

# Environmental Chemistry Compound Identification Using High Resolution Mass Spectrometry Data Integrated to the EPA Chemistry Dashboard

Antony J. Williams, Andrew McEachran, Jon Sobus, Seth Newton, Elin Ulrich, Chris Grulke, Kamel Mansouri, Jennifer Smith and Jeff Edwards

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

### Who is NCCT?



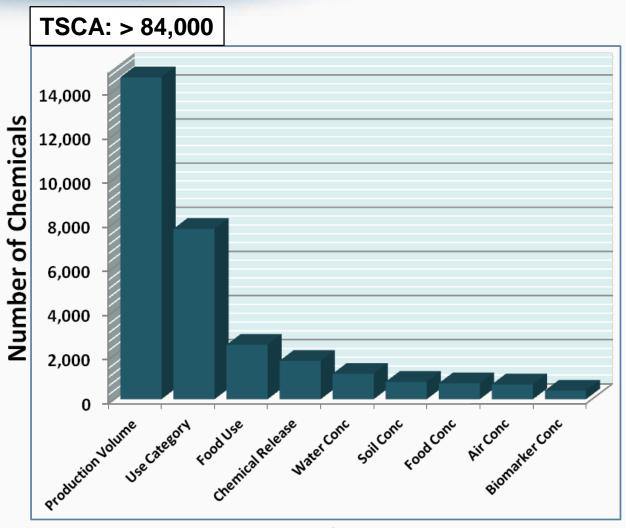
- National Center for Computational Toxicology part of EPA's Office of Research and Development
- Research driven by EPA's Chemical Safety for Sustainability Research Program
  - Develop new approaches to evaluate the safety of chemicals

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.

# Exposure Data Cannot Keep Pace with Regulatory Needs

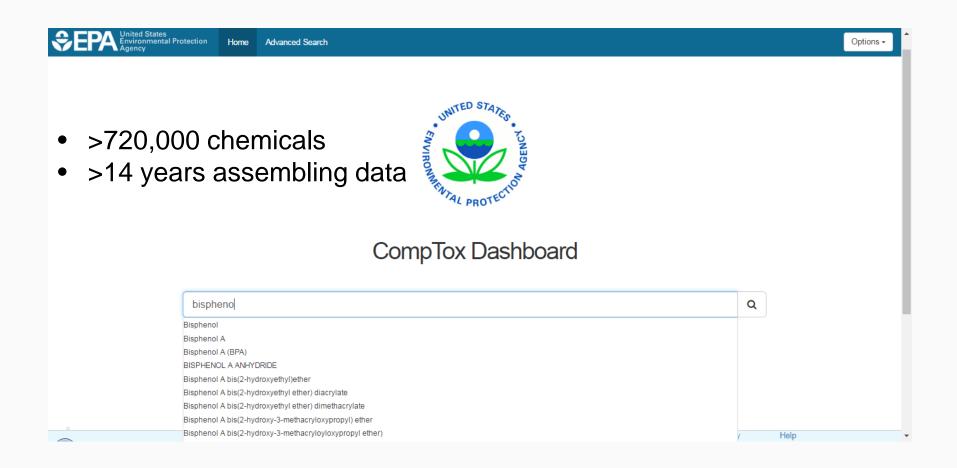




**Data Category** 

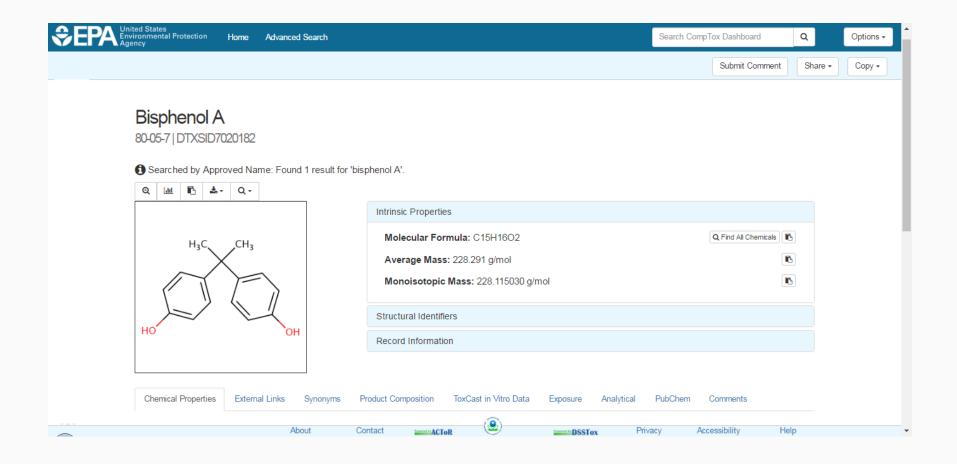
# Introducing Our Latest Dashboard https://comptox.epa.gov





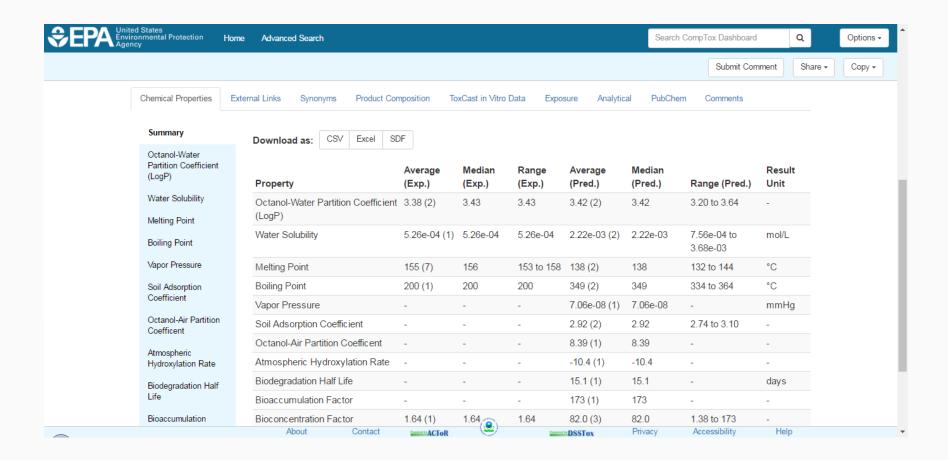
### Bisphenol A





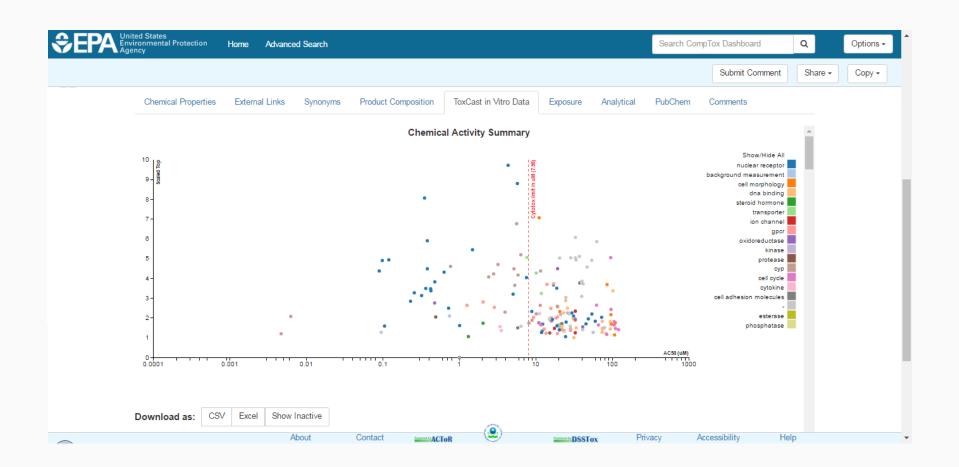
#### Physicochemical Properties





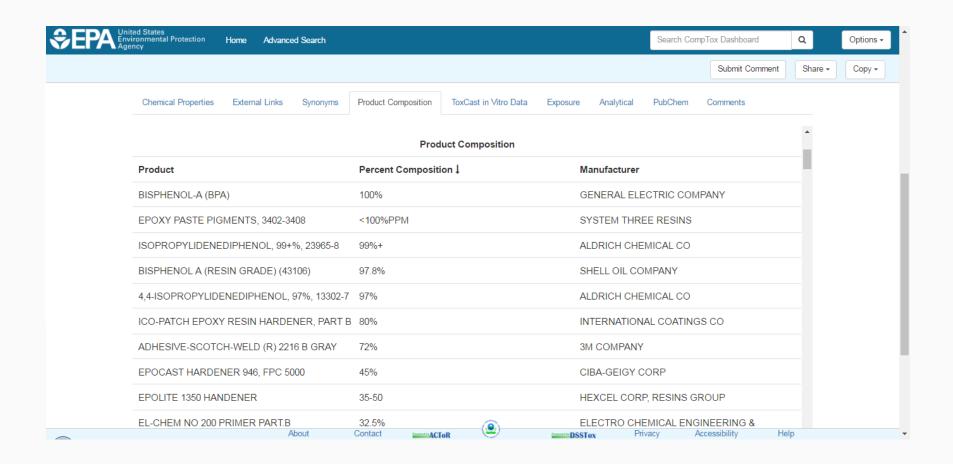
### ToxCast Bioassay Screening Data Useful Meta Data





### Functional Use and Composition VERY Useful Meta Data





# Dashboard: External Links to Analytical Methods and Data



Chemical Properties E	External Links	Synonyms Pr	roduct Composition	ToxCast in Vitro Data Exposure	PubChem Comments	
General		Toxicology		Publications	Analytical	Prediction
<ul> <li>● EPA Substance Registry</li> <li>NIST Chemistry Webbo</li> <li>         ← Household Products Da</li> <li>         ← PubChem</li> <li>● CPCat</li> <li>         ← MSDS Lookup</li> <li>         ← ToxPlanet</li> </ul>	ook atabase	<ul> <li>ACToR</li> <li>DrugPortal</li> <li>CCRIS</li> <li>ChemView</li> <li>CTD</li> <li>eChemPortal</li> <li>EDSP Dashbo</li> </ul>		Toxline Environmental Health Perspecti NIEHS National Toxicology Program Google Books Google Scholar Google Patents	Q National Environmental Method  RSC Analytical Abstracts  MONA: MassBank North America  Tox21 Analytical Data  FOR-IDENT	☑ Proton NMR Prediction ☑ Carbon-13 NMR Prediction ☑ 2D NMR HSQC/HMBC Prediction ☑ ChemRTP Predictor
		Gene-Tox  HSDB  ToxCast Dash	nboard 2	■ PubMed  Q BioCaddie DataMed		

#### National Environmental Methods Index





#### Previous Work with Suspect-Screening



Environment International 88 (2016) 269-280



Contents lists available at ScienceDirect

#### Environment International





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



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

- \* 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
- <sup>c</sup> 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

# ONE ASPECT of the dashboard is to support Non-targeted Analysis

### Rank-Ordering of "Known-Unknowns" using ChemSpider



#### 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 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	

### Some history...



2007 A Hobby Project



2009 ChemSpider Acquired

#### RSC acquires ChemSpider

11 May 200

The Royal Society of Chemistry announced today that it has acquired ChemSpider, heralding a breakthrough investment for the organisation and for the Chemistry Community.

This acquisition reflects RSC's commitment to providing access to rich resources of chemistry data and information. This complements RSC's existing leading role in online chemistry, including award-winning semantic mark-up technology and the release of the InChI resolver, recently launched in partnership with ChemSpider.

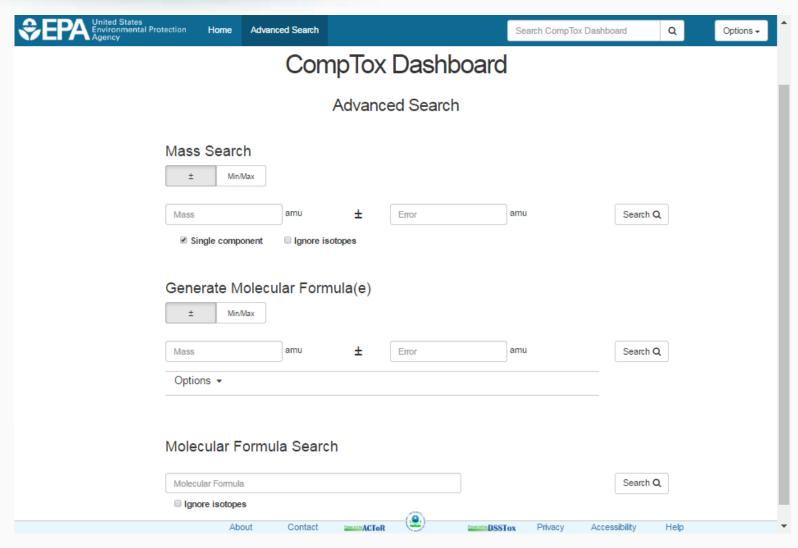


"What originally started as a hobby project to give back something to the chemistry community has become one of the primary internet resources for Chemistry. And this from home built computers in a basement, with no funding and a team of volunteers. With the resources, reputation and vision of the RSC to support Chem Spider our long term goal is to deliver the

 May 2015 Joined EPA – what we are showing is very new

#### Advanced MS Searches





#### Formula Search



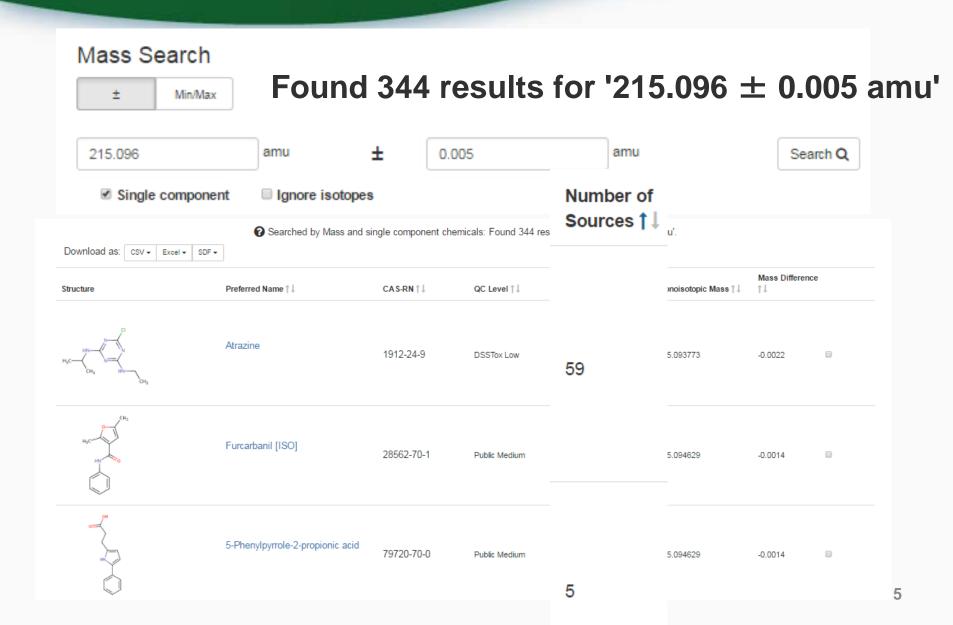
#### Molecular Formula Search

#### Found 8 results for 'C8H14CIN5'

C8H14CIN5		. 00		Number of Sources †↓			Search Q
					Hide Isotopes	Hide Multicomp	onent Chemicals
H <sub>S</sub> C——CH <sub>3</sub> CH <sub>3</sub> CH <sub>2</sub>	Atrazine	1912-24-9	DSSTox Low	59	215.093773	€	
H <sub>3</sub> C CH <sub>3</sub> N CH <sub>3</sub>	GS 18183	34333-27-2	Public Medium	3	215.093773	€	
H <sub>2</sub> C CH <sub>2</sub>	6-Chloro-N <sup>2</sup> -ethyl-N <sup>4</sup> -pr	90952-64-0	Public Low		215.093773	€	
	Contact	Powered by ACT o R	<b>9</b>	1	Privacy	Accessibility	Help

#### Monoisotopic Mass Search



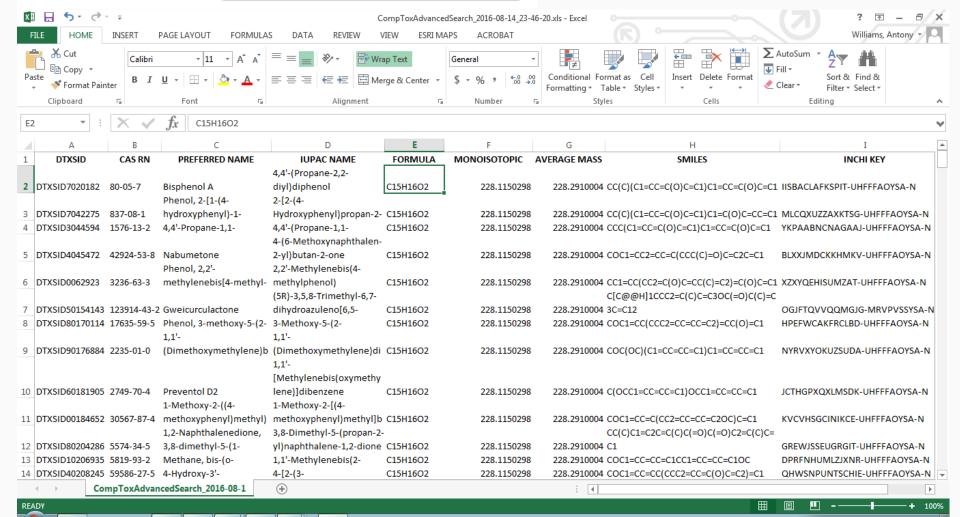


#### Download to Excel



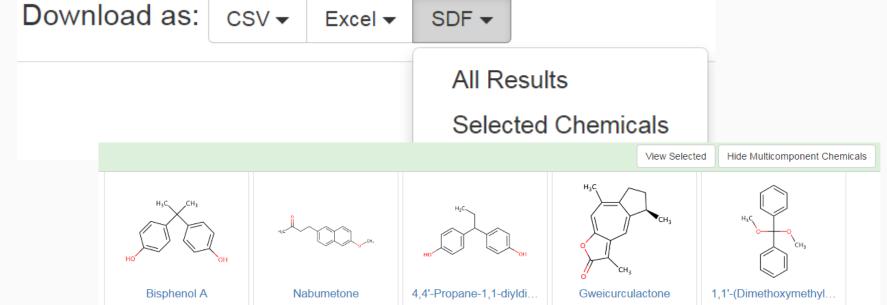


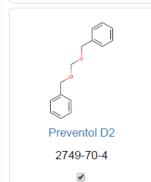




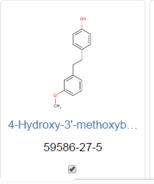
#### Download as SDF file



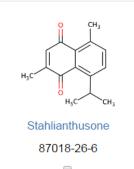




80-05-7



42924-53-8



1576-13-2

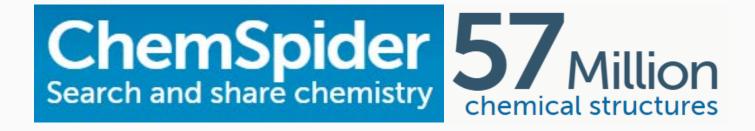


123914-43-2



#### Does the Dashboard Add Value?







#### Does the Dashboard Add Value?



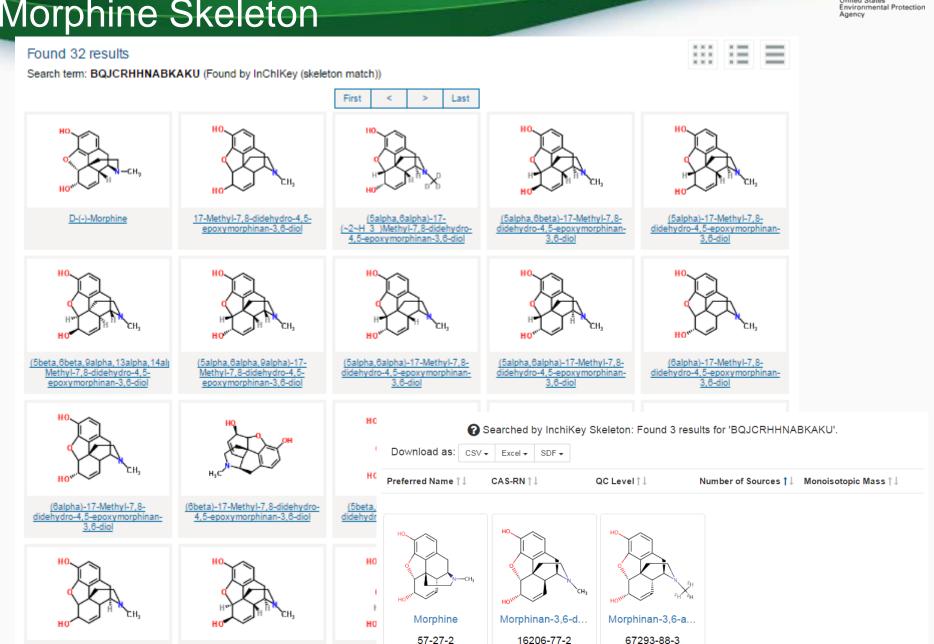
#### Remember:

- Focus on high quality data and curation
- Data sources include EPA data sources and a focus on environmental chemistry
- No "dilution" by chemical vendors

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

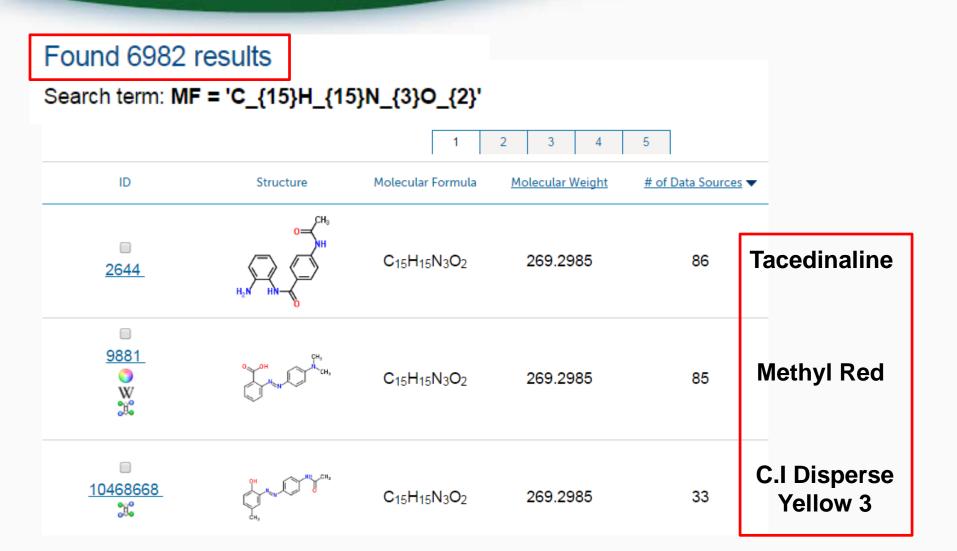
# Dilution Example... Morphine Skeleton





### ChemSpider 6982 Results!!! Search for C15H15N3O2





#### Same top hits – different ranking 90 hits only versus 6926 hits



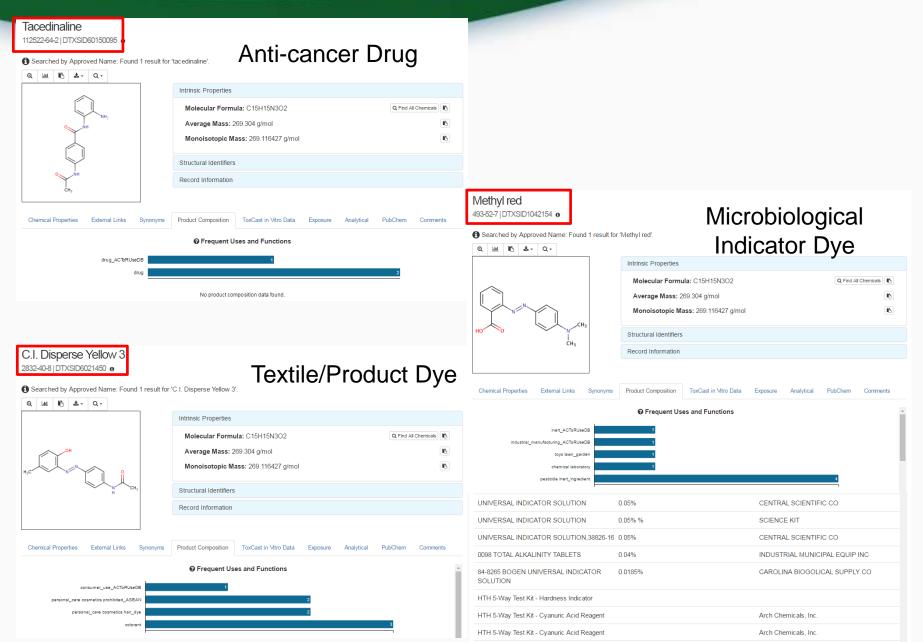
#### Search Results

② Searched by Molecular Formula, ignoring isotopes: Found 90 results for 'C15H15N3O2'.

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
HO CH <sub>3</sub>	Methyl Red	493-52-7	DSSTox Low	17	269.116427
	Tacedinaline	112522-64-2	Public Medium	4	269.116427

#### Using Meta-Data to Sort Candidates





### Chemical Identification Dashboard vs ChemSpider



#### Monoisotopic Mass (+/- 0.005 amu) Search

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

				1	Positio	on of co	mpour	nd sort	ed
Source of List	# of Compounds	Search Tool	Mean Position	Median Position	#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

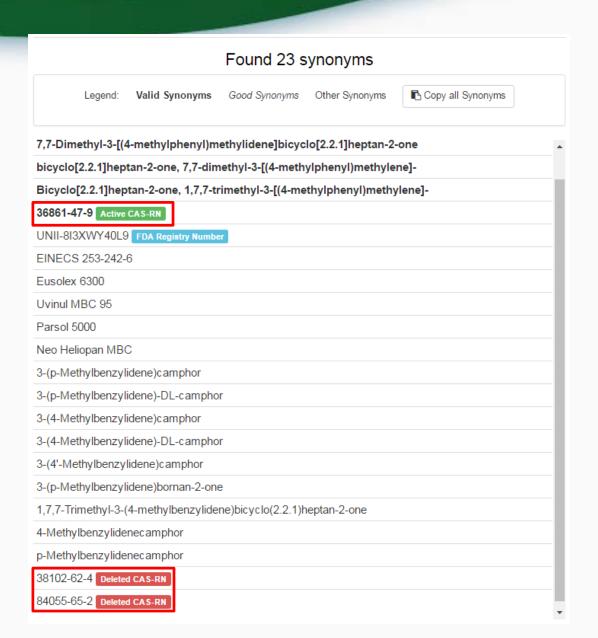


- Selected peer-reviewed publications
- 162 total individual chemicals in search

	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%		

#### Active vs Deleted CASRN





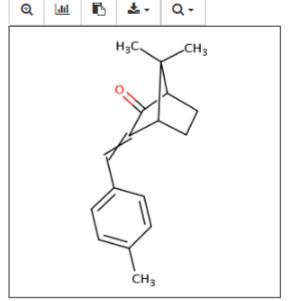
#### Collisions in CAS Numbers



#### Enzacamene

36861-47-9 | DTXSID8047896 •

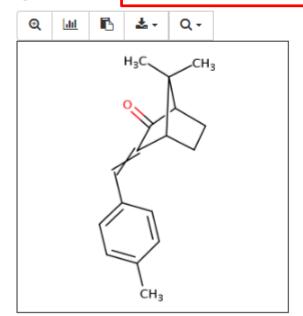
Searched by CAS-RN: Found 1 result for '36861-47-9'.



#### Enzacamene

36861-47-9 | DTXSID8047896 •

Searched by Deleted CAS-RN: Found 1 result for '38102-62-4'



#### But there are MANY CASRNs!



### http://web.stanford.edu/group/swain/cinf/c asreg/snumber.html

#### 4.1 Other Registry Number Fields

- A. Deleted Registry Number
- B. Replacing Registry Numbers
- C. Preferred Registry Number
- D. Alternate Registry Number

An answer may contain other Registry Number fields: Deleted Registry Number (DR), Replacing Registry Number (RR), Alternate Registry Number (AR), and Preferred Registry Number (PR). Their definitions follow.

- **A. Deleted Registry Number** A Registry Number assigned to a substance but later changed to another Registry Number. The record for this deleted number contains only the replacing Registry Number, RR. All of the substance information is contained in the replacing Registry Number record. The deleted Registry Number displays in the replacing Registry Number record in the DR display field. Examples of substances that are deleted include trade names that appear in the literature with no structural information but are later elucidated to a registered substance.
- B. Replacing Registry Number The Registry Number for a substance. Other Registry Numbers have been deleted to this number. These deleted Registry Numbers display in the DR display field. This Registry Number displays in the RR field for the deleted Registry Number record.
- C. Preferred Registry Number The Registry Number assigned to the more preferred structure of a substance that has more than one structural representation. The Registry Number of the alternate structure appears in the AR display field. The preferred Registry Number displays in the PR display field of the substance record of the alternate structure.
- D. Alternate Registry Number A second Registry Number generated for a second structural representation of a substance. These records have a more preferred structure that is the structure usually encountered in the literature. The record of this less preferred structure will have the Registry Number of the more preferred substance in the PR display field. The record of the preferred substance has the Registry Number searched as if it were an RN. Examples of substances that have AR's include fluorescein dyes which have the spiro form as the preferred registration and the closed form as the alternate registration.

All of these Registry Numbers are searched when SEARCH NNNNNN-N/RN or DISPLAY ACC is specified. If you SEARCH with a Deleted Registry Number, the Replacing Registry Number record is automatically located. If you SEARCH with an Alternate Registry Number, the Preferred Registry Number record is automatically located. If you DISPLAY ACC on a Deleted Registry Number, its record contains only the Deleted Registry Number in the RN field and the Replacing Registry Number in the RR field.

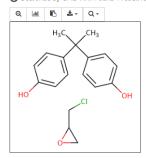
#### How Bad Can It Get??



#### Bisphenol A/ Epichlorohydrin resin

25068-38-6 | DTXSID0050479 6

1 Searched by CAS-RN: Found 1 result for '25068-38-6'.



Intrinsic Properties

Molecular Formula: C18H21CIO3

Average Mass: 320.81 g/mol

Monoisotopic Mass: 320.117922 g/mol

Structural Identifiers

Record Information

Found 664 synonyms											
Lege	end:	Valid Synonyms	Good Synonyms	Other Synonyms	<b>I</b> Copy all Synonyms						
4000470.07.0											
1096473-97-0											
110158-22-0											
11097-80-8											
11098-40-1											
11100-23-5 D											
11108-41-1											
11120-31-3 D											
11121-19-0											
11126-36-6											
1114797-08-8											
111517-59-0											
114013-37-5	Delet	ed CAS-RN									
115902-32-4	Delet	ed CAS-RN									
117216-90-7	Delet	ed CAS-RN									
117786-92-2	Delet	ed CAS-RN									
118340-04-8	Delet	ed CAS-RN									
1189565-70-5	Dele	eted CAS-RN									
1190235-62-1	Dele	eted CAS-RN									
1190729-68-0	Dele	eted CAS-RN									
1192045-32-1	Dele	eted CAS-RN									

### How Bad Can It Get? This one is 316 Deleted CASRN



CAS Registry Number: 25068-38-6

(C<sub>15</sub> H<sub>16</sub> O<sub>2</sub> . C<sub>3</sub> H<sub>5</sub> CI O)<sub>X</sub>

Phenol, 4,4'-(1-methylethylidene)bis-, polymer with 2-(chloromethyl)oxirane

Polymer

Polymer Class Terms: Epoxy resin

Alternate CAS Registry Numbers: 26402-79-9

Deleted CAS Registry Numbers: 1336-88-5, 1337-15-1, 8000-31-5, 9015-99-0, 9049-54-1, 9050-21-9, 9081-91-8, 9083-76-5, 9084-94-0, 9086-62-8, 9087-26-7, 9087-76-7, 11097-80-6, 11098-13-8, 11098-40-1, 11100-23-5, 11108-41-1, 11120-31-3, 11121-19-0, 11126-36-6, 20232-24-0, 35038-60-9, 7, 37270-82-9, 37291-75-1, 37293-07-5, 37294-18-1, 37305-82-1, 37307-45-2, 37317-45-6, 37325-21-6, 37338-63-9, 37342-17-9, 37345-34-9, 37348-56-4. 37348-57-5. 37357-73-6. 37360-93-3. 39277-59-3. 39288-99-8. 39296-08-7. 39296-09-8. 39296-11-2. 39296-15-6. 39315-77-0. 39349-91-2. 39354-86-4, 39362-25-9, 39362-45-3, 39373-81-4, 39378-29-5, 39378-55-7, 39389-49-6, 39405-18-0, 39412-57-2, 39419-66-4, 39453-22-0, 39454-54-1, 39454-69-8, 39470-62-7, 42612-34-0, 42618-03-1, 50642-36-9, 50642-55-2, 50642-78-9, 51158-20-4, 51273-81-5, 51329-73-8, 51393-99-8, 51394-03-7, 51553-00-5, 52011-87-7, 52038-45-6, 52051-70-4, 52051-82-8, 52052-16-1, 52232-05-0, 52232-75-4, 52276-55-8, 52365-33-0, 52519-66-1, 52519-67-2. 52627-94-8. 52907-38-7. 53027-88-6. 53127-14-3. 53200-30-9. 53238-86-1. 53238-87-2. 53239-67-1. 53239-68-2. 53570-97-1. 53570-98-2. 53681-78-0. 53858-93-8. 54018-73-4. 54352-05-5. 55464-96-5. 55584-55-9. 55585-07-4. 55818-73-0. 56258-35-6. 56449-43-5. 56509-48-9. 57107-66-1, 57284-90-9, 57534-21-1, 57693-04-6, 58052-05-4, 58128-38-4, 58392-89-5, 58392-92-0, 58516-14-6, 58572-71-7, 59029-19-5, 59459-14-2, 59473-30-2. 59948-36-6. 60202-19-9. 60267-31-4. 60382-89-0. 60606-56-6. 60800-54-6. 60831-77-8. 60894-16-8. 61036-82-6. 61287-42-1. 61356-27-2. 61711-38-4. 61763-30-2. 61991-18-2. 62169-28-2. 62169-29-3. 62601-75-6. 62601-76-7. 62887-23-4. 63055-40-3. 63172-55-4. 63799-24-6. 63993-57-7. 63993-58-8. 64086-14-2. 64086-16-4. 64176-52-9. 64176-61-0. 64176-66-5. 64177-03-3. 65233-49-0. 65931-38-6. 65931-39-7. 66995-96-8. 67185-62-0. 68821-97-6. 69899-40-7. 70179-83-8. 70213-44-4. 70726-45-3. 71965-91-8. 72514-40-0. 73413-19-1. 74504-20-4. 74564-76-4. 75831-44-6, 78564-77-9, 79585-43-6, 80702-61-0, 81458-12-0, 81843-57-4, 81843-58-5, 81855-87-0, 82197-12-4, 82197-46-4, 83202-85-1, 84286-97-5, 84683-04-5, 84931-29-3, 85537-69-5, 86090-60-0, 88385-37-9, 88528-19-2, 88651-18-7, 89750-00-5, 91727-28-5, 91727-29-6, 92481-37-3, 95327-25-6, 96420-31-4, 96510-68-8, 97568-16-6, 97709-01-8, 99400-50-7, 101027-12-7, 102256-87-1, 103599-13-9, 103599-14-0, 104364-97-8, 104491-99-8, 105521-57-1, 106207-08-3, 106856-89-7, 107991-47-9, 108556-05-4, 108728-21-8, 110158-22-0, 111367-08-9, 111517-59-0, 114013-37-5, 115902-32-4, 117216-90-7, 117313-45-8, 117786-92-2, 118340-04-8, 120146-74-9, 120797-43-5, 121181-85-9, 121273-37-8, 121547-73-7, 123939-44-6, 125147-87-7, 127176-80-1, 127176-81-2, 128281-71-0, 132822-20-9, 132893-73-3, 135976-90-8, 137545-29-0, 138157-20-7, 138361-18-9, 139554-29-3. 142540-11-2. 144046-24-2. 144046-25-3, 144855-66-3, 149013-58-1, 150825-32-4, 157321-42-1, 157481-46-4, 158725-45-2, 160674-45-3, 161937-12-8. 162031-55-2. 167972-06-7. 168042-08-8. 179607-24-0. 183581-68-2. 183890-12-2. 187619-11-0. 188448-56-8. 189282-49-3. 191606-83-4, 220090-06-2, 222835-65-6, 222835-66-7, 222835-68-9, 222835-69-0, 222835-70-3, 222835-72-5, 222835-74-7, 222835-77-0, 309945-96-8, 339530-81-3, 353239-57-3, 367523-08-8, 383889-26-7, 383889-27-8, 395069-05-3, 470462-49-8, 681001-41-2, 848887-61-6, 913745-83-2, 917483-69-3, 922728-11-8, 934588-09-7, 945610-97-9, 950907-45-6, 1033821-54-3, 1034342-45-4, 1068160-75-7, 1082736-74-0, 1096473-97-0, 1114797-08-8. 1189565-70-5. 1190235-62-1. 1190729-68-0. 1192045-32-1. 1195324-26-5. 1196030-95-1. 1198291-96-1. 1199811-18-1. 1203835-26-0. 1206700-05-1, 1228639-00-6, 1245563-83-0, 1271727-39-9, 1300093-58-6, 1300102-07-1, 1305321-17-8, 1338071-08-1, 1446691-72-0, 1450839-98-1. 1620807-39-7. 1641551-32-7. 1807886-28-7. 1815624-46-4. 1815624-47-5

#### Our OPEN Data is available...



 Various types of data at FTP download site: <u>ftp://newftp.epa.gov/COMPTOX/Sustainable\_Chemistry\_Data/Chemistry\_Dashboard</u>

#### Index of /COMPTOX/Sustainable\_Chemistry\_Data/Chemistry

Name	Size	Date Modified
1 [parent directory]		
PHYSPROP_Analysis/		9/6/16, 10:30:00 AM
DSSTOX_MS_Ready_Chemical_Structures.zip	280 MB	10/7/16, 1:32:00 PM
DSSTox_Mapping_20160701.zip	32.9 MB	8/25/16, 9:31:00 AM
DSSTox_Predicted_NCCT_Model.zip	343 MB	9/29/16, 10:13:00 AM
DSSTox_SDF_File_20160720.sdf.gz	162 MB	8/25/16, 9:34:00 AM
DSSTox_Synonyms_20161018.zip	145 MB	10/21/16, 3:31:00 PM
■ Dsstox_CAS_number_name.xlsx	32.1 MB	8/25/16, 9:36:00 AM
■ PubChem_DTXSID_mapping_file.txt	23.1 MB	9/27/16, 9:46:00 AM

### Coming December 2016 Batch Searching Names/CASRNs

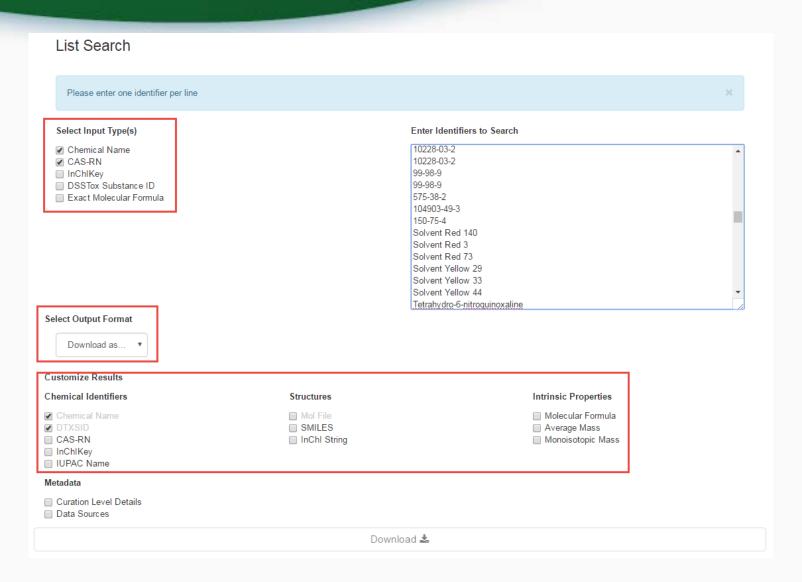


#### What are these chemicals?

533-73-3 95-54-5 108-73-6 108-45-2 2051-85-6 106-50-3 128-95-0 83-56-7 27581-07-3 5697-02-9 90-15-3 89-25-8 92-44-4 15086-94-9 615-05-4 5862-77-1 5862-77-1 5862-77-1 95-86-3 95-80-7 1004-75-7 141-86-6 141-86-6 84540-47-6 149330-25-6 606-20-2 823-40-5 582-17-2 16867-03-1 16867-03-1 1603-85-0 96-91-3 95-85-2	1,5-Naphthalenediol 2,7-Naphthalenediol 4-Amino-o-cresol Disperse Blue 3 Disperse Blue 377 3,4-Dihydroxyaniline 2,6-Dihydroxyethylaminotoluene Catechol 1-Acetoxy-2-methylnaphthalene Fast Green FCF Acid Red 14 6-Nitro-2,5-pyridinediamine Pyridine HC Orange 2 m-Aminophenol p-Aminophenol 2-Methyl-4-nitroaniline o-Phenylenediamine Acid Blue 1 2,3-Indolinedione 1,2,4-Trihydroxybenzene Solvent Red 23 2-Chloro-4-phenylenediamine 2-Chloro-p-phenylenediamine p-Phenylenediamine Solvent Red 43 Acid Violet 9 Direct Blue 86 4-Hydroxyindole 2,5,6-Triamino-4-pyrimidinol Solvent Red 1 HC Yellow 10 Oxazine 1 2,4-Diaminotoluene m-Phenylenediamine
---	--

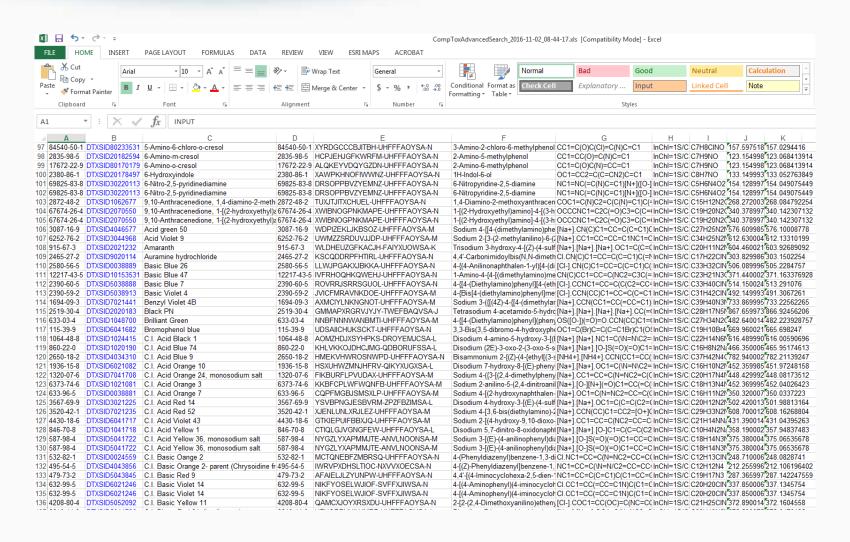
# Coming December 2016 Batch Searching...





### Coming December 2016 Download to Excel





### Batch Searching of Molecular Formula



#### Enter Identifiers to Search Select Input Type(s) Include top hits Chemical Name C12H25NO2 CAS-RN C11H21NO3 InChlKey C10H9N5O DSSTox Substance ID C11H21NOS Exact Molecular Formula C10H14CINO2 C12H9NO3 C12H13N3O C9H18N3OP C8H14CIN5 Select Output Format Excel ₹ Customize Results Intrinsic Properties Chemical Identifiers Structures Chemical Name Mol File Molecular Formula DTXSID SMILES Average Mass CAS-RN InChl String Monoisotopic Mass InChlKey IUPAC Name Metadata Curation Level Details Data Sources

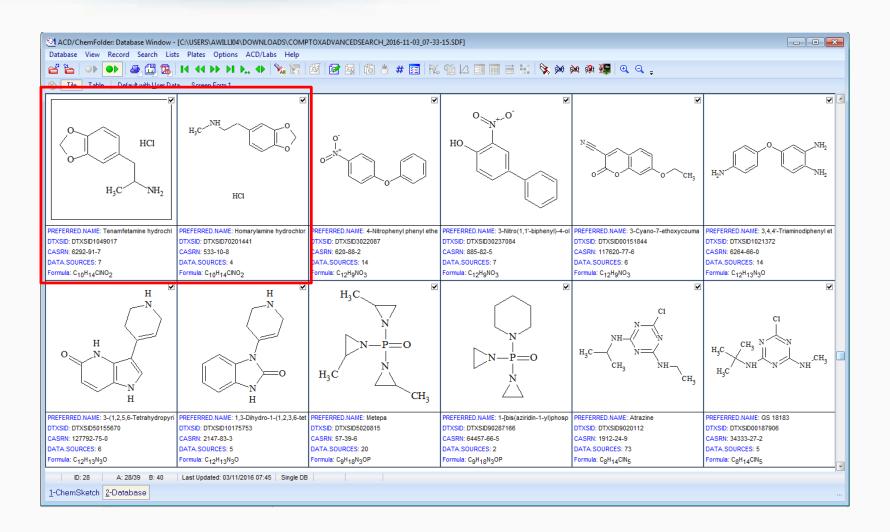
### Metadata included for Ranking



4		U	U	U	L	I	G
1	INPUT	DTXSID	CASRN	PREFERRED NAME	<b>IUPAC NAI</b>	SMILES	DATA SOURCES
31	C12H9NO3	DTXSID3022087	620-88-2	4-Nitrophenyl phenyl ether	1-Nitro-4-ph	[O-][N+](=O)C1=CC=C(OC2	14
32	C12H9NO3	DTXSID30237084	885-82-5	3-Nitro(1,1'-biphenyl)-4-ol	3-Nitro[1,1'-	OC1=C(C=C(C=C1)C1=CC	7
33	C12H9NO3	DTXSID00151844	117620-77-6	3-Cyano-7-ethoxycoumarin	7-Ethoxy-2-	CCOC1=CC2=C(C=C1)C=C	6
34	C12H13N3O	DTXSID1021372	6264-66-0	3,4,4'-Triaminodiphenyl ether	4-(4-Aminor	NC1=CC=C(OC2=CC=C(N)	14
35	C12H13N3O	DTXSID50155670	127792-75-0	3-(1,2,5,6-Tetrahydropyrid-4-yl)p	3-(1,2,3,6-T	O=C1NC2=C(NC=C2C2=C0	6
36	C12H13N3O	DTXSID10175753	2147-83-3	1,3-Dihydro-1-(1,2,3,6-tetrahydro	1-(1,2,3,6-T	O=C1NC2=CC=CC=C2N1C	5
37	C9H18N3OP	DTXSID5020815	57-39-6			CC1CN1P(=0)(N1CC1C)N1	
38	C9H18N3OP	DTXSID90287166	64457-66-5	1-[bis(aziridin-1-yl)phosphoryl]pi	1-[Bis(azirid	O=P(N1CC1)(N1CC1)N1CC	2
39	C8H14CIN5	DTXSID9020112	1912-24-9			CCNC1=NC(NC(C)C)=NC(C	
10	C8H14CIN5	DTXSID00187906	34333-27-2	GS 18183	N~2~-tert-B	CNC1=NC(NC(C)(C)C)=NC(	5
11							

#### Need for "MS-Ready Structures"





### "QSAR-Ready Structures"



- For the purpose of building QSAR Models we already "standardize" structures
  - Desalt/Neutralize
  - Desolvate
  - Remove stereochemistry

 Some minor tweaks gets us "MS-ready Structures". ALREADY in our database.

### "QSAR-Ready Structures"



 Mass and Formula-based searches will be based on MS-ready structures but connected to the original chemical (with name, CAS, rank ordering)

$\overline{}$	U	V	U	L	1	U	11
INPUT	DTXCID	DTXSID	CASRN	PREFERRED_NAME	MOL_FORMULA	MONOISOTOPI	DATA_SOURCES
C6H6O	DTXCID01234	DTXSID5021124	108-95-2	Phenol	C6H6O	94.04186481	77
C6H6O	DTXCID01234	DTXSID4027072	139-02-6	Sodium phenolate	C6H5NaO	116.023809	19
C6H6O	DTXCID01234	DTXSID10206632	5793-84-0	Calcium phenoxide	C12H10CaO2	226.030671	5
C6H6O	DTXCID01235	DTXSID60179347	24599-57-3	2.4-Cyclohexadienone	C6H6O	94 04186481	4

 MS-ready structures and substance mappings will be available as Open Data

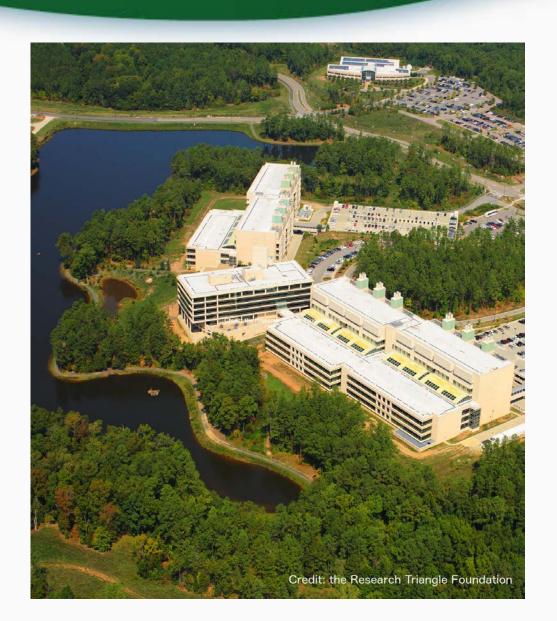
#### **Future Work**



- Continue to research rank-ordering approaches
- Working on "retention time prediction"
- Search for adducts (+Na, +K, +NH4) and handle decarboxylation, loss of water etc
- Expand link outs to Mass Spec databases Thermo's mzCloud, Massbank, etc.
- Predicting metabolites and degradants
- Optimize web services for the community

### Acknowledgements





EPA NCCT
Chris Grulke
Jeff Edwards
Ann Richard
Jennifer Smith
Andrew McEachran\*

EPA NERL
Jon Sobus
Seth Newton
Elin Ulrich

\* = ORISE Participant