# Data Aggregation, Curation and Modeling Approaches to Deliver Prediction Models to Support Computational Toxicology at the US EPA

**Antony Williams**†
Kamel Mansouri
Todd Martin
Chris Grulke
John Wambaugh
Richard Judson
Grace Patlewicz
Imran Shah
Ann Richard

*NCCT, U.S. EPA*


American Chemical Society Meeting, Fall 2016

21-25 August 2016, Philadelphia, PA

# 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

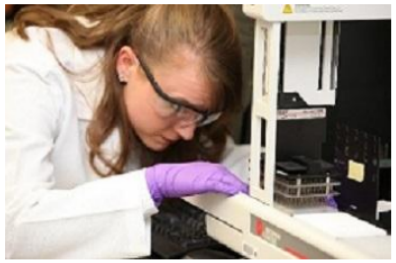
- Goal - To identify **chemical exposures** that may disrupt biological processes and cause adverse outcomes.

# Data, models, algorithms, …

- Our outputs include a lot of data, models, algorithms and software applications
- We produce Open Data – we want people to interrogate it, learn from it, develop understanding

## Toxicity Forecasting

### Advancing the Next Generation of Chemical Evaluation



EPA needs rapid and efficient methods to prioritize, screen and evaluate thousands of chemicals. EPA's Toxicity Forecaster (ToxCast) generates data and predictive models on thousands of chemicals of interest to the EPA. ToxCast uses high-throughput screening methods and computational toxicology approaches to rank and prioritize chemicals. In fact, EPA's Endocrine Disruption Screening Program (EDSP) is working to use ToxCast to rank and prioritize chemicals.

- ToxCast has data on over 1,800 chemicals from a broad range of sources including industrial and consumer products, food additives, and potentially "green" chemicals that could be safer alternatives to existing chemicals.
- ToxCast screens chemicals in over 700 high-throughput assays that cover a range of high-

## Downloadable Computational Toxicology Data

EPA's computational toxicology research efforts evaluate the potential health effects of thousands of chemicals. The process of evaluating potential health effects involves generating data that investigates the potential harm, or hazard of a chemical, the degree of exposure to chemicals as well as the unique chemical characteristics.

As part of EPA's commitment to share data, all of the computational toxicology data is publicly available for anyone to access and use.

**High–throughput Screening Data**

EPA researchers use rapid chemical screening (called high-throughput screening assays) to limit the number of laboratory animal tests while quickly and efficiently testing thousands of chemicals for potential health effects.

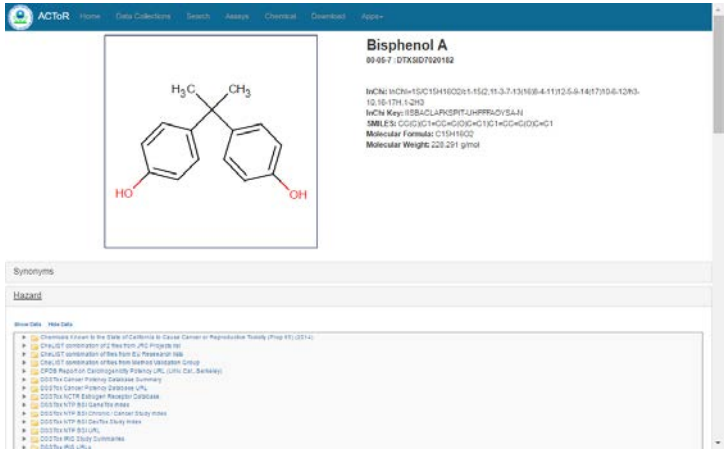
- ToxCast Data: High-throughput screening data on thousands of chemicals.

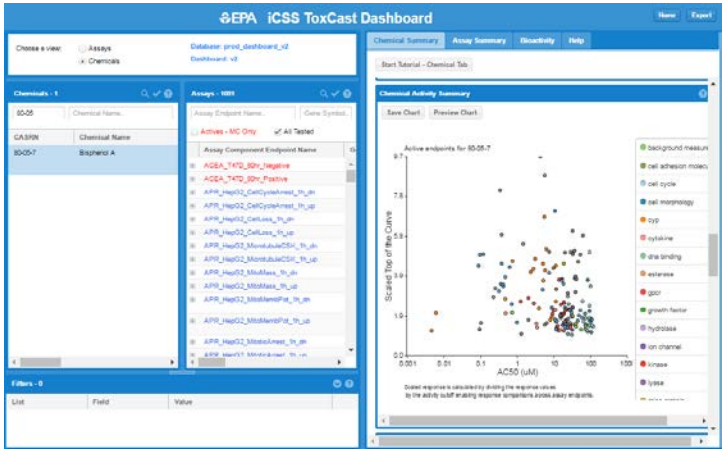
**Rapid Exposure and Dose Data**

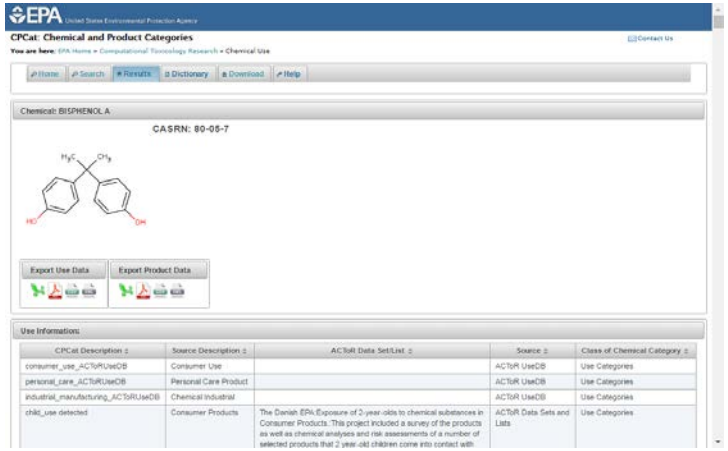
EPA researchers develop and use rapid exposure estimates to predict potential exposure for thousands of chemicals.

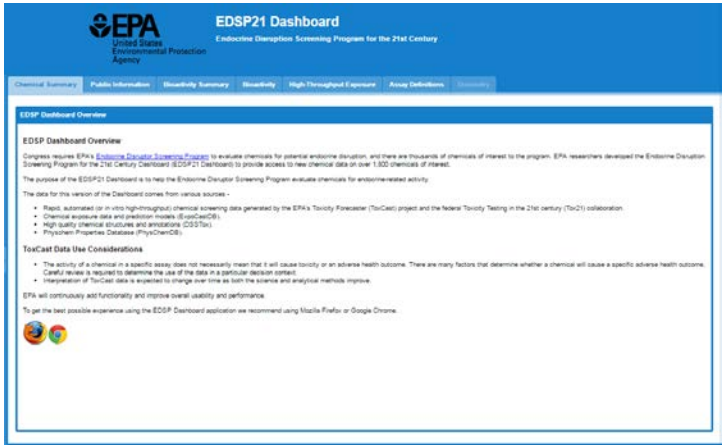
- High-throughput toxicokinetics data: It is important to link the external dose of a chemical to an internal blood or tissue concentration, this process is called toxicokinetics. EPA researchers measure the critical factors that determine the distribution

# Sharing via Dashboard Apps



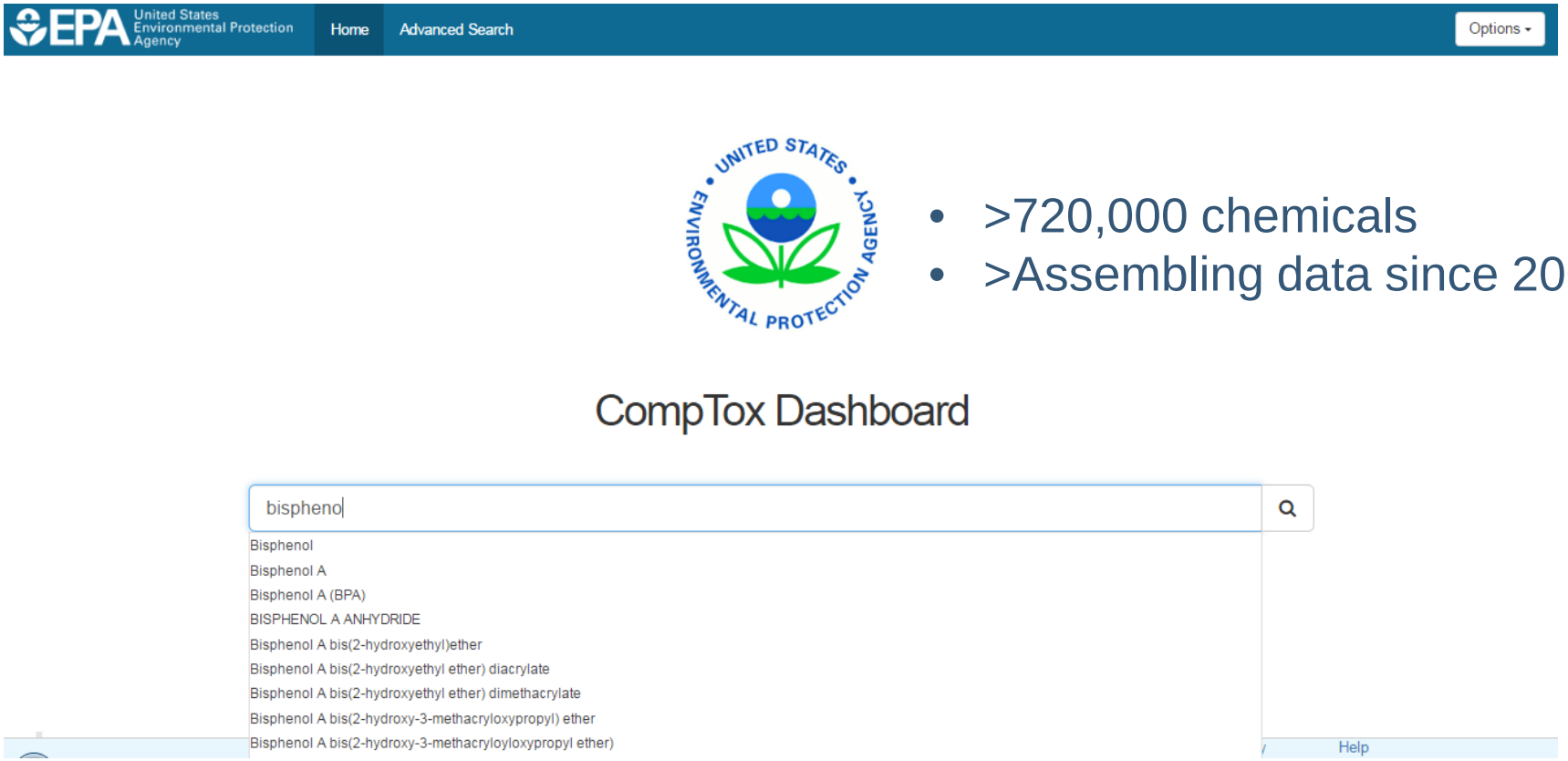



# Recent Cheminformatics Developments

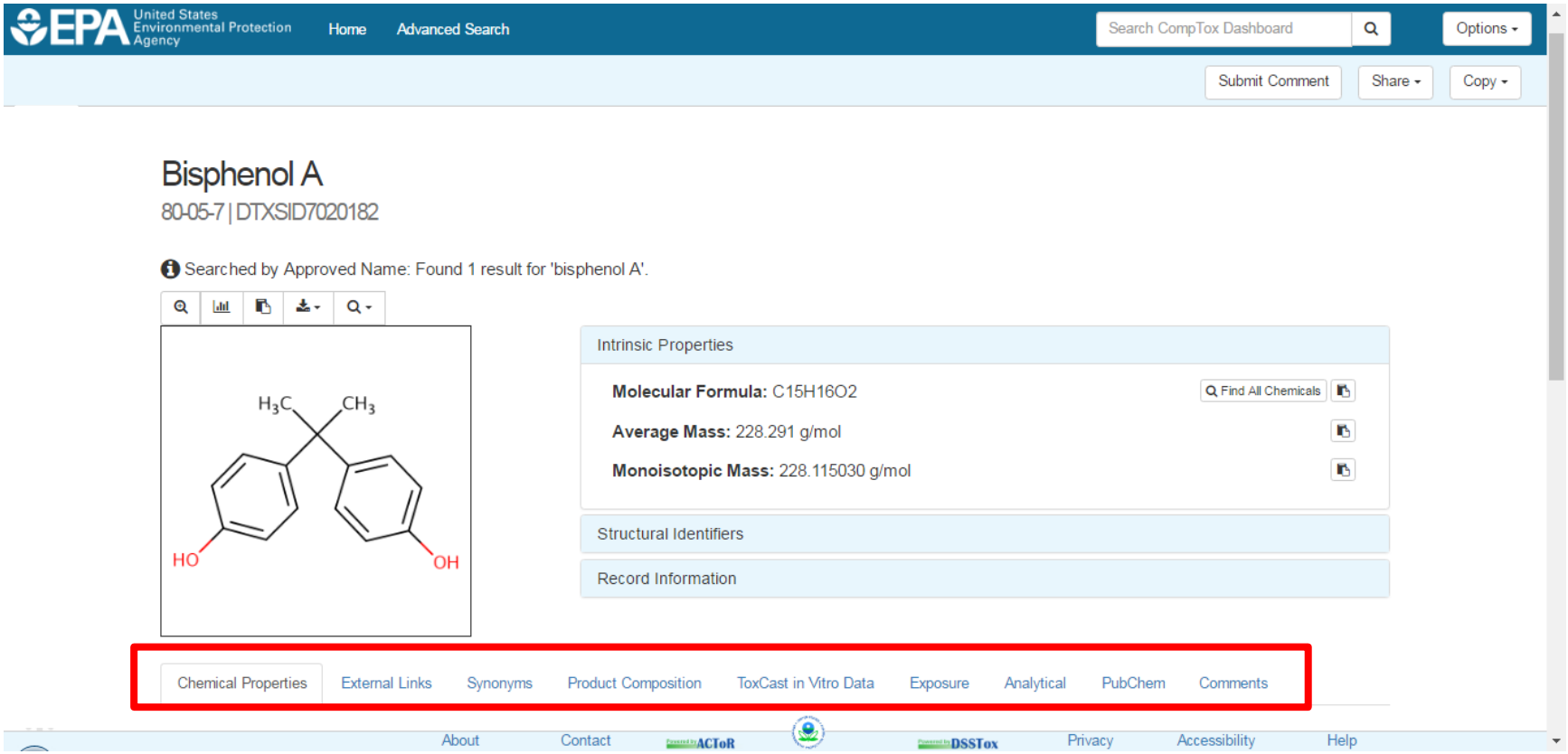
- We are building a new cheminformatics architecture
- PUBLIC dashboard gives access to “curated chemistry”
- Focus on integrating EPA ***and*** external resources
- Aggregating and curating data, visualization elements and “services” to underpin other efforts
  - QSAR
  - Read-across
  - Non-targeted screening

# Introducing Our Latest Dashboard https://comptox.epa.gov

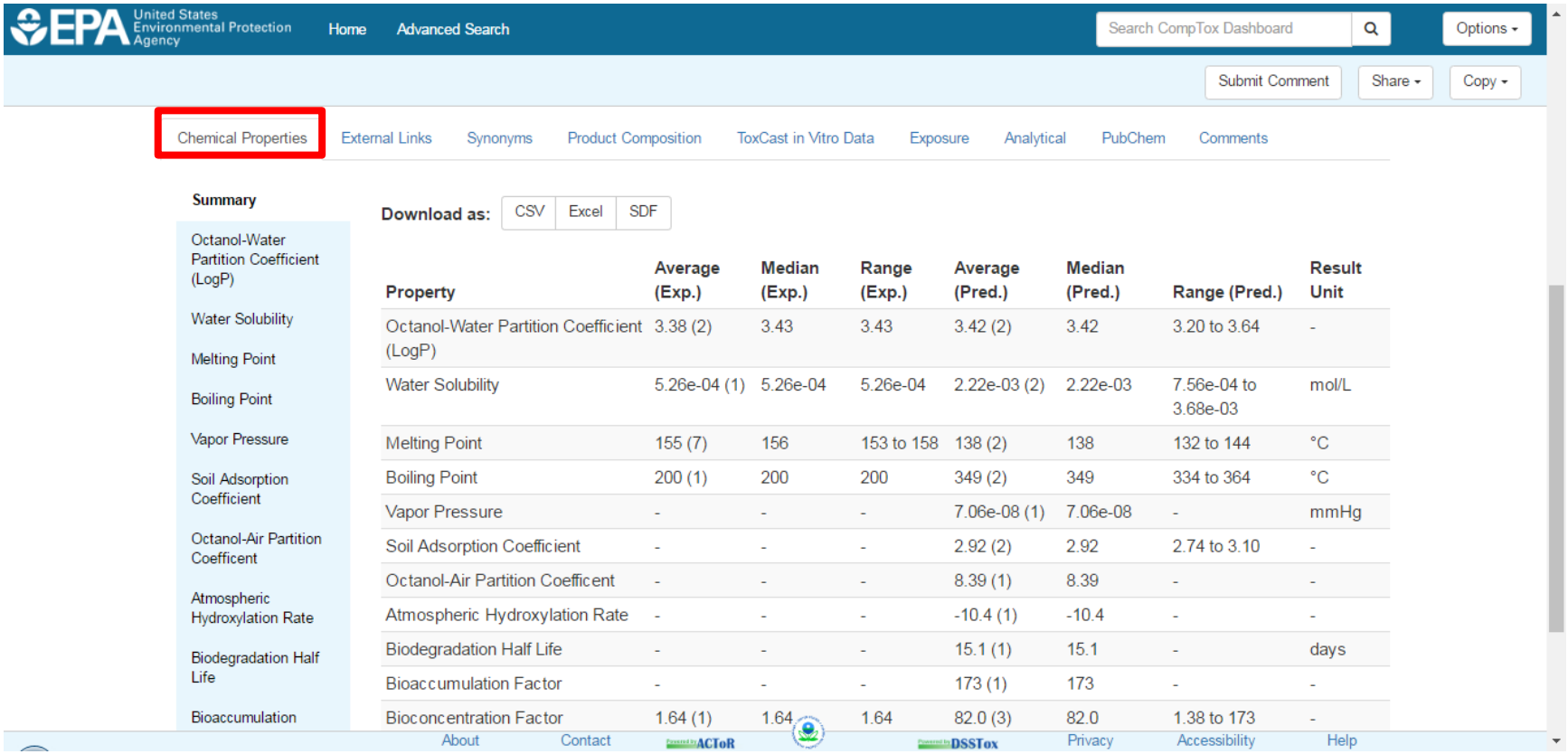



- >720,000 chemicals
- >Assembling data since 2002

# Bisphenol A


United States Environmental Protection Agency
Home
Advanced Search
Search CompTox Dashboard
Options
Submit Comment
Share
Copy
Bisphenol A
80-05-7 | DTXSID7020182
Searched by Approved Name: Found 1 result for 'bisphenol A'.
H3C
CH3
HO
OH
Intrinsic Properties
Molecular Formula: C15H16O2
Find All Chemicals
Average Mass: 228.291 g/mol
Monoisotopic Mass: 228.115030 g/mol
Structural Identifiers
Record Information
Chemical Properties
External Links
Synonyms
Product Composition
ToxCast in Vitro Data
Exposure
Analytical
PubChem
Comments
About
Contact
ACToR
DSSTox
Privacy
Accessibility
Help

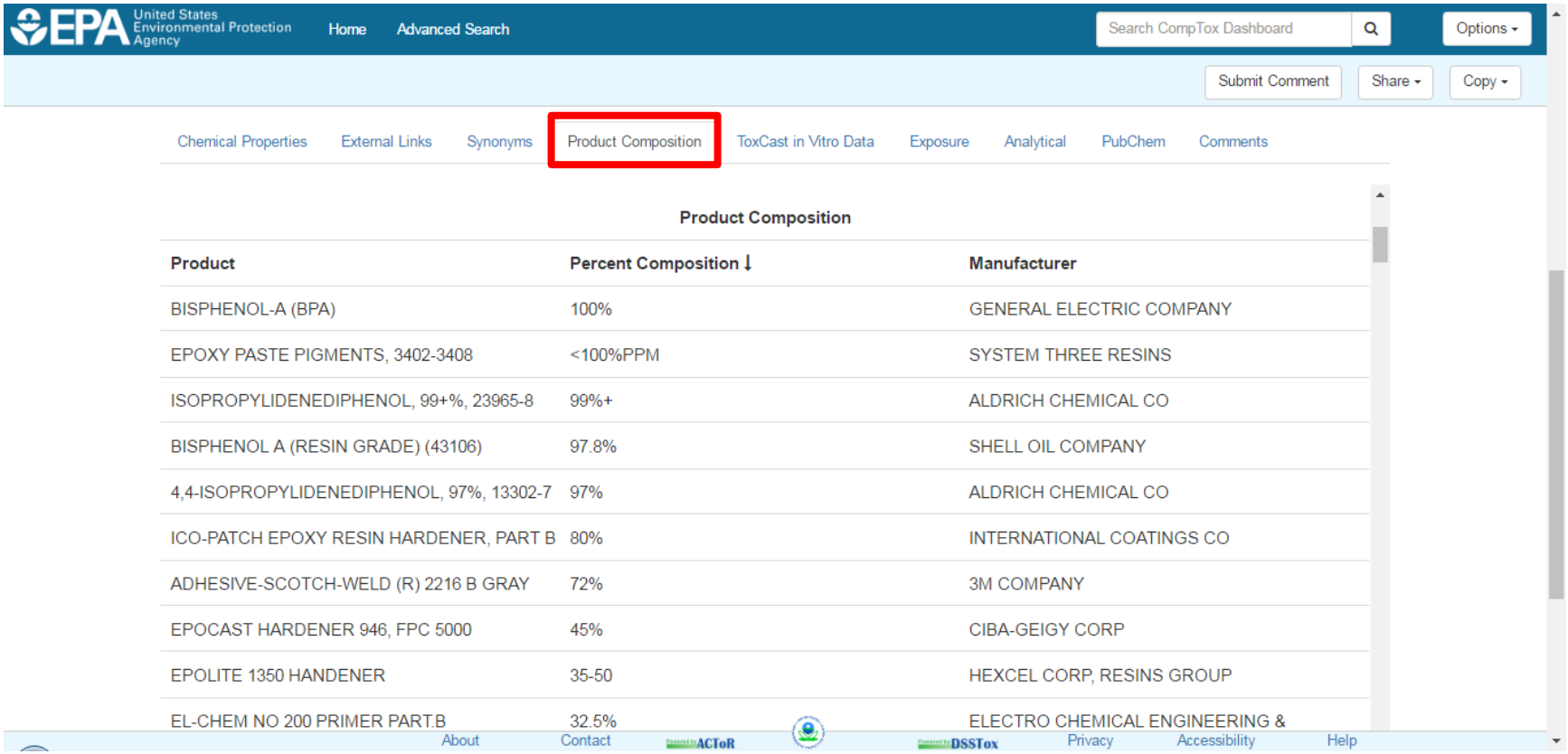
# Physicochemical Properties




| Property | Average (Exp.) | Median (Exp.) | Range (Exp.) | Average (Pred.) | Median (Pred.) | Range (Pred.) | Result Unit |
|---|---|---|---|---|---|---|---|
| Octanol-Water Partition Coefficient (LogP) | 3.38 (2) | 3.43 | 3.43 | 3.42 (2) | 3.42 | 3.20 to 3.64 | - |
| Water Solubility | 5.26e-04 (1) | 5.26e-04 | 5.26e-04 | 2.22e-03 (2) | 2.22e-03 | 7.56e-04 to 3.68e-03 | mol/L |
| Melting Point | 155 (7) | 156 | 153 to 158 | 138 (2) | 138 | 132 to 144 | °C |
| Boiling Point | 200 (1) | 200 | 200 | 349 (2) | 349 | 334 to 364 | °C |
| Vapor Pressure | - | - | - | 7.06e-08 (1) | 7.06e-08 | - | mmHg |
| Soil Adsorption Coefficient | - | - | - | 2.92 (2) | 2.92 | 2.74 to 3.10 | - |
| Octanol-Air Partition Coefficent | - | - | - | 8.39 (1) | 8.39 | - | - |
| Atmospheric Hydroxylation Rate | - | - | - | -10.4 (1) | -10.4 | - | - |
| Biodegradation Half Life | - | - | - | 15.1 (1) | 15.1 | - | days |
| Bioaccumulation Factor | - | - | - | 173 (1) | 173 | - | - |
| Bioconcentration Factor | 1.64 (1) | 1.64 | 1.64 | 82.0 (3) | 82.0 | 1.38 to 173 | - |

# Bioassay Screening Data

# Functional Use and Composition




**Product Composition**

| Product | Percent Composition | Manufacturer |
|---|---|---|
| BISPHENOL-A (BPA) | 100% | GENERAL ELECTRIC COMPANY |
| EPOXY PASTE PIGMENTS, 3402-3408 | <100%PPM | SYSTEM THREE RESINS |
| ISOPROPYLIDENEDIPHENOL, 99+%, 23965-8 | 99%+ | ALDRICH CHEMICAL CO |
| BISPHENOL A (RESIN GRADE) (43106) | 97.8% | SHELL OIL COMPANY |
| 4,4-ISOPROPYLIDENEDIPHENOL, 97%, 13302-7 | 97% | ALDRICH CHEMICAL CO |
| ICO-PATCH EPOXY RESIN HARDENER, PART B | 80% | INTERNATIONAL COATINGS CO |
| ADHESIVE-SCOTCH-WELD (R) 2216 B GRAY | 72% | 3M COMPANY |
| EPOCAST HARDENER 946, FPC 5000 | 45% | CIBA-GEIGY CORP |
| EPOLITE 1350 HANDENER | 35-50 | HEXCEL CORP, RESINS GROUP |
| EL-CHEM NO 200 PRIMER PART.B | 32.5% | ELECTRO CHEMICAL ENGINEERING & |

# Dashboard: External Links

Chemical Properties | External Links | Synonyms | Product Composition | ToxCast in Vitro Data | Exposure | Analytical | PubChem | Comments

**General**

- EPA Substance Registry...
- NIST Chemistry Webbook
- Household Products Dat...
- PubChem
- Chemspider
- CPCat
- DrugBank
- HMDB
- Wikipedia
- MSDS Lookup
- ToxPlanet
- ChemHat: Hazards and ...

**Toxicology**

- ACToR
- DrugPortal
- CCRIS
- ChemView
- CTD
- eChemPortal
- EDSP Dashboard
- Gene-Tox
- HSDB
- ToxCast Dashboard 2
- LactMed
- International Toxicity Esti...

**Publications**

- Toxline
- Environmental Health P...
- NIEHS
- National Toxicology Prog...
- Google Books
- Google Scholar
- Google Patents
- PubMed

**Analytical**

- National Environmental ...
- RSC Analytical Abstracts

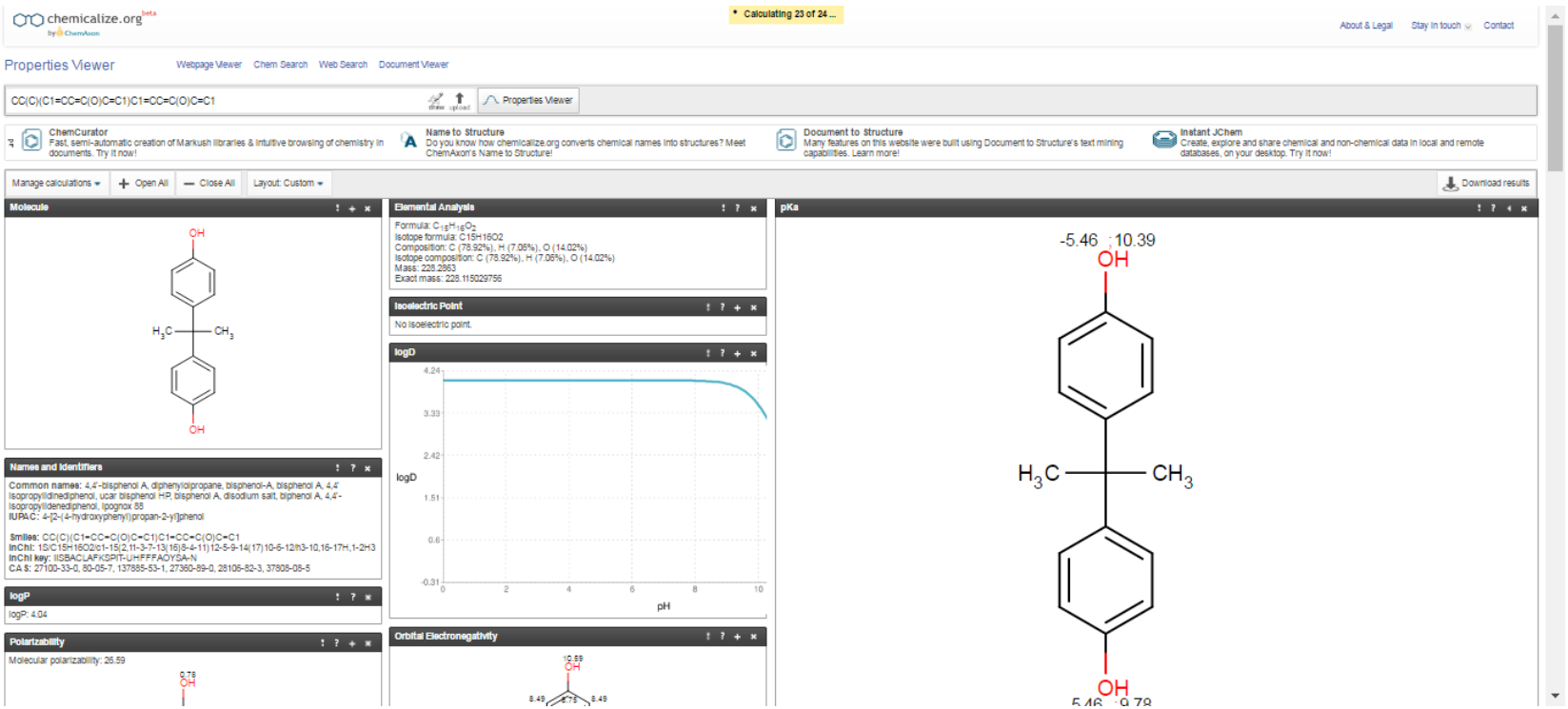
**Prediction**

- Chemicalize
- Proton NMR Prediction
- Carbon-13 NMR Prediction
- 2D NMR HSQC/HMBC ...
- ChemRTP Predictor

External Prediction Integration

Take Advantage of Online Resources and Stop Rework!

# External Integrations: Chemicalize

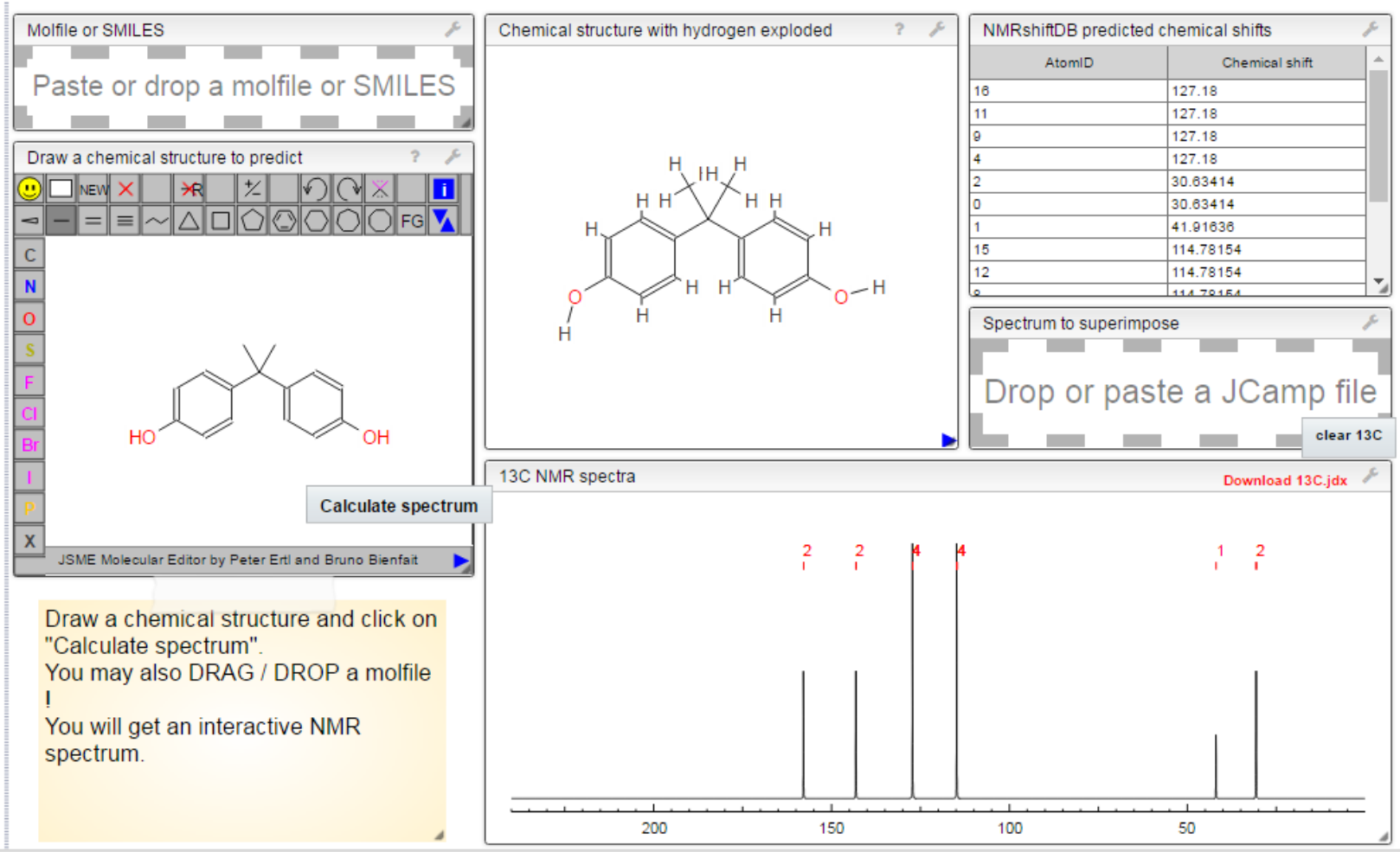


# External Integrations: MolInstincts

ChemRTP | Chemical Real-Time Predictor for Extensive Chemical Properties

| Property | Value | Unit | Accuracy |
|---|---|---|---|
| Absolute Entropy of Ideal Gas at 298.15K and 1bar | 129.0611 | cal/mol/K | |
| Acentric Factor | 0.922278 | dimensionless | |
| Critical Compressibility Factor | 0.271771 | dimensionless | - |
| Critical Pressure | 29.3031 | bar | |
| Critical Temperature | 890.3134 | K | |
| Critical Volume | 6.8654e-4 | m3/mol | |
| Enthalpy of Formation for Ideal Gas at 298.15K | -48.6730 | kcal/mol | |
| Liquid Molar Volume at 298.15K | 2.0095e-4 | m3/mol | |
| Molecular Weight | 228.2863 | g/mol | - |
| Net Standard State Enthalpy of Combustion at 298.15K | -1786.142 | kcal/mol | - |
| Normal Boiling Point | 653.6313 | K | |
| Melting Point | 446.0177 | K | |
| Refractive Index | 1.6036 | dimensionless | |
| Standard State Absolute Entropy at 298.15K and 1bar | 75.2883 | cal/mol/K | |
| Standard State Enthalpy of Formation at 298.15K and 1bar | -88.3085 | kcal/mol | |
| Magnetic Susceptibility | 149.2036 | ppm | |

# External Integrations: NMRDB.org




Draw a chemical structure and click on "Calculate spectrum".
You may also DRAG / DROP a molfile !
You will get an interactive NMR spectrum.

# Developing "NCCT Models"

- Interest in physicochemical properties to include in exposure modeling, augmented with ToxCast HTS *in vitro* data etc.

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

- Work has been initiated using available **physicochemical data**

# PHYSPROP Data: Available from:

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

## EPI Suite Data

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

**Basic Instructions:**

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

**WSKOWWIN Program Methodology & Validation Documents (includes Training & Validation datasets)** - Download file is: WSKOWWIN_Datasets.zip (180 KB)

**Click here to download WSKOWWIN_Datasets.zip**

**WATERNT (Water Solubility Fragment) Program Methodology & Validation Documents (includes Training & Validation datasets)** - Download file is: WaterFragmentDataFiles.zip (511 KB)

**Click here to download WaterFragmentDataFiles.zip**

**MPBPWIN (Melting Pt, Boiling Pt, Vapor Pressure) Program Test Sets** - Download file is: MP-BP-VP-TestSets.zip (1983 KB)

**Click here to download MP-BP-VP-TestSets.zip**

**BCFBAF Excel spreadsheets of BCF and kM data used in training & validation ... (includes the Jon Arnot Source BCF DB with multiple BCF values)** - Download file is: Data_for_BCFBAF.zip (1.4 MB)

**Click here to download Data_for_BCFBAF.zip**

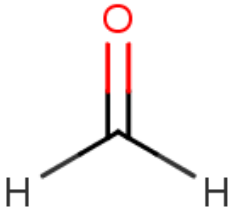
**HENRYWIN Data files used in training & validation ... (includes Meylan and Howard (1991) Data document)** - Download file is: HENRYWIN_Data_EPI.zip (531 K )

**Click here to download HENRYWIN_Data_EPI.zip**

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

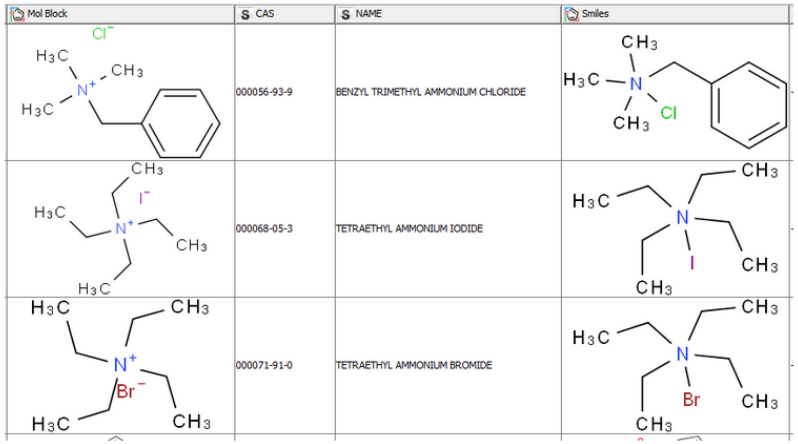
# Check and Curate Public Data

- Public data should always be checked and curated prior to modeling. This dataset was no different.
- The data files have **FOUR** representations of a chemical, plus the property value.

| SDF Molecule | Mol Mol Block | S Smiles | S CAS | S NAME | D Kow |
|---|---|---|---|---|---|
| -ISIS- 09141018452D<br><br>4 3 0 0 0 0 0 0 0 0999 V2000<br>2.4667 -0.0833 0.0000 O 0 0 ...<br>2.4667 -0.9125 0.0000 C 0 0 ...<br>1.7500 -1.3292 0.0000 H 0 0 ...<br>3.1833 -1.3292 0.0000 H 0 0 ...<br>2 1 2 0 0 0 0<br>3 2 1 0 0 0 0<br>4 2 1 0 0 0 0<br>M END<br>> <CAS> (000050-00-0)<br>000050-00-0<br><br>> <NAME> (000050-00-0)<br>FORMALDEHYDE<br><br>> <Kow> (000050-00-0)<br>3.500000000000000e-001 |  | O=C | 000050-00-0 | FORMALDEHYDE | 0.35 |

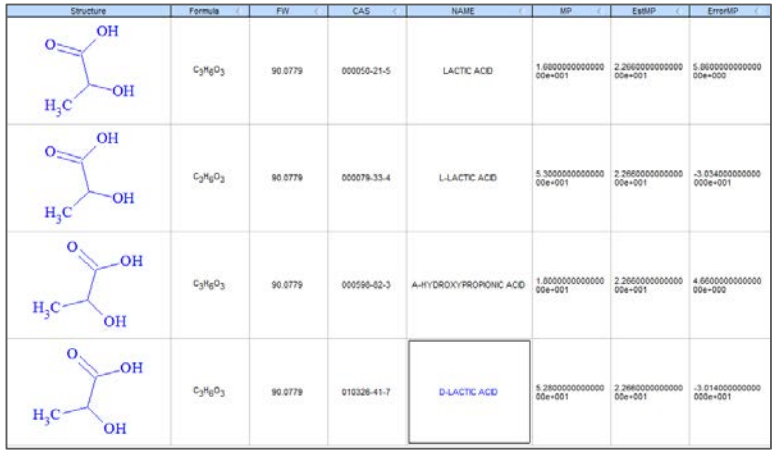
# Check and Curate Public Data

- Public data should always be checked and curated prior to modeling. This dataset was no different.



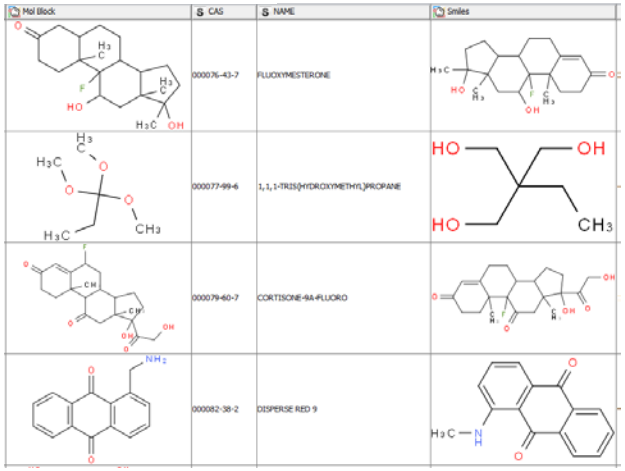
| Mol Block | CAS | NAME | Smiles |
|---|---|---|---|
| | 000056-93-9 | BENZYL TRIMETHYL AMMONIUM CHLORIDE | |
| | 000068-05-3 | TETRAETHYL AMMONIUM IODIDE | |
| | 000071-91-0 | TETRAETHYL AMMONIUM BROMIDE | |

Covalent Halogens



| Structure | Formula | FW | CAS | NAME | MP | EstMP | ErrorMP |
|---|---|---|---|---|---|---|---|
| | $C_3H_6O_3$ | 90.0779 | 000050-21-5 | LACTIC ACID | 1.6000000000000000e+001 | 2.2660000000000000e+001 | 5.6600000000000000e+000 |
| | $C_3H_6O_3$ | 90.0779 | 000079-33-4 | L-LACTIC ACID | 5.3000000000000000e+001 | 2.2660000000000000e+001 | -3.0340000000000000e+001 |
| | $C_3H_6O_3$ | 90.0779 | 000598-82-3 | A-HYDROXYPROPIONIC ACID | 1.8000000000000000e+001 | 2.2660000000000000e+001 | 4.6600000000000000e+000 |
| | $C_3H_6O_3$ | 90.0779 | 010326-41-7 | D-LACTIC ACID | 5.2800000000000000e+001 | 2.2660000000000000e+001 | -3.0140000000000000e+001 |

Identical Chemicals



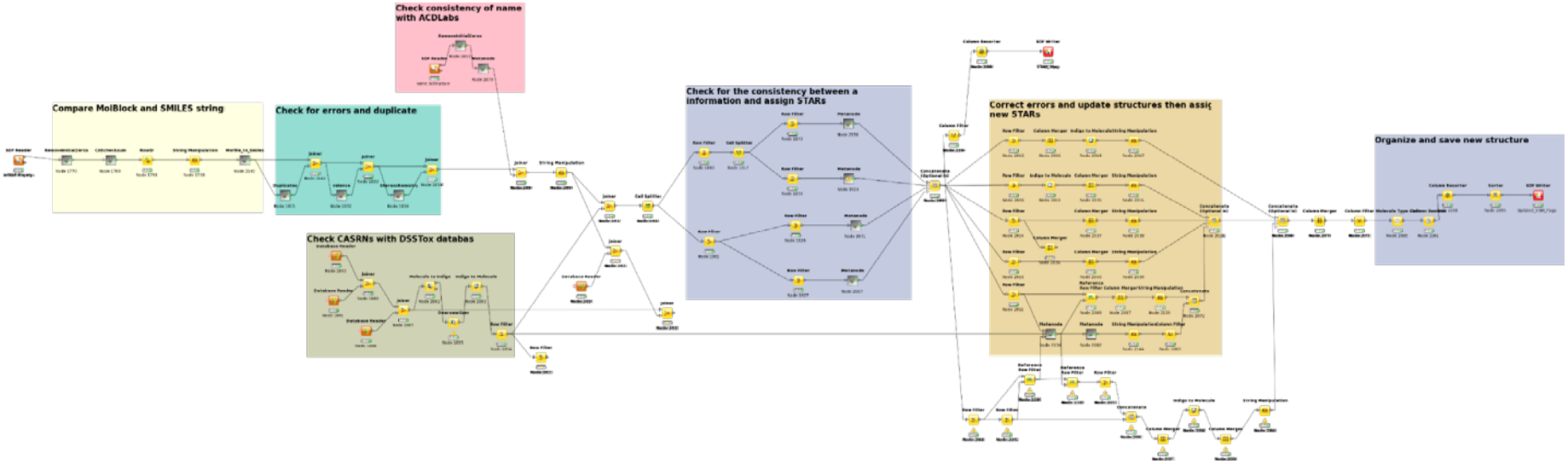
| Mol Block | CAS | NAME | Smiles |
|---|---|---|---|
| | 000076-43-7 | FLUOXYMESTERONE | |
| | 000077-99-6 | 1,1,1-TRIS(HYDROXYMETHYL)PROPANE | |
| | 000079-60-7 | CORTISONE-9A-FLUORO | |
| | 000082-38-2 | DISPERSE RED 9 | |

Mismatches

# The Approach

- Our curation process
  - Decide on the “chemical” by checking levels of consistency
  - We did NOT validate each measured property value
  - Perform initial analysis manually to understand how to clean the data (chemical structure and ID)
  - Automate the process (and test iteratively)
  - Process all datasets using final method

# KNIME Workflow to Evaluate the Dataset



# 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%)**

# 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 |

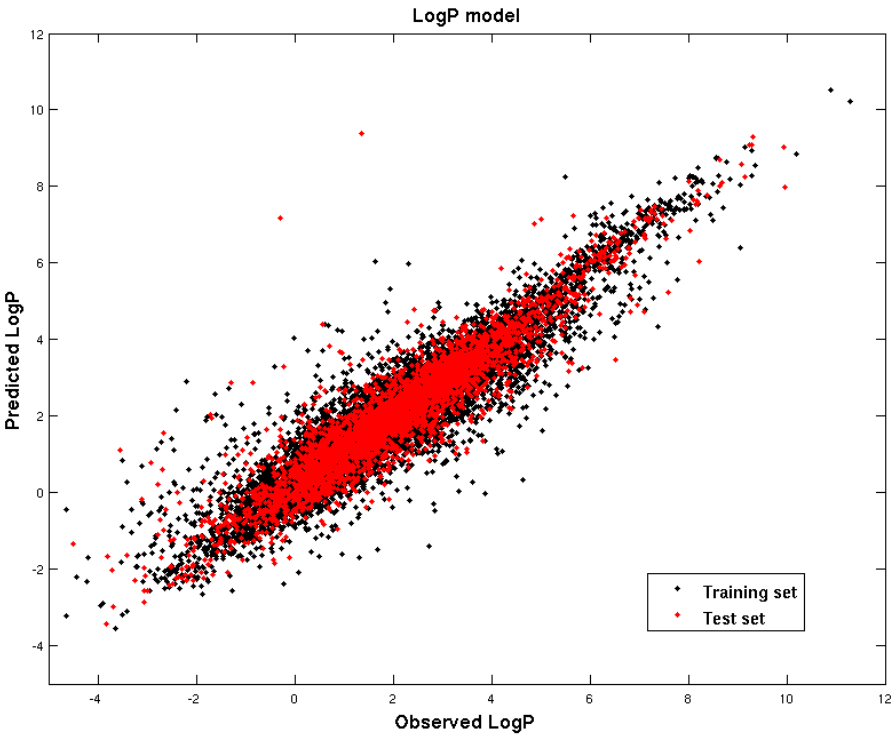
# Following the 5 OECD Principles*

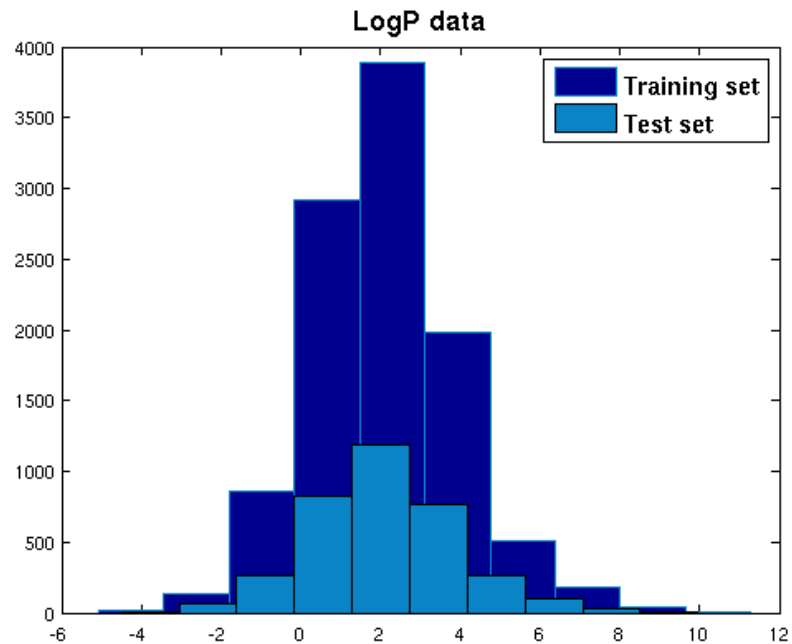
| Principle | Description |
|---|---|
| 1) A defined endpoint | Any **physicochemical, biological or environmental** effect that can be measured and therefore modelled. |
| 2) An unambiguous algorithm | **Ensure transparency** in the description of the model algorithm. |
| 3) A defined domain of applicability | **Define limitations** in terms of the types of **chemical structures**, physicochemical properties and mechanisms of action for which the models can generate **reliable predictions**. |
| 4) Appropriate measures of goodness-of-fit, robustness and predictivity | a) The internal **fitting** performance of a model<br>b) the **predictivity** of a model, determined by using an appropriate **external test set**. |
| 5) Mechanistic interpretation, if possible | Mechanistic **associations** between the **descriptors** used in a model and the **endpoint being predicted**. |



http://www.oecd.org/chemicalsafety/risk-assessment/37849783.pdf

# LogP Model: Weighted kNN Model, 9 descriptors




Weighted 5-nearest neighbors

**9 Descriptors**

Training set: 10531 chemicals

Test set: 3510 chemicals

| | |
|---|---|
| 5 fold CV: | Q2=0.85,RMSE=0.69 |
| Fitting: | R2=0.86,RMSE=0.67 |
| Test: | R2=0.86,RMSE=0.78 |

# NCCT Models

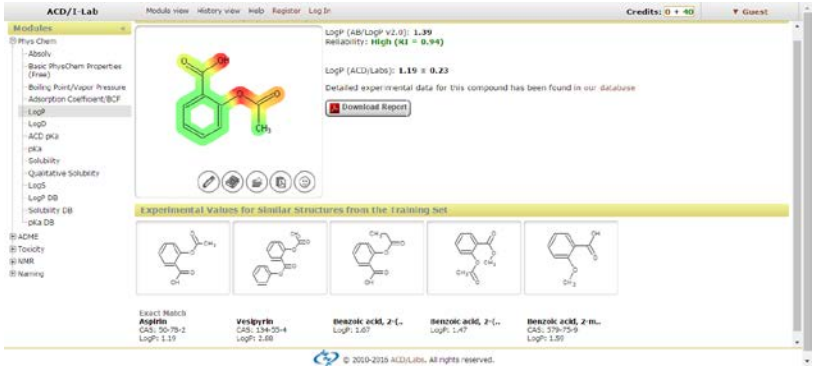
## What you would report in a paper

| Prop | Vars | 5-fold CV (75%) | | Training (75%) | | | Test (25%) | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Q2 | RMSE | N | R2 | RMSE | N | R2 | RMSE |
| **BCF** | 10 | 0.84 | 0.55 | 465 | 0.85 | 0.53 | 161 | 0.83 | 0.64 |
| **BP** | 13 | 0.93 | 22.46 | 4077 | 0.93 | 22.06 | 1358 | 0.93 | 22.08 |
| **LogP** | 9 | 0.85 | 0.69 | 10531 | 0.86 | 0.67 | 3510 | 0.86 | 0.78 |
| **MP** | 15 | 0.72 | 51.8 | 6486 | 0.74 | 50.27 | 2167 | 0.73 | 52.72 |
| **VP** | 12 | 0.91 | 1.08 | 2034 | 0.91 | 1.08 | 679 | 0.92 | 1 |
| **WS** | 11 | 0.87 | 0.81 | 3158 | 0.87 | 0.82 | 1066 | 0.86 | 0.86 |
| **HL** | 9 | 0.84 | 1.96 | 441 | 0.84 | 1.91 | 150 | 0.85 | 1.82 |

# Communicating Transparency in Models to Users of an App

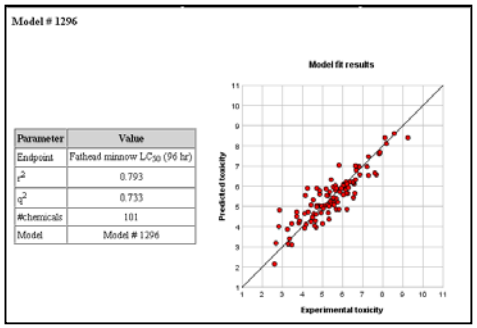
- Too often predicted values just give “numbers”
- Users have no real understanding of model performance
- There are good examples though! ACD/Ilab, T.E.S.T, OCHEM

ACD/ILab




OCHEM


EPA T.E.S.T




More on this later

# What about our Property Data?

Chemical Properties | External Links | Synonyms | Product Composition | ToxCast in Vitro Data | Exposure | Analytical | PubChem | Comments

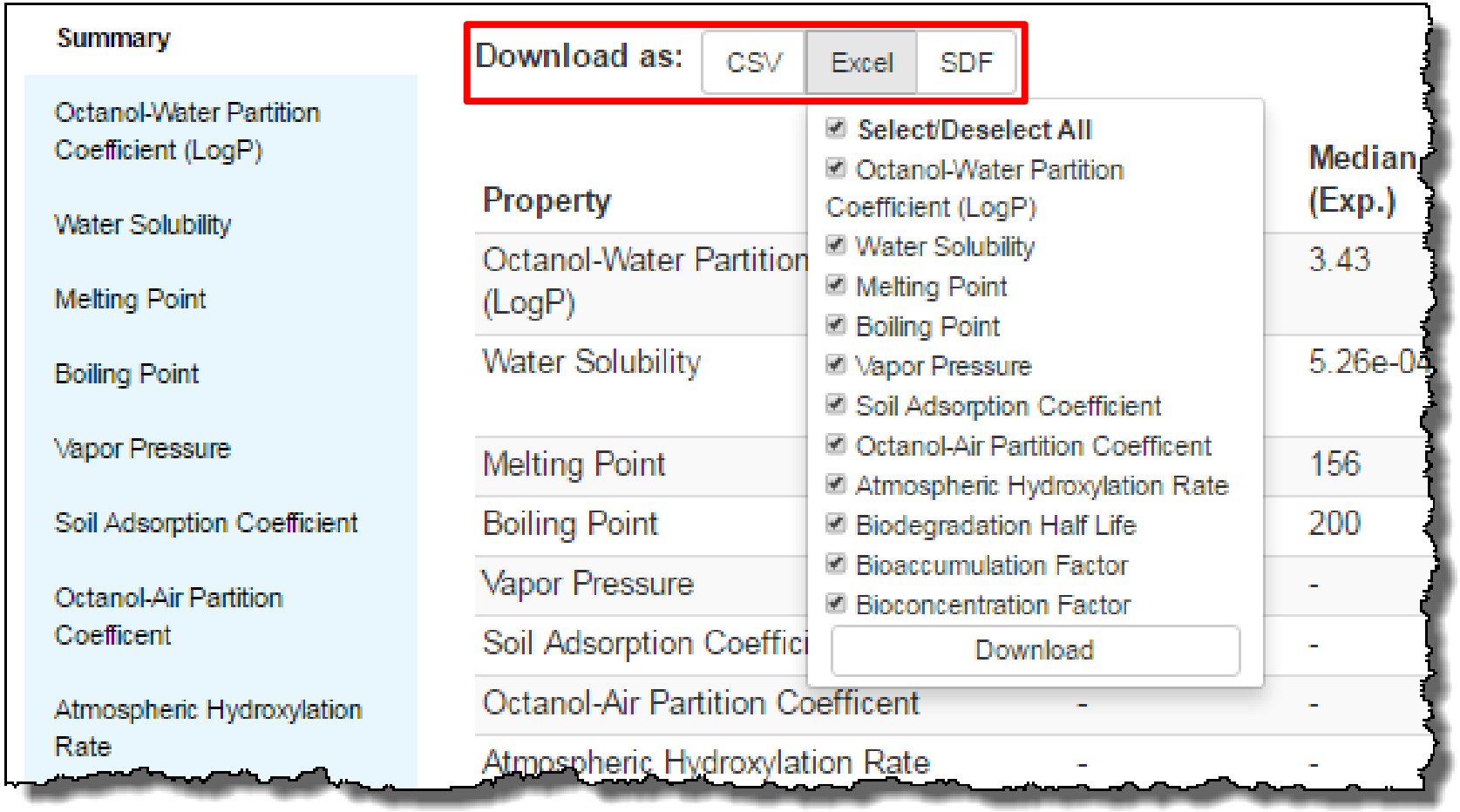
**Summary**

- Octanol-Water Partition Coefficient (LogP)
- Water Solubility
- Melting Point
- Boiling Point
- Vapor Pressure
- Soil Adsorption Coefficient
- Octanol-Air Partition Coefficent
- Atmospheric Hydroxylation Rate
- Biodegradation Half Life
- Bioaccumulation Factor
- Bioconcentration Factor

**Download as:** CSV | Excel | SDF

| Property | Average (Exp.) | Median (Exp.) | Range (Exp.) | Average (Pred.) | Median (Pred.) | Range (Pred.) | Result Unit |
|---|---|---|---|---|---|---|---|
| Octanol-Water Partition Coefficient (LogP) | 3.38 (2) | 3.43 | 3.43 | 3.42 (2) | 3.42 | 3.20 to 3.64 | - |
| Water Solubility | 5.26e-04 (1) | 5.26e-04 | 5.26e-04 | 2.22e-03 (2) | 2.22e-03 | 7.56e-04 to 3.68e-03 | mol/L |
| Melting Point | 155 (7) | 156 | 153 to 158 | 138 (2) | 138 | 132 to 144 | °C |
| Boiling Point | 200 (1) | 200 | 200 | 349 (2) | 349 | 334 to 364 | °C |
| Vapor Pressure | - | - | - | 7.06e-08 (1) | 7.06e-08 | - | mmHg |
| Soil Adsorption Coefficient | - | - | - | 2.92 (2) | 2.92 | 2.74 to 3.10 | - |
| Octanol-Air Partition Coefficent | - | - | - | 8.39 (1) | 8.39 | - | - |
| Atmospheric Hydroxylation Rate | - | - | - | -10.4 (1) | -10.4 | - | - |
| Biodegradation Half Life | - | - | - | 15.1 (1) | 15.1 | - | days |
| Bioaccumulation Factor | - | - | - | 173 (1) | 173 | - | - |
| Bioconcentration Factor | 1.64 (1) | 1.64 | 1.64 | 82.0 (3) | 82.0 | 1.38 to 173 | - |

# Data Downloads

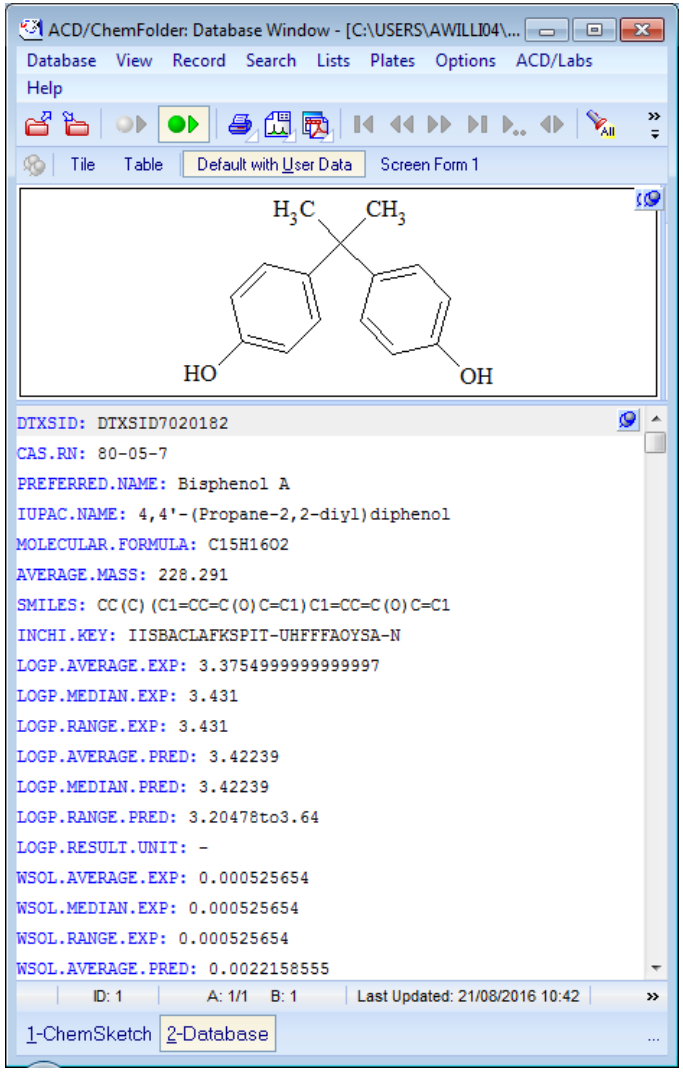


# Data Download: Excel

A1 | Property

| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | Property | Average (Exp.) | Median (Exp.) | Range (Exp.) | Average (Pred.) | Median (Pred.) | Range (Pred.) | Result Unit |
| 2 | Octanol-Water Partition Coefficient (LogP) | 3.38 (2) | 3.43 | 3.43 | 3.42 (2) | 3.42 | 3.20 to 3.64 | - |
| 3 | Water Solubility | 5.26e-04 (1) | 5.26E-04 | 5.26E-04 | 2.22e-03 (2) | 2.22E-03 | 7.56e-04 to 3.68e-03 | mol/L |
| 4 | Melting Point | 155 (7) | 156 | 153 to 158 | 138 (2) | 138 | 132 to 144 | °C |
| 5 | Boiling Point | 200 (1) | 200 | 200 | 349 (2) | 349 | 334 to 364 | °C |
| 6 | Vapor Pressure | - | - | - | 7.06e-08 (1) | 7.06E-08 | - | mmHg |
| 7 | Soil Adsorption Coefficient | - | - | - | 2.92 (2) | 2.92 | 2.74 to 3.10 | - |
| 8 | Octanol-Air Partition Coefficent | - | - | - | 8.39 (1) | 8.39 | - | - |
| 9 | Atmospheric Hydroxylation Rate | - | - | - | -10.4 (1) | -10.4 | - | - |
| 10 | Biodegradation Half Life | - | - | - | 15.1 (1) | 15.1 | - | days |
| 11 | Bioaccumulation Factor | - | - | - | 173 (1) | 173 | - | - |
| 12 | Bioconcentration Factor | 1.64 (1) | 1.64 | 1.64 | 82.0 (3) | 82 | 1.38 to 173 | - |
| 13 | | | | | | | | |
| 14 | | | | | | | | |

# Data Download: SDF



# Access to Experimental Data

| Property | Average (Exp.) | Median (Exp.) | Range (Exp.) |
|---|---|---|---|
| Octanol-Water Partition Coefficient (LogP) | 3.38 (2) | 3.43 | 3.43 |
| Water Solubility | 5.26e-04 (1) | 5.26e-04 | 5.26e-04 |
| Melting Point | 155 (7) | 156 | 153 to 158 |
| Boiling Point | | | |
| Vapor Pressure | | | |

**Melting Point**

| | Average | Median | Range |
|---|---|---|---|
| **Experimental** | 155 (7) | 156 | 153 to 158 |
| **Predicted** | 138 (2) | 138 | 132 to 144 |

**Download as:** CSV Excel SDF

**Experimental**

| Source | Result |
|---|---|
| PhysPropNCCT | 153 °C |
| 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 |

# Predictions for >720,000 Chemicals

- NCCT_Model predictions were built on curated training sets
- All chemicals in DSSTox, accessed via the CompTox Dashboard, were pushed through all predictive models
- Predicted data made available, with detailed **MODEL REPORTS**

# Predicted Data

Predicted

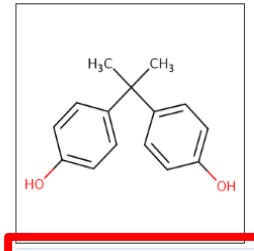
| Source | Result | Calculation Details |
|---|---|---|
| EPISUITE | 132 °C | Not Available |
| NCCT | 144 °C | NCCT Model Report |

EPA United States Environmental Protection Agency | Home | Advanced Search | 20182

NCCT Models: Melting Point

Bisphenol A

80-05-7 | DTXSID7020182




Model Results

**Predicted value:** 144 °C

**Global applicability domain:** Inside

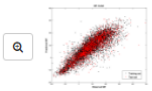
**Local applicability domain index:** 0.91

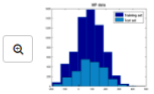
**Confidence level:** 0.65

Calculation Result for a chemical

Model Performance with full QMRF

Model Performance





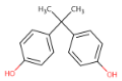
Weighted KNN model

QMRF

| 5-fold CV (75%) | | Training (75%) | | Test (25%) | |
|---|---|---|---|---|---|
| Q2 | RMSE | R2 | RMSE | R2 | RMSE |
| 0.72 | 51.8 | 0.74 | 50.3 | 0.73 | 52.7 |

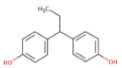
Nearest Neighbors from Training Set

Nearest Neighbors from Training Set



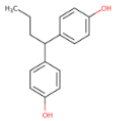
Bisphenol A

**Measured:** 153
**Predicted:** 144



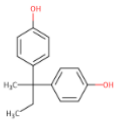
4,4'-Propane-1,1-diyldiphenol

**Measured:** 132
**Predicted:** 133



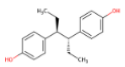
phenol, 4,4'-butylidenebis-

**Measured:** 137
**Predicted:** 142



Bisphenol B

**Measured:** 121
**Predicted:** 140



meso-Hexestrol

**Measured:** 187
**Predicted:** 157

# QMRF Reports

(Q)SAR Model Reporting Format Inventory

Log in Register

Home | Search documents | Search structures

All published QMRF documents **(109)** are available for download and can be searched either through free text queries or by several predefined fields.
All substances, available in the QMRF Database, can be searched by exact or similar structure.

## What is QMRF Database?

Do you need to register to use the QMRF Database?

Please register only if you wish to submit a QMRF. Registration is not necessary if you only wish to search the database and access information on QMRFs.

Help

## How to create an QMRF Document?

- log in into QMRF Database and use the *New document* tab;
- by QMRF editor : once started, it will create shortcut on your desktop and can be started later even offline.

## Most recent QMRF documents

| # | QMRF# | Title | Last updated | View | Download |
|---|---|---|---|---|---|
| 1 | Q50-54-55-501 | BIOVIA toxicity prediction model – Ames Mutagenicity | 2016-6-17 14:58 | | |
| 2 | Q51-54-55-502 | BIOVIA toxicity prediction model – rat oral LD50 | 2016-6-17 14:58 | | |

# Prediction Details and QMRF Report

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

QMRF_NCCT_MP_08212016 - Adobe Acrobat Pro

File Edit View Window Help

Create Customize

1 / 10 143% Tools Fill & Sign Comment

| | |
|---|---|
| QMRF | *QMRF identifier (JRC Inventory):* **To be entered by JRC** |
| | *QMRF Title:* **MP:** Melting point prediction **from the** NCCT Models Suite. |
| | *Printing Date:* **May 4, 2016** |

**1.QSAR identifier**

**1.1.QSAR identifier (title):**

**MP:** Melting point prediction

from the NCCT_Models Suite.

**1.2.Other related models:**

No related models

**1.3.Software coding the model:**

NCCT_models V1.02

Suite of QSAR models to predict physicochemical properties and environmental fate of organic chemicals

# EPA T.E.S.T

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

EPA US Environmental Protection Agency

Español | 中文: 繁體版 | 中文: 简体版 | Tiếng Việt | 한국어

Learn the Issues | Science & Technology | Laws & Regulations | About EPA

Search EPA.gov

Related Topics: Safer Chemicals Research

Contact Us Share

## Toxicity Estimation Software Tool (TEST)

On this page:

- QSAR Methodologies
- What's New in Version 4.2?
- Prior Version History
- System Requirements
- Installation Instructions
- Publications
- Get Email Alerts

The Toxicity Estimation Software Tool (TEST) was developed to allow users to easily estimate the toxicity of chemicals using Quantitative Structure Activity Relationships (QSARs) methodologies. QSARs are mathematical models used to predict measures of toxicity from the physical characteristics of the structure of chemicals (known as molecular descriptors). Simple QSAR models calculate the toxicity of chemicals using a simple linear function of molecular descriptors:

**Ask a Technical Expert**

Got a question about our research model? Want to give us feedback? Contact a technical expert about TEST.

# Physical properties in TEST

| Endpoint | Definition |
|---|---|
| Normal boiling point | Temperature ($^{o}C$) at which a chemical boils at atmospheric pressure |
| Density | Density ($g/cm^3$) for chemicals which have a normal boiling point greater than $25^{o}C$ |
| Flash point | The lowest temperature ($^{o}C$) at which it can vaporize to form an ignitable mixture in air |
| Thermal conductivity | The property of a material (mW/mK) reflecting its ability to conduct heat |

# Physical properties in TEST, cont.

| 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 |

# Test set of predictions available…

- A test set of predictions already performed
- This initial set of data already available
- 1,000 chemicals done, 720,000 to go…

New Properties from T.E.S.T

| | |
|---|---|
| Octanol-Water Partition Coefficient (LogP) | Vapor Pressure |
| Water Solubility | Viscosity |
| Density | Soil Adsorption Coefficient |
| Flash Point | Octanol-Air Partition Coefficent |
| Melting Point | Atmospheric Hydroxylation Rate |
| Boiling Point | Biodegradation Half Life |
| Surface Tension | Bioaccumulation Factor |
| Thermal Conductivity | Bioconcentration Factor |

# EPA T.E.S.T

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

EPA US Environmental Protection Agency

Español | 中文: 繁體版 | 中文: 简体版 | Tiếng Việt | 한국어

Learn the Issues | Science & Technology | Laws & Regulations | About EPA

Search EPA.gov

Related Topics: Safer Chemicals Research

Contact Us Share

## Toxicity Estimation Software Tool (TEST)

On this page:

- QSAR Methodologies
- What's New in Version 4.2?
- Prior Version History
- System Requirements
- Installation Instructions
- Publications
- Get Email Alerts

From physicochemical property endpoints to toxicity endpoints

The Toxicity Estimation Software Tool (TEST) was developed to allow users to easily estimate the toxicity of chemicals using Quantitative Structure Activity Relationships (QSARs) methodologies. QSARs are mathematical models used to predict measures of toxicity from the physical characteristics of the structure of chemicals (known as molecular descriptors). Simple QSAR models calculate the toxicity of chemicals using a simple linear function of molecular descriptors:

**Ask a Technical Expert**

Got a question about our research model? Want to give us feedback? Contact a technical expert about TEST.

# 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 |

# Endpoints in TEST, cont.

| Endpoint | Definition |
|---|---|
| Bioaccumulation factor | Ratio of the chemical concentration in fish as a result of absorption via the respiratory surface to that in water at steady state |
| 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 |

# For T.E.S.T >800 Descriptors used

- Estate values and E-state counts
- Constitutional descriptors
- Topological descriptors
- Walk and path counts
- Connectivity
- Information content
- 2d autocorrelation
- Burden eigenvalue
- Molecular property (such as Kow)
- Kappa
- Hydrogen bond acceptor/donor counts
- Molecular distance edge
- Molecular fragment counts

# Full transparency for each prediction

United States Environmental Protection Agency

## Toxicity prediction results for 333-41-5 for Hierarchical clustering method

Prediction results

| Endpoint | Experimental value CAS: 333-41-5 Source: ECOTOX | Predicted value[a] | Prediction interval |
|---|---|---|---|
| Fathead minnow $LC_{50}$ (96 hr) -Log(mol/L) | 4.81 | 5.39 | $4.54 \leq Tox \leq 6.24$ |
| Fathead minnow $LC_{50}$ (96 hr) mg/L | 4.70 | 1.23 | $0.17 \leq Tox \leq 8.71$ |

[a]Note: the test chemical was present in the external test set.

Cluster model predictions and statistics

| Cluster model | Test chemical descriptor values | Prediction interval -Log(mol/L) | $r^2$ | $q^2$ | #chemicals |
|---|---|---|---|---|---|
| 1296 | Descriptors | 6.010 ± 1.136 | 0.793 | 0.733 | 101 |
| 1300 | Descriptors | 5.458 ± 1.312 | 0.729 | 0.645 | 111 |
| 1301 | Descriptors | 5.136 ± 1.169 | 0.747 | 0.718 | 294 |
| 1302 | Descriptors | 4.922 ± 1.182 | 0.774 | 0.751 | 641 |

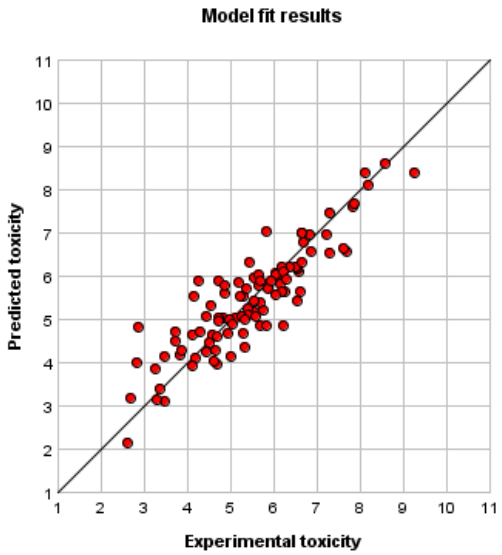
**Model # 1296**

Cluster models with violated constraints

| Cluster Model | $r^2$ | $q^2$ | # chemicals | Message |
|---|---|---|---|---|
| 1121 | 0.810 | 0.576 | 10 | Rmax constraint not met |
| 1209 | 0.799 | 0.574 | 11 | Fragment constraint not met |
| 1247 | 0.919 | 0.647 | 20 | Fragment constraint not met |
| 1264 | 0.869 | 0.781 | 22 | Fragment constraint not met |
| 1268 | 0.675 | 0.553 | 24 | Fragment constraint not met |

Descriptor values for test chemical

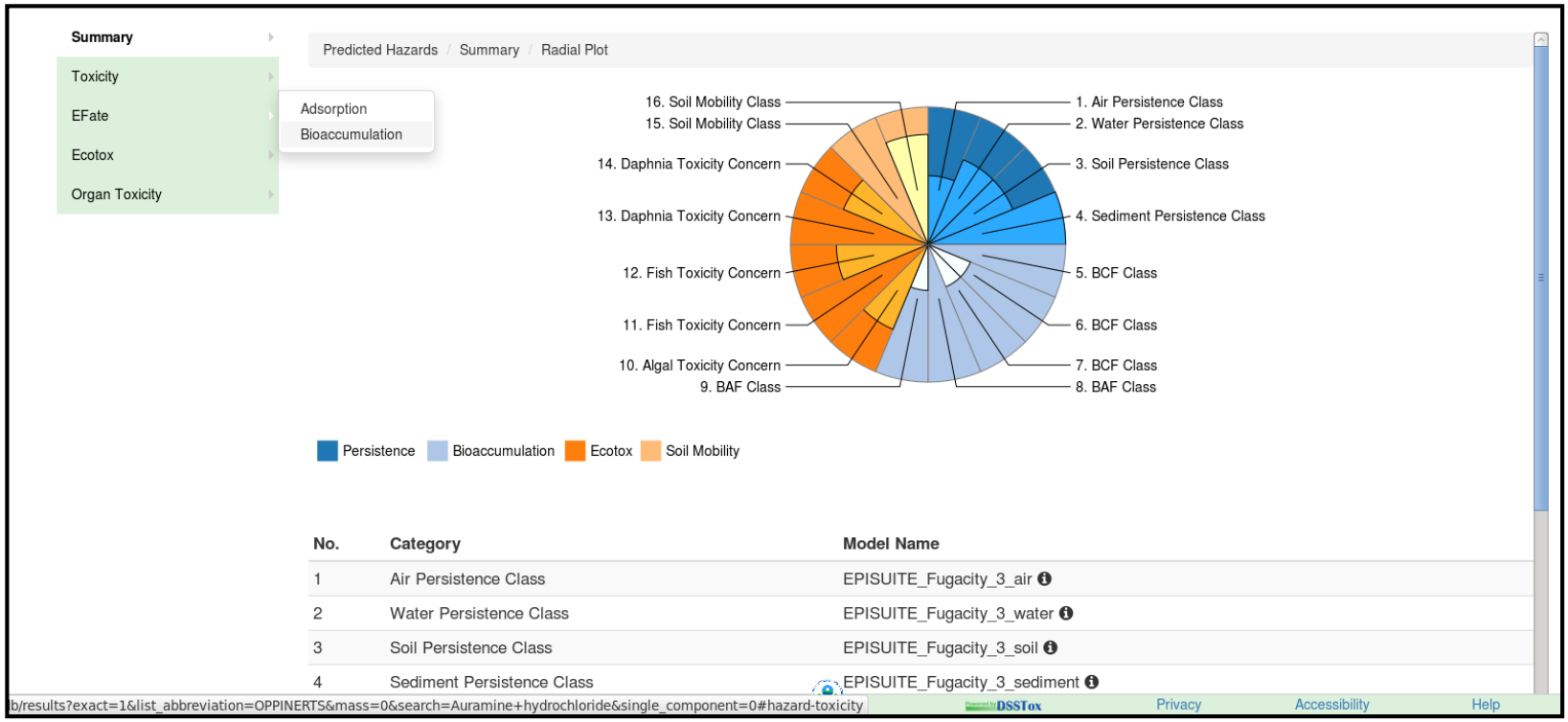
| Parameter | Value |
|---|---|
| Endpoint | Fathead minnow $LC_{50}$ (96 hr) |
| $r^2$ | 0.793 |
| $q^2$ | 0.733 |
| #chemicals | 101 |
| Model | Model # 1296 |



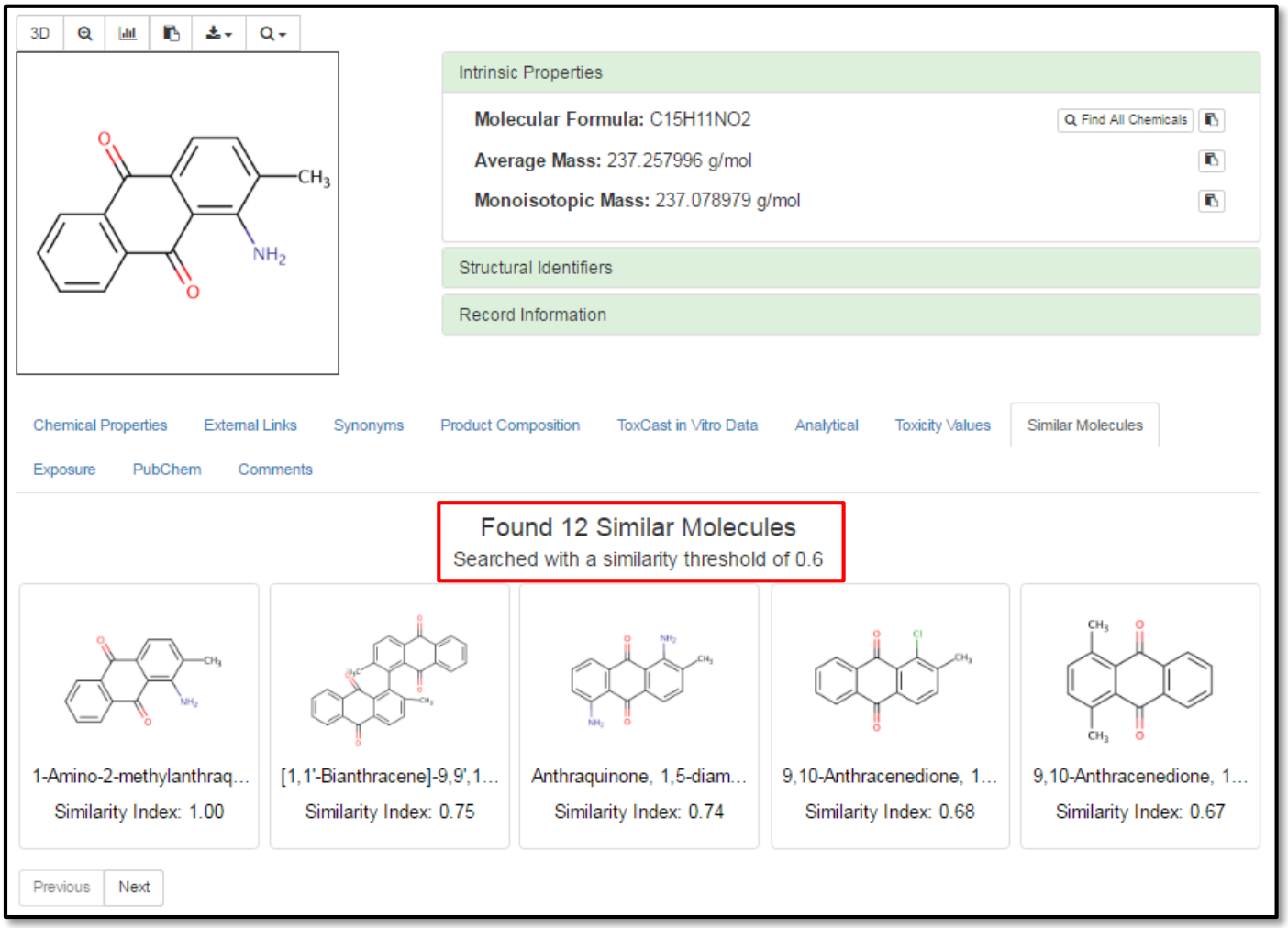
# Future Work

- Curation manuscript presently going through clearance – all data and models to be made available as Open Data
- “Real-time prediction” using NCCT_Models: single or list-based calculations (SDF/Excel)
- Access to data via API/web services
- Complete T.E.S.T. physchem predictions
- Integrate environmental fate and toxicity predictions

# Work in Progress: Environmental Fate, Transport and Toxicity



# Work in Progress: Analog Identification and Similarity Search



# Conclusion

- The CompTox dashboard is an entry point for curated physchem data and the resulting NCCT_Models (>720k chemicals)
- Inclusion of properties from other EPA prediction modules (i.e. T.E.S.T, is under way)
- Full performance statistics available for all models
- The dashboard will become a data dissemination hub for experimental and predicted data as well as direct access to various types of prediction models

# Acknowledgements



EPA NCCT

Imran Shah
Chris Grulke
Jeff Edwards
Ann Richard
Jordan Foster
Jennifer Smith
Richard Judson
Grace Patlewicz
John Wambaugh
Michelle Krzyzanowski

EPA NRMRL

Todd Martin