

# Structure identification approaches using the US EPA CompTox Chemicals Dashboard to support mass spectrometry analyses

#### **Antony Williams and John Sloop**

Center for Computational Toxicology and Exposure, US-EPA, 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

#### Outline

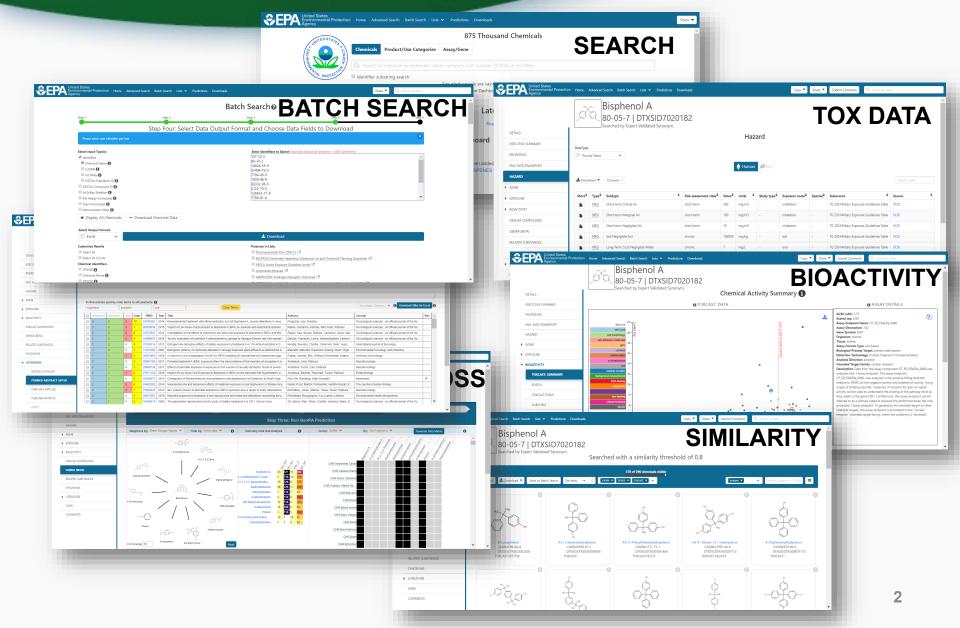


- Quick overview of the dashboard
- Support for Mass Spectrometry
- Data quality in the public domain
- Work in progress proof-of-concepts (PoCs)
- Applications of the Dashboard and PoCs in the hands of a mass spec scientist

#### CompTox Chemicals Dashboard







#### How we are supporting metabolomics



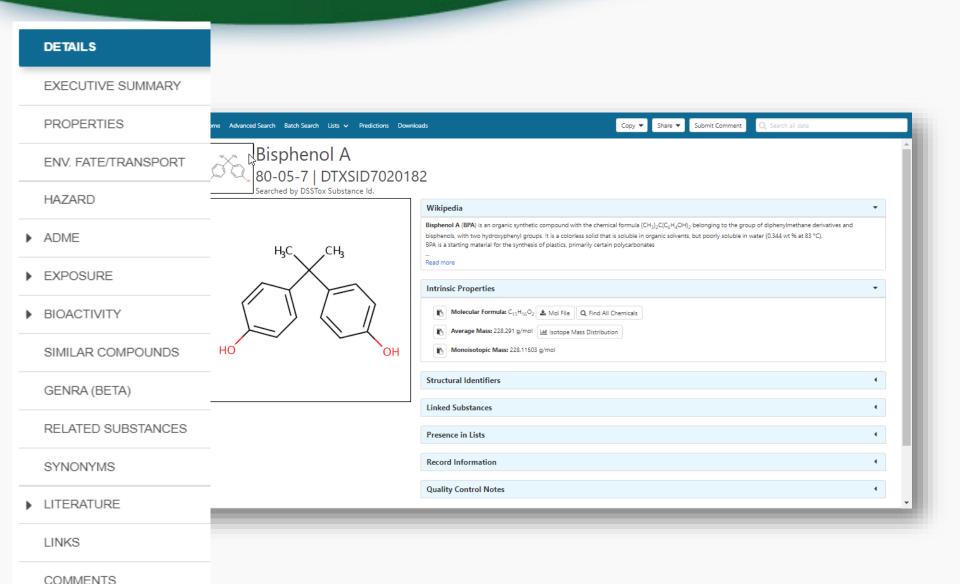
- The Dashboard as a "first-stop shop" for toxicity, exposure and predictions data
- Highly curated data sets with dedicated curation team led by scientists



- Data set development is project focused based on specific activities of interest
- Collaborating with multiple teams around data sets and functionality development

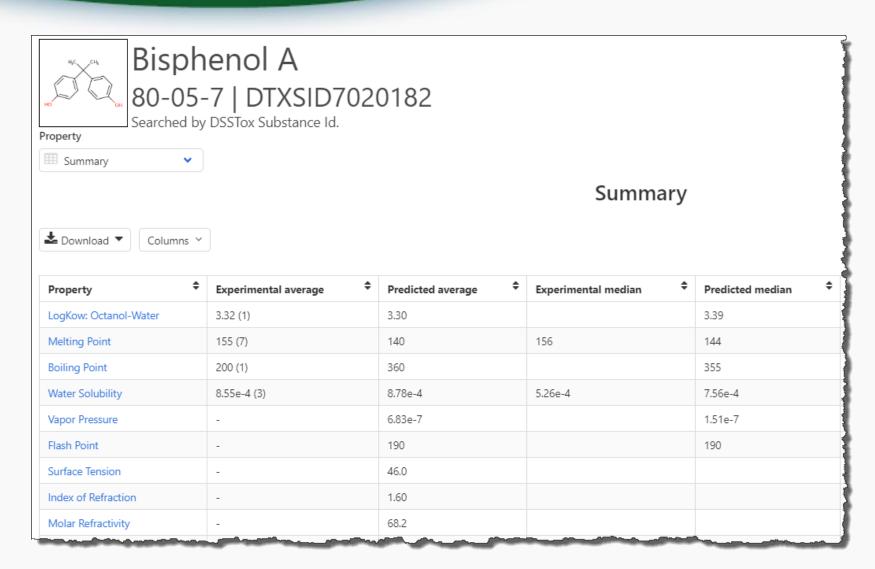
#### **Detailed Chemical Pages**





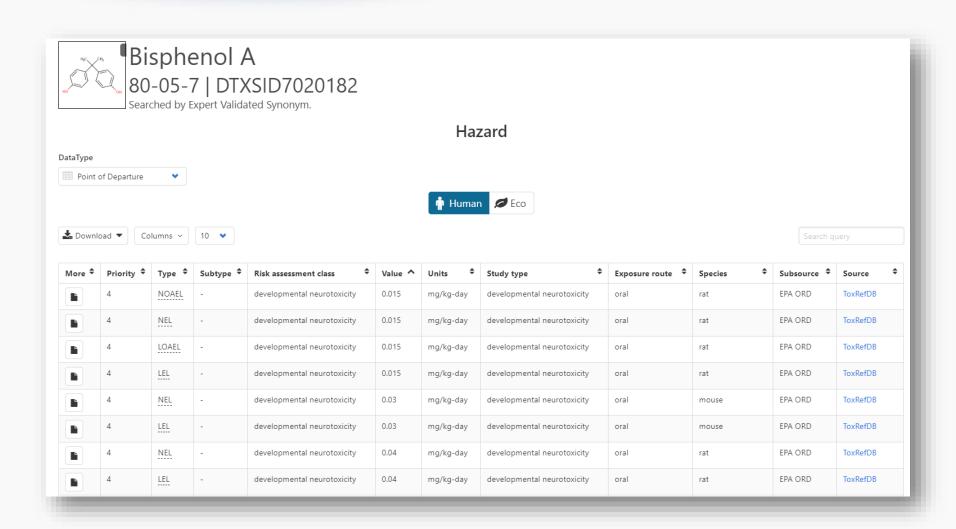
#### Properties, Fate and Transport





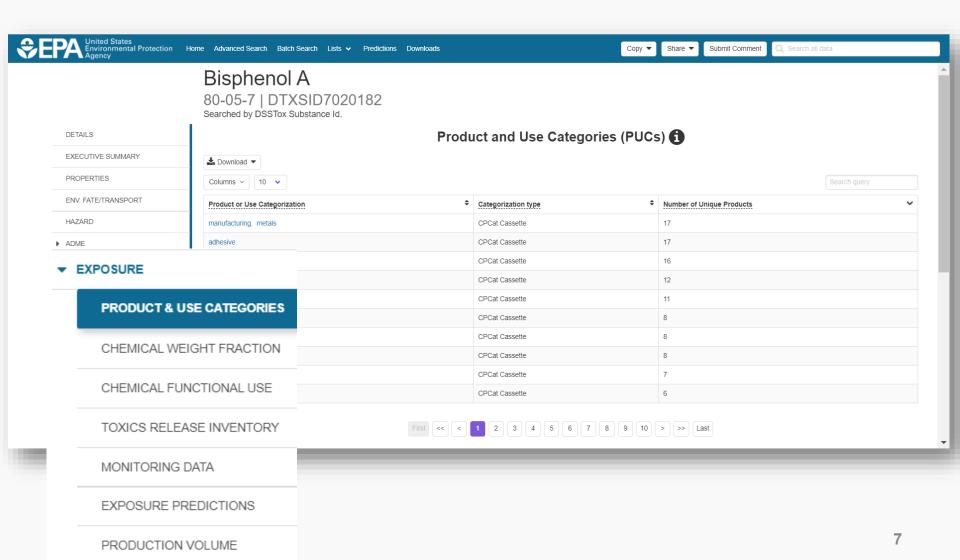
#### Hazard





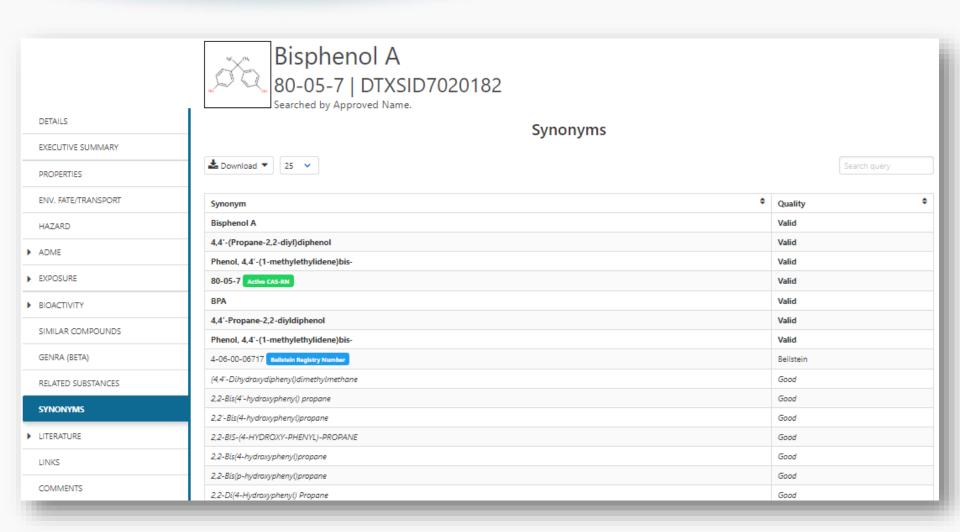
#### Sources of Exposure to Chemicals





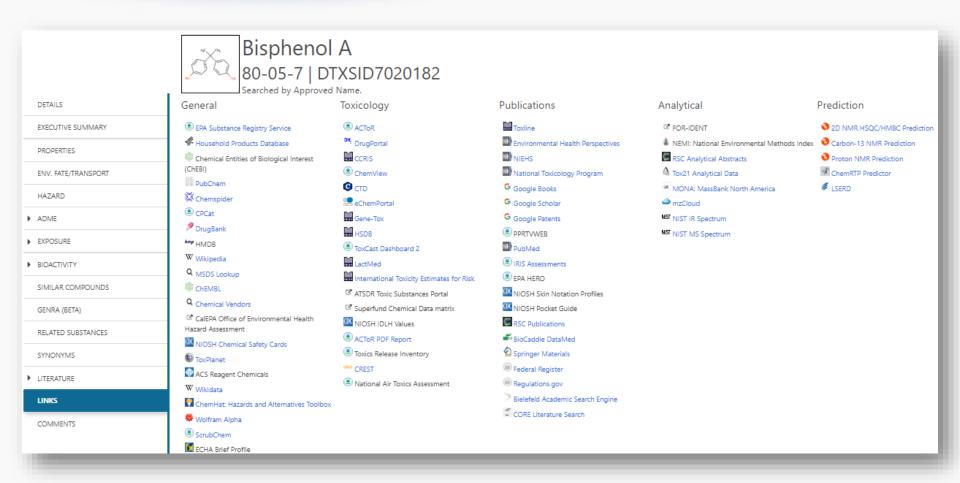
#### Identifiers to Support Searches





#### Link Access





#### Mass Spec Links

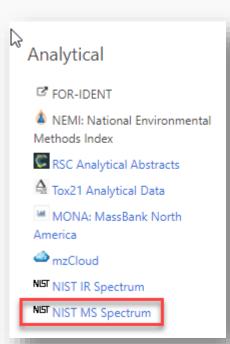


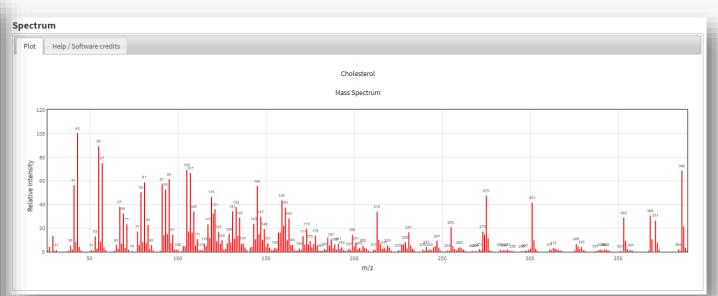
#### Analytical

- RSC Analytical Abstracts
- ♠ Tox21 Analytical Data
- MONA: MassBank North America
- mzCloud
- NIST IR Spectrum
- NIST MS Spectrum
- MassBank
- NEMI: National Environmental Methods Index
- NIST Antoine Constants
- IR Spectra on PubChem
- NIST Kovats Index values

## NIST WebBook https://webbook.nist.gov/chemistry/

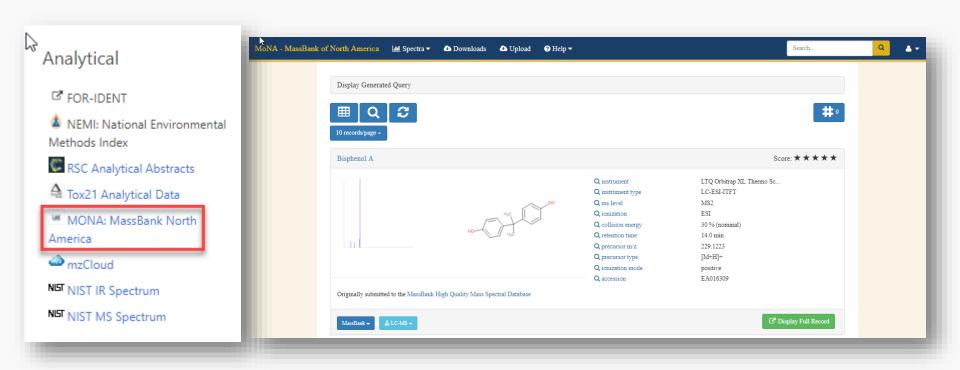






## MassBank of North America https://mona.fiehnlab.ucdavis.edu







# Batch Searching

#### Aggregate data for a list of chemicals





## Trends in Environmental Analytical Chemistry



Volume 20, October 2018, e00059

## Opioid occurrence in environmental water samples—A review

Marina Celia Campos-Mañas ³, Imma Ferrer <sup>b</sup> △ 🖾, E.Michael Thurman <sup>b</sup>, Ana Agüera ³

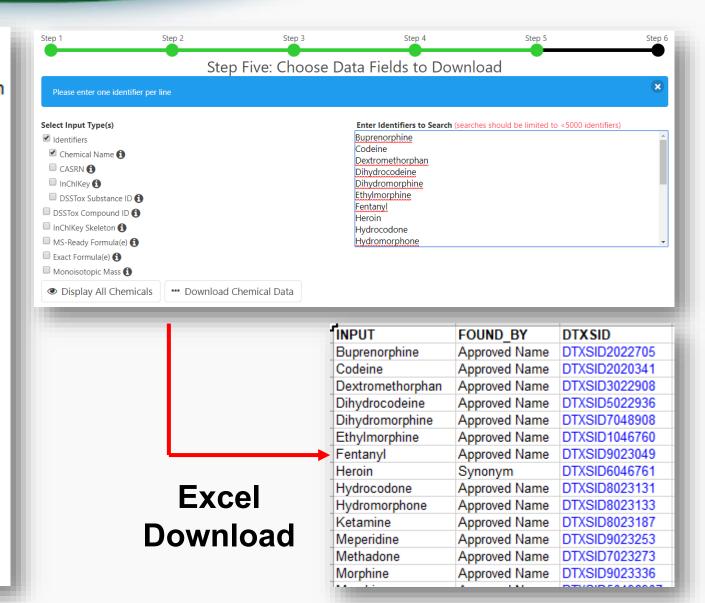
https://doi.org/10.1016/j.teac.2018.e00059

Get rights and content

#### **Batch Search Names**



Buprenorphine Codeine Dextromethorphan Dihydrocodeine Dihydromorphine Ethylmorphine Fentanyl Heroin Hydrocodone Hydromorphone Ketamine Meperidine Methadone Morphine Morphinone Naloxone Naltriben 0xycodone Oxymorphone Propoxyphene Sufentanil Tramadol



#### Add Other Data of Interest



#### Chemical Identifiers

- ✓ DTXSID
- Chemical Name
- ☐ DTXCID **(**)
- ✓ CAS-RN
- ✓ InChlKey <a>f</a>
- ☐ IUPAC Name **(1)**

#### Structures

- ☐ Mol File 🚯
- SMILES 1
- InChl String
- ✓ MS-Ready SMILES
- QSAR-Ready SMILES (1)

#### **Intrinsic And Predicted Properties**

- Molecular Formula 6
- Average Mass < 1</p>
- ✓ Monoisotopic Mass 

  ⑤
- TEST Model Predictions
- OPERA Model Predictions

INPUT	DTXSID	CASRN	MOLECULAR_FO	MONOISOTOPIC	MS_READY_SMI
Buprenorph	DTXSID202	52485-79-7	C29H41NO4	467.3035588	[H]C12CC3=C4C
Codeine	DTXSID202	76-57-3	C18H21NO3	299.1521435	[H]C12CC3=C4C
Dextrometh	DTXSID302	125-71-3	C18H25NO	271.1936144	[H]C12CC3=C(C=
Dihydrocod	DTXSID502	125-28-0	C18H23NO3	301.1677936	[H]C12CC3=C4C
Dihydromor	DTXSID704	509-60-4	C17H21NO3	287.1521435	[H]C12CC3=C4C
Ethylmorph	DTXSID104	76-58-4	C19H23NO3	313.1677936	[H]C12CC3=C4C
Fentanyl	DTXSID902	437-38-7	C22H28N2O	336.2201635	CCC(=O)N(C1CC
Heroin	DTXSID604	561-27-3	C21H23NO5	369.1576228	[H]C12CC3=C4C
Hydrocodon	DTXSID802	125-29-1	C18H21NO3	299.1521435	[H]C12CC3=C4C
Hydromorph	DTXSID802	466-99-9	C17H19NO3	285.1364935	[H]C12CC3=C4C
Ketamine	DTXSID802	6740-88-1	C13H16CINO	237.0920418	CNC1(CCCCC1=
Meperidine	DTXSID902	57-42-1	C15H21NO2	247.1572289	CCOC(=0)C1(CC
Methadone	DTXSID702	76-99-3	C21H27NO	309.2092645	CCC(=O)C(CC(C)
Morphine	DTXSID902	57-27-2	C17H19NO3	285.1364935	[H]C12CC3=C4C
Morphinone	DTXSID501	467-02-7	C17H17NO3	283.1208434	[H]C12CC3=C4C
Naloxone	DTXSID802	465-65-6	C19H21NO4	327.1470582	[H]C12CC3=C4C
Naltriben	-	-	-	-	-
Oxycodone	DTXSID502	76-42-6	C18H21NO4	315.1470582	[H]C12CC3=C4C
Oxymorpho	DTXSID502	76-41-5	C17H19NO4	301.1314081	[H]C12CC3=C4C
Propoxyphe	DTXSID102	469-62-5	C22H29NO2	339.2198292	CCC(=O)OC(CC1
Sufentanil	DTXSID602	56030-54-7	C22H30N2O2S	386.2027994	CCC(=O)N(C1=C
Tramadol	DTXSID908	27203-92-5	C16H25NO2	263.188529	COC1=CC=CC(=
l.					



# Chemical Lists of Interest...

## >300 Chemical Lists (and growing)



Home	Advanced Search	Batch Search	Lists 🕶	Predictions	Downloads
			Lists of Ch		
			List of Assa	ays O	



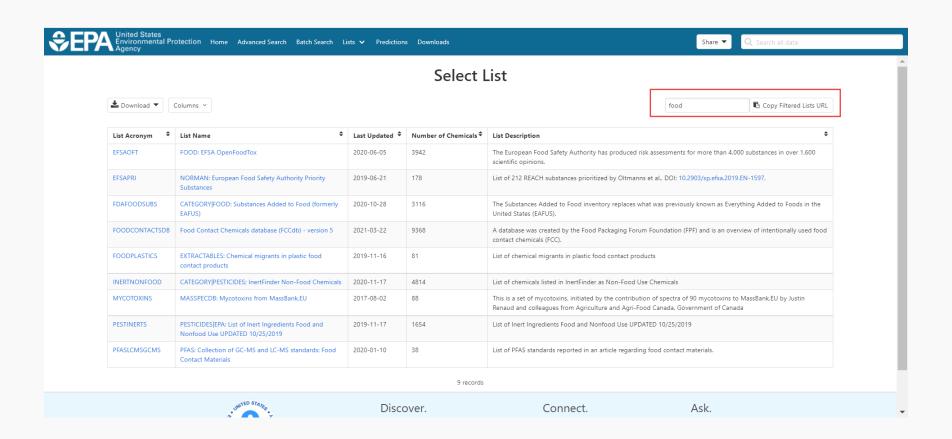
Columns V

mass Copy Filtered Lists URL

List Acronym 🕏	List Name	Last Updated 🕏	Number of Chemicals *	List Description
HDXEXCH	MASSPECDB: Hydrogen Deuterium Exchange Standard Set - Under HDX Conditions	2018-11-07	592	Observed species (deuterated and undeuterated) from the HDXNOEX list under hydrogen deuterium exchange conditions (Ruttkies, Schymanski et al. in prep.)
HDXNOEX	MASSPECDB: Hydrogen Deuterium Exchange Standard Set - No Exchange	2018-11-07	765	Environmental standard set used to investigate hydrogen deuterium exchange in small molecule high resolution mass spectrometry (Ruttkies, Schymanski et al. in prep.)
MASSBANKEUSP	MASSPECDB: MassBank.EU Collection: Special Cases	2017-07-16	263	The MassBank.EU list contains curated chemicals (Schymanski/Williams) associated with the literature/tentative/unknown/SI spectra available on MassBank.EU that are not available as part of the full MassBank collection of reference standard spectra.
MASSBANKREF	MASSPECDB: MassBank Reference Spectra Collection	2017-07-13	1267	This MassBank list contains chemicals associated with the full MassBank collection of reference standard spectra available on MassBank.EU, MassBank.JP and MassBank of North America as well as the Open Data collection, curated by Williams/Schymanski.
MYCOTOXINS	MASSPECDB: Mycotoxins from MassBank.EU	2017-08-02	88	This is a set of mycotoxins, initiated by the contribution of spectra of 90 mycotoxins to MassBank.EU by Justin Renaud and colleagues from Agriculture and Agri-Food Canada, Government of Canada

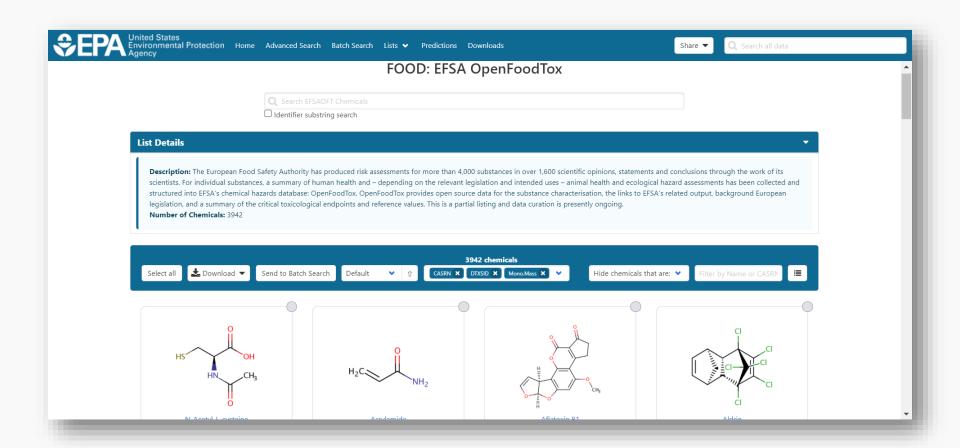
#### Nine "Food" Lists





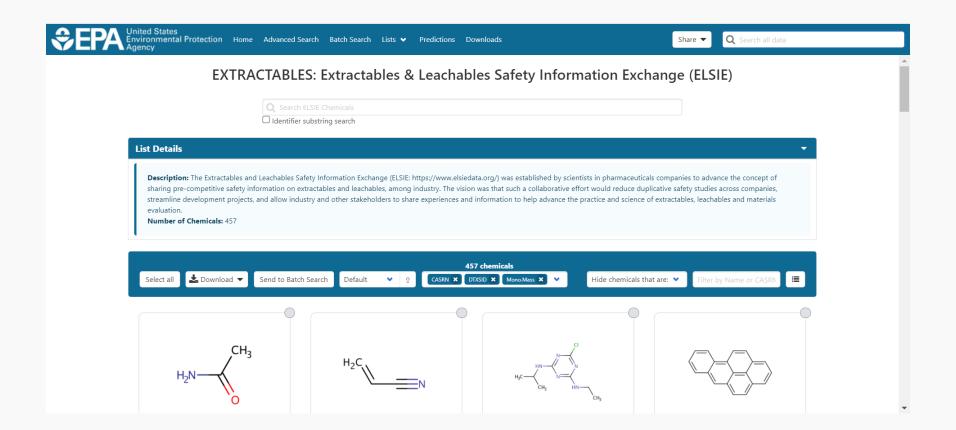
#### EFSA OpenFoodTox





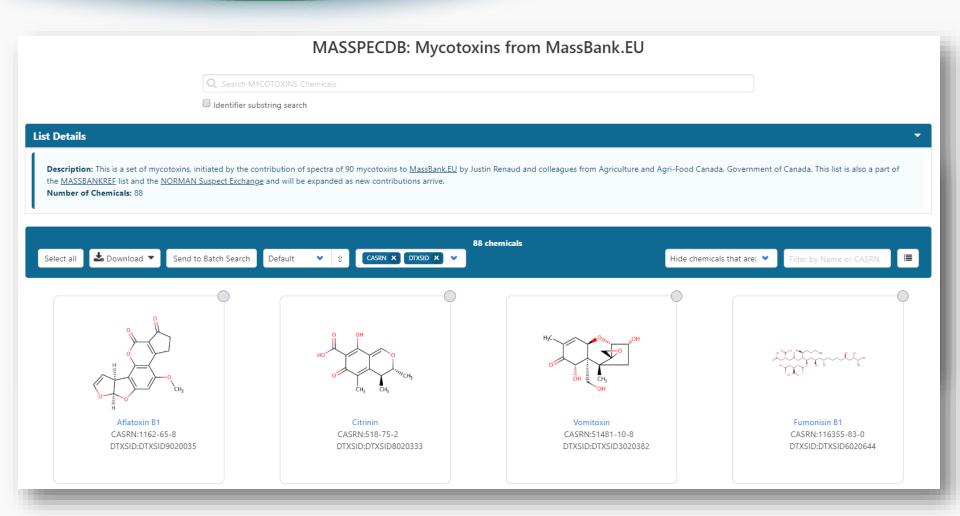
#### ELSIE - Extractables and Leachables





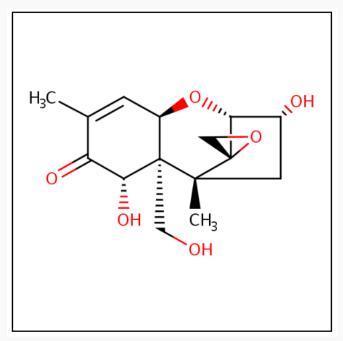
### Mycotoxins

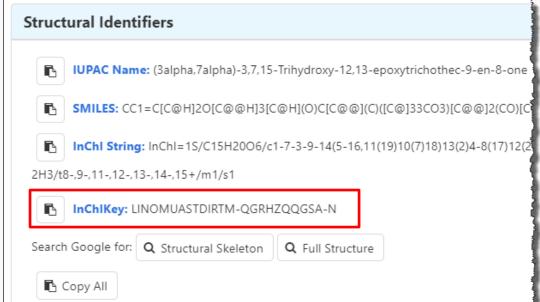




#### Vomitoxin



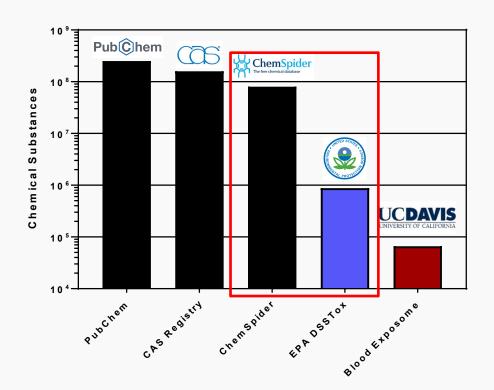




#### **BIG** databases are GREAT!



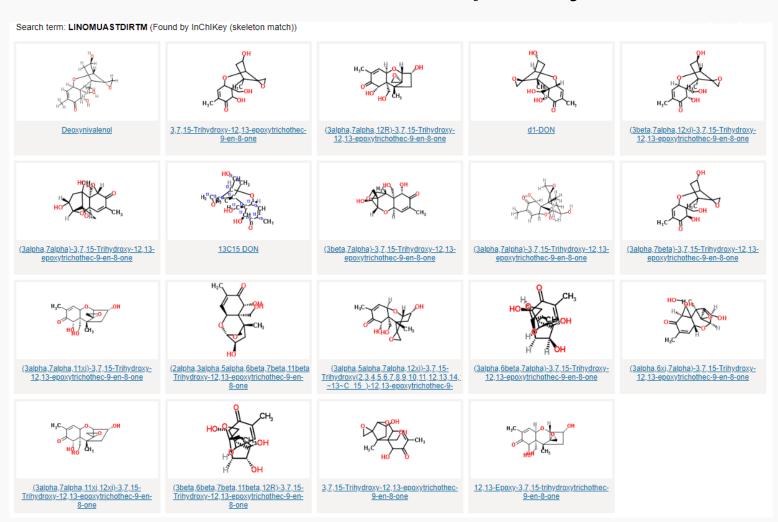
- Thanks to all of the public database efforts
- So much benefit from what's been done
- There are hundreds of them at this point...



#### Vomitoxin - ChemSpider



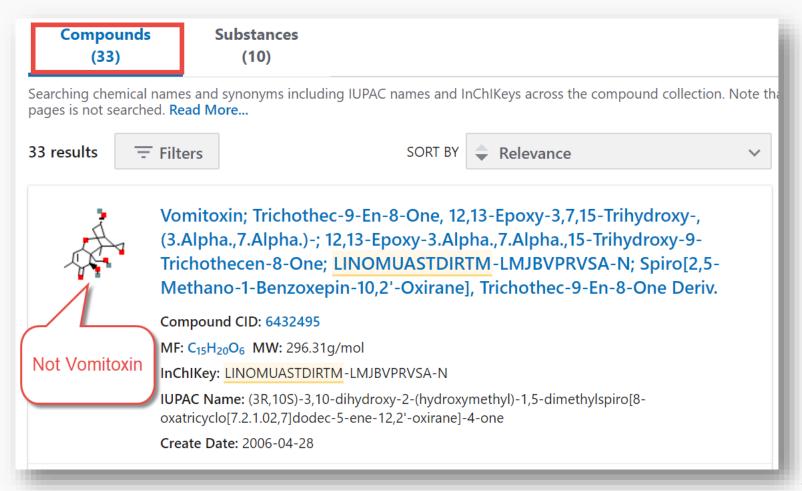
### 19 "Vomitoxins" – 3 isotopically labeled



#### Vomitoxin – PubChem



#### 33 unique InChl Keys



#### ChemSpider – lots of virtuals???



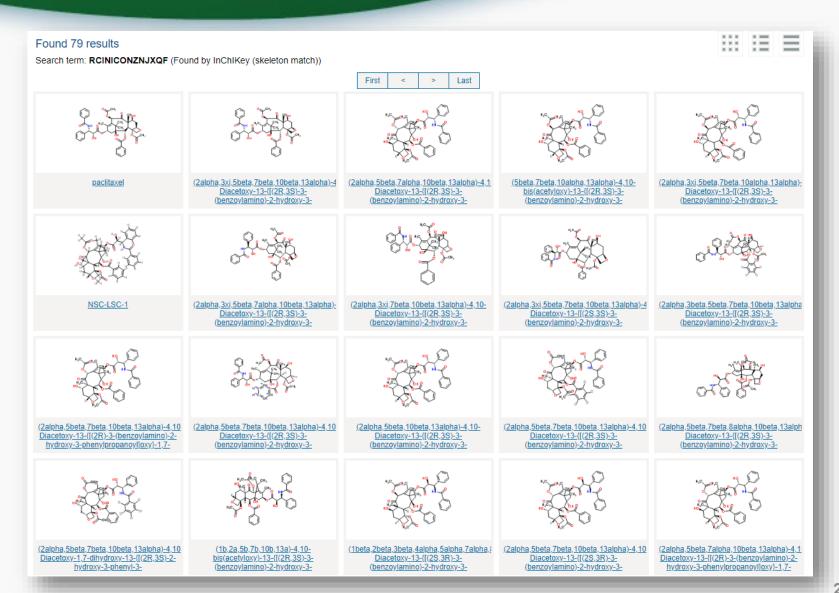


 52 million chemicals from one vendor

<b>Data Sources</b>			
Data Source	Count	<u>Date</u> <u>Created</u>	<u>Last</u> <u>Updated</u>
Aurora Fine Chemicals	<u>51885566</u>	13/04/2009	09/01/2020
<u>Chemspace</u>	14283313	30/11/2016	04/12/2018
AKos	12326374	15/04/2008	09/10/2017
<u>Mcule</u>	9299739	21/01/2014	26/10/2018
Molport	8200357	09/02/2010	09/01/2020
<u>Enamine</u>	3056649	15/04/2008	15/10/2019

#### Taxol: 79 Results





### Data Quality is important



Data quality in free web-based databases!



Drug Discovery Today

Volume 17, Issues 13-14, July 2012, Pages 685-701



Review Keynote

Towards a gold standard: **ELSEVIER** 

quality in public domain

databases and approaches

**⊞** Show



Drug Discovery Today

Volume 16, Issues 17-18, September 2011, Pages 747-750



Machines first, humans second: on the importance Antony). of algorithmic interpretation of open chemistry data

Alex M Clark M, Antony J Williams and Sean Ekins

Journal of Cheminformatics 2015 7:9

https://doi.org/10.1186/s13321-015-0057-7 © Clark et al.; licensee Springer. 2015

Received: 24 November 2014 | Accepted: 23 February 2015 | Published: 22 March 2015

and content



# "MS-ready" structures

McEachran et al. J Cheminform (2018) 10:45 https://doi.org/10.1186/s13321-018-0299-2 Journal of Cheminformatics

#### **METHODOLOGY**

**Open Access** 

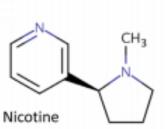
"MS-Ready" structures for non-targeted high-resolution mass spectrometry screening studies

Andrew D. McEachran<sup>1,2\*</sup>, Kamel Mansouri<sup>1,2,3</sup>, Chris Grulke<sup>2</sup>, Emma L. Schymanski<sup>4</sup>, Christoph Ruttkies<sup>5</sup> and Antony J. Williams<sup>2\*</sup>

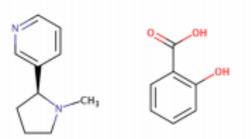
#### Overview of MS-Ready Structures



- All structure-based chemical substances are algorithmically processed to
  - Split multicomponent chemicals into individual structures
  - Desalt and neutralize individual structures
  - Remove stereochemical bonds from all chemicals
- MS-Ready structures are then mapped to original substances to provide a path between chemicals detected by mass spectrometry to original substances



CN1CCC[C@H]1C1=CN=CC=C1 DTXSID1020930| SNICXCGAKADSCV 54-11-5 | **162.1157**| 0.929| **72** Tox: **yes**| Expo: **yes**| Bioassay: **yes** 

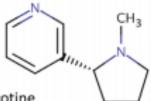


Benzoic acid, 2-hydroxy-, compd. with 3-[(2S)-1-methyl-2-pyrrolidinyl]pyridine (1:1)

OC(=0)C1=C(0)C=CC=C1.CN1CCC[C@H]1C1=CN=CC=C1

DTXSID5075319| AIBWPBUAKCMKNS

29790-52-1 | **300.1474** | 0.929 | **6** Tox: **no** | Expo: **yes** | Bioassay: **no** 



D-Nicotine

CN1CCC[C@@H]1C1=CN=CC=C1 DTXSID004635| SNICXCGAKADSCV 25162-00-9 | **162.1157**| 0.929| **20** 

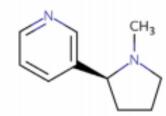
Tox: no | Expo: yes | Bioassay: yes



CN1CCCC1C1=CN=CC=C1
DTXSID3048154 | SNICXCGAKADSCV
22083-74-5 | **162.1157** | 0.953 | 9
Tox: **yes** | Expo: **no** | Bioassay: **yes** 

LEGEND: Name, SMILES DTXSID | InChIKey 1st Block

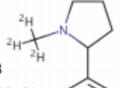
CAS | Monoiso. Mass | logP | Sources Data on: Toxicity | Exposure | Bioassays



HCI

Nicotine hydrochloride

CI.CN1CCC[C@H]1C1=CN=CC=C1 DTXSID602093| HDJBTCAJIMNXEW 2820-51-1 | **198.0924** | 0.929| **9** Tox: **no**| Expo: **yes**| Bioassay: **yes** 



DL-Nicotine-d3

[2H]C([2H])([2H])N1CCCC1C1=CN=CC=C1 DTXSID80442666| SNICXCGAKADSCV 69980-24-1| **165.1345**| 0.929| **1** 

Tox: no | Expo: no | Bioassay: no





## MS-Ready Mappings from Details Page

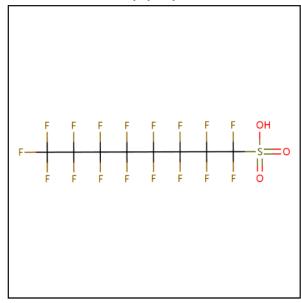




ed search/index

## Perfluorooctanesulfonic acid 1763-23-1 | DTXSID3031864

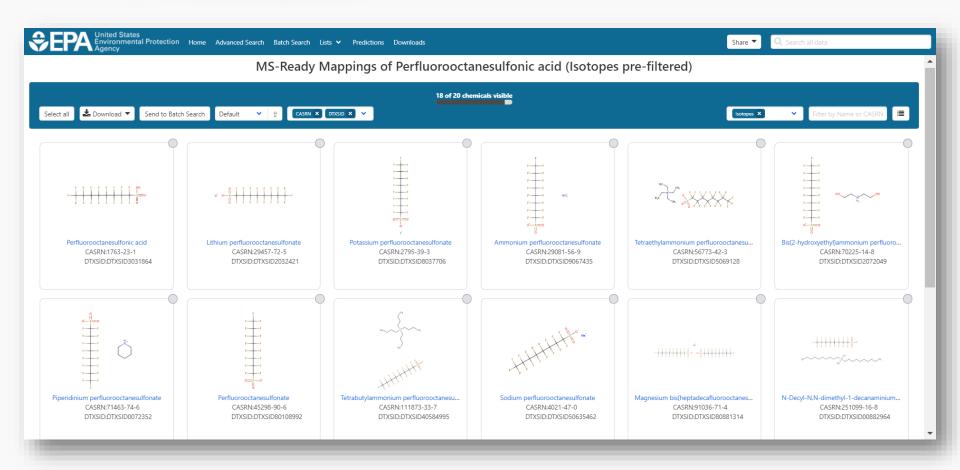
Searched by Synonym from Valid Source.



Wikipedia ▼	,				
Perfluorooctanesulfonic acid (conjugate base perfluorooctanesulfonate) (PFOS) is an anthropogenic fluorosurfactant and global pollutant. PFOS was the key ingredient in Scotchgard, a fabric protector made by 3M, and numerous stain repellents. It was added to Annex B of the Stockholm Convention on Persistent Organic Pollutants in May 2009. PFOS can be synthesized in industrial production or result from the degradation of precursors. PFOS levels that have been detected in wildlife					
Read more					
Quality Control Notes					
Intrinsic Properties					
Structural Identifiers					
Linked Substances	,				
Same Connectivity: 4 records (based on first layer of InChI)  Mixtures, Components and Neutralized Forms: 9 records (based on QSAR ready mappings and with the compound as a component of a mixture)					
MS-Ready Mappings: DTXCID1011864: 18 records;					
Similar Compounds: 83 records (based on Tanimoto coefficient > 0.8)					

## MS-Ready Mappings Set of 20 substances for "PFOS"



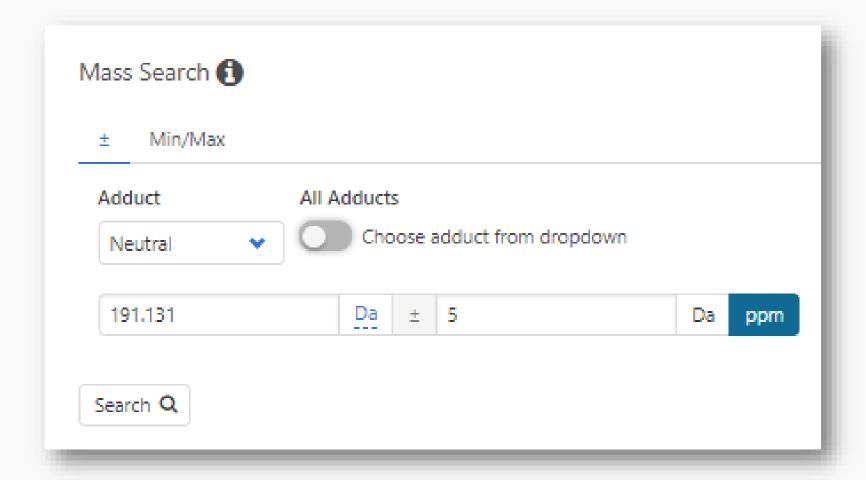




# Mass and Formula Searching

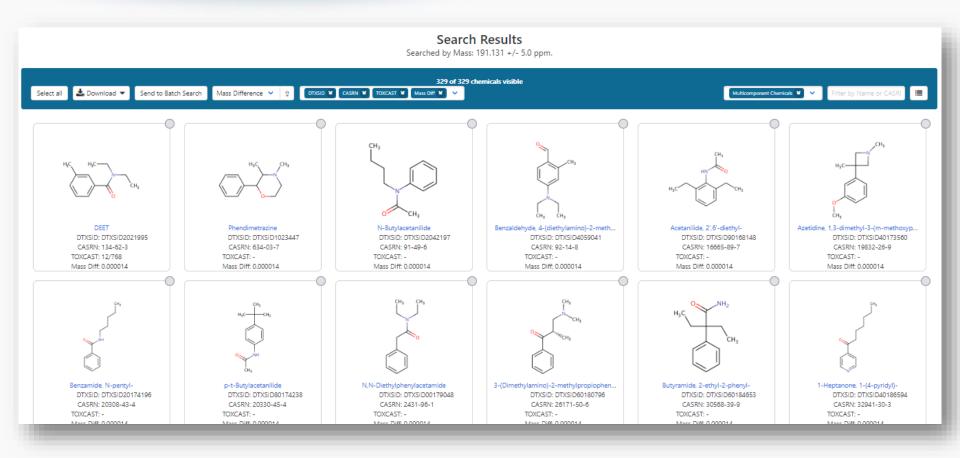
## Advanced Searches Mass Search





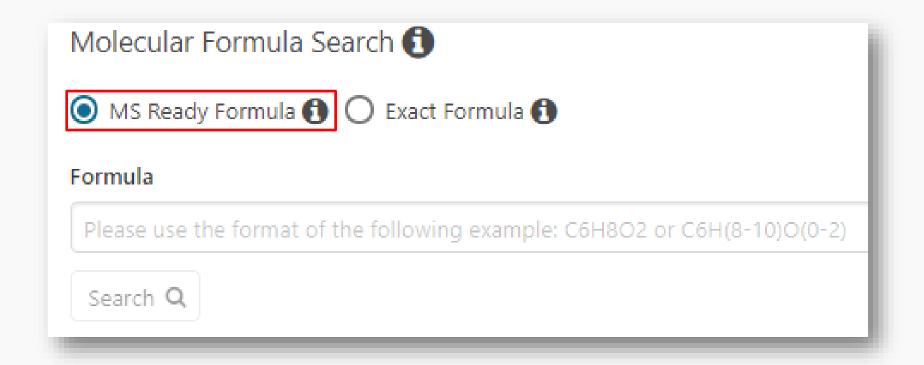
## Advanced Searches Mass Search





## MS-Ready Structures for Formula Search



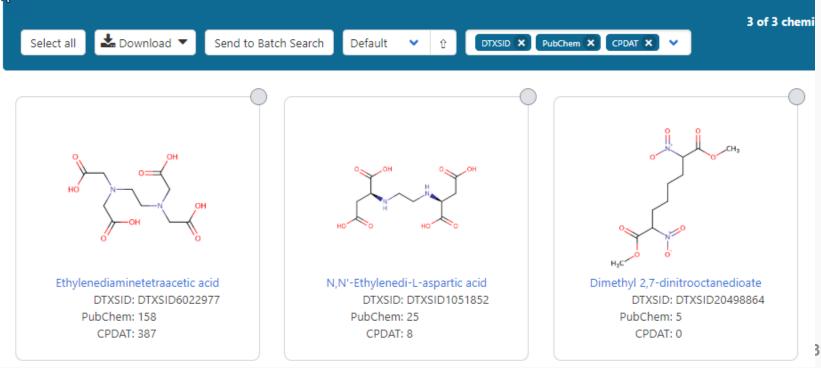


#### MS-Ready Mappings



EXACT Formula: C10H16N2O8: 3 Hits

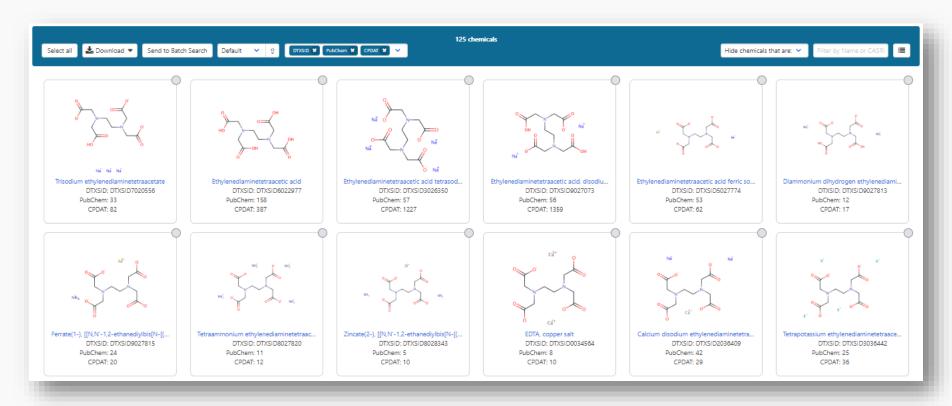




#### MS-Ready Mappings



- Same Input Formula: C10H16N2O8
- MS Ready Formula Search: 125 Chemicals



#### MS-Ready Mappings



- 125 chemicals returned in total
  - 8 of the 125 are single component chemicals
  - 3 of the 8 are isotope-labeled
  - 3 are neutral compounds and 2 are charged
- Multiple components, stereo, isotopes and charge all collapsed and mapped through MS-Ready



## Batch Searching mass and formula

#### Batch Searching



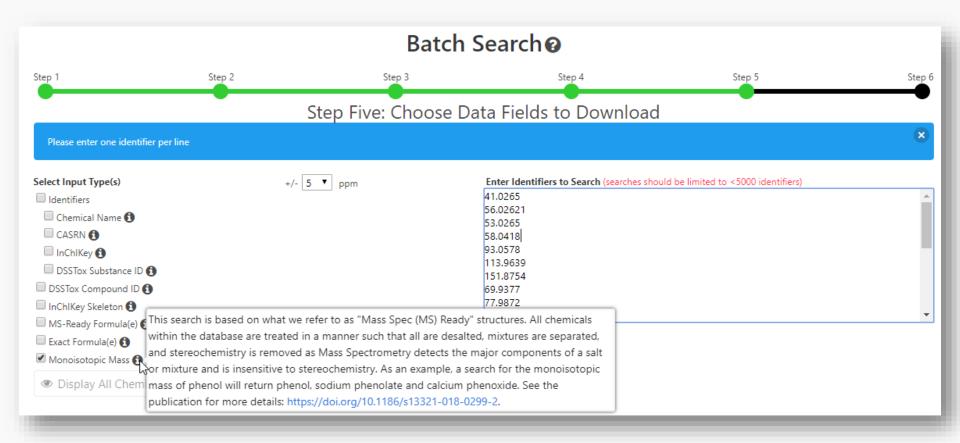
 Singleton searches are useful but we work with thousands of masses and formulae!

#### Typical questions

- What is the list of chemicals for the formula C<sub>x</sub>H<sub>y</sub>O<sub>z</sub>
- What is the list of chemicals for a mass +/- error
- Can I get chemical lists in Excel files? In SDF files?
- Can I include properties in the download file?

#### Batch Searching Formula/Mass





## Searching batches using MS-Ready Formula (or mass) searching



Torrida (or made) deardining								
		Α	В	С	D	E	F	G
		INPUT	DTXSID	CASRN	PREFERRED NAME	MOL FORMULA	MONOISOTOPIC MASS	
		C14H22N2O3	DTXSID2022628	29122-68-7		C14H22N2O3	266.163042576	46
	3	C14H22N2O3	DTXSID0021179	6673-35-4		C14H22N2O3	266.163042576	32
	4	C14H22N2O3	DTXSID4048854	841-73-6		C14H22N2O3		20
		C14H22N2O3	DTXSID1045407	13171-25-0	Trimetazidine dihydrochloride	C14H24Cl2N2O3		19
	6	C14H22N2O3	DTXSID0045753	56715-13-0	R-(+)-Atenolol	C14H22N2O3		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	C14H23CIN2O3	302.1397203	6
	11	C14H22N2O3	DTXSID1068693	51963-82-7	Benzenamine, 2,5-diethoxy-4-(4-morpholinyl)-	C14H22N2O3	266.163042576	5
			DTXSID3023215	154-21-2	Lincomycin	C18H34N2O6S	406.213757997	35
	13	C18H34N2O6S	DTXSID7047803	859-18-7	Lincomycin hydrochloride	C18H35CIN2O6S		22
	14	C18H34N2O6S	DTXSID20849438	1398534-62-7		C18H35CIN2O6S	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	C10H13CIN2O	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		C10H12N2O	176.094963014	9
	21	C10H12N2O	DTXSID90185693			C10H12N2O	176.094963014	7
	22	C10H12N2O	DTXSID40178777	2403-66-9		C10H12N2O	176.094963014	7
	23	C10H12N2O	DTXSID80157026	13140-86-8		C10H12N2O	176.094963014	6
	24	C10H12N2O	DTXSID30205607	570-14-9	4-Hydroxytryptamine	C10H12N2O	176.094963014	6
		C14H18N4O3	DTXSID5023900	17804-35-2	Benomyl	C14H18N4O3	290.137890456	68
	26	C14H18N4O3	DTXSID3023712	738-70-5		C14H18N4O3		51
	27	C14H18N4O3	DTXSID40209671	60834-30-2	•	C14H19CIN4O3	326.1145682	8
		C14H18N4O3	DTXSID70204210		Benzenemethanol, 4-((2,4-diamino-5-pyrimidinyl)methyl)-2,		290.137890456	5
		C14H18N4O3	DTXSID20152671		6-Methoxy-4-(3-(N,N-dimethylamino)propylamino)-5,8-quina		290.137890456	4
	30	C14H18N4O3	DTXSID30213742		1H-1,2,4-Benzotriazepine-3-carboxylic acid, 4,5-dihydro-4-		290.137890456	3
		C14H18N4O3	DTXSID30219608		2,4-Pyrimidinediamine, 5-((3,4,5-trimethoxyphenyl)methyl)-		308.14845514	3
		C14H18N4O3	DTXSID20241155		L-Aspartic acid, compound with 5-((3,4,5-trimethoxyphenyl	C18H25N5O7	423.175398165	3
		C14H18N4O3	DTXSID80241156		L-Glutamic acid, compound with 5-((3,4,5-trimethoxypheny	C19H27N5O7	437.191048229	3
		C14H18N4O3	DTXSID20143781		1H-Pyrido(2,3-e)-1,4-diazepine-2,3,5-trione, 4-(2-(diethylam		290.137890456	3
		C12H11N7	DTXSID6021373	396-01-0		C12H11N7	253.107593382	52
		C12H11N7	DTXSID00204465			C12H11N7	253.107593382	7
		C12H11N7	DTXSID5064621	7300-26-7		C12H9N7	251.091943318	4
		C12H11N7	DTXSID00848025			C12H13N7O4S	351.074973101	1
		C12H11N7	DTXSID50575293			C12H11N7	253.107593382	1
		C8H9NO2	DTXSID2020006	103-90-2		C8H9NO2	151.063328534	75
		CSHONOS	DTYSID6026667	13/1 20/3	Mothyl 2 aminohonzoato	C8H0NIO3	151 063338534	60

#### Batch Search in specific lists

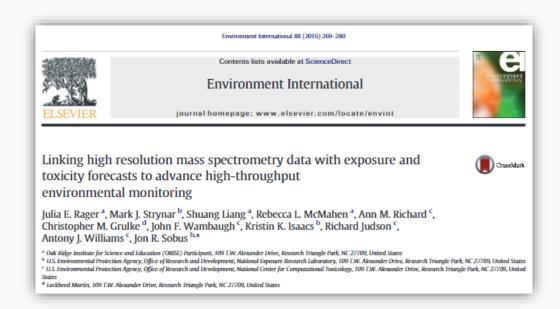


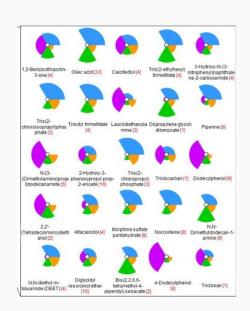
	INPUT	DTXSID	MASSBANKREF	NEMILIST	WRTMSD	NORMANPRI	SUSDAT
	Buprenorphi	DTXSID202	-	-	Υ	-	Υ
ΠË		DTXSID202		Υ	Υ	Υ	Υ
	Dextrometh	DTXSID302:	Υ	Υ	Y	-	Υ
	Dihydrocode			-	Υ	Υ	Υ
l N	Dihydromor	DTXSID704	-	-	-	-	Υ
	Ethylmorphi			_	Υ	-	Υ
_ ::	Fentanyl	DTXSID902:	Υ	_	Υ	-	Υ
<b>✓</b> N		DTXSID604		-	Υ	Υ	Υ
<b>₽</b> \	Hydrocodor			Υ	Υ	Υ	Υ
	Hydromorph			-	Υ	-	Υ
$\square$ $\overline{N}$	Ketamine			-	Υ	-	Υ
<b>₩</b> N-	Meperidine			-	Υ	-	Υ
_ ::	Methadone			Υ	Υ	-	Υ
M V	Morphine			Υ	Υ	Υ	Υ
₩ <u> </u>	Morphinone			-	-	-	Υ
		DTXSID802:	-	-	Y	-	Y
$\square$ $\wedge$	Naltriben	-	-	-	-	-	-
	Oxycodone			Υ	Υ	Υ	Υ
_ <u> </u>	Oxymorpho	DTXSID502:	-	-	Υ	-	Υ
	Propoxyphe			Υ	Υ	-	Υ
	Sufentanil			-	Υ	-	Υ
□ <u>V</u>	Tramadol	DTXSID908	Υ	Υ	Υ	Υ	Υ

#### Benefits of bringing it all together



- The true dashboard benefit is integration
- Rank potential candidates for toxicity using available data – hazard, exposure, in vitro







# Candidate ranking using metadata



C American Society for Mass Spectrometry, 2011

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

#### RESEARCH ARTICLE

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

## Data Source Ranking of "known unknowns"



 A mass and/or formula search is for an *unknown* chemical but it is a *known* chemical contained within a reference database

 Most likely candidate chemicals have the most associated data sources, most associated literature articles or both C14H22N2O3 266.16304



Chemical Reference Database



Sorted candidate structures

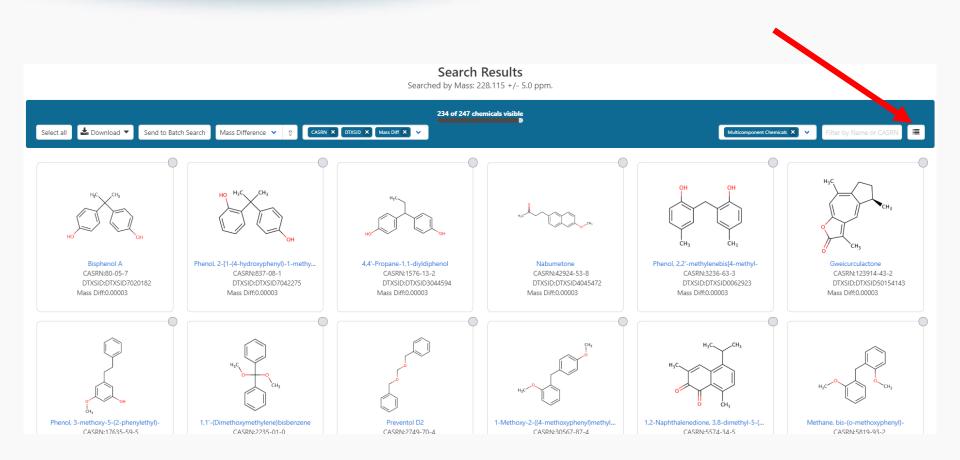
#### Data Streams for Ranking



- CompTox Dashboard Data Sources
- Pub©hem Data Source Count
- Publ@ed.gov Reference Count
- Toxcast in vitro bioactivity
- Presence in CPDat database
- OPERA PhysChem Properties
- Other possibilities predicted media occurrence, frequency of InChls online

## Search 228.115 +/- 5.0 ppm 234 single component chemicals





## Search 228.115 +/- 5.0 ppm 234 single component chemicals



CASRN	QC Level	CPDat Count	Number of Sources	PubChem Data Sources	PubMed Ref. Counts
80-05-7	Level 1	326	170	161	3850
42924-53-8	Level 2	14	45	138	342
87619-52-1	Level 5	0	2		0
87607-32-7	Level 5	0	2		0

#### The original ChemSpider work



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			

#### Is a bigger database better?



chemical structures

- ChemSpider was 26 million chemicals for the original work
- Much BIGGER today
- Is bigger better??
- Are there other metadata to use for ranking?

#### Comparing Search Performance



Anal Bioanal Chem (2017) 409:1729–1735 DOI 10.1007/s00216-016-0139-z



#### RAPID COMMUNICATION

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

Andrew D. McEachran<sup>1</sup> · Jon R. Sobus<sup>2</sup> · Antony J. Williams<sup>3</sup>

- When dashboard contained 720k chemicals
- Only 3% of ChemSpider size
- What was the comparison in performance?

#### **SAME** dataset for comparison



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
Sterbid he mones Perfluorochemicals		SAME	7 5	)A	TA	SE	T
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			

#### How did performance compare?



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

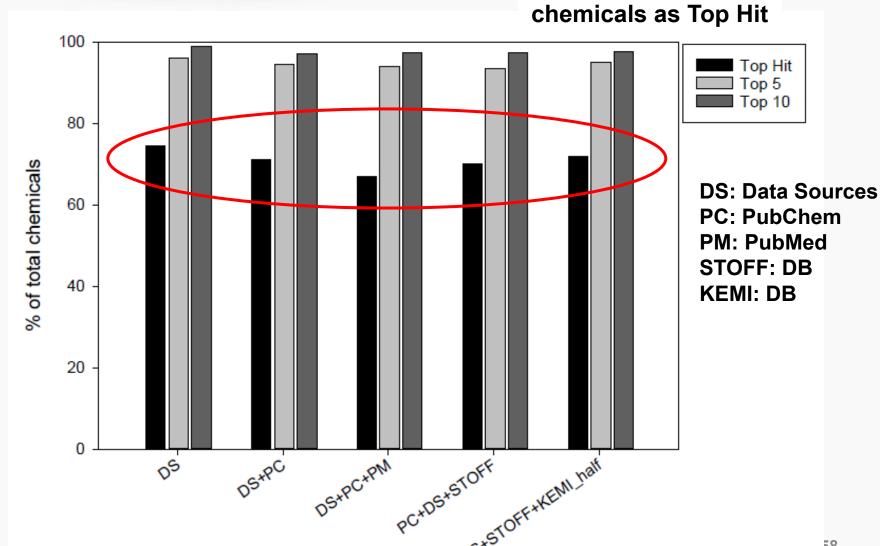
<sup>&</sup>lt;sup>a</sup> Average rank in ChemSpider shown here does not include an outlier where the rank was 201, when added the average rank position is 3.5

# For the same 162 chemicals, Dashboard outperforms ChemSpider for both Mass and Formula Ranking

#### Identification ranks for 1783 chemicals using multiple data streams



**Data Sources alone** rank ~75% of the



#### Prototype Work in Progress



- CFM-ID
  - Viewing and Downloading pre-predicted spectra
  - Search spectra against the database
- Structure/substructure/similarity search
- Access to API and web services
- Integration to EPA "Chemical Transformation Simulator"

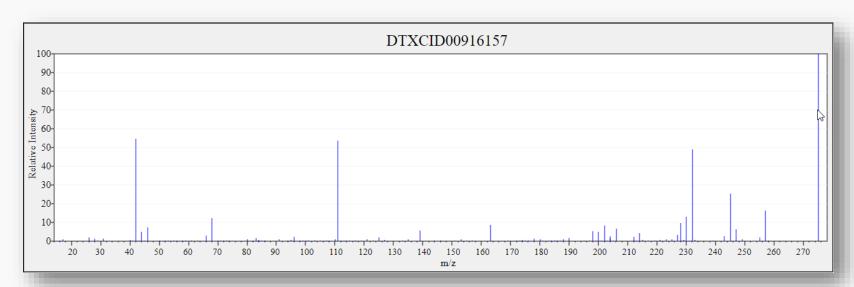
#### Predicted Mass Spectra

http://cfmid.wishartlab.com/



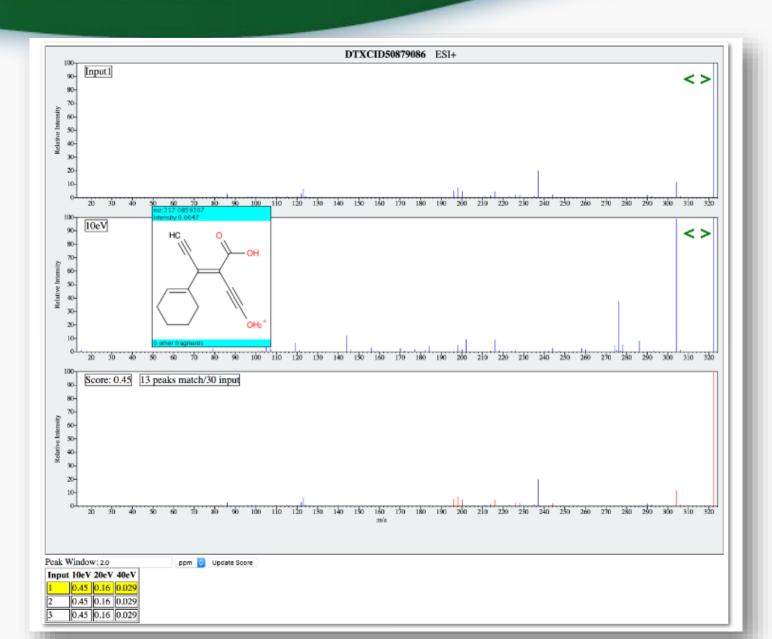


- MS/MS spectra prediction for ESI+, ESI-, and EI
- Predictions generated and stored for >800,000 structures, to be accessible via Dashboard



#### Spectral Viewer Comparison





#### Published: Alex Chao et al



Analytical and Bioanalytical Chemistry

#### **RESEARCH PAPER**

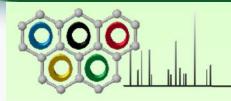
## In silico MS/MS spectra for identifying unknowns: a critical examination using CFM-ID algorithms and ENTACT mixture samples

Alex Chao <sup>1,2</sup> • Hussein Al-Ghoul <sup>1,2</sup> • Andrew D. McEachran <sup>1,3</sup> • Ilya Balabin <sup>4</sup> • Tom Transue <sup>4</sup> • Tommy Cathey <sup>4</sup> • Jarod N. Grossman <sup>2,3</sup> • Randolph Singh <sup>1,5</sup> • Elin M. Ulrich <sup>2</sup> • Antony J. Williams <sup>6</sup> • Jon R. Sobus <sup>2</sup>

Received: 4 October 2019 / Revised: 27 November 2019 / Accepted: 11 December 2019 © The Author(s) 2019

#### CASMI 2012-2017 revisited







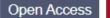
## Critical Assessment of Small Molecule Identification

The experimental and computational mass spectrometry communities are invited to participate in the fifth round of an open contest on the identification of small molecules from mass spectrometry data.

This year the contest will test the applicability of MS and MS/MS on natural products chemistry identifications. With 45 (Category 1) and up to 243 (Categories 2&3) natural products challenges - including a few tricky ones - there's something for everyone!

 Application of metadata candidate ranking and CFM-ID to all five years of CASMI data







## Revisiting Five Years of CASMI Contests with EPA Identification Tools

```
by ♠ Andrew D. McEachran <sup>1,*</sup> ☑ ♠, ♠ Alex Chao <sup>1</sup> ☑ ♠, ♠ Hussein Al-Ghoul <sup>1</sup> ☑ ♠, ♠ Charles Lowe <sup>2</sup> ☑ ♠, ♠ Christopher Grulke <sup>2</sup> ☑ ♠, ♠ Jon R. Sobus <sup>2</sup> ☑ ♠ and ♠ Antony J. Williams <sup>2,*</sup> ☑ ♠
```

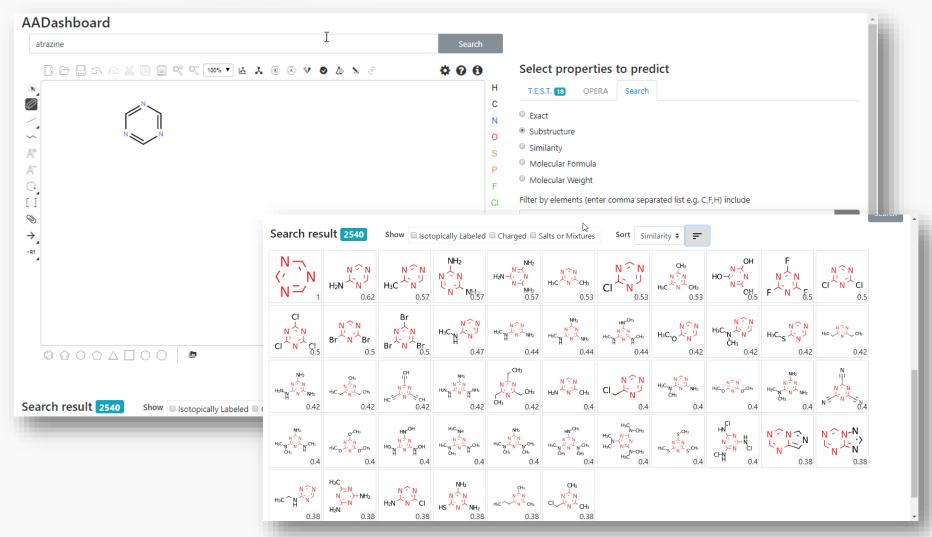
- Oak Ridge Institute for Science and Education (ORISE) Participant, 109 T.W. Alexander Drive, Research Triangle Park, NC 27709, USA
- Center for Computational Toxicology and Exposure, Office of Research and Development, U.S. Environmental Protection Agency, 109 T.W. Alexander Drive, Research Triangle Park, NC 27709, USA
- \* Authors to whom correspondence should be addressed.

Metabolites 2020, 10(6), 260; https://doi.org/10.3390/metabo10060260

Received: 3 May 2020 / Revised: 3 June 2020 / Accepted: 17 June 2020 / Published: 23 June 2020

#### Prototype Development

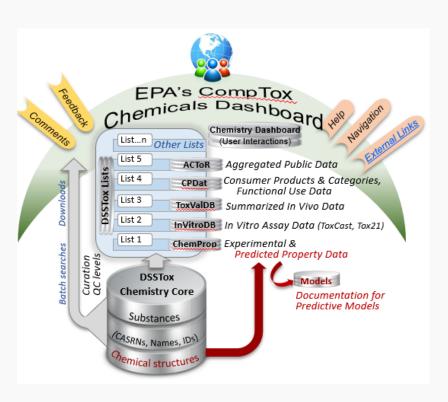




#### Conclusion

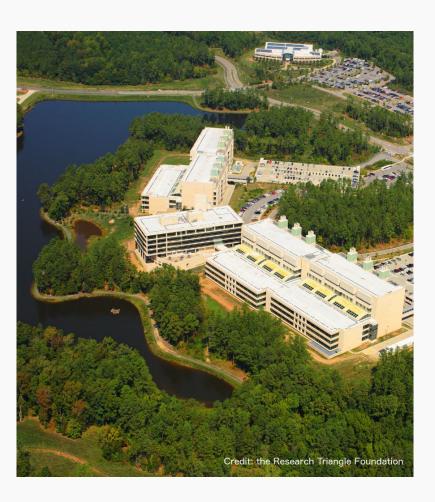


- Dashboard access to data for ~883,000 chemicals
   MS-Ready data facilitates structure identification
- Related metadata facilitates candidate ranking
- Relationship mappings and chemical lists of great utility
- Curation and mutual sharing of chemical lists is important (e.g. NORMAN)



#### Acknowledgements





#### **EPA ORD**

Ann Richard Chris Grulke Jeremy Dunne Jeff Edwards Grace Patlewicz Alex Chao Kristin Isaacs Charles Lowe James McCord Seth Newton Katherine Phillips Jon Sobus Mark Strynar Elin Ulrich Joach Pleil

#### **TEAMS**

IT Development Team Curation Team

#### **ILS**

Kamel Mansouri

#### **GDIT**

Ilya Balabin Tom Transue Tommy Cathey

#### **Collaborators**

Emma Schymanski NORMAN Network Andrew McEachran

#### Contact



#### **Antony Williams**

CCTE, US EPA Office of Research and Development,

Williams.Antony@epa.gov

ORCID: https://orcid.org/0000-0002-2668-4821



https://doi.org/10.1186/s13321-017-0247-6