

Sharing chemical structures with peer-reviewed publications. Are we there yet?

Antony Williams

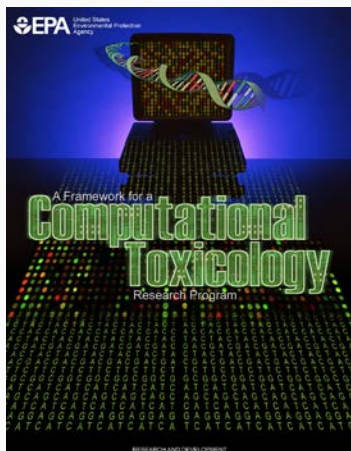
National Center for Computational Toxicology, U.S. Environmental Protection Agency, RTP, NC

The views expressed in this presentation are those of the author and do not necessarily reflect the views or policies of the U.S. EPA

*March 2018
ACS Spring Meeting, New Orleans*

- Mention of or referral to commercial products or services, and/or links to non-EPA sites **does not imply official EPA endorsement** of or responsibility for the opinions, ideas, data, or products presented at those locations, or guarantee the validity of the information provided.























National Center for Computational Toxicology



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

We publish a lot...

And lots of chemistry...

 United States Environmental Protection Agency		Home	About Us	Contact Us
Publications	Research Project	Altmetric	PlumX	Kudos
Tal, T., C. Kilty, A. Smith, C. LaLone, B. Kennedy, A. Tennant, C. McCollum, M. Bondesson, T. Knudsen, S. Padilla, and N. Kleinstreuer. Screening for angiogenic inhibitors in zebrafish to evaluate a predictive model for developmental vascular toxicity. REPRODUCTIVE TOXICOLOGY. Elsevier Science Ltd, New York, NY, USA. (2016). doi:10.1016/j.reprotox.2016.12.004	Virtual Tissues, vEmbryo			
 Abstract Views 9  Citations 6  Downloads N/A				
Mceachran, A., J. Sobus, and A. Williams. (Analytical and Bioanalytical Chemistry) Identifying known unknowns using the US EPAs CompTox Chemistry Dashboard. Analytical and Bioanalytical Chemistry. Springer, New York, NY, USA, 1-7, (2016). doi:10.1007/s00216-016-0139-z				
 Abstract Views 19  Citations 11  Downloads N/A				
Patlewicz, Grace; Casati, Silvia; Basketter, David A.; Asturiol, David; Roberts, David W.; Lepoittevin, Jean-Pierre; Worth, Andrew P.; Aschberger, Karin. Can currently available non-animal methods detect pre and pro-haptens relevant for skin sensitization?. REGULATORY TOXICOLOGY AND PHARMACOLOGY. 82(), (2016). doi:10.1016/j.yrtph.2016.08.007				
 Abstract Views 184  Citations 9  Downloads N/A				
Cowden, J., and J. Lee. Relationships between Arsenic Concentrations in Drinking Water in U.S. Counties and Lung and Bladder Cancer Incidence: Supplemental Material. ENVIRONMENTAL HEALTH PERSPECTIVES. National Institute of Environmental Health Sciences (NIEHS), Research Triangle Park, NC, USA, TBD, (2016). doi:10.1038/ehp.2016.58	Human Health Risk Assessment			

The project I work on...

<https://comptox.epa.gov>

Chemistry Dashboard

Submit Comment

Share ▾

Copy ▾

Aa ▾

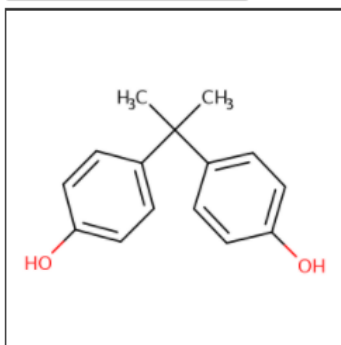
Aa

Aa ▴

Bisphenol A

80-05-7 | DTXSID7020182

© Searched by Approved Name: Found 1 result for 'bisphenol A'.



Wikipedia

Bisphenol A (BPA) is an organic synthetic compound with the chemical formula $(\text{CH}_3)_2\text{C}(\text{C}_6\text{H}_4\text{OH})_2$ belonging to the group of diphenylmethane derivatives and bisphenols, with two hydroxyphenyl groups. It is a colorless solid that is soluble in organic solvents, but poorly soluble in water. It has been in commercial use since 1957. BPA is employed to make certain plastics and epoxy resins. BPA-based plastic is clear and tough...[Read more](#)

Intrinsic Properties

Structural Identifiers

Linked Substances

Presence in Lists

Record Information

Quality Control Notes

Executive Summary (Beta)

Chemical Properties

Env. Fate/Transport

Hazard

ADME (Beta)

Exposure

Bioassays

Similar Compounds

Related Substances

Synonyms

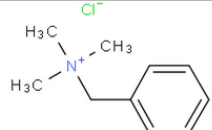
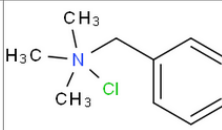
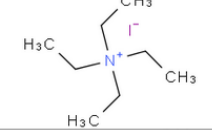
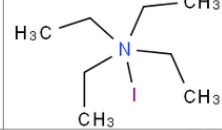


Literature

Links

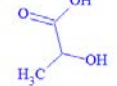
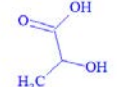
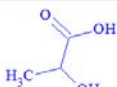
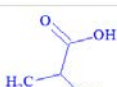
Comments

We Curated Public Data to Create the Models – STANDARDS!

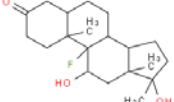
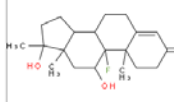
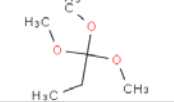

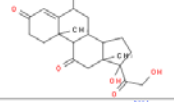
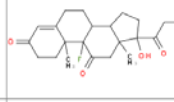
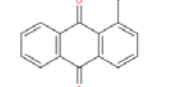
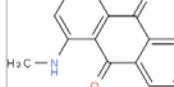
Public data should be curated prior to modeling

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	EpiMP	ErrorMP
	C ₃ H ₅ O ₃	90.0779	000050-21-5	LACTIC ACID	1.6000000000000000e+001	2.2600000000000000e+001	5.8000000000000000e+000
	C ₃ H ₅ O ₃	90.0779	000079-33-4	L-LACTIC ACID	5.3000000000000000e+001	2.2600000000000000e+001	-3.0340000000000000e+001
	C ₃ H ₅ O ₃	90.0779	000590-02-3	2-HYDROXYPROPIONIC ACID	1.6000000000000000e+001	2.2600000000000000e+001	4.6000000000000000e+000
	C ₃ H ₅ O ₃	90.0779	010328-41-7	D-LACTIC ACID	5.3000000000000000e+001	2.2600000000000000e+001	-3.0140000000000000e+001

Identical Chemicals

Mol Block	CAS	NAME	Smiles
	000076-43-7	FLUCYMESTERONE	
	000077-99-6	1,1,1-TRIS(4-HYDROXYETHYL)PROPANE	
	000079-60-7	CORTISONE-4A-FLUORO	
	000082-38-2	DISPERSE RED 9	

Mismatches



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

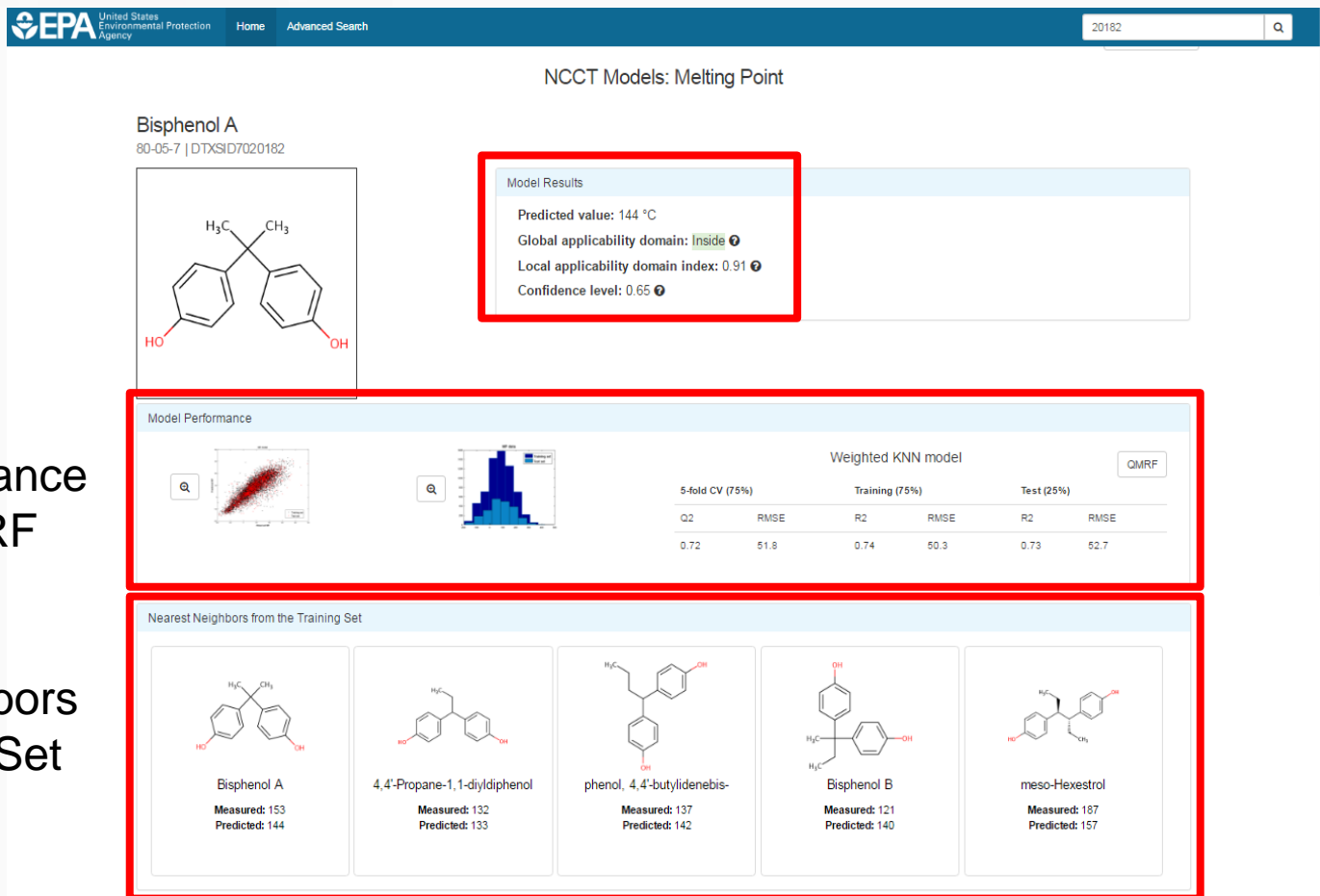
 Download citation

 <http://dx.doi.org/10.1080/1062936X.2016.1253611>

 Check for updates

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

OPEN Models for Prediction



Model Performance
with full QMRP

Nearest Neighbors
from Training Set

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

 Journal of Cheminformatics

RESEARCH ARTICLE

Open Access




OPERA models for predicting physicochemical properties and environmental fate endpoints


Kamel Mansouri^{1,2,3*} , Chris M. Grulke¹, Richard S. Judson¹ and Antony J. Williams¹

- The best way to publish our data as Supplementary Information Files?

- Supplementary information files generally Word Docs and Spreadsheets to PDF
- PDF conversions can be problematic. A recent example of interest



ELSEVIER



DRUG DISCOVERY
TODAY
TECHNOLOGIES

Drug Discovery Today: Technologies Vol. 10, No. 1 2013

Editors-in-Chief
Kelvin Lam – Simplex Pharma Advisors, Inc., Arlington, MA, USA
Henk Timmerman – Vrije Universiteit, The Netherlands

Metabolites: structure determination and prediction

**High-throughput, computer assisted,
specific MetID. A revolution for drug
discovery**

Chemical structures as text

Table 1. Metabolites found in the different incubations tested

Name	RT	m/z	Formula	m/z Diff (ppm)	Mass score	SMILES
Parent	3.13	455.2926	C ₂₇ H ₃₈ N ₂ O ₄	-3.48		<chem>N#CC(CCCN(C)CCc1ccc(OC)c(OC)c1)(C(C)C)c2ccc(OC)c(OC)c2</chem>
M6 -164	2.26	291.2077	C ₁₇ H ₂₆ N ₂ O ₂	-1.37	429	<chem>N(C)CCCC(C#N)(C(C)C)c1ccc(OC)c(OC)c1</chem>
M16 -14	3.06	441.2743	C ₂₆ H ₃₆ N ₂ O ₄	2.41	534	<chem>c1cc(CCNCCCC(C#N)(C(C)C)c2ccc(OC)c(OC)c2)cc(OC)c1OC</chem>
M14 +16	2.92	471.2866	C ₂₇ H ₃₈ N ₂ O ₅	-1.59	476	<chem>N#CC(CCCN(C)CC(O)c1ccc(OC)c(OC)c1)(C(C)C)c2ccc(OC)c(OC)c2</chem>
M9 -14	2.78	441.2761	C ₂₆ H ₃₆ N ₂ O ₄	-1.7	590	<chem>C(#N)C(CCCN(C)CCc1ccc(OC)c(OC)c1)(C(C)C)c2ccc(O)c(OC)c2</chem>
M11 -14	2.84	441.2742	C ₂₆ H ₃₆ N ₂ O ₄	2.57	473	<chem>Oc1ccc(CCN(C)CCCC(C#N)(C(C)C)c2ccc(OC)c(OC)c2)cc1OC</chem>
M12 +2	2.87	457.2707	C ₂₆ H ₃₆ N ₂ O ₅	-0.93	570	<chem>O(C)c1cc(ccc1OC)C(C#N)(CCCNCC(O)c2ccc(OC)c(OC)c2)C(C)C</chem>
M5 -178	2.2	277.1894	C ₁₆ H ₂₄ N ₂ O ₂	7.84	419	<chem>C(C)(C)C(C#N)(CCCN)c1ccc(OC)c(OC)c1</chem>
M8 +2	2.67	457.2708	C ₂₆ H ₃₆ N ₂ O ₅	-1.18	581	<chem>OC(CN(C)CCCC(C#N)(C(C)C)c1ccc(OC)c(OC)c1)c2ccc(O)c(OC)c2</chem>
M15 -14	2.92	441.2743	C ₂₆ H ₃₆ N ₂ O ₄	2.23	614	<chem>Oc1ccc(CCN(C)CCCC(C#N)(C(C)C)c2ccc(OC)c(OC)c2)cc1OC</chem>
M2 -259	0.73	196.1326	C ₁₁ H ₁₇ NO ₂	5.91	618	<chem>c1(CCNC)ccc(OC)c(c1)OC</chem>
M10 -28	2.8	427.2617	C ₂₅ H ₃₄ N ₂ O ₄	-4.79	487	<chem>c1(OC)cc(ccc1OC)C(C#N)(CCCNCCc2ccc(O)c(OC)c2)C(C)C</chem>
M7 +2	2.46	457.2717	C ₂₆ H ₃₆ N ₂ O ₅	-3.23	492	<chem>c1cc(CCN(O)CCCC(C#N)(C(C)C)c2ccc(OC)c(OC)c2)cc(OC)c1OC</chem>
M17 +16	3.21	471.2853	C ₂₇ H ₃₈ N ₂ O ₅	1.19	534	<chem>N#CC(CCCN(C)CCc1ccc(OC)c(OC)c1)(c2ccc(OC)c(OC)c2)C(C)(C)O</chem>
M4 -178	1.86	277.1927	C ₁₆ H ₂₄ N ₂ O ₂	-3.8	444	<chem>COc1cc(ccc1O)C(C#N)(CCCN)C(C)C</chem>
M1 -289	0.44	166.0858	C ₉ H ₁₁ NO ₂	5.93	136	<chem>c1(CCNC)ccc(=O)c(c1)=O</chem>
M13 -16	2.88	439.2603	C ₂₆ H ₃₄ N ₂ O ₄	-1.43	367	<chem>c1cc(CC=NCCCC(C#N)(C(C)C)c2ccc(OC)c(OC)c2)cc(OC)c1OC</chem>

How do I extract structures?

- Copy and paste into Excel as a start point
- Assume no loss of formatting!
- Convert SMILES to structures

How do I extract structures?

- Copy and paste into Excel as a start point
- Assume no loss of formatting !
- Convert SMILES to structures
- But Copy-Paste doesn't work

```
c1(CCNC)ccc(=O)c(c1)=O
```

```
c1cc(CC=NCCCC(C#N)(C(C)C)c2ccc(OC)c(OC)c2)cc(OC)c1OC
```

```
c1(CCNC)ccc( O)c(c1) O
```

```
c1cc(CC NCCCC(C#N)(C(C)C)c2ccc(OC)c(OC)c2)cc(OC)c1OC
```

How do I extract structures?

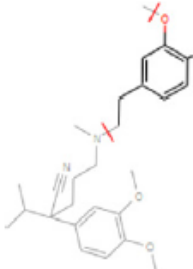
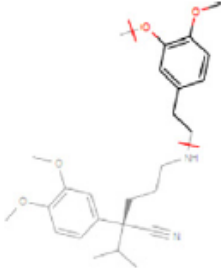
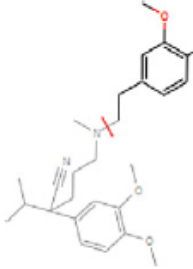
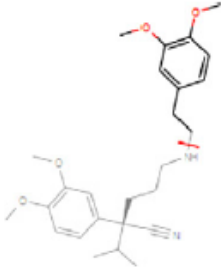
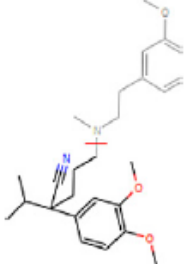
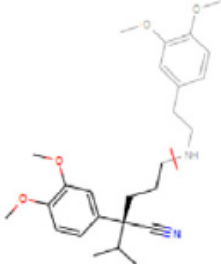
- Copy and paste into Excel as a start point
- Assume no loss of formatting !
- Convert SMILES to structures
- But Copy-Paste doesn't work

<chem>c1(CCNC)ccc(=O)c(c1)=O</chem>
<chem>c1cc(CC=NCCCC(C#N)(C(C)C)c2ccc(OC)c(OC)c2)cc(OC)c1OC</chem>

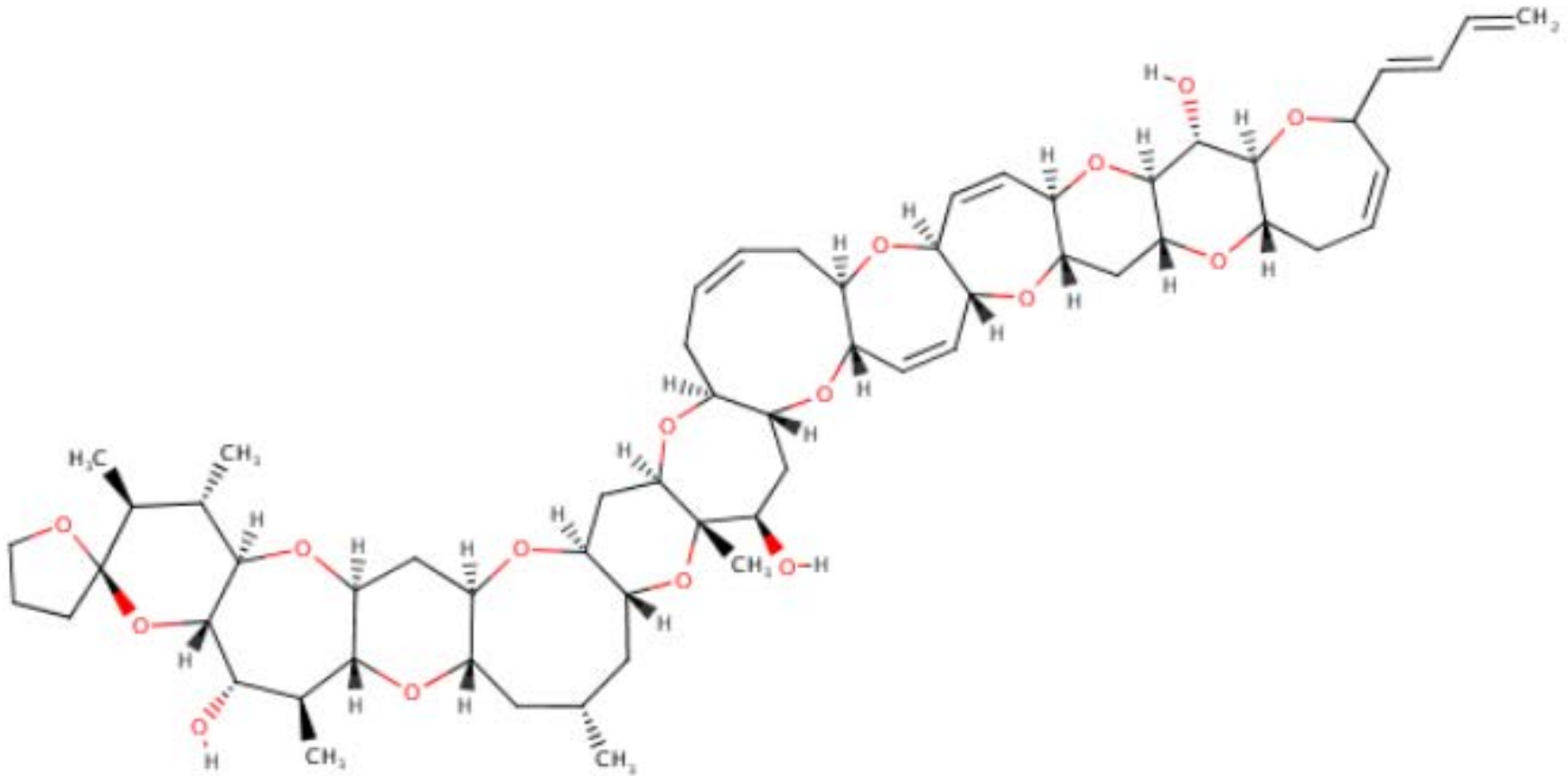
```
c1(CCNC)ccc( O)c(c1) O
c1cc(CC NCCCC(C#N)(C(C)C)c2ccc(OC)c(OC)c2)cc(OC)c1OC
```

Extract Structure Drawings???

Table 2. Selection of fragments that help in the M16-16 metabolite structure elucidation

Sub. obs. m/z	Sub. cal. m/z	Sub. m/z diff. ppm	Substrate	Metabolite	Δ	Met. obs. m/z	Met. calc. m/z	Met. m/z diff. ppm
150.0664	150.0681	11.42			+0	150.0670	150.0681	7.25
165.0869	165.0916	28.22			+0	165.0892	165.0916	14.30
260.1637	260.1651	5.33			+0	260.1652	260.1651	-0.50

Try hand-drawing Algal Toxins!



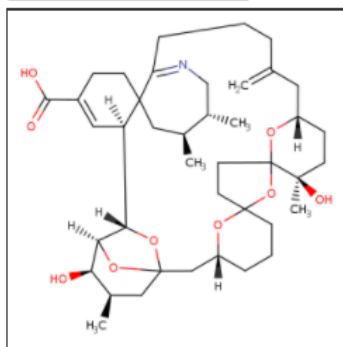
Think of files in multiple formats!

- SMILES are hyper-dependent on good layout algorithms. It's not easy!

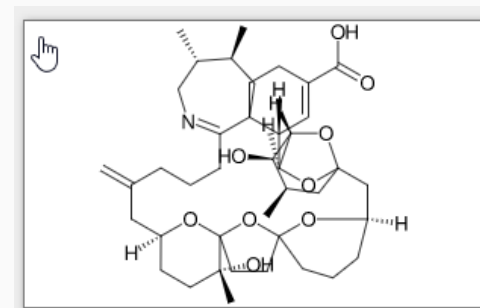
Pinnatoxin A

160759-36-4 | DTXSID40880101

© Searched by DSSTox_Substance_Id: Found 1 result for



[H][C@]12OC3(C[C@
@H](C)[C@H]1O)C[C
@]1([H])CCCC4(CC
C5(O4)O[C@]([H])(
CC[C@]5(C)O)CC(=
C)CCCC4=NC[C@H](
C)[C@]@H](C)CC44C
CC(=C[C@]4([H])[C@]
2([H])O3)C(O)=O)O1



Names and CASRNs are **NOT** structures

- In our domain most chemicals are text – chemical names and CAS Numbers

Attachment D (Method 3)

SIM quantitation ions and qualifiers for internal standards, references method analysis, and surrogates

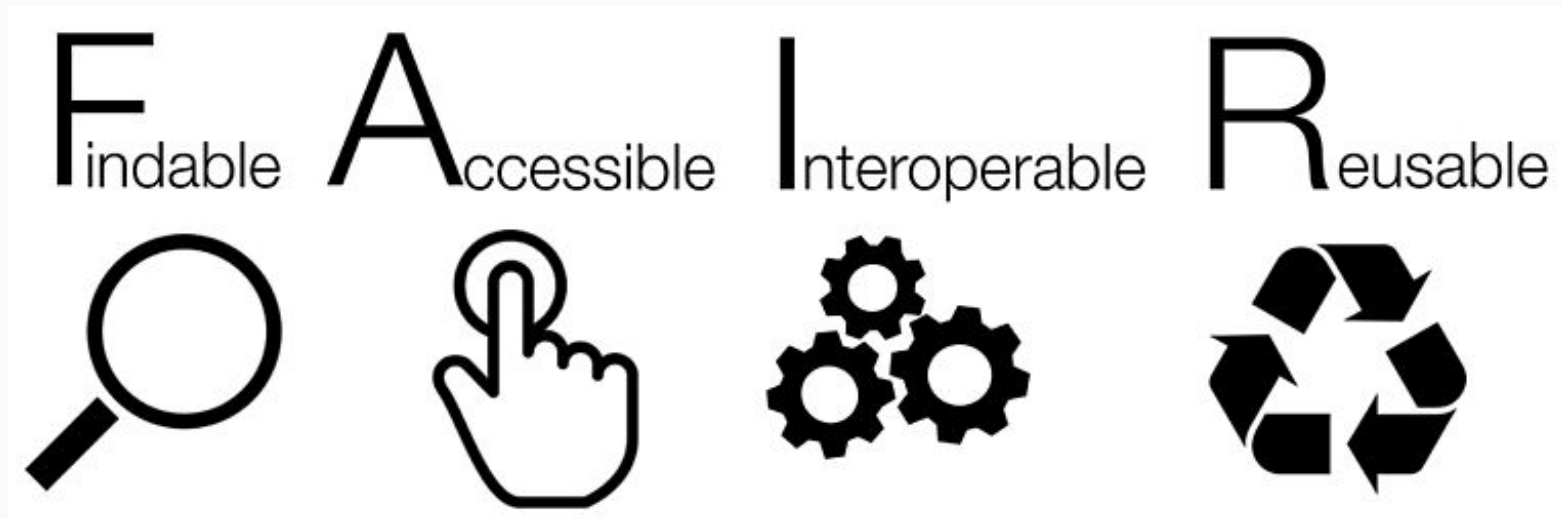
<u>Name of Compound</u>	<u>CAS No.</u>	<u>Quantitation Ion</u>	<u>Qualifier Ions</u>
Phenol-d6 (SS)	13187-88-3	99	71, 42
Phenol	108-95-2	94	66
1,4-Dichlorobenzene	106-46-0	146	111, 75, 50
Acetophenone	98-86-2	105	77, 51, 120
Acenaphthene-d10 (IS)	15067-26-2	162	160, 80
p-Cresol	106-44-5	107	108, 77
Isophorone	78-59-1	82	138, 54
Camphor	76-22-2	95	81, 108, 152
Isoborneol	124-76-5	95	110, 121, 136
Menthol	89, 78, 1	71	81, 123, 138
Naphthalene	91-20-3	128	102, 51
Methyl salicilate	119-36-8	120	92, 152, 65

And generally problematic...

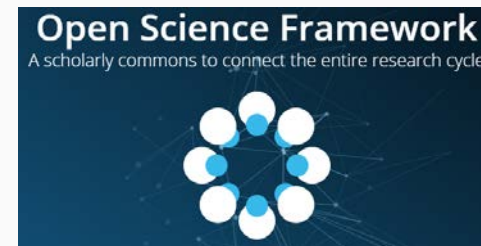
<u>Name of Compound</u>	<u>CAS No.</u>
Phenol-d6 (SS)	13187-88-3
Phenol	108-95-2
1,4-Dichlorobenzene	106-46-0
Acetophenone	98-86-2
Acenaphthene-d10 (IS)	15067-26-2
p-Cresol	106-44-5
Isophorone	78-59-1
Camphor	76-22-2
Isoborneol	124-76-5
Menthol	89, 78, 1
Naphthalene	91-20-3
Methyl salicylate	119-36-8

When publishing data consider..


- Licenses should be an important part of your data distribution considerations
- Digital Object Identifiers for versioning
- Consider your greater contribution to science outside of the publication



Consider Distribution Repositories: Examples



Repositories cover it all...

 [June 30, 2017 \(5.2\)](#) [Software](#) [Open Access](#) [View](#)

CMAQ


US EPA Office of Research and Development;

The Community Multiscale Air Quality (CMAQ) model is an active open-source development project of the U.S. EPA that consists of a suite of programs for conducting air quality model simulations. CMAQ combines current knowledge in atmospheric science and air quality modeling, multi-proc

Uploaded on February 6, 2018

[4 more version\(s\) exist for this record](#)

VERSIONING

 **Publication date:**
June 30, 2017

DOI:
[DOI 10.5281/zenodo.1167892](#) **DOIs**

Keyword(s):
[EPA](#) [National Exposure Research Laboratory](#)
[air quality model](#) [ozone](#) [particulate matter](#)
[acid deposition](#) [toxics](#)

License (for files):
[Creative Commons Attribution 4.0](#) **LICENSING**

- Generally store files on our FTP site PLUS copies in a repository (or two)
- Multiple formats of data as appropriate
 - Can be as supplementary data or DOI'd data files
- DOI'd data gives altmetrics also..

Mapping file of InChIStrings, InChIKeys and DTXSIDs for the EPA CompTox Dashboard
12.08.2016, 18:38 by Antony Williams


The foundation of chemical safety testing relies on chemistry information such as high-quality chemical structures and physical chemical properties. This information is used by scientists to predict the potential health risks of chemicals.

The iCSS CompTox Dashboard is part of a suite of dashboards developed by EPA to help evaluate the safety of chemicals. The dashboard provides access to a variety of information on over 700,000 chemicals currently in use.

Within the 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.

This data are compiled from sources including the EPA's computational toxicology research databases, and public domain databases such as the National Center for Biotechnology Information's PubChem database.

699 views | 161 downloads | 2 citations



CATEGORIES


- Cheminformatics
- Computational Chemistry
- Environmental Chemistry

KEYWORD(S)

EPA CompTox Dashboard DTXSID

InChIKeys Environmental Chemistry


EPA

[Browse](#)[Upload](#)[Sign up](#)[Log in](#)

File(s) stored somewhere else

ftp://newftp.epa.gov/COMPTOX/NCCT_Publication_Data/Williams_A/Opera_Model_Paper/

Please note: Linked content is NOT stored on figshare and we can't guarantee its availability, quality, security or accept any liability.

 This item is shared privately

OPERA Model Paper Data

08.03.2018, 19:33

Data associated with the OPERA Model paper authored by Kamel Mansouri and Tony Williams.


CATEGORIES

- Toxicology

KEYWORD(S)

- OPERA
- Computational Toxicology
- NCCT

LICENCE

 CC0

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

 Journal of Cheminformatics

RESEARCH ARTICLE

Open Access



OPERA models for predicting physicochemical properties and environmental fate endpoints

Kamel Mansouri^{1,2,3*} , Chris M. Grulke¹, Richard S. Judson¹ and Antony J. Williams¹

Availability of data and materials

ftp://newftp.epa.gov/COMPTOX/Sustainable_Chemistry_Data/Chemistry_Dashboard/OPERA. The dataset supporting the conclusions of this article are available within the article and its additional file as well as OPERA's Github repository (<https://github.com/kmansouri/OPERA>) and the EPA's ftp site [150]:
<https://figshare.com/s/6fa1babbc9a0e9560317>.

- What has changed?
 - Cheminformatics has progressed
 - The internet proliferates data access
 - Standards have progressed (InChI) and are improving
 - Anybody can access/download millions of structures!
- What hasn't changed?
 - Minor progress presenting structures in publications
 - Delivery via PDFs still dominates
 - Mandates on scientists are very unlikely

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>