http://www.orcid.org/0000-0002-2668-4821
@ChemConnector on Twitter

Structure Identification Using High Resolution Mass Spectrometry Data and the EPA Chemistry Dashboard

Antony J. Williams[†], *Andrew McEachran, Jon Sobus, Chris Grulke, Jennifer Smith, Michelle Krzyzanowski, Jordan Foster and Jeff Edwards*

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

August 21-25, 2016
ACS Fall Meeting, Philadelphia, PA

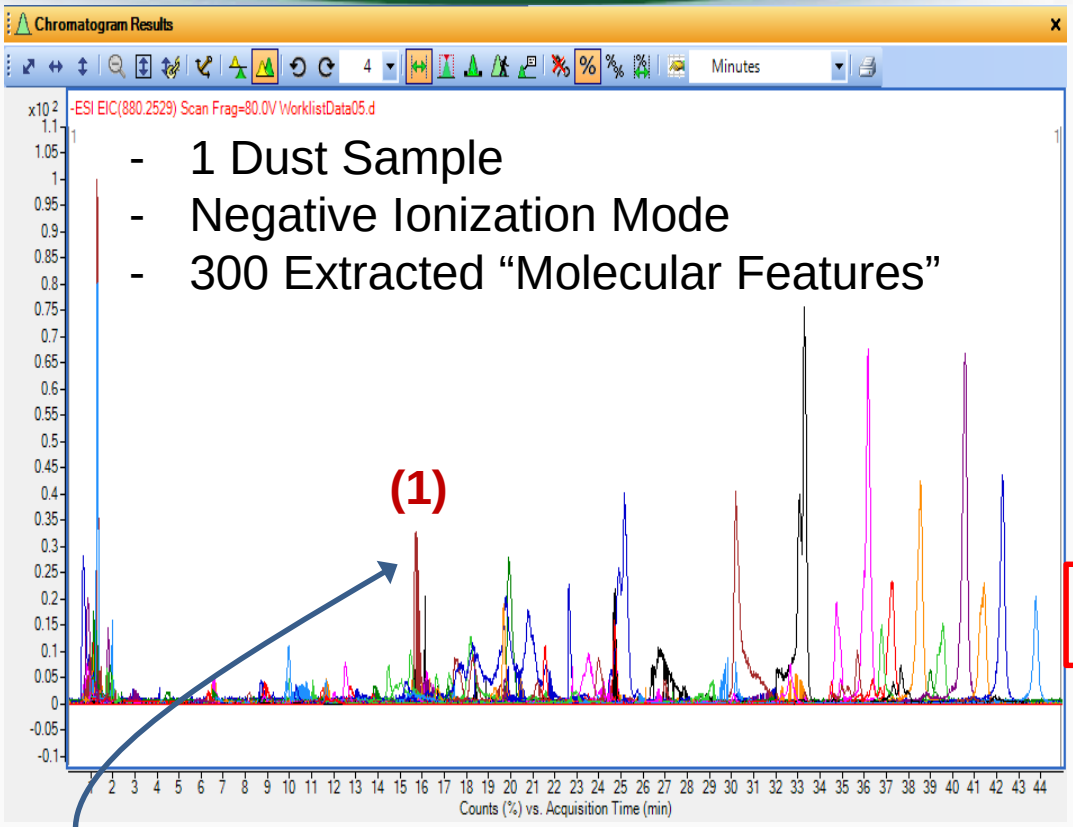
Comparing Analysis Approaches

- Targeted Analysis:
  - We know exactly what we're looking for
  - 10s – 100s of chemicals

- Suspect Screening Analysis (SSA)
  - We have chemicals of interest
  - 100s – 1,000s of chemicals

- Non-Targeted Analysis (NTA):
  - We have no preconceived lists
  - 1,000s – 10,000s of chemicals
  - In dust, soil, food, air, water, products, plants, animals, and…us!!

General Goals of SSA/NTA

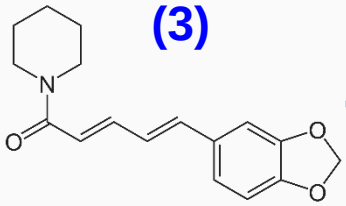


- 1 Dust Sample
- Negative Ionization Mode
- 300 Extracted "Molecular Features"

(1)

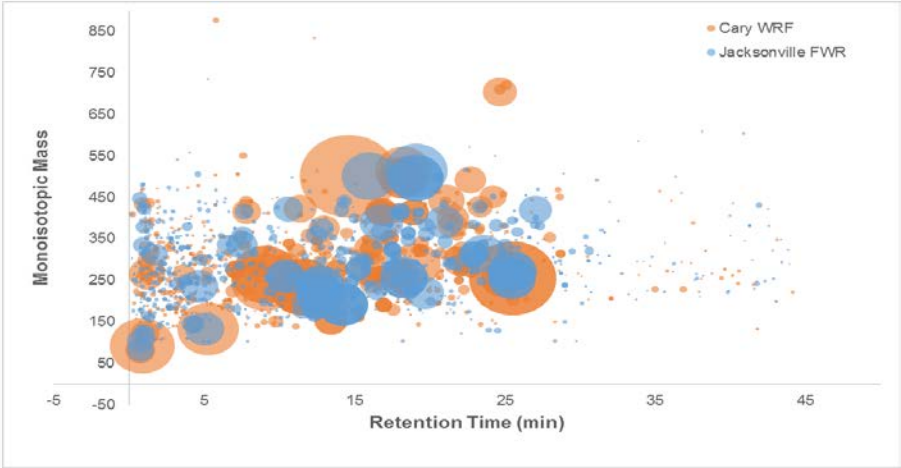
1) **Prioritize "Molecular Features"**
2) **Correctly assign formulas**
3) **Correctly assign structures**
4) **Determine chemical sources**
5) **Predict chemical concentrations**

EXPOSURE

(2) **C17H19NO3** → **(3)**  → **(4)**  → **(5)** **12 µg/g**

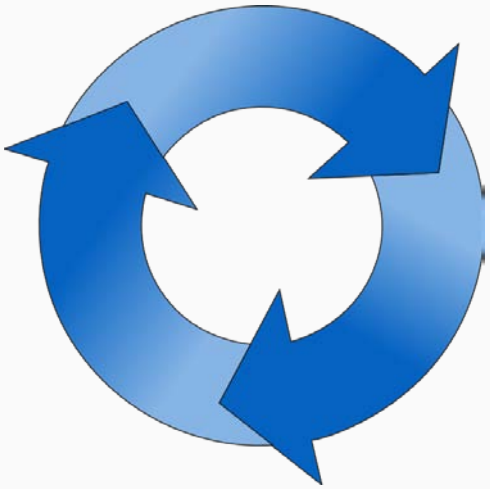
Non-targeted analysis challenges

- 3000-5000 molecular features in a given sample
- Current technologies can identify up to 5%
- How can we improve identification???
  - Simple workflows
  - Reliable formula prediction (Instrument)
  - Accurate ranking of likelihood (Databases)



The General Approach

Analytical Instruments



Comp. Tools & Workflows

FOR IDENT

Xcms Online

MetFrag

Databases

ChemSpider
Search and share chemistry

PubChem

MassBank

Previous Work with Suspect-Screening

Environment International 88 (2016) 269–280

Contents lists available at ScienceDirect

Environment International

journal homepage: www.elsevier.com/locate/envint

Linking high resolution mass spectrometry data with exposure and toxicity forecasts to advance high-throughput environmental monitoring

Julia E. Rager [a], Mark J. Strynar [b], Shuang Liang [a], Rebecca L. McMahen [a], Ann M. Richard [c], Christopher M. Grulke [d], John F. Wambaugh [c], Kristin K. Isaacs [b], Richard Judson [c], Antony J. Williams [c], Jon R. Sobus [b,*]

[a] Oak Ridge Institute for Science and Education (ORISE) Participant, 109 T.W. Alexander Drive, Research Triangle Park, NC 27709, United States
[b] U.S. Environmental Protection Agency, Office of Research and Development, National Exposure Research Laboratory, 109 T.W. Alexander Drive, Research Triangle Park, NC 27709, United States
[c] U.S. Environmental Protection Agency, Office of Research and Development, National Center for Computational Toxicology, 109 T.W. Alexander Drive, Research Triangle Park, NC 27709, United States
[d] Lockheed Martin, 109 T.W. Alexander Drive, Research Triangle Park, NC 27709, United States

We're on the Right Path…

- … but certainly room for improvement
- ~thousands of molecular features (not unique)
- 33 confirmed chemicals
- State-of-the-art SSA yields <5% confirmed IDs
- So what else is in these (and other) samples??

2012: Definitive Study of Known-Unknowns using ChemSpider

J. Am. Soc. Mass Spectrom. (2012) 23:179–185
DOI: 10.1007/s13361-011-0265-y

RESEARCH ARTICLE

Identification of "Known Unknowns" Utilizing Accurate Mass Data and ChemSpider

Table 1. Searching ChemSpider by Elemental Composition then Sorting by Number of Associated References

Class of compounds	Number compounds in class	Position of compound sorted in descending order by number of references					
		#1	#2	#3	#4	#5	>#5
Drugs	45	43	1	1			
Pesticides	8	7	1				
Toxins	2	2					
Polymer antioxidants	15	15					
Polymer UV stabilizers	10	8	1	1			
Polymer clarifying agent (Irgaclear DM)	1						1(14)
Polyurethane additives	4	2	1			1	
Natural products	3	2		1			
Herbicide (clofibric acid)	1	1					
Artificial sweetener (sucralose)	1	1					
Total compounds ChemSpider	90	81	4	3		1	1
Total compounds CAS Registry [1]	90	84	4	1		1	

2012: Definitive Study of Known-Unknowns using ChemSpider

Table 1. Searching ChemSpider by Elemental Composition then Sorting by Number of Associated References

Class of compounds	Number compounds in class	Position of compound sorted in de		
		#1	#2	#3
Drugs	45	43	1	1
Pesticides	8	7	1	
Toxins	2	2		
Polymer antioxidants	15	15		
Polymer UV stabilizers	10	8	1	1
Polymer clarifying agent (Irgaclear DM)	1			
Polyurethane additives	4	2	1	
Natural products	3	2		1
Herbicide (clofibric acid)	1	1		
Artificial sweetener (sucralose)	1	1		
Total compounds ChemSpider	90	81	4	3
Total compounds CAS Registry [1]	90	84	4	1

ChemSpider

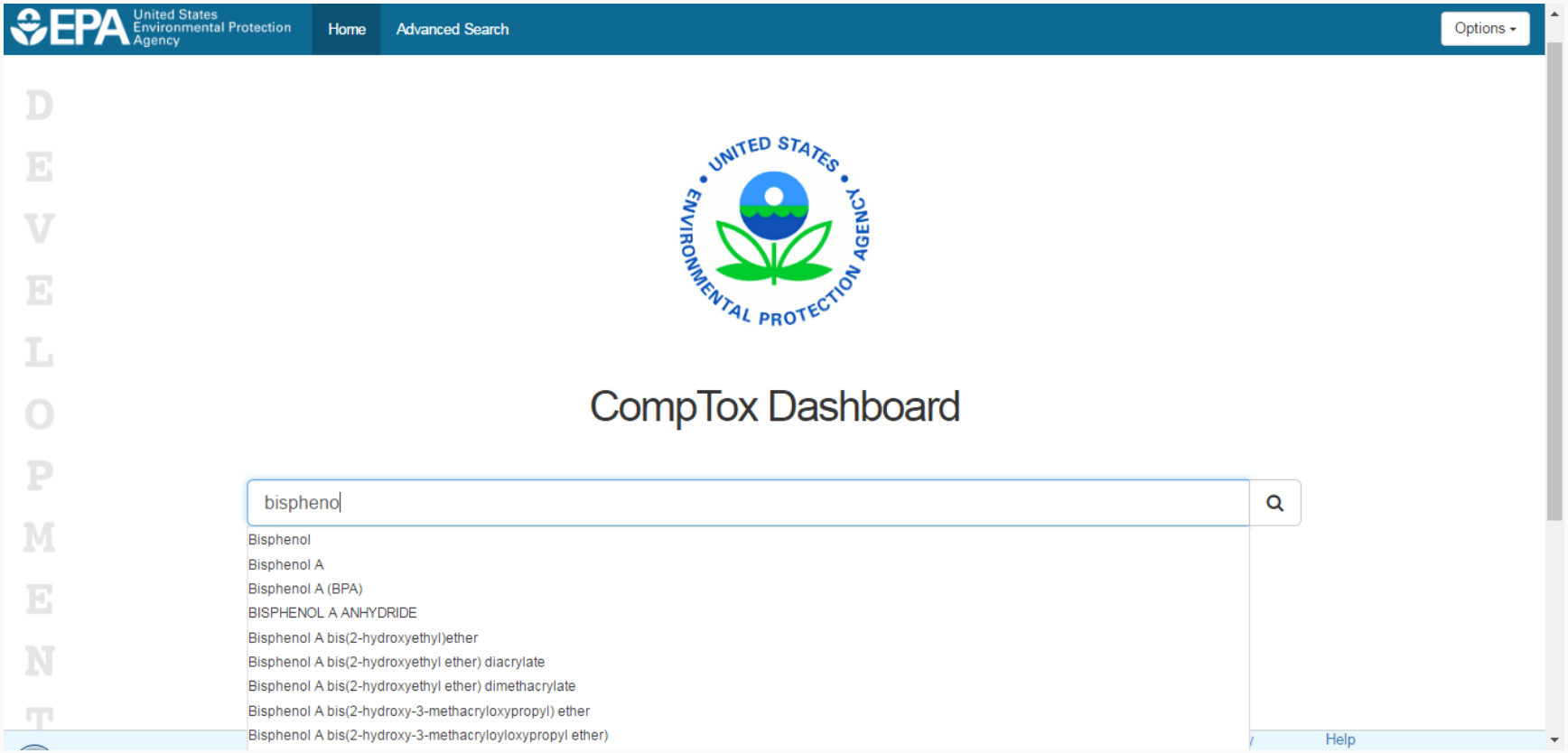
- http://www.chemspider.com
- Grows daily with new depositors and annotations. Example Data Sources

Data Sources

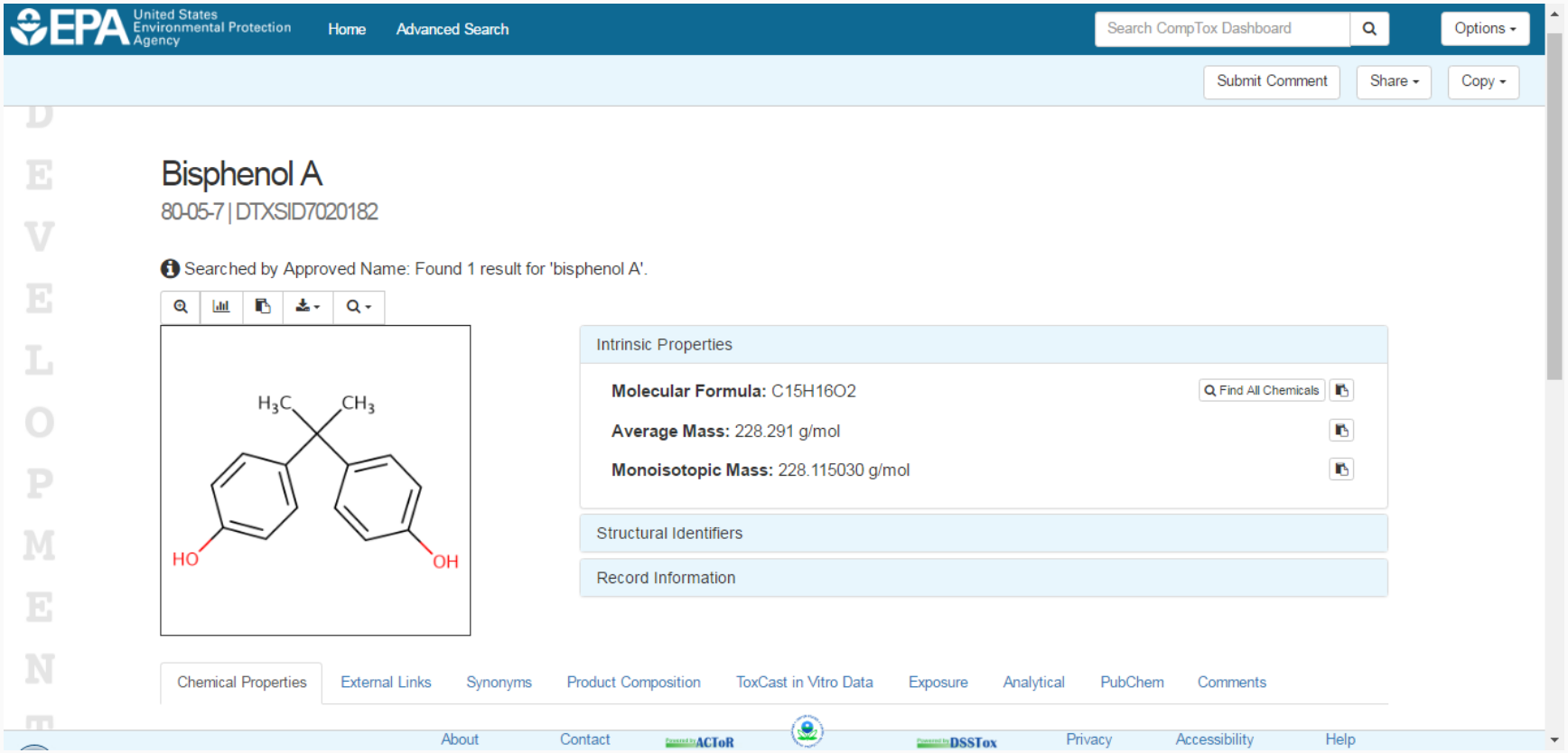
Data Source	Count	Date Created	Last Updated
Aurora Feinchemie	25288289	13/04/2009	12/06/2016
PubChem	10881750	15/04/2008	25/06/2015
AKos	8226932	15/04/2008	15/06/2016
Mcule	5649548	21/01/2014	30/10/2015
Molport	5292029	09/02/2010	02/09/2014
eMolecules	4840366	08/06/2009	30/06/2009

ChemSpider
Search and share chemistry
57 Million
chemical structures

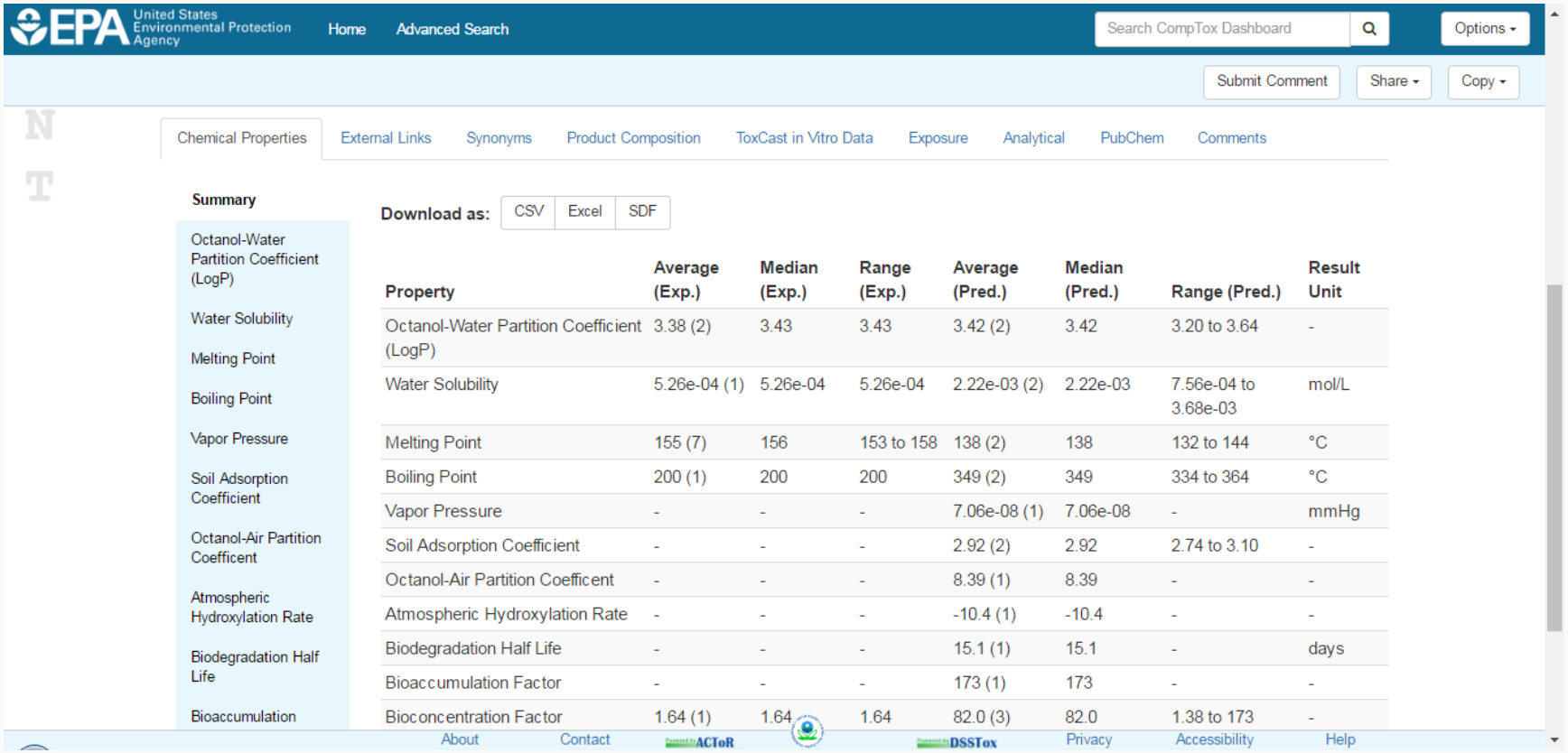
Our **New** Dashboard
https://comptox.epa.gov



Bisphenol A


EPA 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

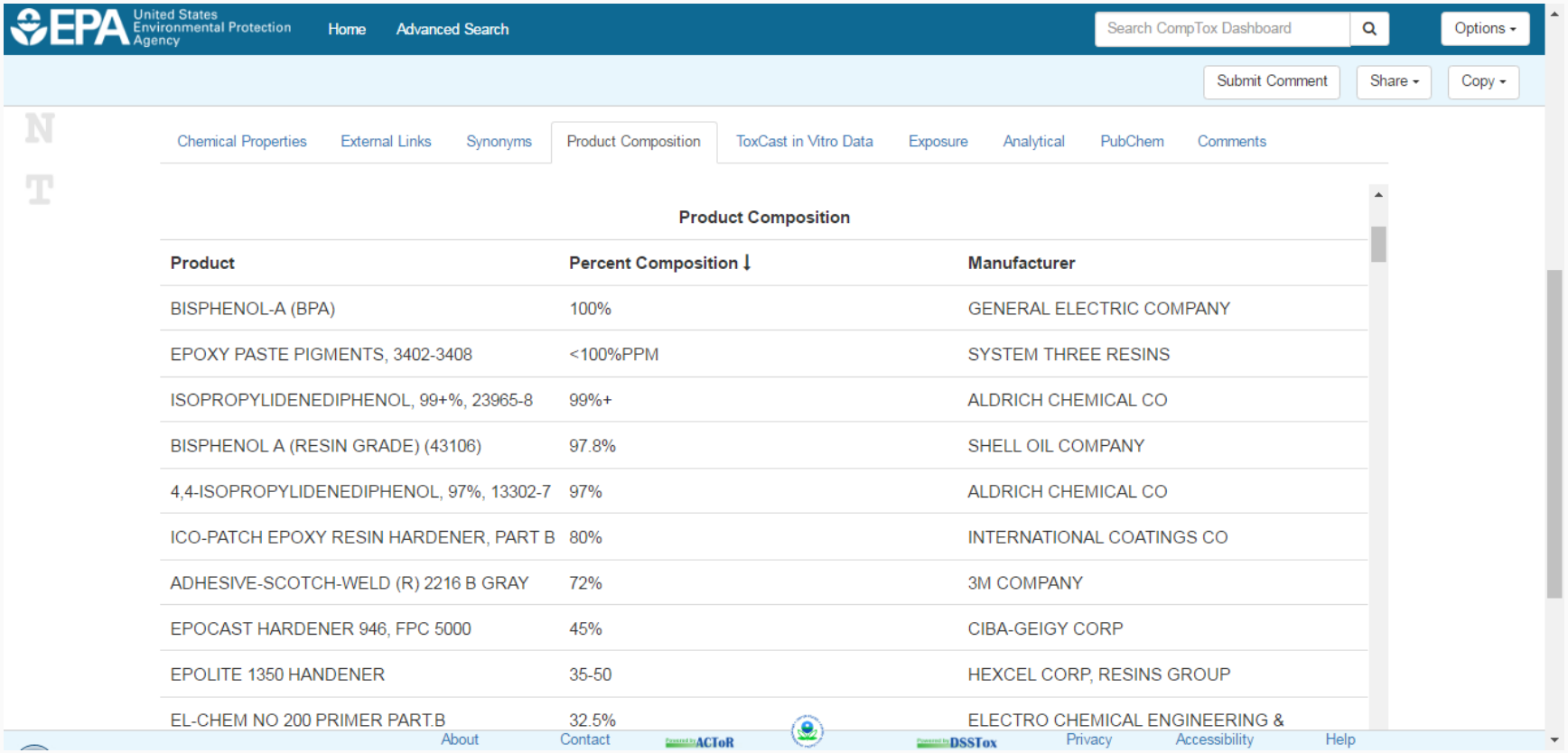
Physicochemical Properties



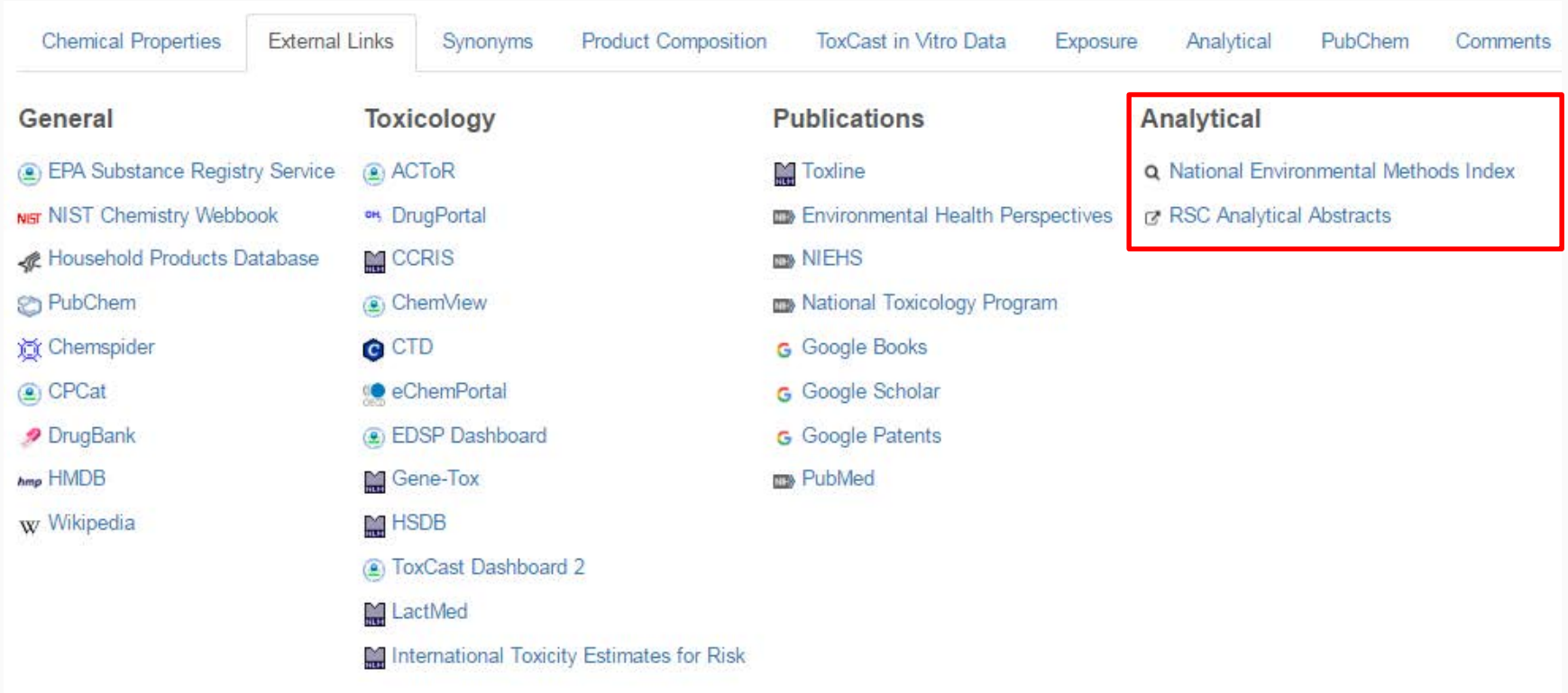
Property	Average (Exp.)	Median (Exp.)	Range (Exp.)	Average (Pred.)	Median (Pred.)	Range (Pred.)	Result Unit
Octanol-Water Partition Coefficient (LogP)	3.38 (2)	3.43	3.43	3.42 (2)	3.42	3.20 to 3.64	-
Water Solubility	5.26e-04 (1)	5.26e-04	5.26e-04	2.22e-03 (2)	2.22e-03	7.56e-04 to 3.68e-03	mol/L
Melting 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

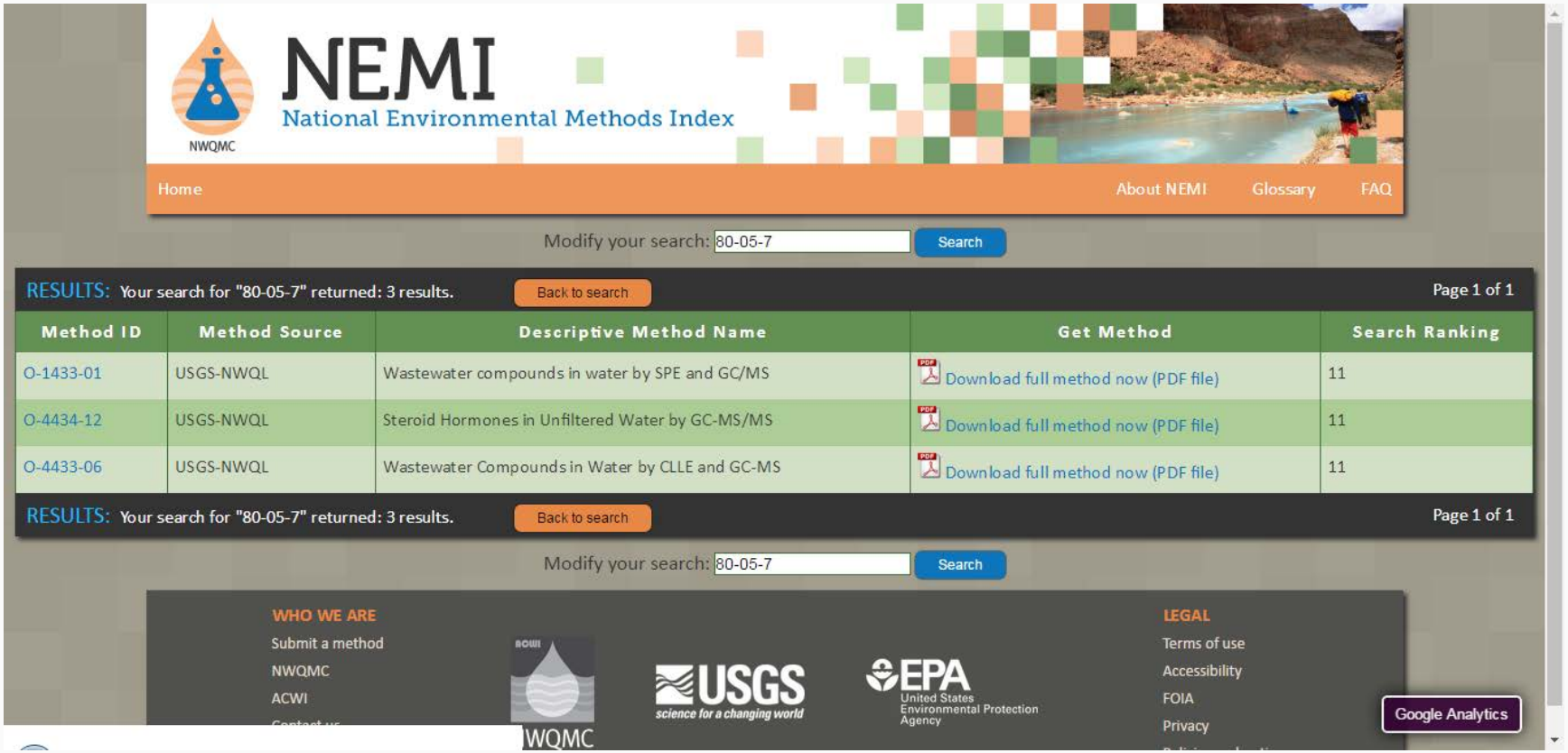
Functional Use and Composition



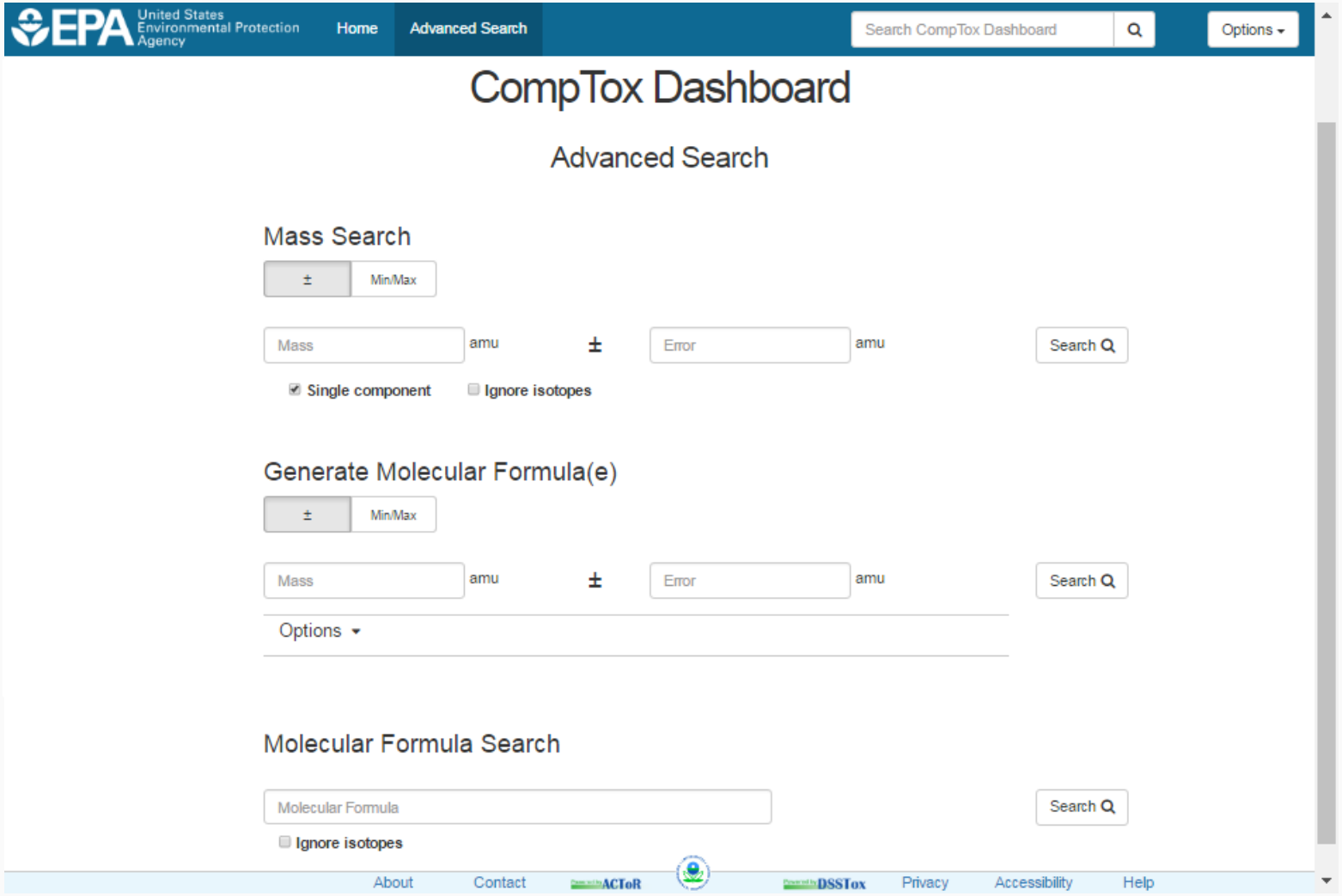
External Links



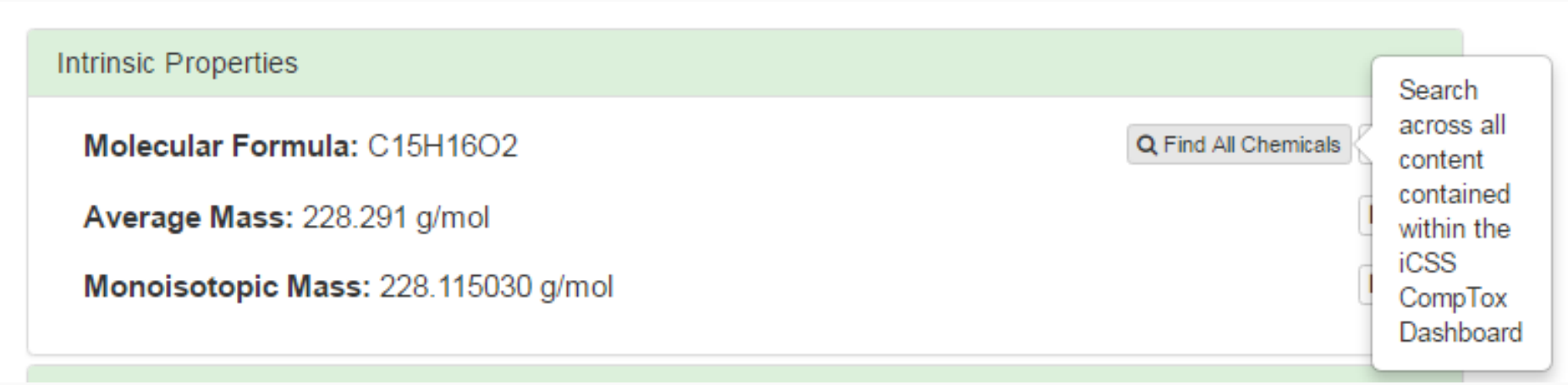
National Environmental Methods Index



Advanced Search



Formula Searching
Formulae matching Bisphenol A



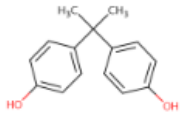
Formula Search Results

Search Results

Searched by Molecular Formul **Found 215 results for 'C15H16O2'.**

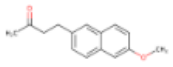
Download as: CSV Excel SDF

Preferred Name	CAS-RN	QC Level	Number of Sources	Monoisotopic Mass



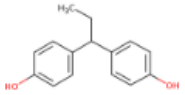
Bisphenol A

80-05-7



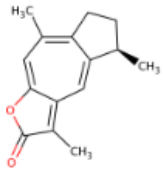
Nabumetone

42924-53-8



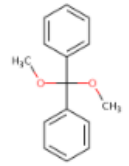
4,4'-Propane-1,1-diyldiphenol

1576-13-2



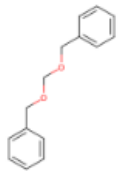
Gweicurculactone

123914-43-2



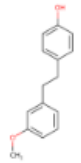
1,1'-(Dimethoxymethylene)bisbe...

2235-01-0



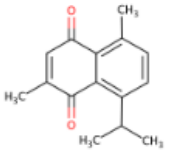
Preventol D2

2749-70-4



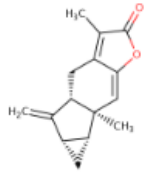
4-Hydroxy-3'-methoxybibenzyl

59586-27-5



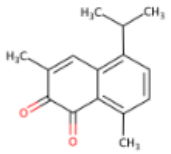
Stahlianthusone

87018-26-6



CHLORANTHALACTONE A

66395-02-6



1,2-Naphthalenedione, 3,8-dime...

5574-34-5

Download to Excel

CompToxAdvancedSearch_2016-08-14_23-46-20.xls - Excel

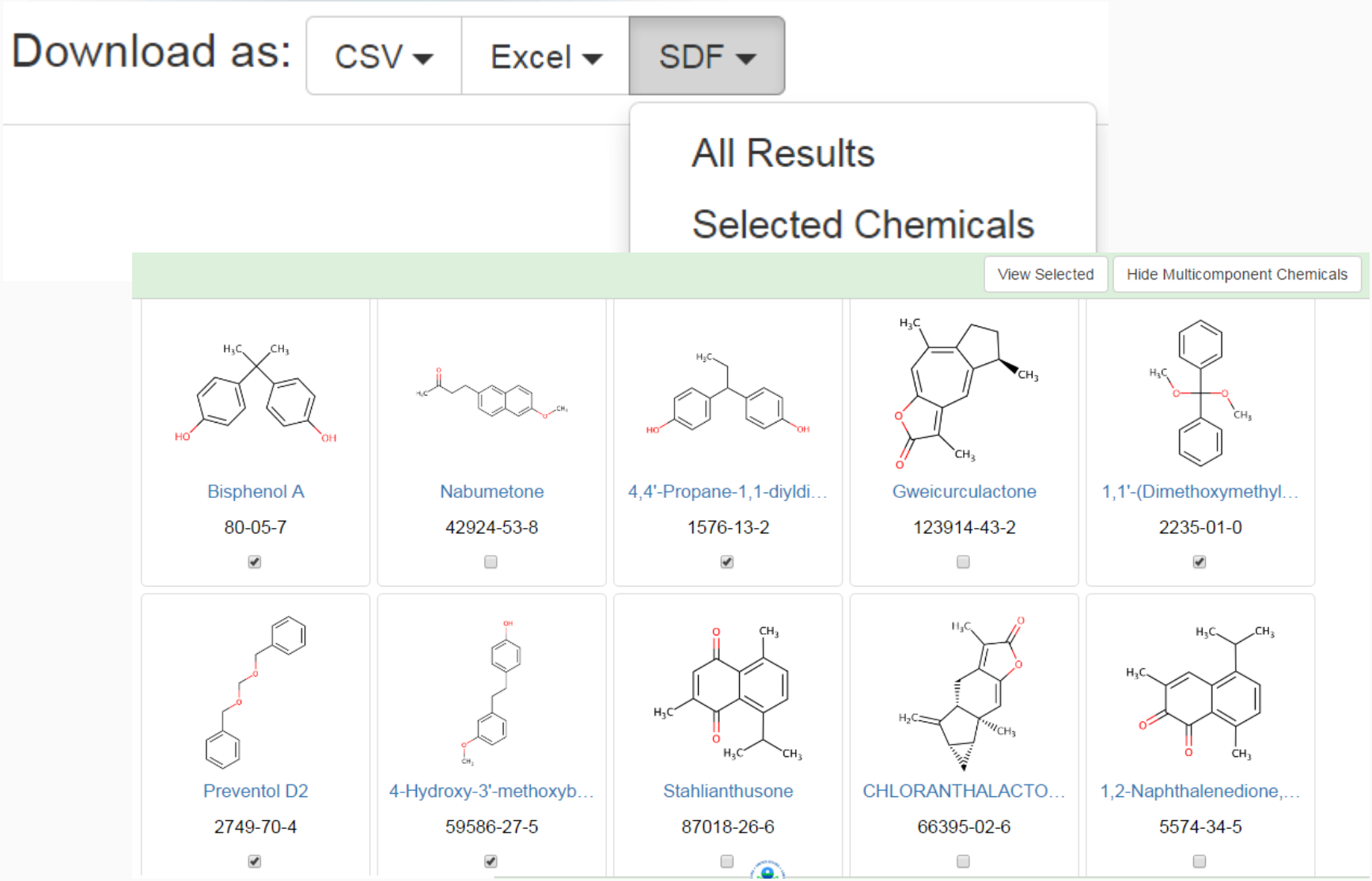
E2: C15H16O2

	A	B	C	D	E	F	G	H	I
1	DTXSID	CAS RN	PREFERRED NAME	IUPAC NAME	FORMULA	MONOISOTOPIC	AVERAGE MASS	SMILES	INCHI KEY
2	DTXSID7020182	80-05-7	Bisphenol A	4,4'-(Propane-2,2-diyl)diphenol	C15H16O2	228.1150298	228.2910004	CC(C)(C1=CC=C(O)C=C1)C1=CC=C(O)C=C1	IISBACLAFKSPIT-UHFFFAOYSA-N
3	DTXSID7042275	837-08-1	Phenol, 2-[1-(4-hydroxyphenyl)-1-	2-[2-(4-Hydroxyphenyl)propan-2-	C15H16O2	228.1150298	228.2910004	CC(C)(C1=CC=C(O)C=C1)C1=C(O)C=CC=C1	MLCQXUZZAXKTSG-UHFFFAOYSA-N
4	DTXSID3044594	1576-13-2	4,4'-Propane-1,1-	4,4'-(Propane-1,1-	C15H16O2	228.1150298	228.2910004	CCC(C1=CC=C(O)C=C1)C1=CC=C(O)C=C1	YKPAABNCNAGAAJ-UHFFFAOYSA-N
5	DTXSID4045472	42924-53-8	Nabumetone	4-(6-Methoxynaphthalen-2-yl)butan-2-one	C15H16O2	228.1150298	228.2910004	COC1=CC2=CC=C(CCC(C)=O)C=C2C=C1	BLXXJMDCKKHMKV-UHFFFAOYSA-N
6	DTXSID0062923	3236-63-3	Phenol, 2,2'-methylenebis[4-methyl-	2,2'-Methylenebis(4-methylphenol)	C15H16O2	228.1150298	228.2910004	CC1=CC(CC2=C(O)C=CC(C)=C2)=C(O)C=C1	XZXYQEHISUMZAT-UHFFFAOYSA-N
7	DTXSID50154143	123914-43-2	Gweicurculactone	(5R)-3,5,8-Trimethyl-6,7-dihydroazuleno[6,5-	C15H16O2	228.1150298	228.2910004	C[C@@H]1CCC2=C(C)C=C3OC(=O)C(C)=C3C=C12	OGJFTQVVQQMGJG-MRVPVSSYSA-N
8	DTXSID80170114	17635-59-5	Phenol, 3-methoxy-5-(2-1,1'-	3-Methoxy-5-(2-1,1'-	C15H16O2	228.1150298	228.2910004	COC1=CC(CCC2=CC=CC=C2)=CC(O)=C1	HPEFWCAKFRCLBD-UHFFFAOYSA-N
9	DTXSID90176884	2235-01-0	(Dimethoxymethylene)b	(Dimethoxymethylene)di 1,1'-	C15H16O2	228.1150298	228.2910004	COC(OC)(C1=CC=CC=C1)C1=CC=CC=C1	NYRVXYOKUZSUDA-UHFFFAOYSA-N
10	DTXSID60181905	2749-70-4	Preventol D2	[Methylenebis(oxymethylene)]dibenzene	C15H16O2	228.1150298	228.2910004	C(OCC1=CC=CC=C1)OCC1=CC=CC=C1	JCTHGPXQXLMSDK-UHFFFAOYSA-N
11	DTXSID00184652	30567-87-4	1-Methoxy-2-((4-methoxyphenyl)methyl)	1-Methoxy-2-[(4-methoxyphenyl)methyl]b	C15H16O2	228.1150298	228.2910004	COC1=CC=C(CC2=CC=CC=C2OC)C=C1	KVCVHSGCINIKCE-UHFFFAOYSA-N
12	DTXSID80204286	5574-34-5	1,2-Naphthalenedione, 3,8-dimethyl-5-(1-	3,8-Dimethyl-5-(propan-2-yl)naphthalene-1,2-dione	C15H16O2	228.1150298	228.2910004	CC(C)C1=C2C=C(C)C(=O)C(=O)C2=C(C)C=C1	GREWJSSEUGRGIT-UHFFFAOYSA-N
13	DTXSID10206935	5819-93-2	Methane, bis-(o-	1,1'-Methylenebis(2-	C15H16O2	228.1150298	228.2910004	COC1=CC=CC=C1CC1=CC=CC=C1OC	DPRFNHUMLZJXNR-UHFFFAOYSA-N
14	DTXSID40208245	59586-27-5	4-Hydroxy-3'-	4-[2-(3-	C15H16O2	228.1150298	228.2910004	COC1=CC=CC(CCC2=CC=C(O)C=C2)=C1	QHWSNPUNTSCHIE-UHFFFAOYSA-N

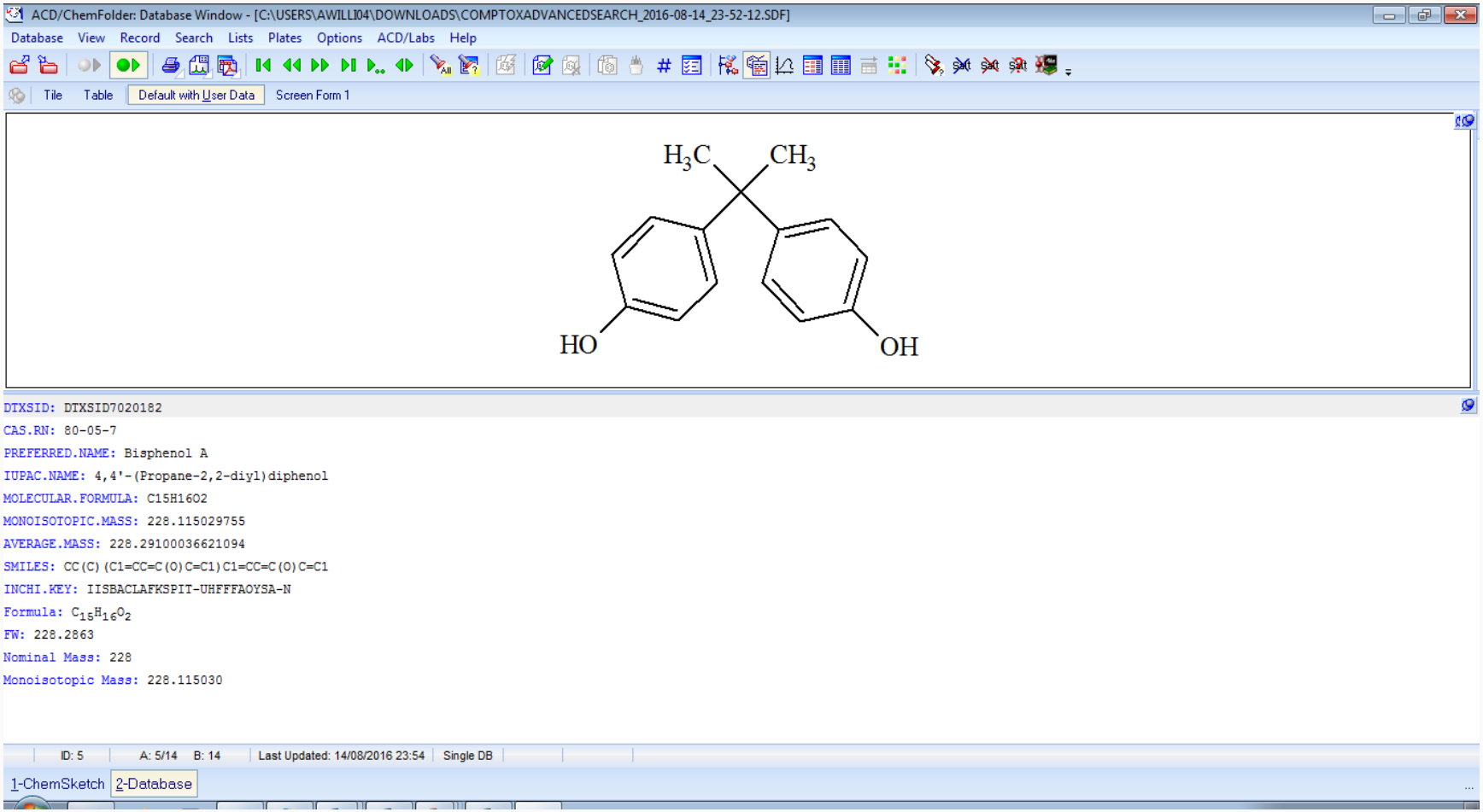
CompToxAdvancedSearch_2016-08-1

READY

Download as SDF file



SDF file opened in ChemFolder



Rank-ordering all of those hits??

- With so many hits how do you rank order based on formulae? Or mass??

Comparing Performance

ChemSpider
Search and share chemistry

57 Million
chemical structures

CompTox Dashboard

721k structures

Bisphenol A as an example
ChemSpider: 1564 Structures

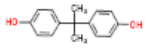
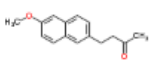
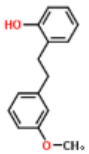
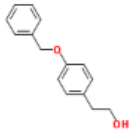
FILTER

Search Hits Limit: 100 CLEAR FORM SEARCH

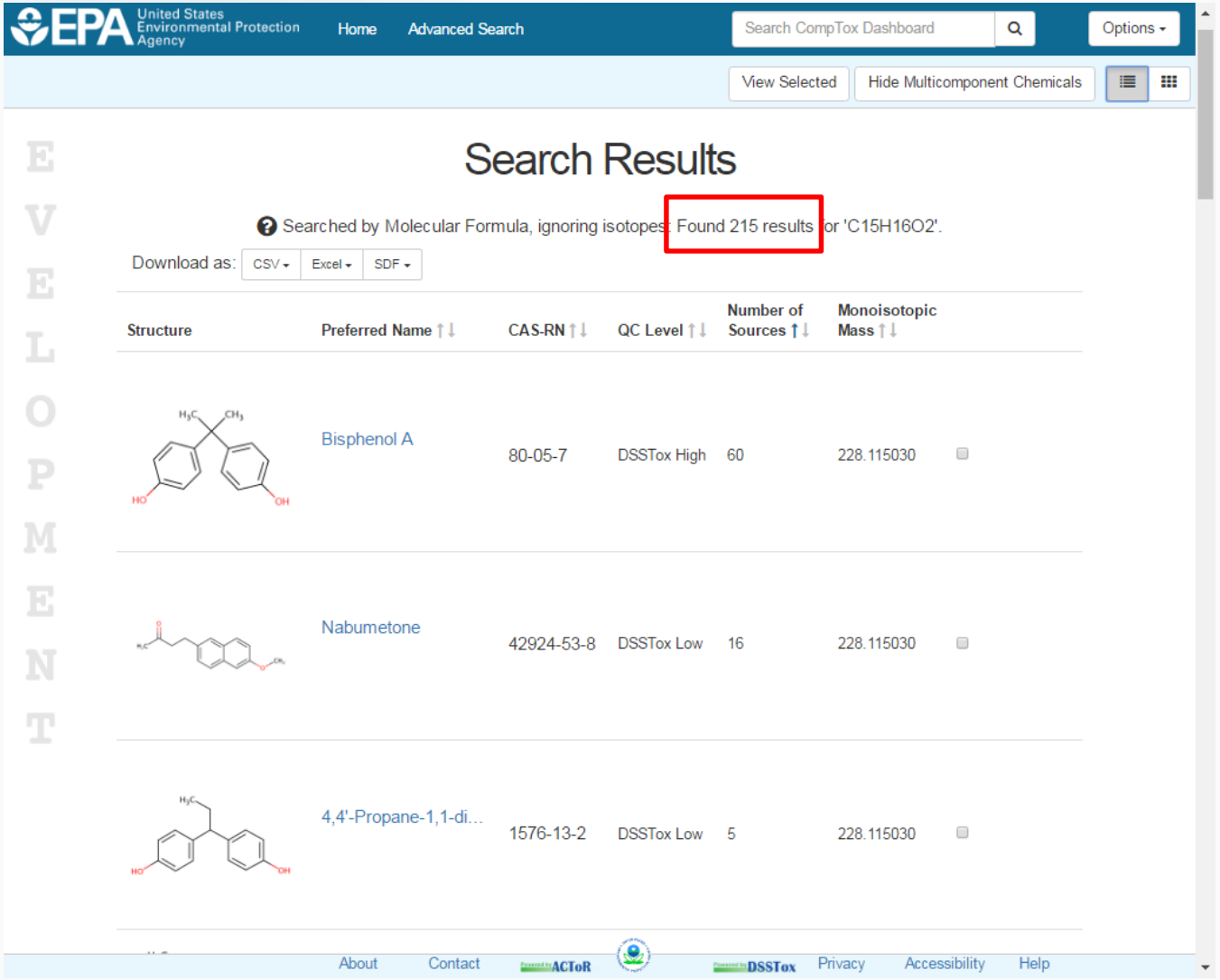
Found 1564 results

Search term: **MF = 'C_{15}H_{16}O_{2}'**

1 2 3 4 5

ID	Structure	Molecular Formula	Molecular Weight	# of Data Sources	# of References	# of PubMed	# of RSC
6371		$C_{15}H_{16}O_2$	228.2863	117	1631	3029	423
4256		$C_{15}H_{16}O_2$	228.2863	110	420	368	23
8760824		$C_{15}H_{16}O_2$	228.2863	58	67	0	1
2285031		$C_{15}H_{16}O_2$	228.28634	48	57	0	0

Bisphenol A as an example
Dashboard: 215 Structures



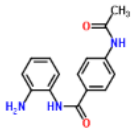
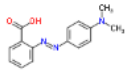
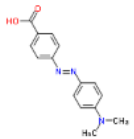
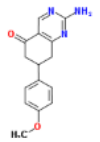
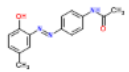
A more pointed example…$C_{15}H_{15}N_3O_2$

6926 results

Found 6926 results

Search term: **MF = 'C_{15}H_{15}N_{3}O_{2}'**

1 2 3 4 5

ID	Structure	Molecular Formula	Molecular Weight	# of Data Sources	# of References	# of PubMed	# of RSC
2644		$C_{15}H_{15}N_3O_2$	269.2985	85	125	4	0
9881		$C_{15}H_{15}N_3O_2$	269.2985	85	164	263	982
21231		$C_{15}H_{15}N_3O_2$	269.2985	32	58	9	10
2052796 - 0/1 defined		$C_{15}H_{15}N_3O_2$	269.2985	32	50	0	0
10468668		$C_{15}H_{15}N_3O_2$	269.2985	32	40	0	0

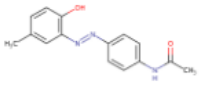
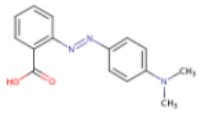
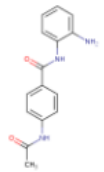
A more pointed example…C15H15N3O2

94 results

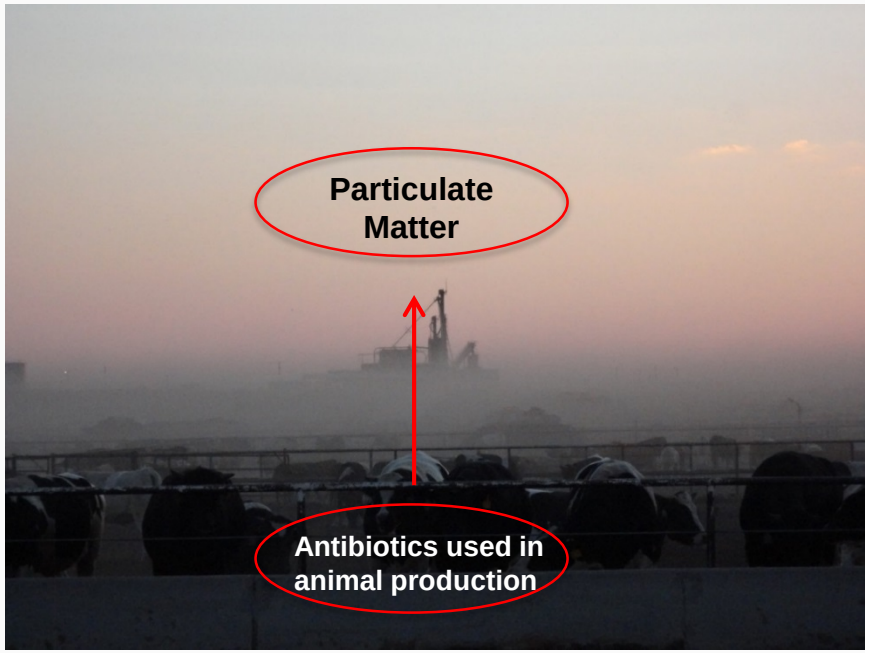
Search Results

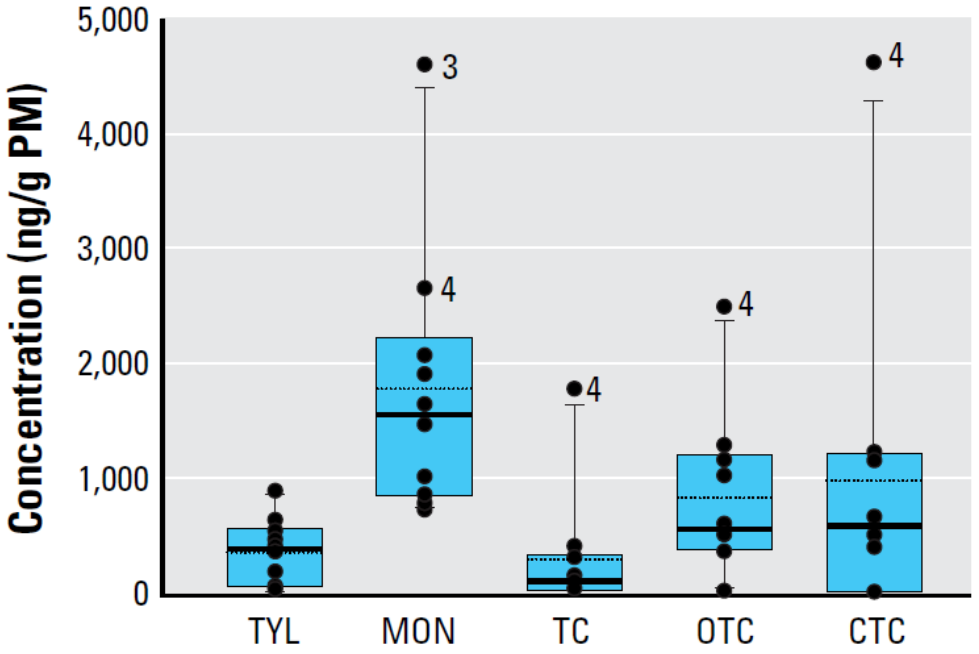
Searched by Molecular Formula, ignoring isotopes: Found 94 results for 'C15H15N3O2'.

Download as: CSV | Excel | SDF

Structure	Preferred Name	CAS-RN	QC Level	Number of Sources	Monoisotopic Mass	
	C.I. Disperse Yellow 3	2832-40-8	DSSTox Low	18	269.116427	☐
	Methyl red	493-52-7	DSSTox Low	17	269.116427	☐
	Tacedinaline	112522-64-2	Public Medium	4	269.116427	☐

Antibiotics in beef commercial feed





TYL= Tylosin
MON= Monensin
TC= Tetracycline
OTC= Oxytetracycline
CTC= Chlortetracycline

McEachran AD, Blackwell BR, Hanson JD, Wooten KJ, Mayer GD, Cox SB, Smith PN. 2015. Antibiotics, bacteria, and antibiotic resistance genes: aerial transport from cattle feed yards via particulate matter. Environ Health Perspect 123:337-343; DOI:10.1289/EHP.1408555

Rank-ordering Comparisons

Rank Position/Total # Results

	Mass-based Search		Formula Based Search	
	Dashboard	ChemSpider	Dashboard	ChemSpider
Tylosin	1/1	1/28	1/1	1/25
Monensin	1/1	1/39	1/1	1/24
Tetracycline	1/ 38	1/4008	1/11	1/355
Oxytetracycline	1/16	1/3271	1/3	1/110
Chlortetracycline	1/23	1/2545	1/3	1/77

Mean Rank Position

		Mass-based Search		Formula-based Search	
Agricultural Source	# Compounds	Dashboard	ChemSpider	Dashboard	ChemSpider
Wastewater land application[1]	34	1.3	1.8	1.1	1.1
Cattle Feedyard[2]	5	1.0	1.0	1.0	1.0

[1]McEachran AD, Shea D, Bodnar W, Nichols EG. 2016. Pharmaceutical Occurrence in groundwater and surface waters in forests land-applied with municipal wastewater. Environ Toxicol Chem 35: 898-905. DOI: 10.1002/etc.3216

[2]McEachran AD, Blackwell BR, Hanson JD, Wooten KJ, Mayer GD, Cox SB, Smith PN. 2015. Antibiotics, bacteria, and antibiotic resistance genes: aerial transport from cattle feed yards via particulate matter. Environ Health Perspect 123:337-343; DOI:10.1289/EHP.1408555

Chemical Identification Dashboard vs ChemSpider

Monoisotopic Mass (+/- 0.005 amu) Search

Sorted by number of references (ChemSpider) or data sources (Dashboard)

Source of List	# of Compounds	Search Tool	Mean Position	Median Position	Position of compound sorted				
					#1	#2	#3	#4	#5+
McEachran et al Wastewater	34	ChemSpider	1.8	1	28	5	0	0	1
		Dashboard	1.3	1	31	2	0	0	1
Misc. NTA Compounds	13	ChemSpider	2	1	7	5	0	0	1
		Dashboard	1.7	1	10	2	0	0	1
Bade et al (2016)	19	ChemSpider	2.1	1	11	2	5	0	1
		Dashboard	1.6	1	12	3	3	1	0
Rager et al (2016)	24	ChemSpider	2.25	1	15	2	1	2	4
		Dashboard	1.08	1	22	2	0	0	0

Dashboard vs ChemSpider Ranking Summary

162 total individual chemicals in search

	Mass-based Searching		Formula Based Searching	
	Dashboard	*ChemSpider*	*Dashboard*	*ChemSpider*
Cumulative Average Position	1.3	2.2	1.2	1.4
% in #1 Position	85%	70%	88%	80%

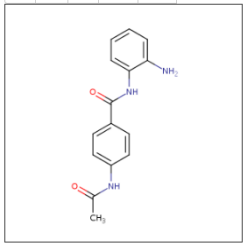
Functional Use to Sort Candidates

Tacedinaline

112522-64-2 | DTXSID60150095

Anti-cancer Drug

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



Intrinsic Properties

Molecular Formula: C15H15N3O2 — Find All Chemicals

Average Mass: 269.304 g/mol

Monoisotopic Mass: 269.116427 g/mol

Structural Identifiers

Record Information

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

Frequent Uses and Functions

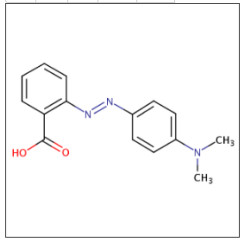
No product composition data found.

Methyl red

493-52-7 | DTXSID1042154

Microbiological Indicator Dye

Searched by Approved Name: Found 1 result for 'Methyl red'.



Intrinsic Properties

Molecular Formula: C15H15N3O2 — Find All Chemicals

Average Mass: 269.304 g/mol

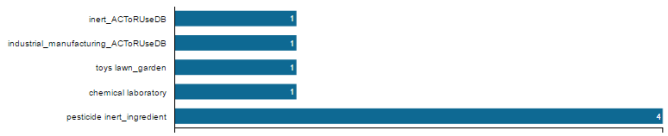
Monoisotopic Mass: 269.116427 g/mol

Structural Identifiers

Record Information

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

Frequent Uses and Functions



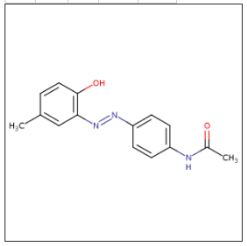
UNIVERSAL INDICATOR SOLUTION	0.05%	CENTRAL SCIENTIFIC CO
UNIVERSAL INDICATOR SOLUTION	0.05% %	SCIENCE KIT
UNIVERSAL INDICATOR SOLUTION,38826-16	0.05%	CENTRAL SCIENTIFIC CO
0098 TOTAL ALKALINITY TABLETS	0.04%	INDUSTRIAL MUNICIPAL EQUIP INC
84-8265 BOGEN UNIVERSAL INDICATOR SOLUTION	0.0185%	CAROLINA BIOGOLICAL SUPPLY CO
HTH 5-Way Test Kit - Hardness Indicator		
HTH 5-Way Test Kit - Cyanuric Acid Reagent		Arch Chemicals, Inc.
HTH 5-Way Test Kit - Cyanuric Acid Reagent		Arch Chemicals, Inc.

C.I. Disperse Yellow 3

2832-40-8 | DTXSID6021450

Textile/Product Dye

Searched by Approved Name: Found 1 result for 'C.I. Disperse Yellow 3'.



Intrinsic Properties

Molecular Formula: C15H15N3O2 — Find All Chemicals

Average Mass: 269.304 g/mol

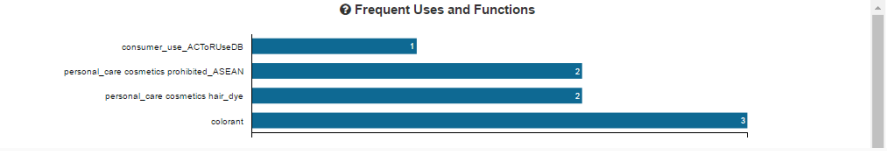
Monoisotopic Mass: 269.116427 g/mol

Structural Identifiers

Record Information

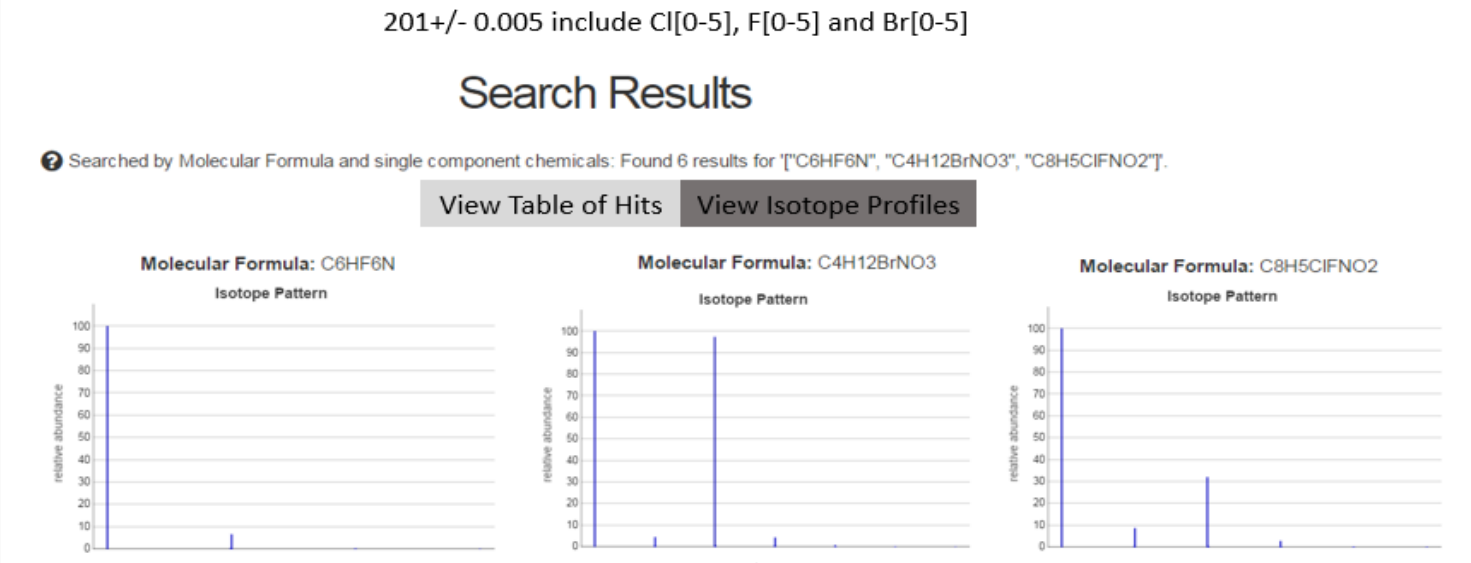
Chemical Properties | External Links | Synonyms | Product Composition | ToxCast in Vitro Data | Exposure | Analytical | PubChem | Comments

Frequent Uses and Functions



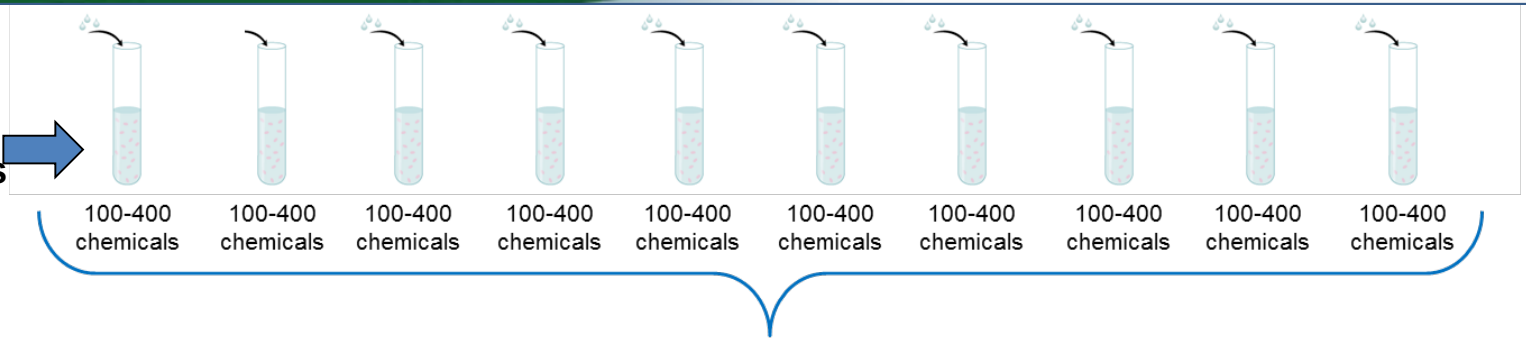
Future Work

- Rank-ordering based on other criteria
- Already testing **QSARs to build retention time models** for ranking
- External links to methods: e.g. CDC NIOSH
- Formula identification using isotope profiles



Engaging the MS Community

ToxCast Chemicals



Why are certain chemicals only found with certain methods?

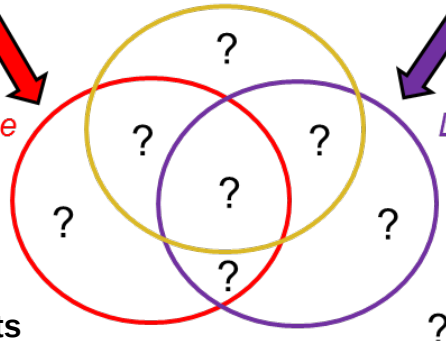
Lab B measurement space

Can we model these behaviors?

Can we expand coverage?

Lab A measurement space

Lab C measurement space



What impurities/ interaction products found?

? *"other" space (missing chemicals)*

Conclusions

- Our NTA research is focused on understanding our **exposure to chemicals**
- **New dashboard with focus on high-quality data** – no large database will be perfect!
- **Specific searches/functionality** are being developed with Non-targeted Analysis in mind
- Dashboard outperforms ChemSpider, a community standard database, in ranking chemicals of environmental concern
- Early work on **new rank-ordering** approaches show that we can improve things even further.

Acknowledgements

EPA NCCT
Chris Grulke
Jeff Edwards
Ann Richard
Jordan Foster
Jennifer Smith
Andrew McEachran*
Michelle Krzyzanowski

EPA NERL
Kathie Dionisio
Katherine Phillips
Jon Sobus
Mark Strynar
Elin Ulrich
Seth Newton

* = ORISE Participant