http://orcid.org/0000-0003-1423-330X

United States Environmental Protection Agency

Identifying known unknowns: A comparison between ChemSpider and the US EPA's CompTox Chemistry Dashboard

Andrew D. McEachran*, *Jon R. Sobus*, *Antony J. Williams*

ANYL 470

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

Office of Research and Development
National Center for Computational Toxicology, RTP, NC

April 2, 2017

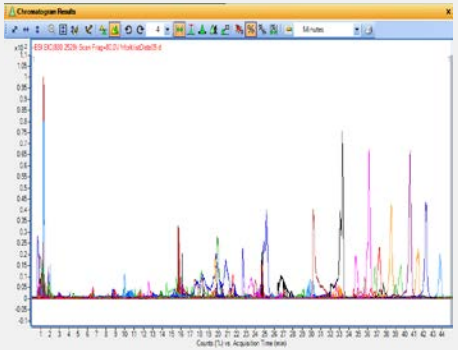
EPA United States Environmental Protection Agency

Comparing Analysis Approaches

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

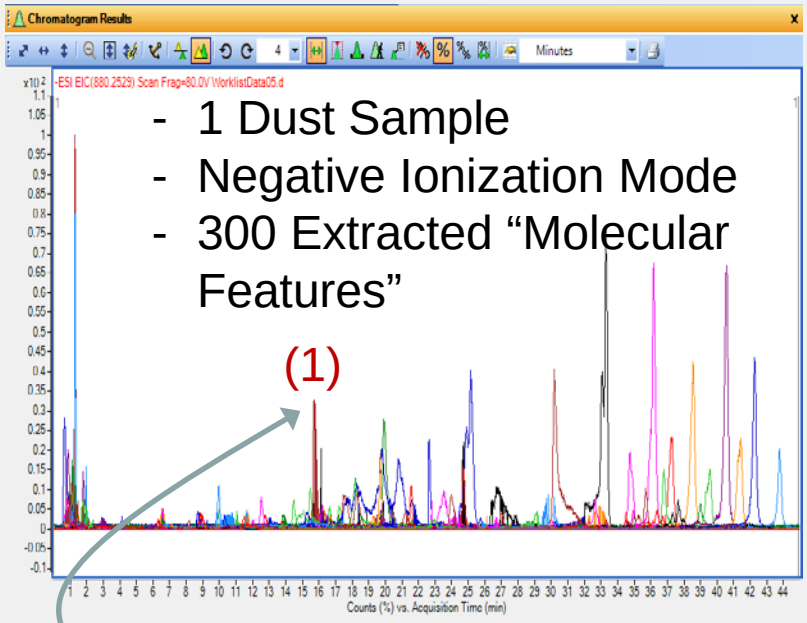
- Suspect Screening Analysis (SSA):
  - We have chemicals of interest
  - 100s – 1,000s of chemicals

- 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!!



Slide from Sobus, Williams

General Goals of SSA/NTA



1) Prioritize "Molecular Features"

2) Correctly assign formulas

3) Correctly assign structures

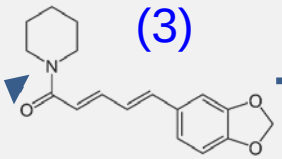
4) Determine chemical sources

5) Predict chemical concentrations

EXPOSURE

(2) C17H19NO3

(3)



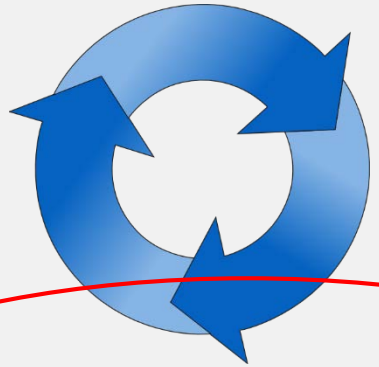
(4)

(5) 12 µg/g

EPA United States Environmental Protection Agency

The General Approach

Analytical Instruments



Comp. Tools & Workflows

FOR IDENT

Xcms Online

MetFrag

Databases

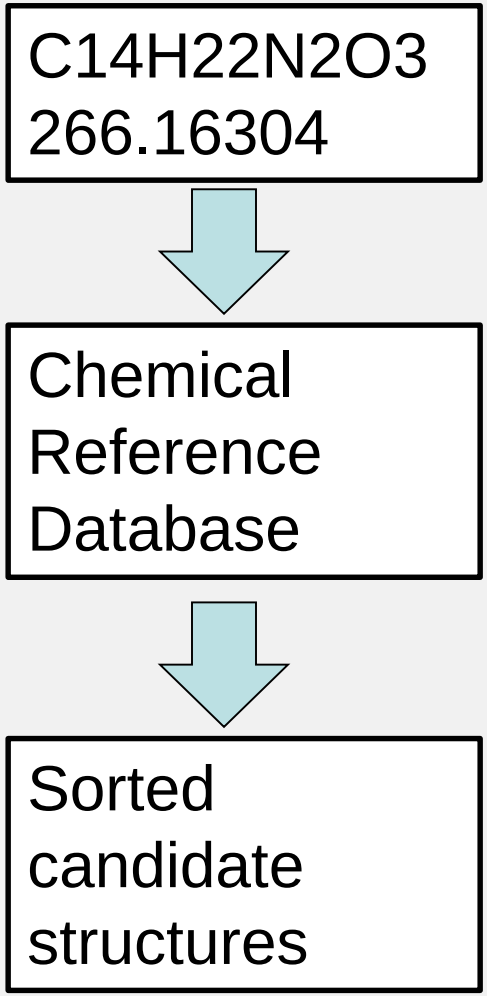
ChemSpider
Search and share chemistry

PubChem

MassBank

Data Source Ranking of “known unknowns”

- Mass and/or formula unknown to a researcher, contained within a reference database
- Most likely candidate chemicals have the most references/sources



Initial Data Source Ranking in ChemSpider

- Adopted by NTA researchers around the world

© American Society for Mass Spectrometry, 2011

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

RESEARCH ARTICLE

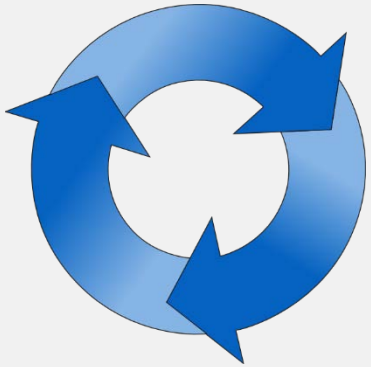
Identification of "Known Unknowns" Utilizing Accurate Mass Data and ChemSpider

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	

The General Approach

Analytical Instruments



Comp. Tools & Workflows

Databases

ChemSpider
Search and share chemistry

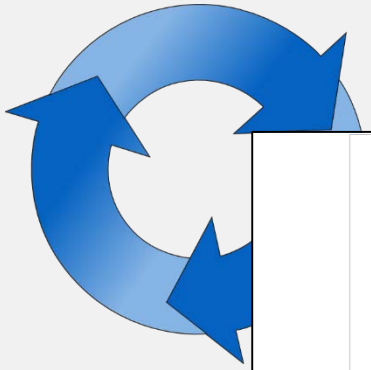
PubChem

MassBank

The General Approach

Analytical Instruments

Comp. Tools & Workflows



FOR IDENT

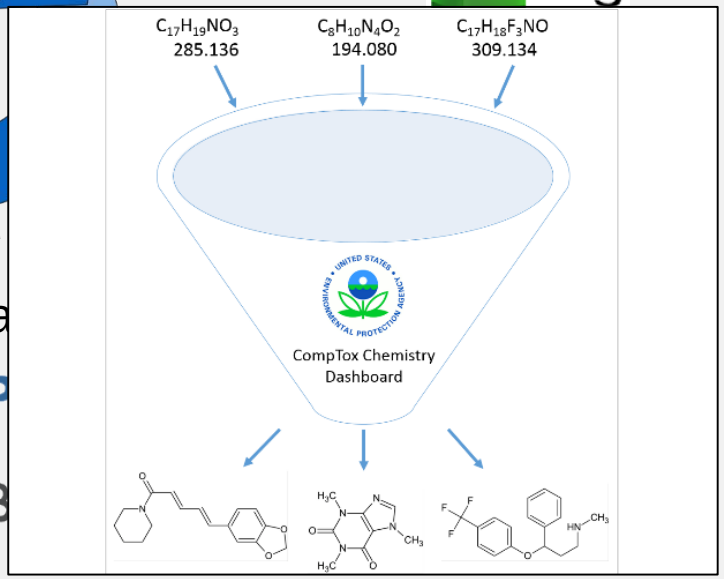
Xcms Online

MetFrag

Databa

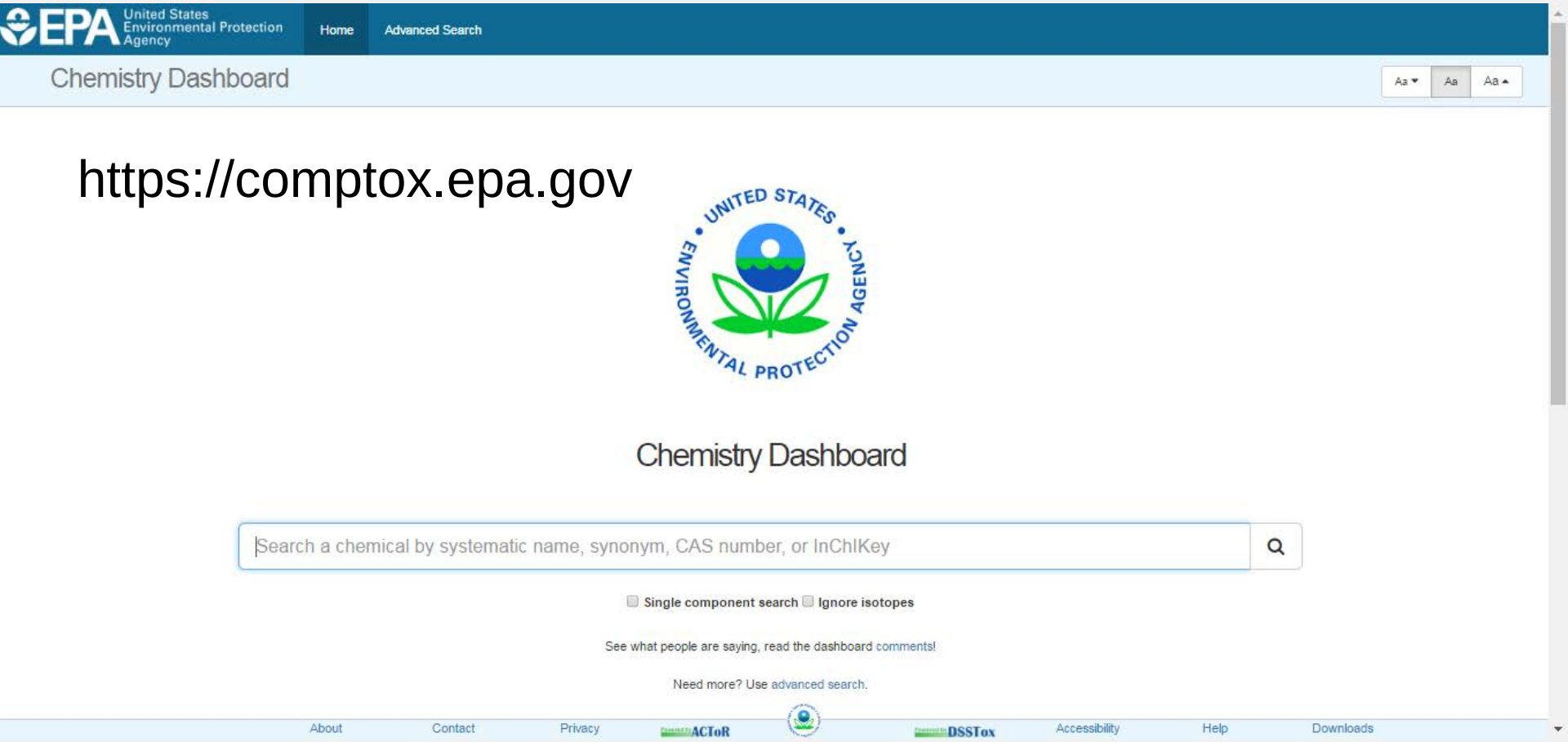
ChemSpider
Search and share chemistry

MassB



EPA United States Environmental Protection Agency

CompTox Chemistry Dashboard



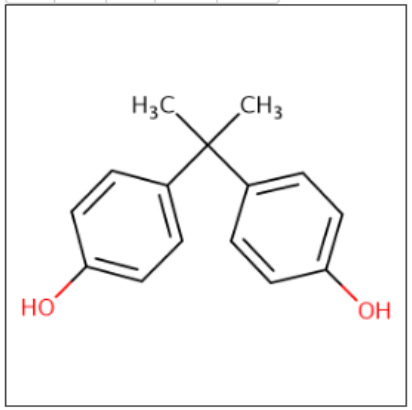
EPA United States Environmental Protection Agency

CompTox Chemistry Dashboard

Bisphenol A

80-05-7 | DTXSID7020182

Searched by Expert Validated Synonym: Found 1 result for 'bpa'.



Wikipedia

Bisphenol A (BPA) is an organic synthetic compound with the chemical formula (CH3)2C(C6H4OH)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
- Related Compounds (Beta)
- Presence in Lists
- Record Information

EPA United States Environmental Protection Agency

CompTox Chemistry Dashboard

Bisphenol A

80-05-7 | DTXSID7020182

Searched by Expert Validated Synonym: Found 1 result for 'bpa'.

Chemical Properties | Env. Fate/Transport | Synonyms | External Links | Toxicity Values (Beta) | Exposure | Bioassays | Similar Molecules (Beta) | Literature | Comments

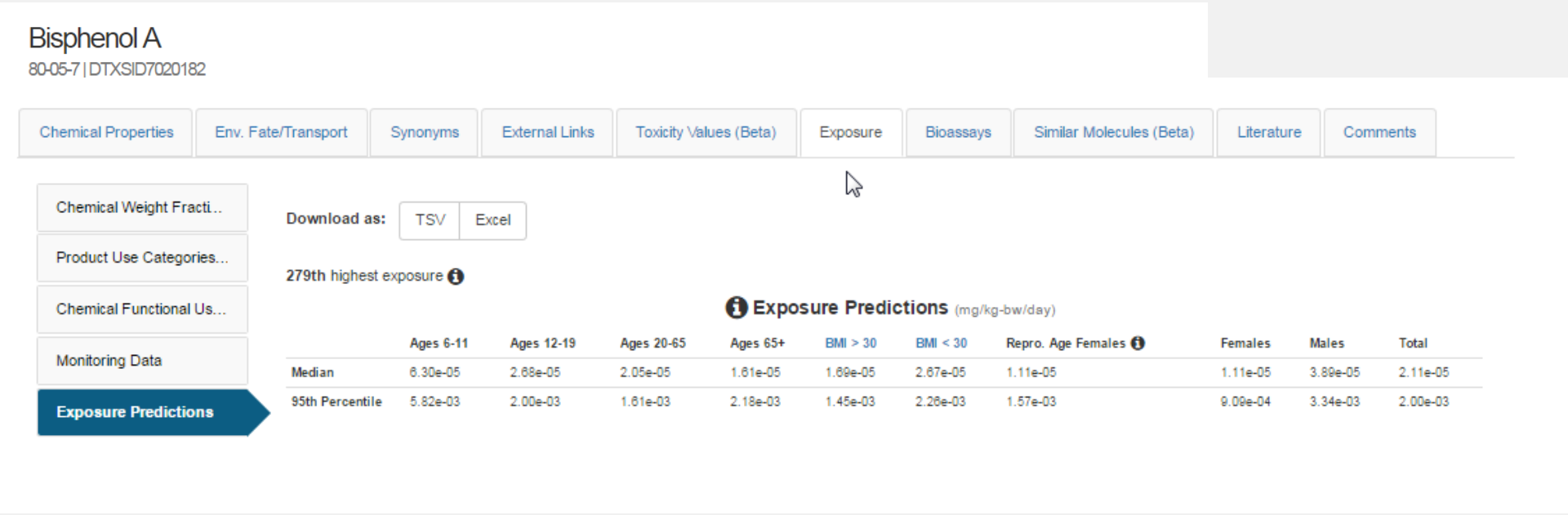
- **Summary**
- LogP: Octanol-Water
- Water Solubility
- Density
- Melting Point
- Boiling Point
- Surface Tension
- Vapor Pressure
- LogKoa: Octanol-Air
- Henry's Law
- Index of Refraction

Download as: TSV | Excel | SDF

Property	Average		Median		Range		Unit
	Experimental	Predicted	Experimental	Predicted	Experimental	Predicted	
LogP: Octanol-Water	3.32 (1)	3.24 (4)	3.32 to 3.32	3.24	3.32	2.40 to 3.73	-
Water Solubility	5.26e-04 (1)	1.58e-03 (4)	5.26e-04 to 5.26e-04	1.58e-03	5.26e-04	5.70e-04 to 3.68e-03	mol/L
Density	-	1.14 (1)	-	1.14	-	-	g/cm^3
Melting Point	155 (7)	144 (3)	153 to 158	144	153 to 158	132 to 157	°C
Boiling Point	200 (1)	349 (3)	200 to 200	349	200	334 to 364	°C
Surface Tension	-	46.0 (1)	-	46.0	-	-	dyn/cm
Vapor Pressure	-	2.52e-07 (3)	-	2.52e-07	-	7.01e-08 to 5.34e-07	mmHg
LogKoa: Octanol-Air	-	8.38 (1)	-	8.38	-	-	-
Henry's Law	-	6.96e-07 (1)	-	6.96e-07	-	-	atm-m3/mole
Index of Refraction	-	1.60 (1)	-	1.60	-	-	-
Molar Refractivity	-	68.2 (1)	-	68.2	-	-	cm^3
pKa Acidic Apparent	-	10.3 (1)	-	10.3	-	-	-
Molar Volume	-	200 (1)	-	200	-	-	cm^3

EPA United States Environmental Protection Agency

CompTox Chemistry Dashboard



	Ages 6-11	Ages 12-19	Ages 20-65	Ages 65+	BMI > 30	BMI < 30	Repro. Age Females	Females	Males	Total
Median	6.30e-05	2.68e-05	2.05e-05	1.61e-05	1.69e-05	2.67e-05	1.11e-05	1.11e-05	3.89e-05	2.11e-05
95th Percentile	5.82e-03	2.00e-03	1.61e-03	2.18e-03	1.45e-03	2.26e-03	1.57e-03	9.09e-04	3.34e-03	2.00e-03

EPA United States Environmental Protection Agency

CompTox Chemistry Dashboard

- Chemical Weight Fraction (Beta)
- Product Use Categories (Beta)
- **Chemical Functional Use (Beta)**
- Monitoring Data
- Exposure Predictions

Predicted Probability of Associated Functional Use

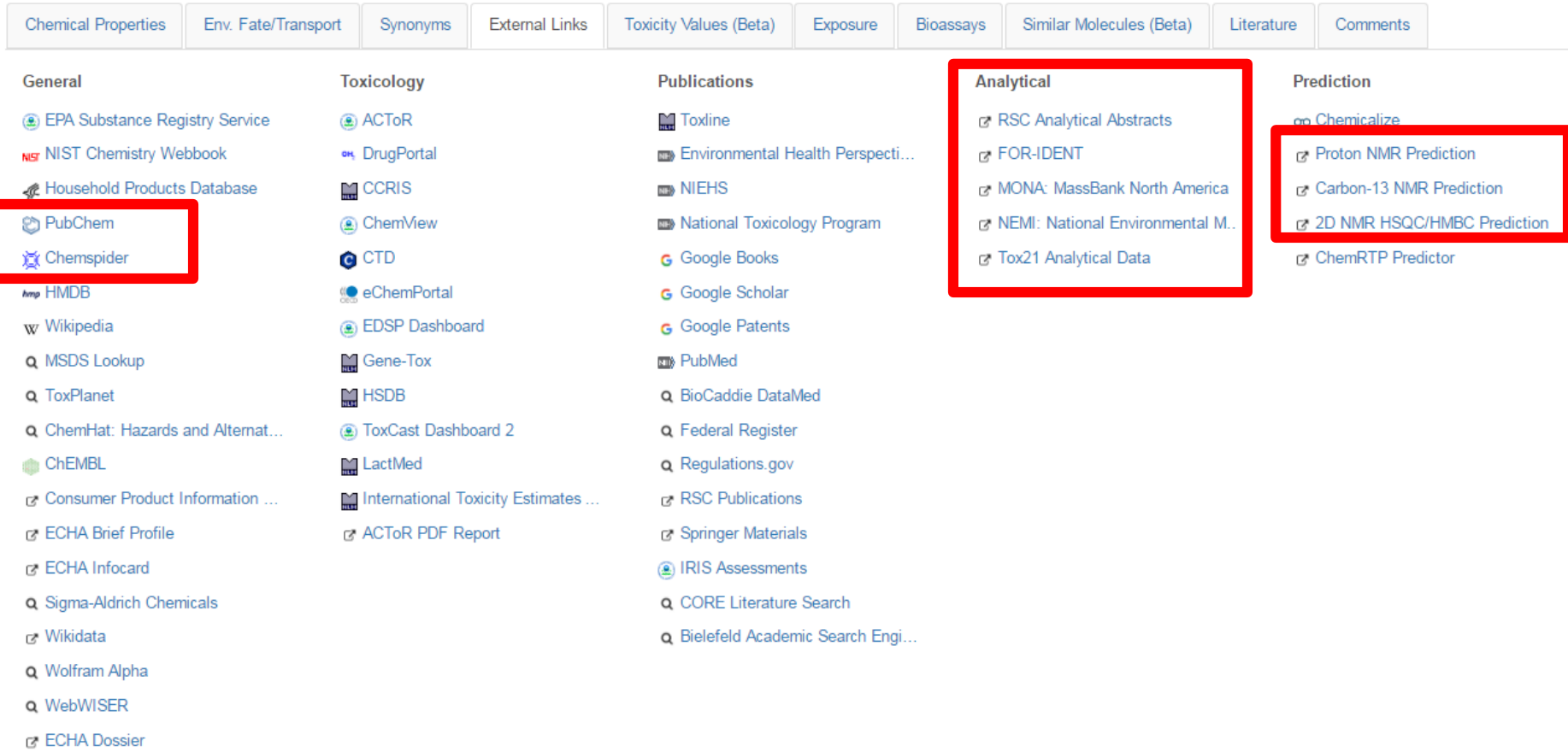
QSAR Version/Date: 2015-11-06

Download as: TSV | Excel

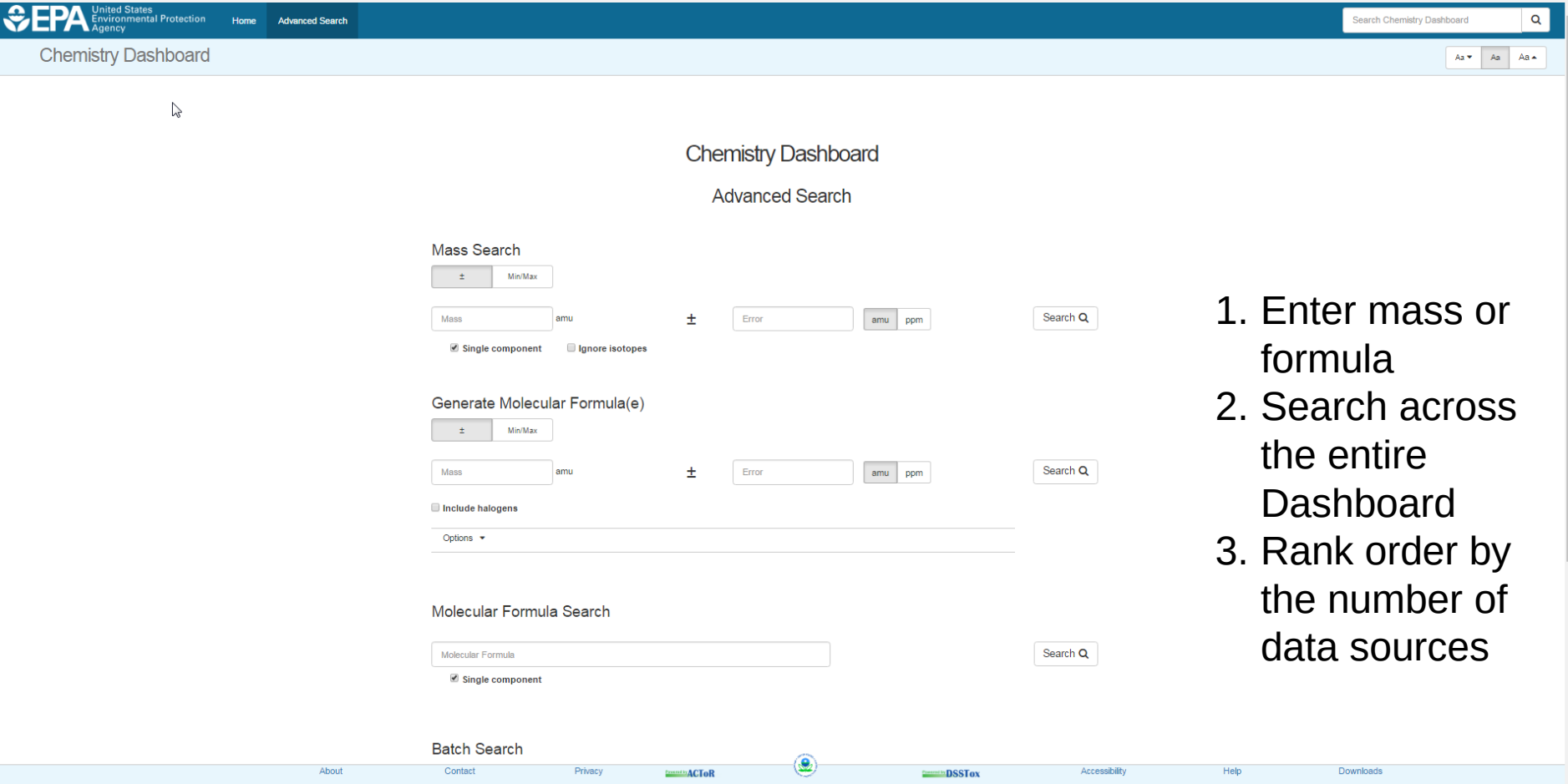
Harmonized Functional Use	Probability
skin_protectant	0.66
hair_conditioner	0.58
antimicrobial	0.57
masking_agent	0.55
skin_conditioner	0.53
antioxidant	0.51
colorant	0.50
uv_absorber	0.44
soluble_dye	0.43
hair_dye	0.42
crosslinker	0.42
photoinitiator	0.41
humectant	0.41
additive_for_rubber	0.41

EPA United States Environmental Protection Agency

CompTox Chemistry Dashboard



Data Source Ranking in the Dashboard



1. Enter mass or formula
2. Search across the entire Dashboard
3. Rank order by the number of data sources

Anal Bioanal Chem
DOI 10.1007/s00216-016-0139-z

RAPID COMMUNICATION

Identifying known unknowns using the US EPA's CompTox Chemistry Dashboard

Andrew D. McEachran[1] · Jon R. Sobus[2] · Antony J. Williams[3]

- On same 162 chemicals, Dashboard outperforms ChemSpider

	Mass-based searching		Formula-based searching	
	Dashboard	ChemSpider	Dashboard	ChemSpider
Average rank position	1.3	2.2[a]	1.2	1.4
Percent in #1 position	85%	70%	88%	80%

[a] Average rank in ChemSpider shown here does not include an outlier where the rank was 201, when added the average rank position is 3.5

Dashboard

Compound class	Number in class	Average rank	Number of compounds in each position rank-ordered				
			#1	#2	#3	#4	#5+
Pharmaceutical drug	72	1.3	59	8	3	2	
Industrial chemicals	42	1.2	38	1	1	2	
Personal care products	8	2.6	6				2
Steroid hormones	7	1.0	7				
Perfluorochemicals	6	1.3	5	1			
Pesticides	12	1.3	10	1	1		
Veterinary drugs	3	1.0	3				
Dyes	2	1.0	2				
Food product/natural compounds	4	1.5	3		1		
Illicit drugs	2	1.5	1	1			
Misc. molecules	3[a]	1.0	3				

[a] One organic molecule (tephrosin) not present in the Dashboard

ChemSpider

Compound class	Number in class	Average rank	Number of compounds in each position rank-ordered				
			#1	#2	#3	#4	#5+
Pharmaceutical drug	72	1.4	55	9	6	2	
Industrial chemicals	42	5.5	28	6	3		5
Personal care products	8	6.1	3	1			4
Steroid hormones	7	1.0	7				
Perfluorochemicals	6	1.2	5	1			
Pesticides	12	2.3	6	2	3		1
Veterinary drugs	3	1.3	2	1			
Dyes	2	1.0	2				
Food product/natural compounds	4	3.8	2			1	1
Illicit drugs	2	2.0	1		1		
Misc. molecules	3 [a]	1.3	2	1			

[a] Tephrosin was removed from average rank calculations as it was not present in a Dashboard search

Original Test Set

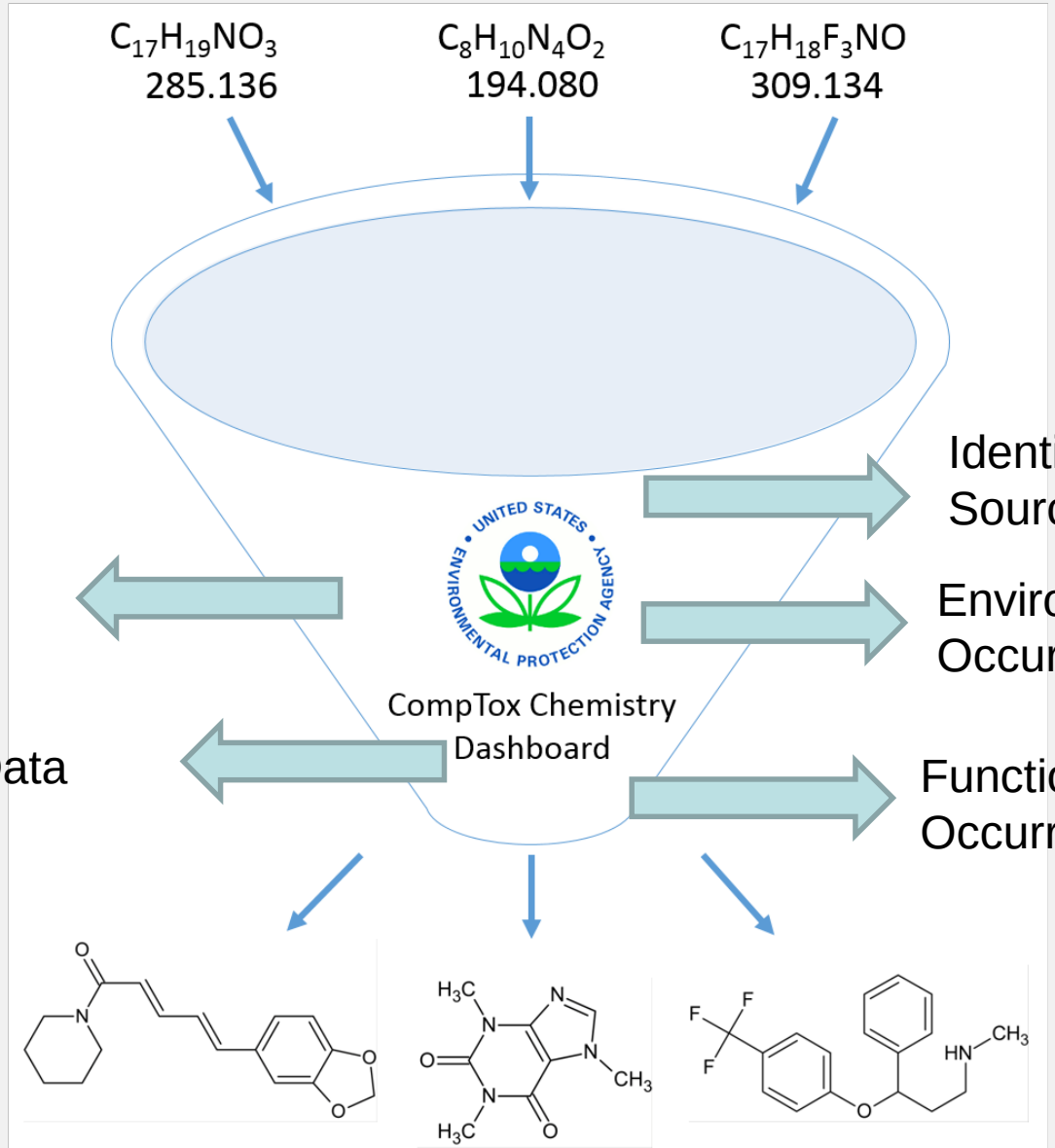
		Average rank (±SD)	Number in each position rank-ordered				
			#1	#2	#3	#4	#5+
Mass-based	Dashboard	1.2 ± 0.7	77[a]	5	3	3	
	ChemSpider	2.2 ± 6.1[b]	68	8	7	1	5
Formula-based	Dashboard	1.1 ± 0.4	78[a]	8	2		
	ChemSpider	1.3 ± 1.0	77	8	2	1	2

[a] One chemical (tephrosin) not present in the Dashboard

[b] Average rank in ChemSpider shown here does not include an outlier where the rank was 201, when added the average rank position is 4.4

McEachran, A.D., Sobus, J.R. & Williams, A.J. Anal Bioanal Chem (2016). doi:10.1007/s00216-016-0139-z

Dashboard in NTA Workflows



Retention Time Prediction

MS/MS Data

Identification by Data Source Ranking

Environmental Media Occurrence

Functional Use/Product Occurrence

MS-Ready structures

- De-salted, de-solvated, no stereochemistry, separation of mixtures and multi-component structures

	A	B	C	D	E	F
1	INPUT	DTXSID	CASRN	PREFERRED NAME	MOL FORMULA	MONOISOTOPIC MASS
2	C6H6O	DTXSID5021124	108-95-2	Phenol	C6H6O	94.041864813
3	C6H6O	DTXSID4027072	139-02-6	Sodium phenolate	C6H5NaO	116.02380906
4	C6H6O	DTXSID8073261	1487-18-9	Furan, 2-ethenyl-	C6H6O	94.041864813
5	C6H6O	DTXSID10219242	6921-27-3	2-Propynyl ether	C6H6O	94.041864813
6	C6H6O	DTXSID10183353	291-70-3	Oxepin	C6H6O	94.041864813
7	C6H6O	DTXSID7064073	5973-17-1	Phenol, ammonium salt	C6H9NO	111.068413914
8	C6H6O	DTXSID90363757	6569-83-1	7-oxabicyclo[2.2.1]hepta-2,5-diene	C6H6O	94.041864813
9	C6H6O	DTXSID60179347	24599-57-3	2,4-Cyclohexadienone	C6H6O	94.041864813
10	C6H6O	DTXSID5070109	64601-04-3	Phenol, compd. with 2-aminoethanol (1:1)	C8H13NO2	155.094628663
11	C6H6O	DTXSID9075294	2122-46-5	Phenoxy	C6H5O	93.034039781
12						

Batch Searching of Unknowns

Batch Search

Please enter one identifier per line

Select Input Type(s)

- [ ] Chemical Name
- [ ] CAS-RN
- [ ] InChIKey
- [ ] DSSTox Substance ID
- [x] Exact Molecular Formula

Include top 10 hits in download

Enter Identifiers to Search

C14H22N2O3
C18H34N2O6S
C10H12N2O
C14H18N4O3
C12H11N7
C8H9NO2
C7H8N4O2
C38H72N2O12
C17H21NO
C8H10N4O2

Display All Chemicals | Download Chemical Data

Select Output Format

Excel

Customize Results

- [ ] Select All

Chemical Identifiers

- [x] Chemical Name
- [x] DTXSID
- [x] CAS-RN
- [ ] InChIKey
- [ ] IUPAC Name

Structures

- [ ] Mol File
- [ ] SMILES
- [ ] InChI String

Metadata

- [ ] Curation Level Details
- [x] Data Sources
- [ ] Assay Hit Count
- [ ] NHANES/Predicted Exposure
- [ ] Include ToxVal Data Availability

Intrinsic And Predicted Properties

- [x] Molecular Formula
- [ ] Average Mass
- [x] Monoisotopic Mass
- [ ] OPERA and TEST Model Predictions

Download

United States Environmental Protection Agency

Batch Searching of Unknowns

	A	B	C	D	E	F	G
1	INPUT	DTXSID	CASRN	PREFERRED NAME	MOL FORMULA	MONOISOTOPIC MASS	DATA SOURCES
2	C14H22N2O3	DTXSID2022628	29122-68-7	Atenolol	C14H22N2O3	266.163042576	46
3	C14H22N2O3	DTXSID0021179	6673-35-4	Practolol	C14H22N2O3	266.163042576	32
4	C14H22N2O3	DTXSID4048854	841-73-6	Bucolome	C14H22N2O3	266.163042576	20
5	C14H22N2O3	DTXSID1045407	13171-25-0	Trimetazidine dihydrochloride	C14H24Cl2N2O3	338.116398	19
6	C14H22N2O3	DTXSID0045753	56715-13-0	R-(+)-Atenolol	C14H22N2O3	266.163042576	19
7	C14H22N2O3	DTXSID2048531	5011-34-7	Trimetazidine	C14H22N2O3	266.163042576	14
8	C14H22N2O3	DTXSID10239405	93379-54-5	Esatenolol	C14H22N2O3	266.163042576	12
9	C14H22N2O3	DTXSID50200634	52662-27-8	N-(2-Diethylaminoethyl)-2-(4-hydroxyphenoxy)acetamide	C14H22N2O3	266.163042576	7
10	C14H22N2O3	DTXSID4020111	51706-40-2	dl-Atenolol hydrochloride	C14H23ClN2O3	302.1397203	6
11	C14H22N2O3	DTXSID1068693	51963-82-7	Benzenamine, 2,5-diethoxy-4-(4-morpholinyl)-	C14H22N2O3	266.163042576	5
12	C18H34N2O6S	DTXSID3023215	154-21-2	Lincomycin	C18H34N2O6S	406.213757997	35
13	C18H34N2O6S	DTXSID7047803	859-18-7	Lincomycin hydrochloride	C18H35ClN2O6S	442.1904357	22
14	C18H34N2O6S	DTXSID20849438	1398534-62-7	PUBCHEM_71432748	C18H35ClN2O6S	442.1904357	1
15	C10H12N2O	DTXSID1047576	486-56-6	Cotinine	C10H12N2O	176.094963014	40
16	C10H12N2O	DTXSID8075330	50-67-9	Serotonin	C10H12N2O	176.094963014	22
17	C10H12N2O	DTXSID8044412	2654-57-1	4-Methyl-1-phenylpyrazolidin-3-one	C10H12N2O	176.094963014	18
18	C10H12N2O	DTXSID80165186	153-98-0	Serotonin hydrochloride	C10H13ClN2O	212.0716407	11
19	C10H12N2O	DTXSID2048870	29493-77-4	(4R,5S)-4-methyl-5-phenyl-4,5-dihydro-1,3-oxazol-2-amine	C10H12N2O	176.094963014	10
20	C10H12N2O	DTXSID10196105	443-31-2	6-Hydroxytryptamine	C10H12N2O	176.094963014	9
21	C10H12N2O	DTXSID90185693	31822-84-1	1,4,5,6-Tetrahydro-5-phenoxypyrimidine	C10H12N2O	176.094963014	7
22	C10H12N2O	DTXSID40178777	2403-66-9	2-Benzimidazolepropanol	C10H12N2O	176.094963014	7
23	C10H12N2O	DTXSID80157026	13140-86-8	N-Cyclopropyl-N'-phenylurea	C10H12N2O	176.094963014	6
24	C10H12N2O	DTXSID30205607	570-14-9	4-Hydroxytryptamine	C10H12N2O	176.094963014	6
25	C14H18N4O3	DTXSID5023900	17804-35-2	Benomyl	C14H18N4O3	290.137890456	68
26	C14H18N4O3	DTXSID3023712	738-70-5	Trimethoprim	C14H18N4O3	290.137890456	51
27	C14H18N4O3	DTXSID40209671	60834-30-2	Trimethoprim hydrochloride	C14H19ClN4O3	326.1145682	8
28	C14H18N4O3	DTXSID70204210	55687-49-5	Benzenemethanol, 4-((2,4-diamino-5-pyrimidinyl)methyl)-2,	C14H18N4O3	290.137890456	5
29	C14H18N4O3	DTXSID20152671	120075-57-2	6-Methoxy-4-(3-(N,N-dimethylamino)propylamino)-5,8-quina	C14H18N4O3	290.137890456	4
30	C14H18N4O3	DTXSID30213742	63931-79-3	1H-1,2,4-Benzotriazepine-3-carboxylic acid, 4,5-dihydro-4-	C14H18N4O3	290.137890456	3
31	C14H18N4O3	DTXSID30219608	69449-07-6	2,4-Pyrimidinediamine, 5-((3,4,5-trimethoxyphenyl)methyl)-	C14H20N4O4	308.14845514	3
32	C14H18N4O3	DTXSID20241155	94232-27-6	L-Aspartic acid, compound with 5-((3,4,5-trimethoxyphenyl	C18H25N5O7	423.175398165	3
33	C14H18N4O3	DTXSID80241156	94232-28-7	L-Glutamic acid, compound with 5-((3,4,5-trimethoxypheny	C19H27N5O7	437.191048229	3
34	C14H18N4O3	DTXSID20143781	101204-93-7	1H-Pyrido(2,3-e)-1,4-diazepine-2,3,5-trione, 4-(2-(diethylam	C14H18N4O3	290.137890456	3
35	C12H11N7	DTXSID6021373	396-01-0	Triamterene	C12H11N7	253.107593382	52
36	C12H11N7	DTXSID00204465	5587-93-9	Ampyrimine	C12H11N7	253.107593382	7
37	C12H11N7	DTXSID5064621	7300-26-7	Benzenamine, 4-azido-N-(4-azidophenyl)-	C12H9N7	251.091943318	4
38	C12H11N7	DTXSID00848025	90293-82-6	Sulfuric acid--6-phenylpteridine-2,4,7-triamine (1/1)	C12H13N7O4S	351.074973101	1
39	C12H11N7	DTXSID50575293	92310-83-3	(1E)-N-Phenyl-1,2-bis(1H-1,2,4-triazol-1-yl)ethan-1-imine	C12H11N7	253.107593382	1
40	C8H9NO2	DTXSID2020006	103-90-2	Acetaminophen	C8H9NO2	151.063328534	75
41	C8H9NO2	DTXSID6026567	134-20-3	Methyl 2-aminobenzoate	C8H9NO2	151.063328534	50

Conclusions

- NTA requires resources for accurate structure identification
- CompTox Chemistry Dashboard outperforms ChemSpider in data source ranking
- Dashboard contains diverse data, will enhance NTA workflows

Acknowledgements

EPA NCCT
Tony Williams
Chris Grulke
John Wambaugh
Kamel Mansouri*
Jeff Edwards
Ann Richard
Jennifer Smith

EPA NERL
Katherine Phillips
Kristin Isaacs
Kathie Dionisio
Jon Sobus
Mark Strynar
Elin Ulrich
Seth Newton
Jarod Grossman
Sarah Laughlin-Toth*
Aurelie Marcotte*

NIEHS
Brandy Beverly

*ORISE Research Participant

EPA United States Environmental Protection Agency

Questions?

- mceachran.andrew@epa.gov
- http://orcid.org/0000-0003-1423-330X
- Associated presentations:
  - 4/2 ANYL 76, AJ Williams
  - 4/2 ANYL 77, AD McEachran
  - 4/6 ENVR 1012, K. Phillips
  - 4/6 ANYL 470, AD McEachran