrous condition. Although a NOAEL was not established, the doses employed in this experiment were in excess of those used in chronic feeding studies in which an early onset of mammary gland tumors was noted. These data demonstrate that atrazine can disrupt ovarian function and bring about major changes in the endocrine profile of the female.

# Integrated Literature Searching

EPA United States Environmental Protection Agency

Select Term:

Hazard

Retrieve Articles | 28 Articles (o

Add additional query terms to filter

estrogen

## Edit the Query Before Retrieving Articles

("1912-24-9" OR "Atrazine" OR "Atrazine") AND (NOAEL or NOEL OR LOEL or Rfd OR "reference dose" OR "reference concentration" OR "adverse effect level"[tiab] OR "cancer slope factor"[tiab])

| est... | end... | Ter... | Total | PMID | Pub.. | Title |
|---|---|---|---|---|---|---|
| 3 | 1 | 0 | 4 | 10613397 | 1999 | Estrous cycle patterns of Sprague-Dawley rats during acute and chronic atrazine administration. |
| 2 | 1 | 0 | 3 | 15655806 | 2005 | A risk assessment of atrazine use in California: human health and ecological aspects. |
| 1 | 1 | 0 | 2 | 15590006 | 2004 | Impact of endocrine toxicants on survival, development, and reproduction of the estuarine copepod Eurytemora affinis... |
| 0 | 1 | 0 | 1 | 8829248 | 1996 | Effect of atrazine on ovarian function in the rat. |
| 0 | 0 | 0 | 0 | 28089211 | 2017 | Toxicity effects of an environmental realistic herbicide mixture on the seagrass Zostera noltei. |

Record: 1 of 28

**Title:** Effect of atrazine on ovarian function in the rat.

**Abstract:** The effect of the chlorotriazine herbicide, atrazine, on ovarian function was studied in Long-Evans hooded (LE-hooded) and Sprague-Dawley (SD) rats. Atrazine was administered by gavage for 21 d to females displaying regular 4-d estrous cycles. In both strains, 75 mg/kg/d disrupted the 4-d ovarian cycle; however, no distinct alteration (i.e., irregular cycles but not persistent estrus or diestrus) was apparent at this dose. At 150 mg/kg/d, atrazine induced repetitive pseudopregnancies in females of both strains. The highest dose tested (300 mg/kg/d) also induced repetitive pseudopregnancies in the SD females, while the ovaries of the LE-hooded female appeared regressed and the smear cytology was indicative of the anestrous condition. Although a NOAEL was not established, the doses employed in this experiment were in excess of those used in chronic feeding studies in which an early onset of mammary gland tumors was noted. These data demonstrate that atrazine can disrupt ovarian function and bring about major changes in the endocrine profile of the female.

# Downloadable Data File

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

EPA United States Environmental Protection Agency

## 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

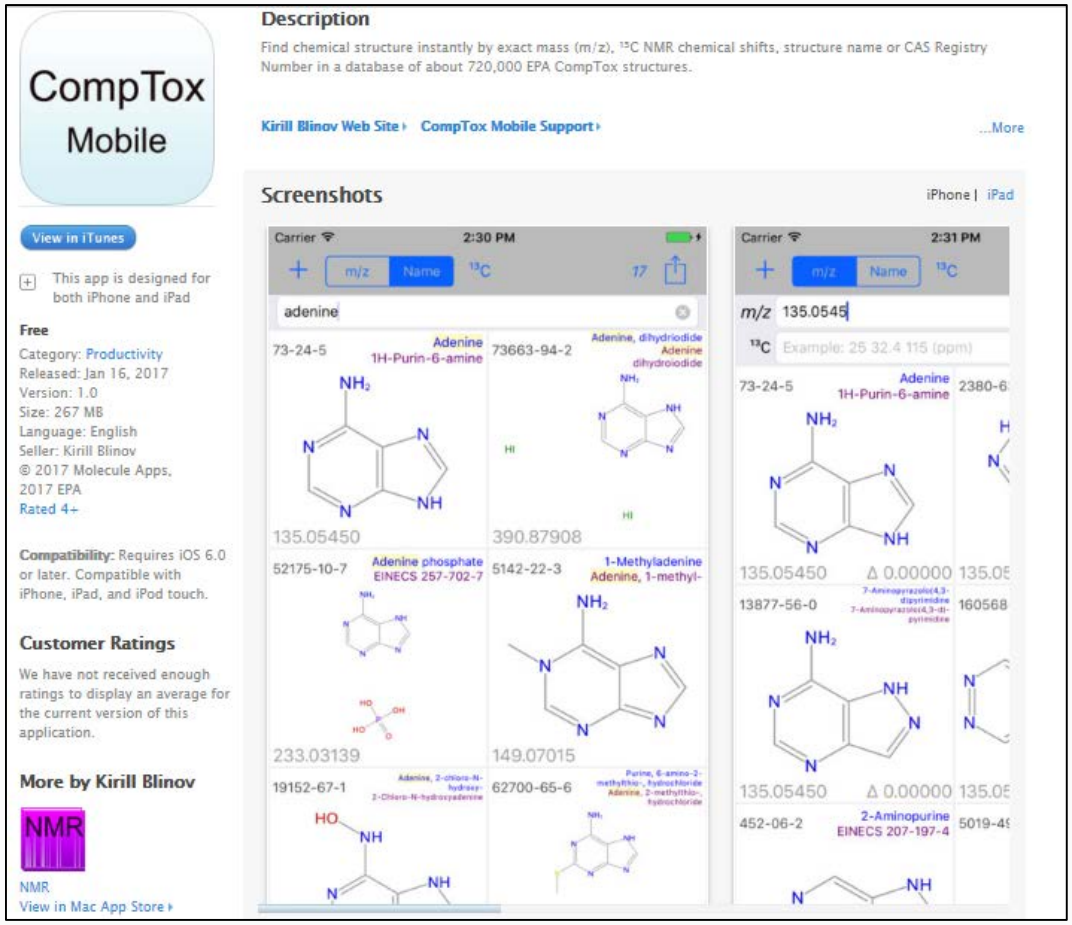
DSSTox MS Ready Mapping File — Posted: 11/14/2016

The CompTox Chemistry Dashboard can be used by mass spectrometrists for the purpose of structure identification. A normal formula search would search the exact formula associated with any chemical, whether it include solvents of hydration, salts or multiple components. However, mass spectrometry detects ionized chemical structures and molecular formulae searches should be based on desalted, and desolvated structures with stereochemistry removed. We refer to these as "MS ready structures" and the MS-ready mappings are delivered as Excel Spreadsheets containing the Preferred Name, CAS-RN. DTXSID, Formula, Formula of the MS-ready structure and associated masses, SMILES and InChI Strings/Keys.
image

DSSTox SDF File — Posted: 12/14/2016

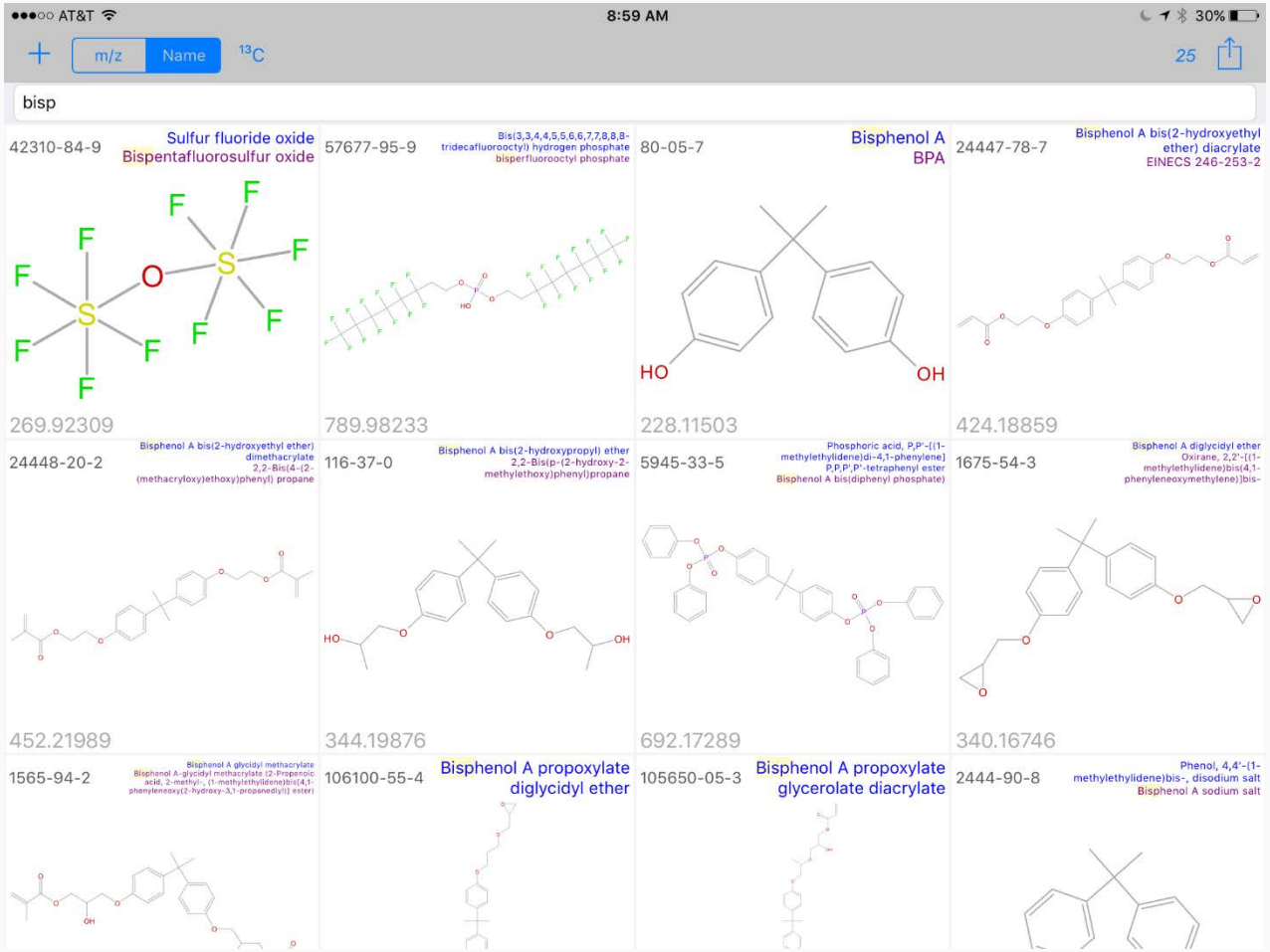
This zip file contains the entire chemical structure collection of over 700,000 chemicals from the DSSTox database contained in one large SDF file. The file contains the structure, The DSSTox Structure Identifier (DTXCID), The DSSTOX Substance Identifier (DTXSID listed as PubChem External Data Source), the associated Dashboard URL, associated synonyms and Quality Control Level details. 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.
image

# ~720,000 structures

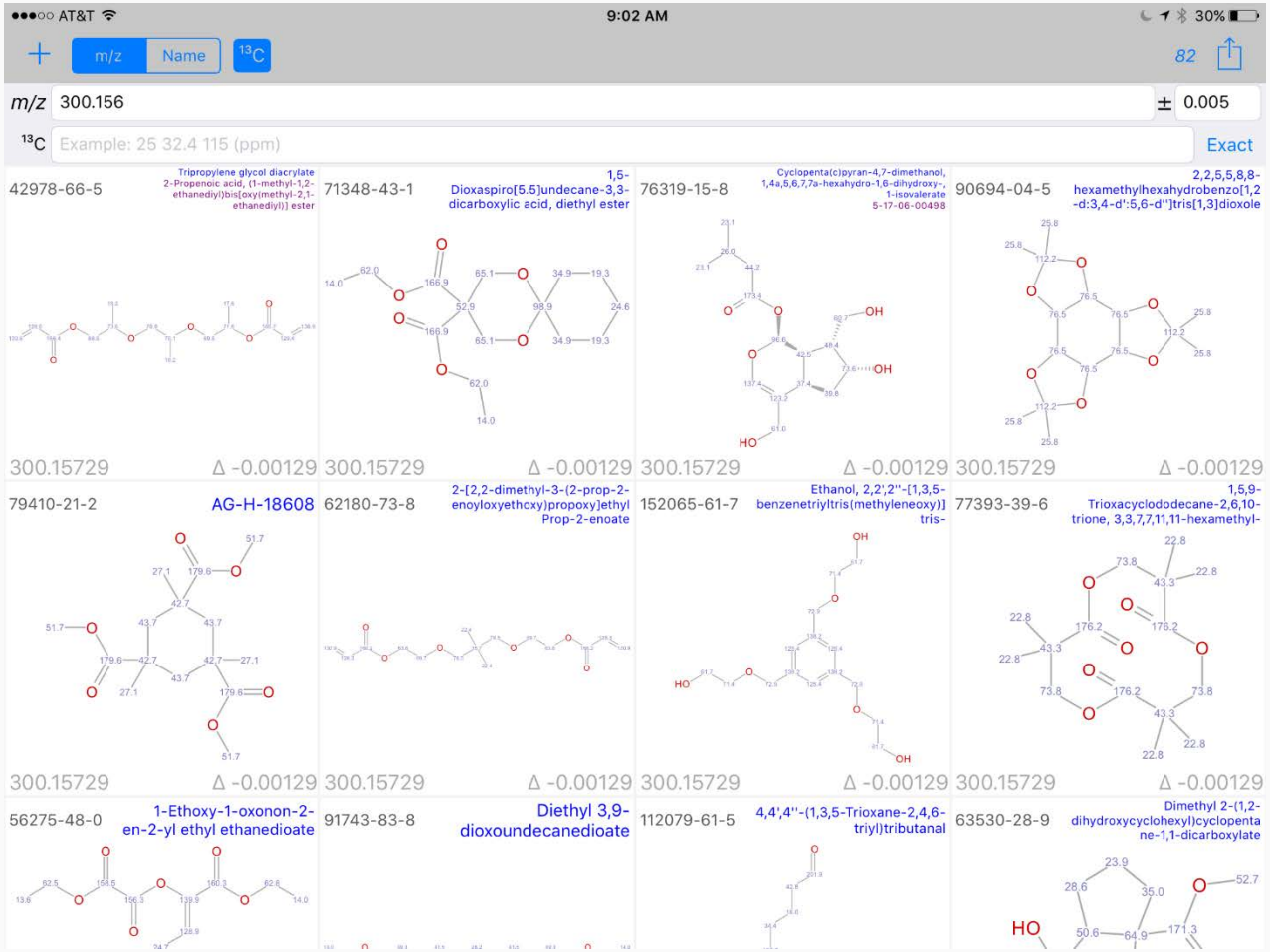



- Open Data for apps
  - Structures
  - CAS Registry Numbers
  - Names
  - Formulae
  - Mass
- iOS app including predicted C13 NMR

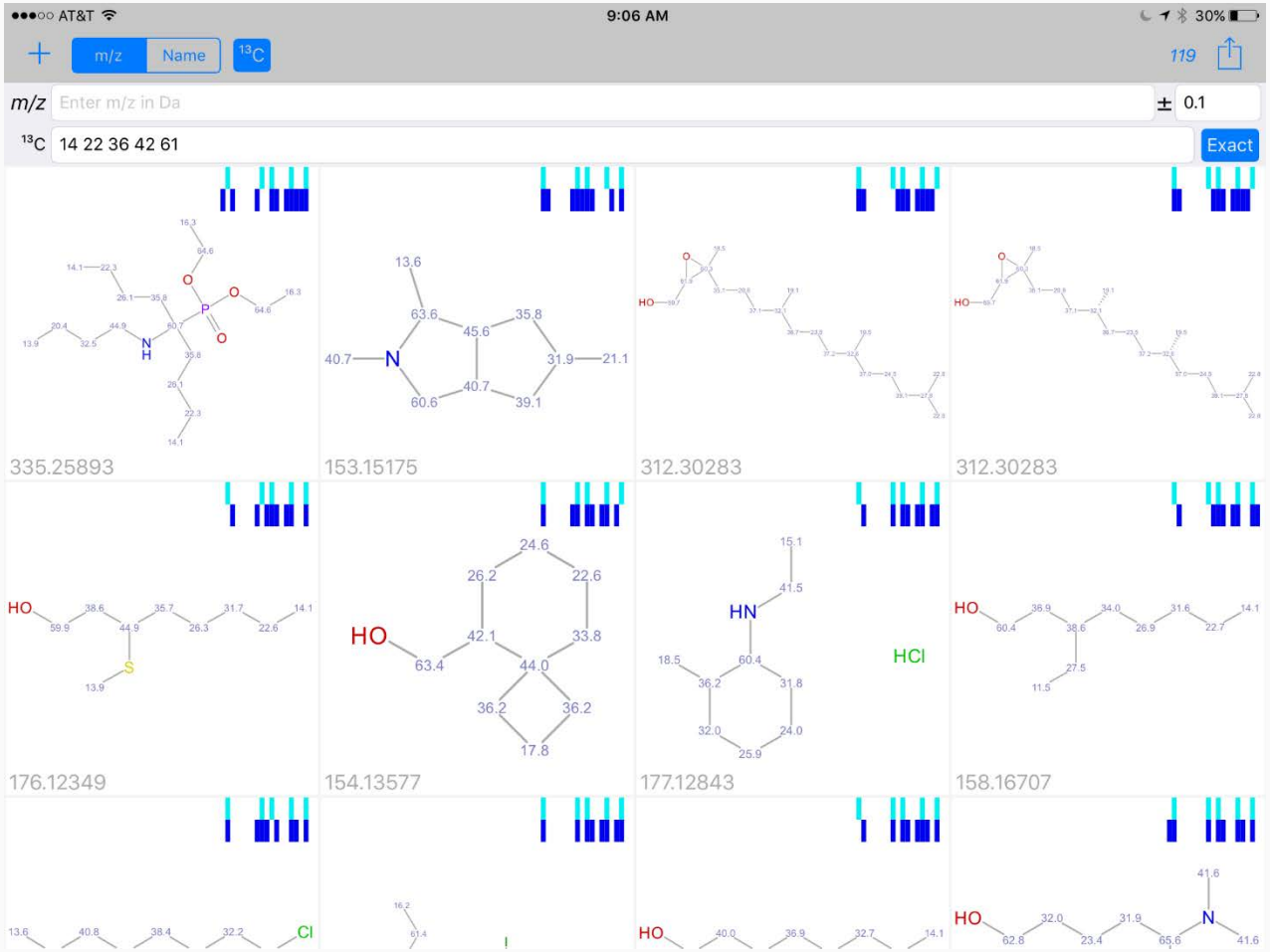
# Name Searching: “Bisp”



# Mass Searching



# Mass and CNMR Searching



# Conclusion

- The CompTox Chemistry Dashboard provides access to data for ~760,000 chemicals
- Data downloads allows for reuse in other systems and integration of resources to support research
  - Mobile applications
  - Integration between resources – PubChem, ChEMBL, ChemSpider, etc.
- Open OPERA Models and curated training and test data freely available
- OUP Open API and prediction web services are imminent

# 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