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

United States Environmental Protection Agency

Consensus ranking and fragmentation prediction for identification of unknowns in high resolution mass spectrometry

Andrew D. McEachran

Hussein Al-Ghoul, Ilya Balabin, Tommy Cathey, Alex Chao, Jon Sobus, and Antony J. Williams

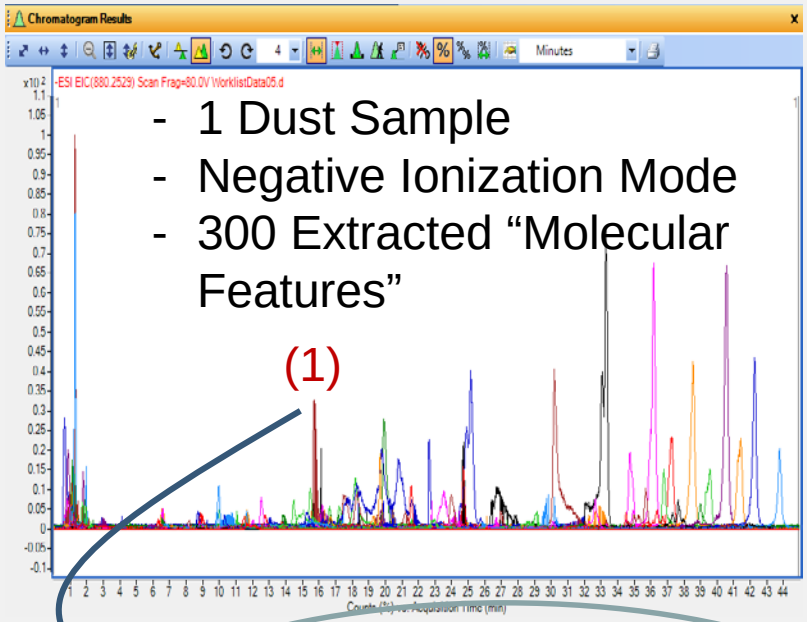
AGRO 107

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

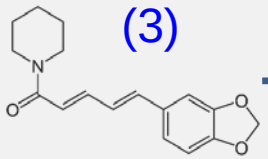
August 20, 2018

General Goals of SSA/NTA



(2)

C17H19NO3



(3)

(4)

(5)

12 µg/g

1) Prioritize "Molecular Features"

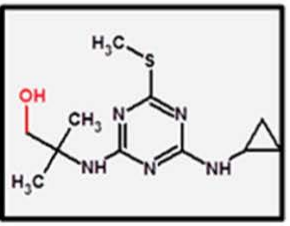
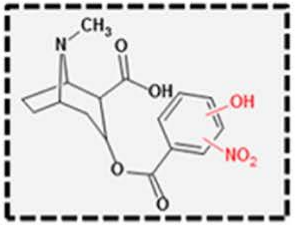
2) Correctly assign formulas

3) Correctly assign structures

4) Identify chemical sources

5) Predict chemical concentrations

EXPOSURE

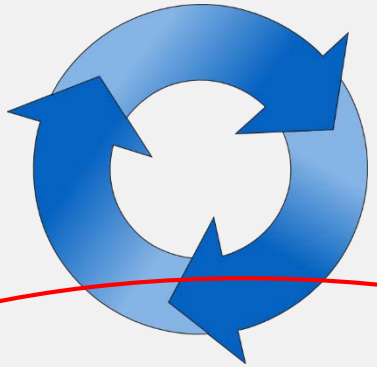
Example	Identification confidence	Minimum data requirements
	***Level 1: Confirmed structure*** by reference standard	MS, MS^2, RT, Reference Std.
	Level 2: Probable structure a) by library spectrum match b) by diagnostic evidence	MS, MS^2, Library MS^2 MS, MS^2, Exp. data
	***Level 3: Tentative candidate(s)*** structure, substituent, class	MS, MS^2, Exp. data
$C_6H_5N_3O_4$	***Level 4: Unequivocal molecular formula***	MS isotope/adduct
192.0757	***Level 5: Exact mass*** of interest	MS

Schymanski, *et al*. 2014

The General Approach

Analytical Instruments

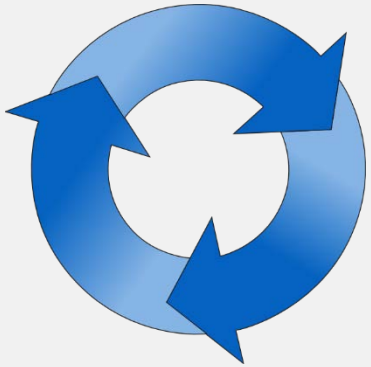
Comp. Tools & Workflows



Databases

The General Approach

Analytical Instruments



Comp. Tools & Workflows

FOR IDENT

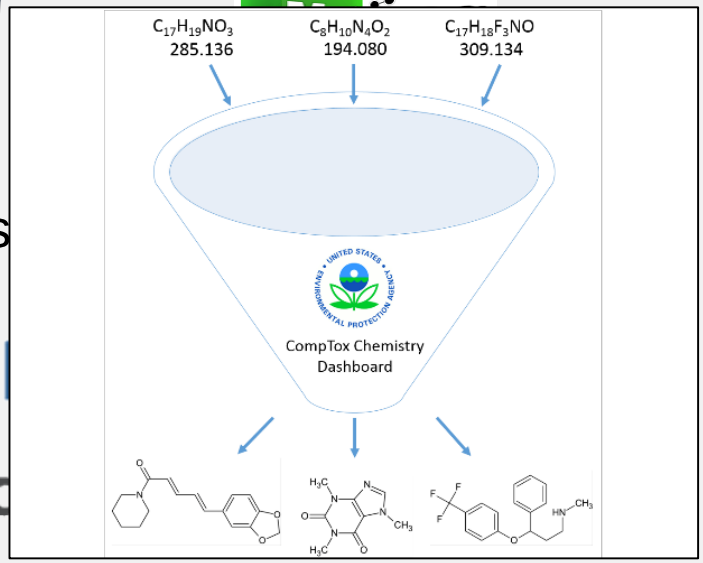
Xcms Online

Databases

ChemSpider
Search and share chemistry

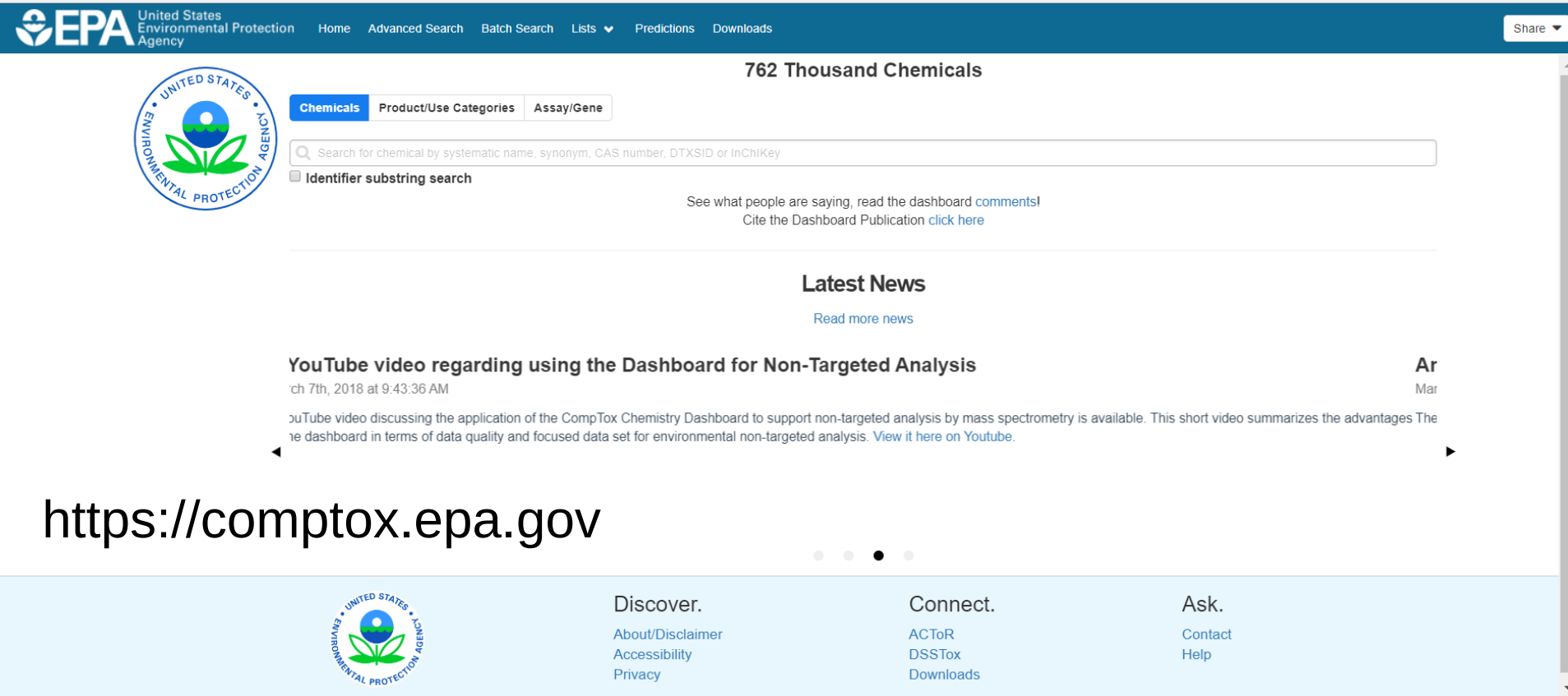
Pu

MassBa



United States Environmental Protection Agency

CompTox Dashboard



https://comptox.epa.gov

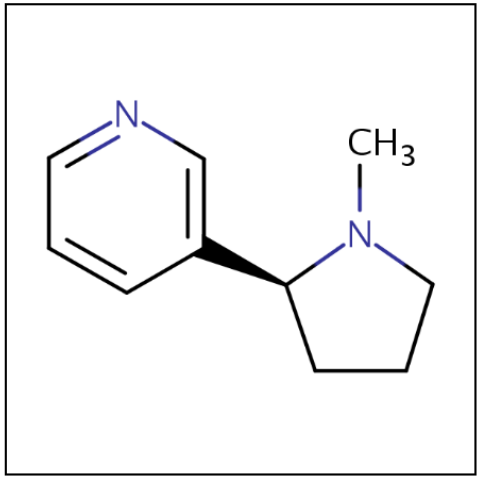
United States Environmental Protection Agency

Nicotine

54-11-5 | DTXSID1020930

Searched by Approved Name.

- DETAILS
- EXECUTIVE SUMMARY
- PROPERTIES
- ENV. FATE/TRANSPORT
- HAZARD
- ADME
- EXPOSURE
- BIOACTIVITY
- SIMILAR COMPOUNDS
- GENRA (BETA)
- RELATED SUBSTANCES
- SYNONYMS
- LITERATURE
- LINKS
- COMMENTS



Wikipedia

Nicotine is a potent parasympathomimetic stimulant and an alkaloid found in the nightshade family of plants. Nicotine acts as an agonist at most nicotinic acetylcholine receptors (nAChRs), except at two nicotinic receptor subunits (nAChRα9 and nAChRα10) where it acts as a receptor antagonist. Nicotine is found in the leaves of *Nicotiana rustica*, in concentrations of 2–14%; in the tobacco plant, *Nicotiana tabacum*; in *Duboisia hopwoodii*; and in *Asclepias syriaca*

...

Read more

Intrinsic Properties

Molecular Formula: $C_{10}H_{14}N_2$ Mol File Find All Chemicals

Average Mass: 162.236 g/mol Isotope Mass Distribution

Monoisotopic Mass: 162.115698 g/mol

Structural Identifiers

Linked Substances

Presence in Lists

Record Information

Quality Control Notes

EPA United States Environmental Protection Agency

EPA United States Environmental Protection Agency

Home | Advanced Search | Batch Search | Lists | Predictions | Downloads | Copy | Share | Submit Comment | Search all data

Nicotine

54-11-5 | DTXSID1020930

Searched by Approved Name.

- DETAILS
- EXECUTIVE SUMMARY
- **PROPERTIES**
- ENV. FATE/TRANSPORT
- HAZARD
- ADME
- EXPOSURE
- BIOACTIVITY
- SIMILAR COMPOUNDS
- GENRA (BETA)
- RELATED SUBSTANCES
- SYNONYMS
- LITERATURE
- LINKS

Property

Summary

Summary

Download

Columns

Search query

Property	Experimental average	Predicted average	Experimental median	Predicted median	Experimental range	Predicted range	Unit
LogP: Octanol-Water	1.17 (1)	0.751		0.821	1.17	3.85e-2 to 1.18	
Melting Point	-79.0 (3)	12.4	-79.0	13.4	-79.0	-34.4 to 57.3	°C
Boiling Point	247 (2)	249	247	248	247	244 to 254	°C
Vapor Pressure	3.80e-2 (1)	1.70e-2		1.76e-2	3.80e-2	2.39e-3 to 3.03e-2	mmHg
Water Solubility	6.16 (1)	3.74		4.51	6.16	8.00e-2 to 6.63	mol/L
Flash Point	-	99.8		99.8	-	97.9 to 102	°C
Surface Tension	-	38.6		38.6	-	37.7 to 39.6	dyn/cm
Index of Refraction	-	1.54			-	1.54	
Molar Refractivity	-	49.3			-	49.3	cm^3
Polarizability	-	19.5			-	19.5	Å^3

United States Environmental Protection Agency

United States Environmental Protection Agency

Home | Advanced Search | Batch Search | Lists | Predictions | Downloads | Copy | Share | Submit Comment | Search all data

Nicotine

54-11-5 | DTXSID1020930

Searched by Approved Name.

- DETAILS
- EXECUTIVE SUMMARY
- PROPERTIES
- ENV. FATE/TRANSPORT
- HAZARD
- ADME
- EXPOSURE
  - PRODUCT & USE CATEGORIES
  - CHEMICAL WEIGHT FRACTION
  - CHEMICAL FUNCTIONAL USE
  - TOXICS RELEASE INVENTORY
  - MONITORING DATA
  - EXPOSURE PREDICTIONS
  - PRODUCTION VOLUME

Exposure Predictions (mg/kg-bw/day)

Download | Columns | Search query

Demographic	Median	95th Percentile
Ages 6-11	1.07e-6	9.90e-5
Ages 12-19	7.88e-7	5.53e-5
Ages 20-65	6.60e-7	4.09e-5
Ages 65+	5.27e-7	2.76e-5
BMI > 30	7.95e-7	4.13e-5
BMI < 30	7.93e-7	5.06e-5
Repro. Age Females	7.35e-7	4.93e-5
Females	8.98e-7	5.34e-5
Males	6.51e-7	5.08e-5
Total	7.08e-7	4.37e-5

United States Environmental Protection Agency

United States Environmental Protection Agency

Home | Advanced Search | Batch Search | Lists | Predictions | Downloads

Search all data

Batch Search

Step 1 — Step 2 — Step 3 — Step 4 — Step 5 — Step 6

Step One: Select Input

Please enter one identifier per line

Select Input Type(s)

- Identifiers
  - Chemical Name
  - CASRN
  - InChIKey
  - DSSTox Substance ID
- InChIKey Skeleton
- MS-Ready Formula(e)
- Exact Formula(e)
- Monoisotopic Mass

Chemical Data

Enter Identifiers to Search (searches should be limited to <5000 identifiers)

Batch Search for SSA/NTA

United States Environmental Protection Agency

Home | Advanced Search | Batch Search | Lists | Predictions | Downloads

Search all data

Batch Search

Step 1 — Step 2 — Step 3 — Step 4 — Step 5 — Step 6

Step One: Select Input

Please enter one identifier per line

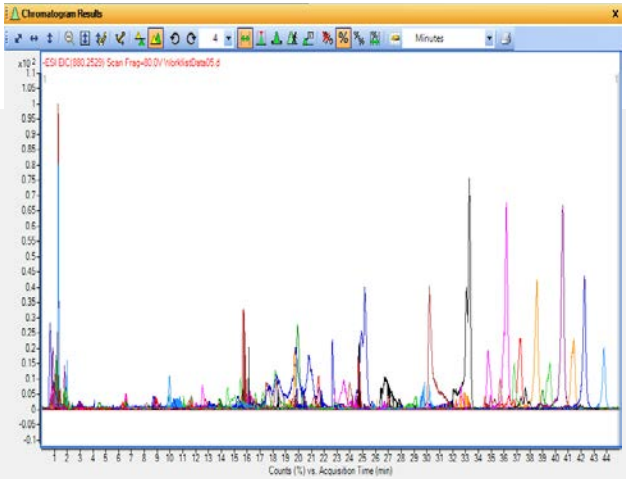
Select Input Type(s)

- Identifiers
  - Chemical Name
  - CASRN
  - InChIKey
  - DSSTox Substance ID
- InChIKey Skeleton
- MS-Ready Formula(e)
- Exact Formula(e)
- Monoisotopic Mass

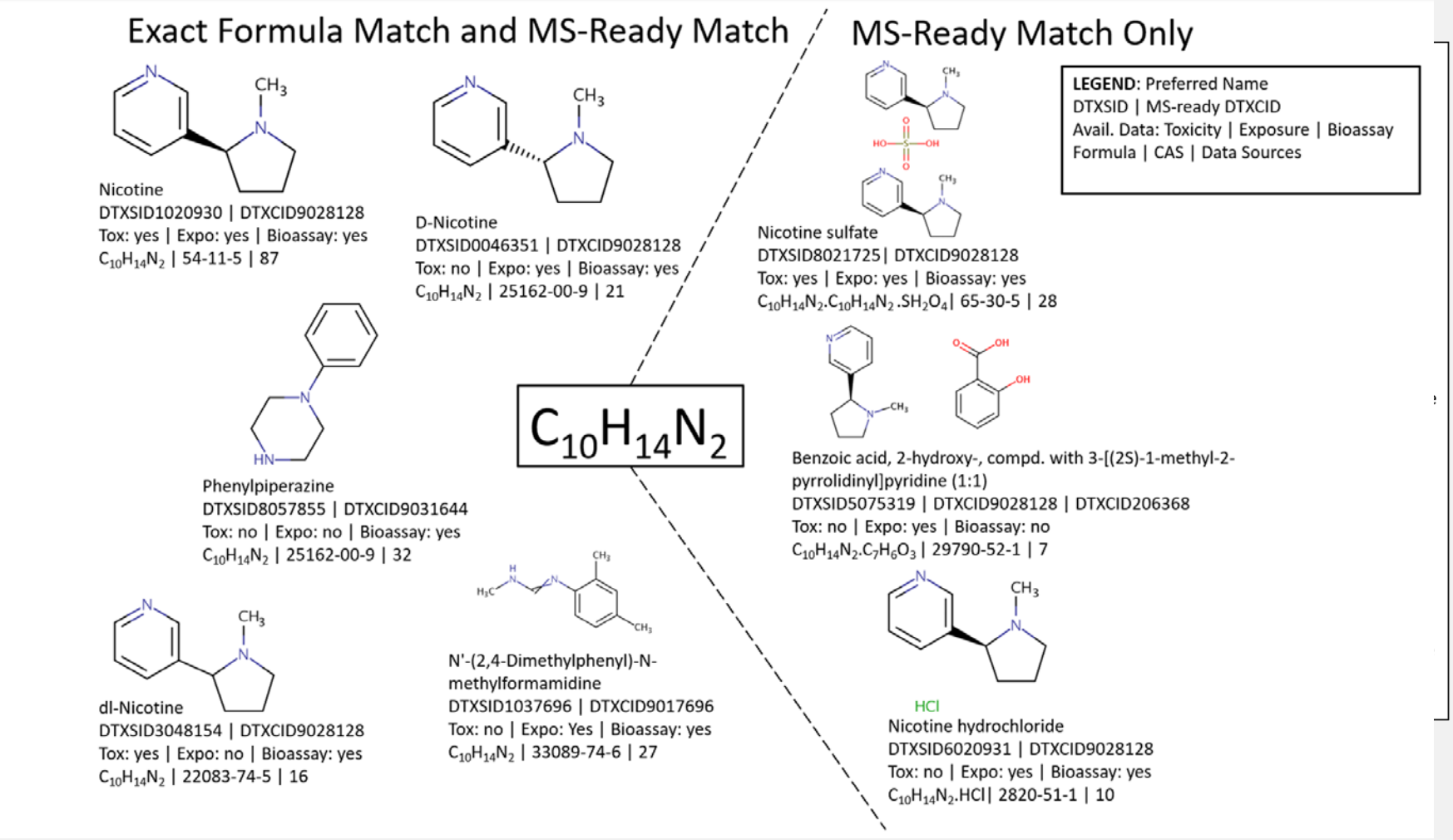
Enter Identifiers to Search (searches should be limited to <5000 identifiers)

Chemical Data

C10H14N2



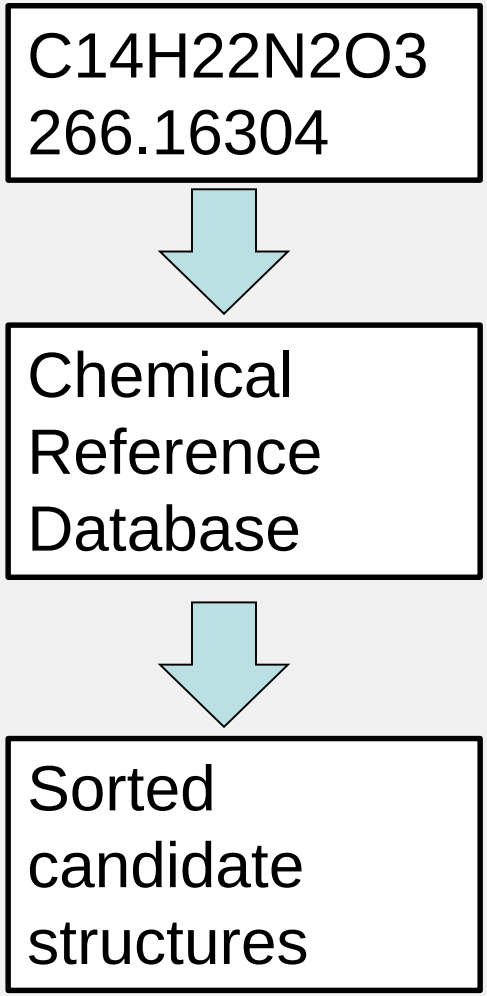
MS-Ready Structures improve database searching



McEachran, *et al*, accepted manuscript

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

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	

United States Environmental Protection Agency

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%

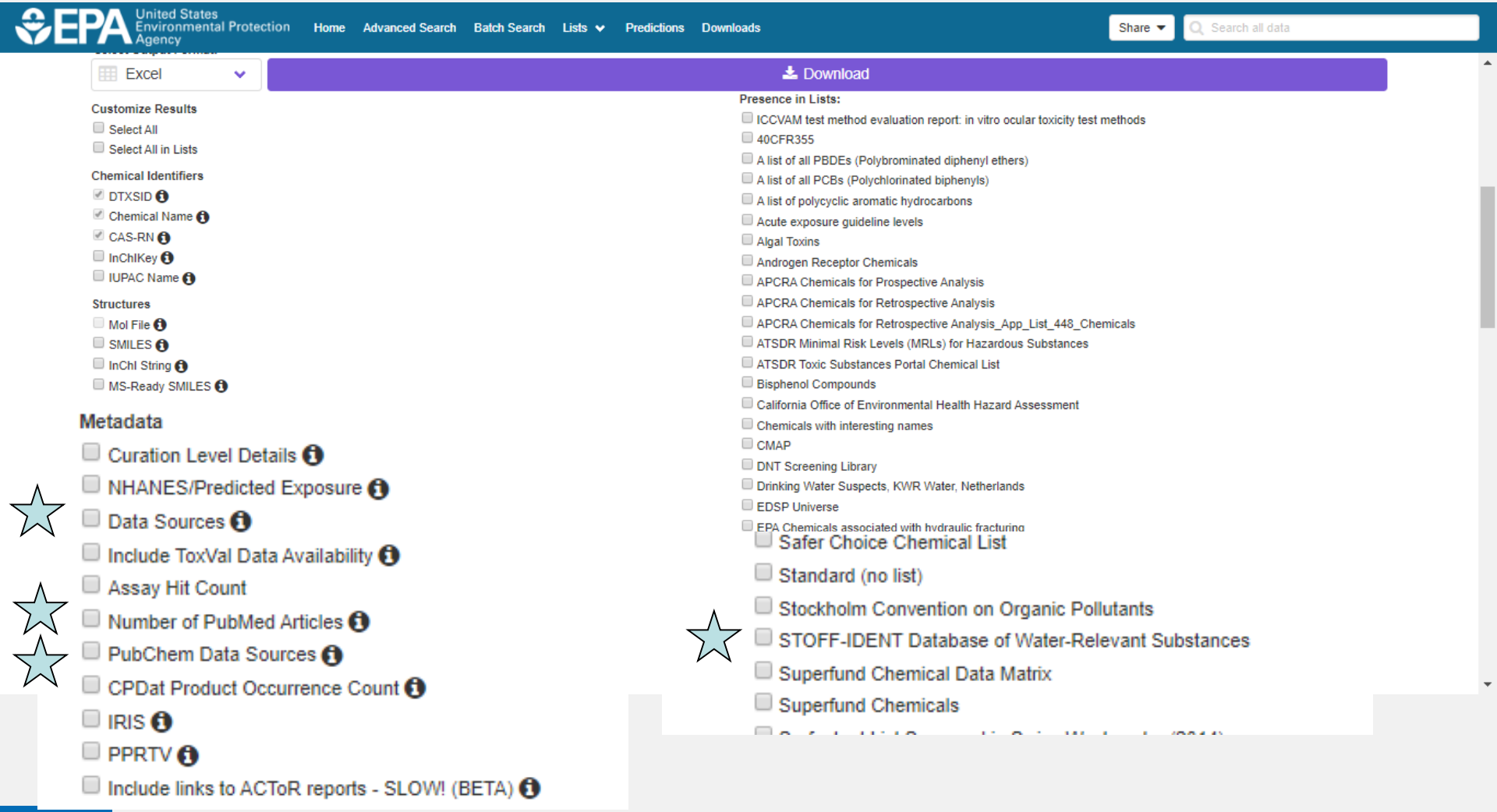
[a] Average rank in ChemSpider shown here does not include an outlier where the rank was 201, when added the average rank position is 3.5

Additional Data Streams to Improve Identifications

- US EPA CompTox Dashboard Data Sources (DS)
- PubChem Data Source Count
- PubMed Reference Count
- Presence in STOFF-IDENT Database
- Predicted Environmental Media Occurrence
- OPERA PhysChem Properties
- NORMAN Network Priority List

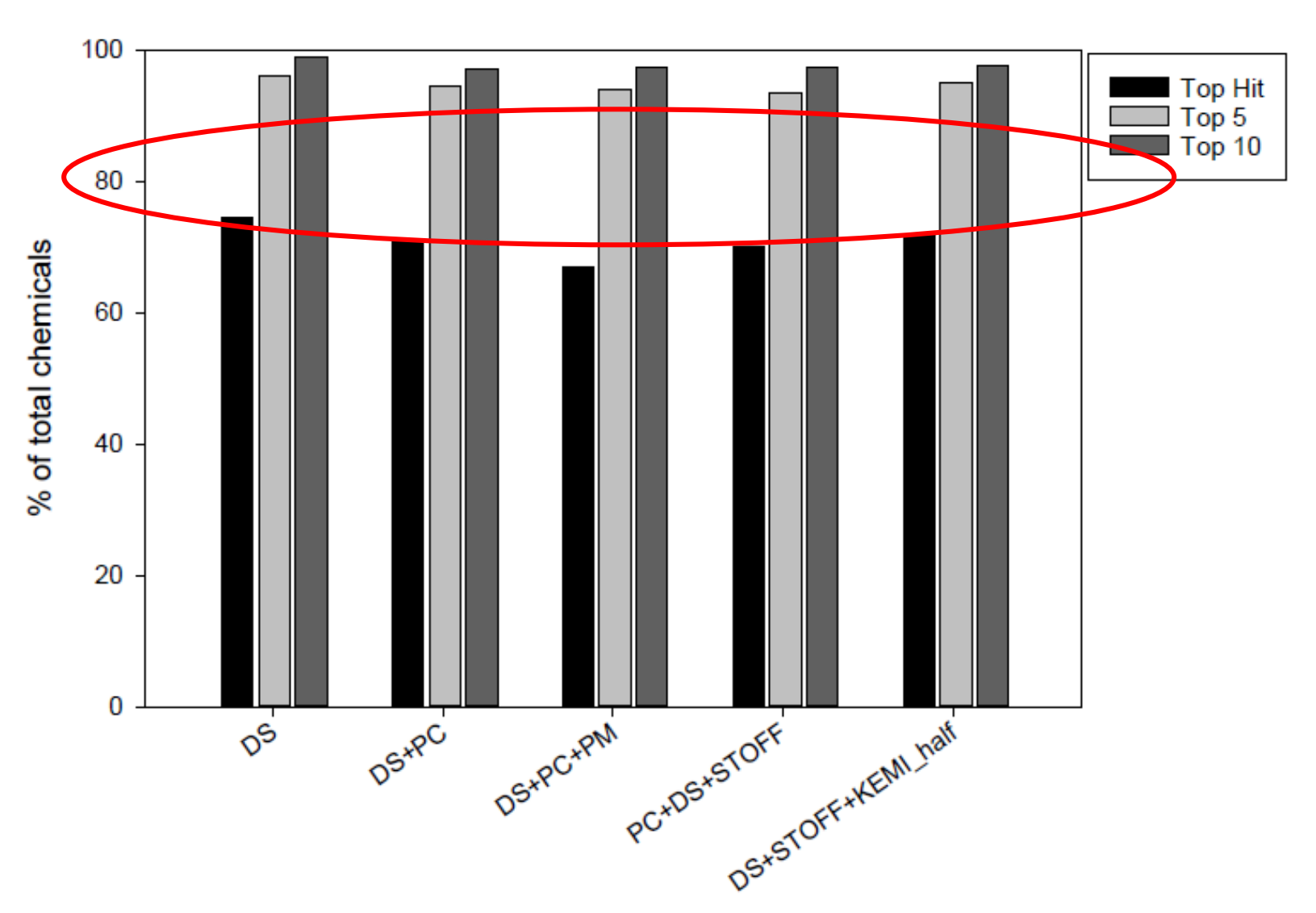
Chemistry Dashboard

All available via Batch Search:

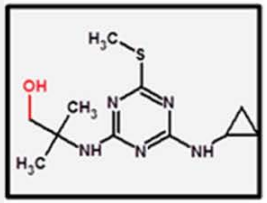
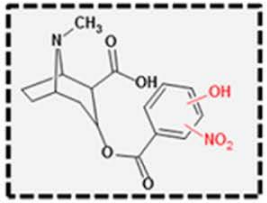


Identification ranks for 1783 chemicals using multiple data streams

$$SC_{TOTAL} = SC_{DS} + SC_{PM} + SC_{RT} + SC_{MO} + \cdots$$



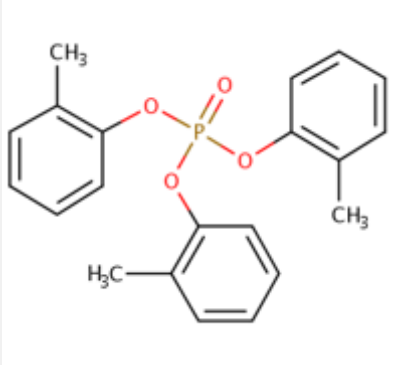
Metadata is critical, but need structural confirmation to increase confidence

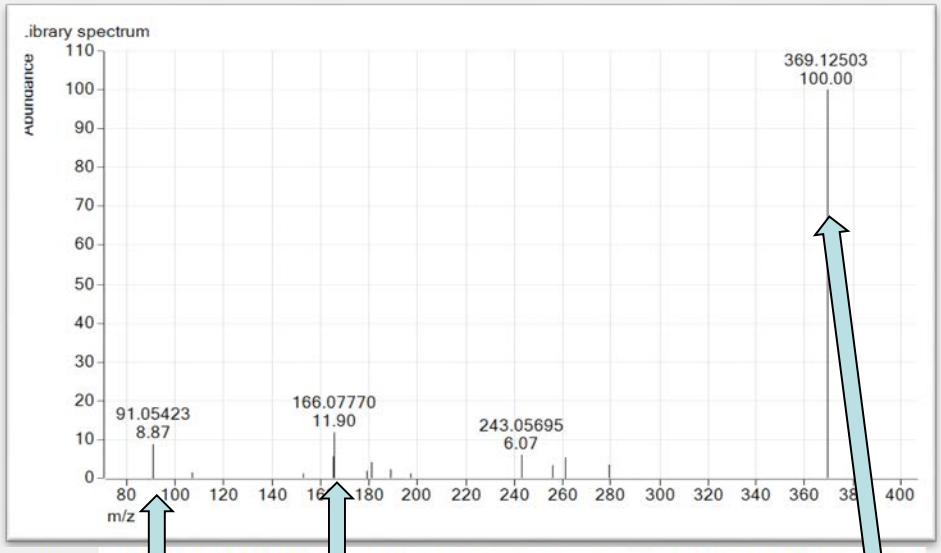
Example	Identification confidence	Minimum data requirements
	***Level 1: Confirmed structure*** by reference standard	MS, MS^2, RT, Reference Std.
	Level 2: Probable structure a) by library spectrum match b) by diagnostic evidence	MS, MS^2, Library MS^2 MS, MS^2, Exp. data
	***Level 3: Tentative candidate(s)*** structure, substituent, class	MS, MS^2, Exp. data
$C_6H_5N_3O_4$	***Level 4: Unequivocal molecular formula***	MS isotope/adduct
192.0757	***Level 5: Exact mass*** of interest	MS

Schymanski, *et al*. 2014

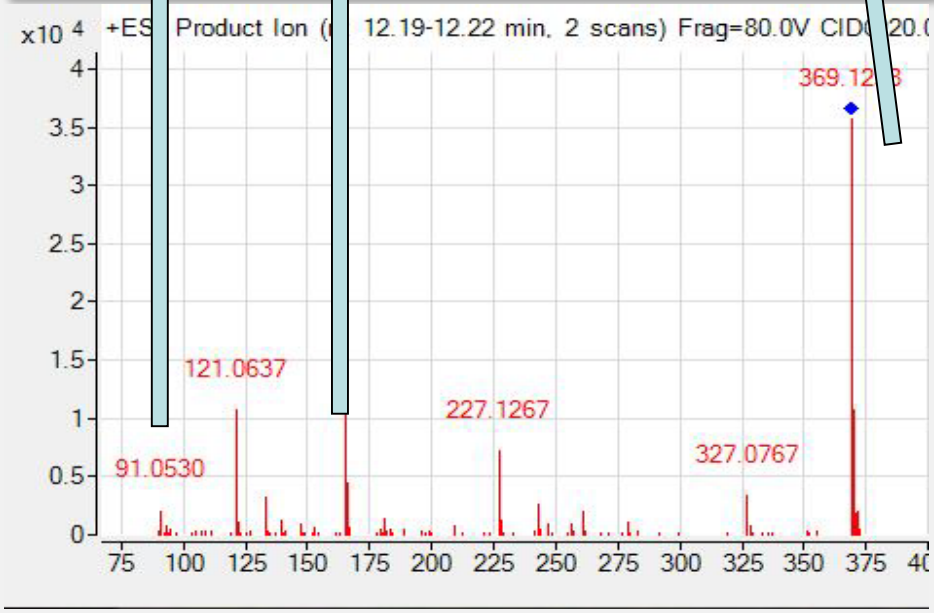
MS/MS Spectral Matching for Identification

Library Fragmentation Spectra (20eV)





Observed Fragmentation Spectra (20eV)



Match Score

EPA United States Environmental Protection Agency

CFM-ID

Metabolomics
February 2015, Volume 11, Issue 1, pp 98–110 | Cite as

Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification

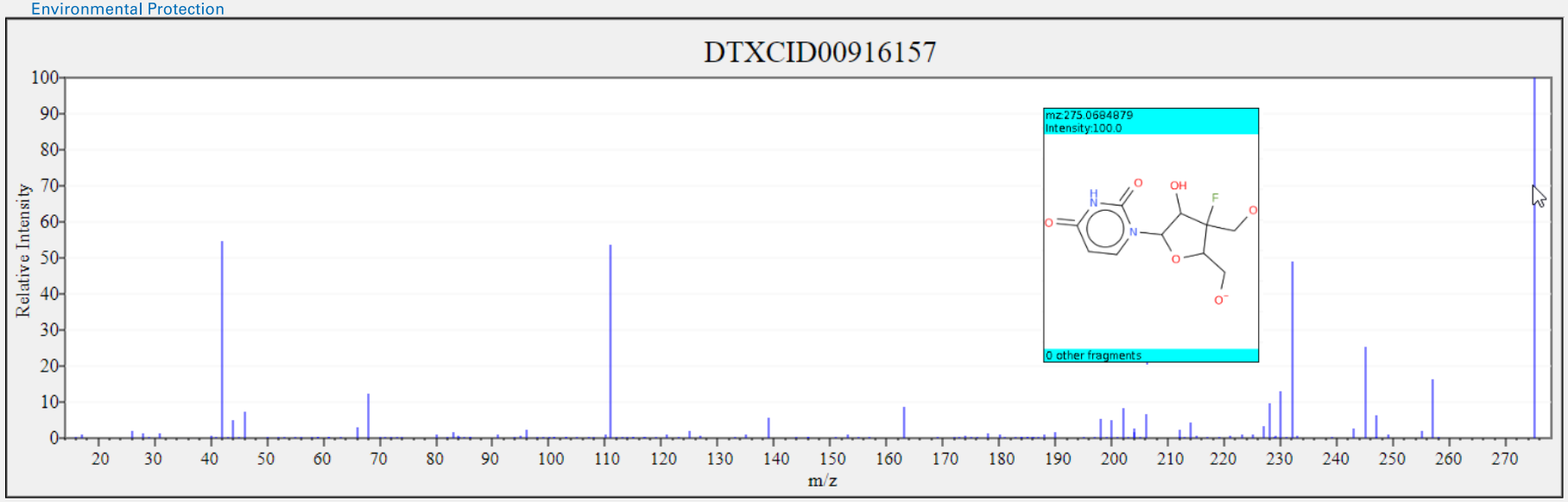
Authors Authors and affiliations

Felicity Allen, Russ Greiner, David Wishart

- Fragmentation prediction for identification in HRMS
- Open source code allows for MS/MS spectra prediction for ESI+, ESI-, and EI
- Predictions generated and stored for >700,000 structures, to be accessible via CompTox Dashboard
- Python code to pull matches and score experimental vs predicted spectra
- Cosine dot product match score calculation

CFM-ID
Competitive Fragmentation Modeling for Metabolite Identification

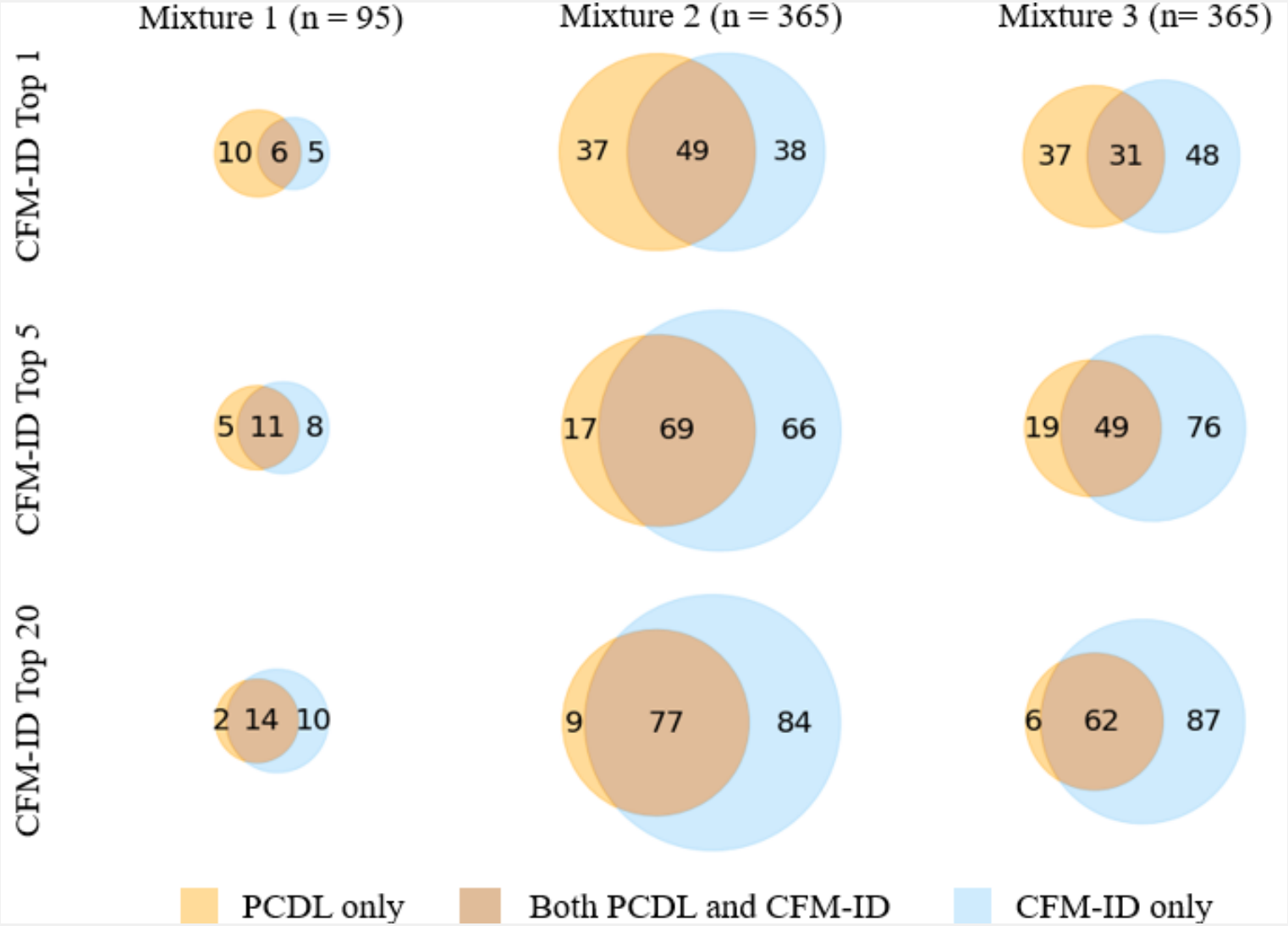
Allen, *et al* 2014, 2015, 2016




```
97 44.04947561 CC=[NH2+]
98 112.0393049 C=C(C([CH2+])=N)C(=O)O

# Date/time: 02/13/18 22:23:52
# CFM-ID version: 2.0 snapshot 10/23/2017
# DTXCID: DTXCID40539667
# SMILES: CCOC(=O)C1=C(C)N=C2C=C(C)C(C)=CC2=C1
# MASS: 243.125928791
# FORMULA: C15H17NO2
# INCHI_KEY: PKWJEOYUFQXCBY-UHFFFAOYSA-N
energy0
15.02292652 0.2213297725 4 (0.22133)
27.02292652 0.1671394176 13 (0.16714)
29.00219107 0.003415204535 26 (0.0034152)
29.03857658 0.7931575846 14 (0.79316)
41.00219107 0.05449899314 52 (0.054499)
42.03382555 0.002436421336 104 (0.0024364)
43.01784114 0.03020513753 53 (0.030205)
44.99710569 0.002435326653 84 (0.0024353)
45.0334912 0.045173947 54 (0.045174)
46.06512568 0.00275766249 105 (0.0027577)
47.01275576 0.001588587881 83 (0.0015886)
47.04914126 0.7597376869 55 (0.75974)
51.02292652 0.0008416543192 41 (0.00084165)
57.0334912 0.002233145503 57 (0.0022331)
65.00219107 0.0008022807615 30 (0.00080228)
65.03857658 0.005921490873 43 (0.0059215)
67.05422664 0.01637516089 112 (0.016375)
68.99710569 0.0308626766 80 (0.030863)
69.06987671 0.07637646561 127 (0.076376)
71.01275576 0.01301877006 81 (0.013019)
73.02840582 0.09231084238 82 (0.092311)
75.04405588 0.03569749353 85 (0.035697)
103.0542266 0.05442462493 70 (0.054425)
105.0698767 0.07415736715 96 (0.074157)
117.0698767 0.04642433142 94 (0.046424)
```

Predicted MS/MS spectra provide greater coverage than empirical libraries



Chao, *et al*, in prep

Evaluating on CASMI 2016

- Critical Assessment of Small Molecule Identification
  - Training data= 312 peak lists (from 285 substances)
    - 234 MS/MS in positive mode
    - 58 in negative mode
  - Challenge Data= 208 peak lists (from 188 substances)
    - 127 in positive mode
    - 81 in negative mode
- Precursor ion search window= 15 ppm
- Fragment ion match threshold= 0.02 Da
- Candidates limited to Dashboard results within precursor ion search window

http://www.casmi-contest.org/2016/index.shtml

CASMI 2016 Contest Challenge Set (n=208)

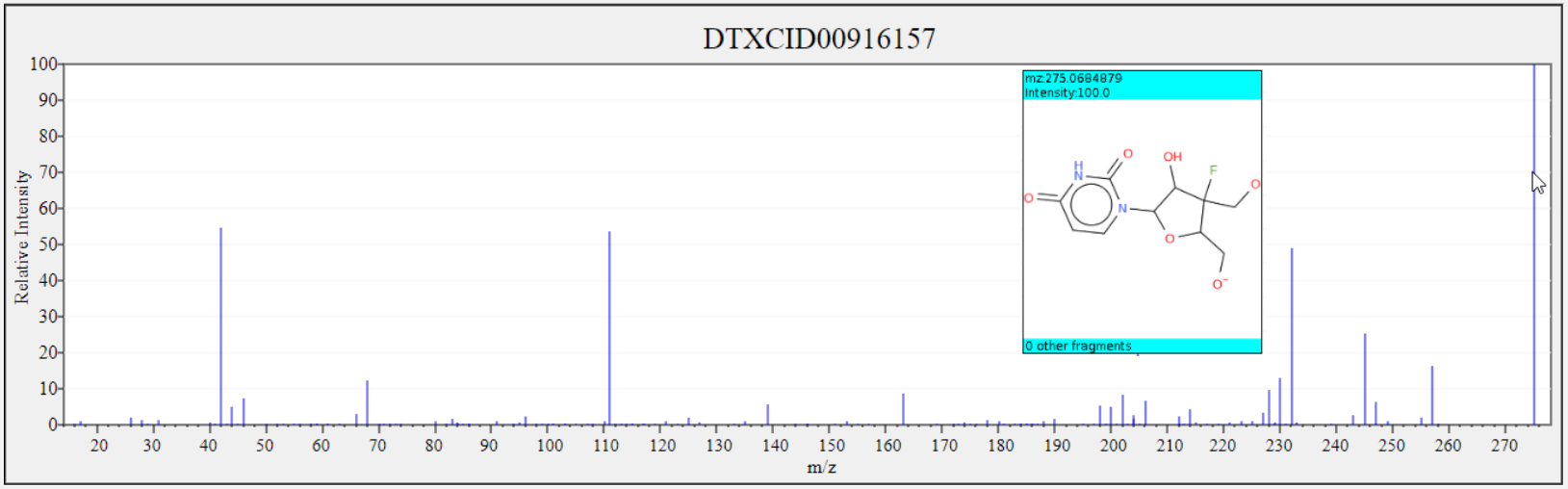
CFM-ID only

	# Identified	% of Total
#1 Hits	89	43%
Top 5	154	74%
Top 10	174	84%
Top 20	190	91%

CFM-ID +DSSTox Data Sources

	# Identified	% of Total
#1 Hits	154	74%
Top 5	195	94%
Top 10	198	95%
Top 20	202	97%

Access via CompTox Dashboard



- Data available for download after publication

McEachran, *et al*, in prep

Mockup, work in progress….

Non Target Analysis Prototype

Mass Search

± | Min/Max

322.144318476 Da ± 5 Da ppm

Molecular Formula Search

Molecular Formula

Mass or Formula must be entered before searching spectrum

Ionization Type

ESI+

Spectra Input

Single Energy | Multiple

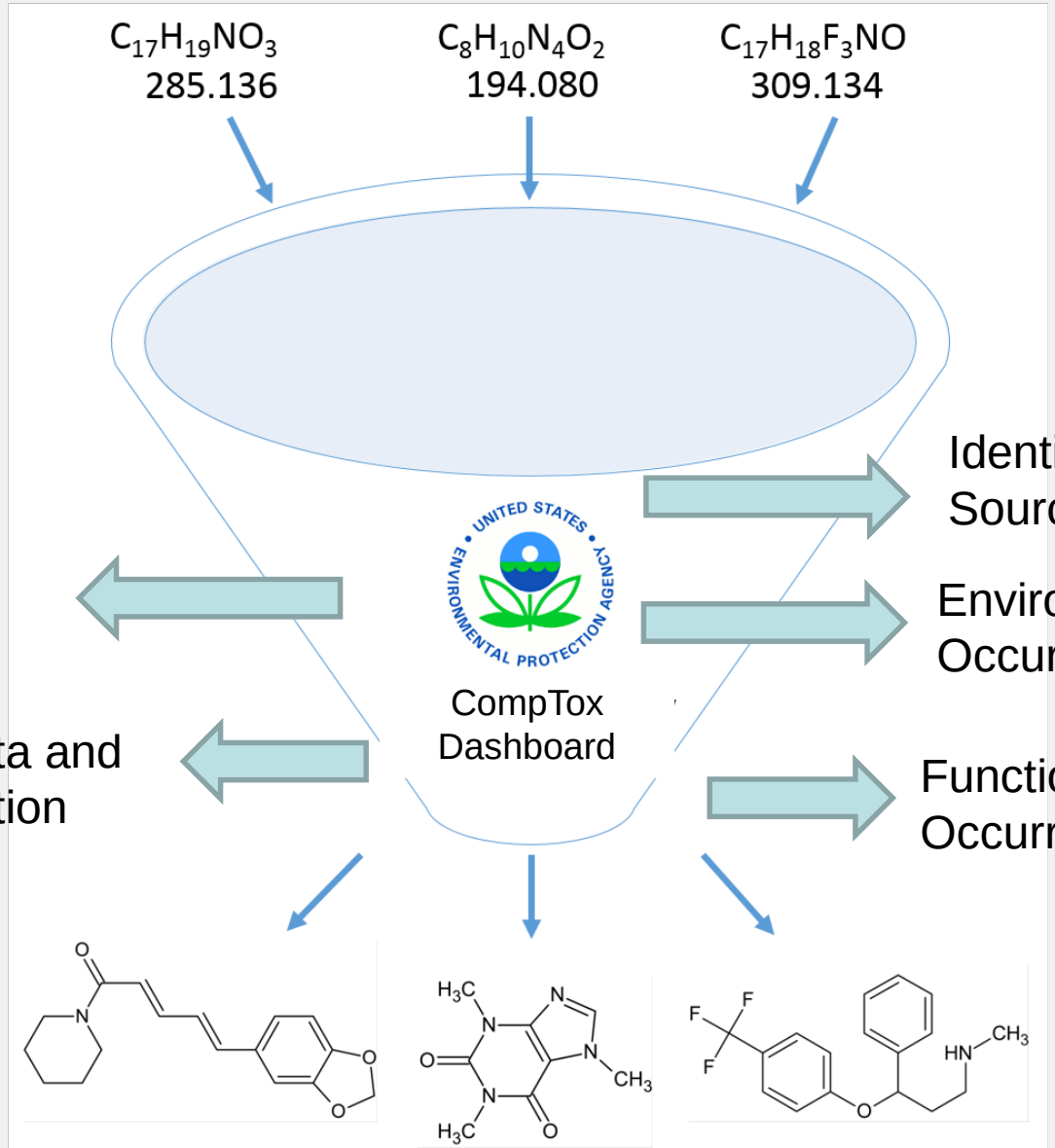
304.1332052 11.6199475
198.0913404 7.306439699
123.0440559 6.538348292
196.0756904 5.269463115
216.1019051 4.700461978

Peak Match Window:

0.02 Da ppm

Search

Dashboard in NTA Workflows



Identification by Data Source Ranking

Environmental Media Occurrence

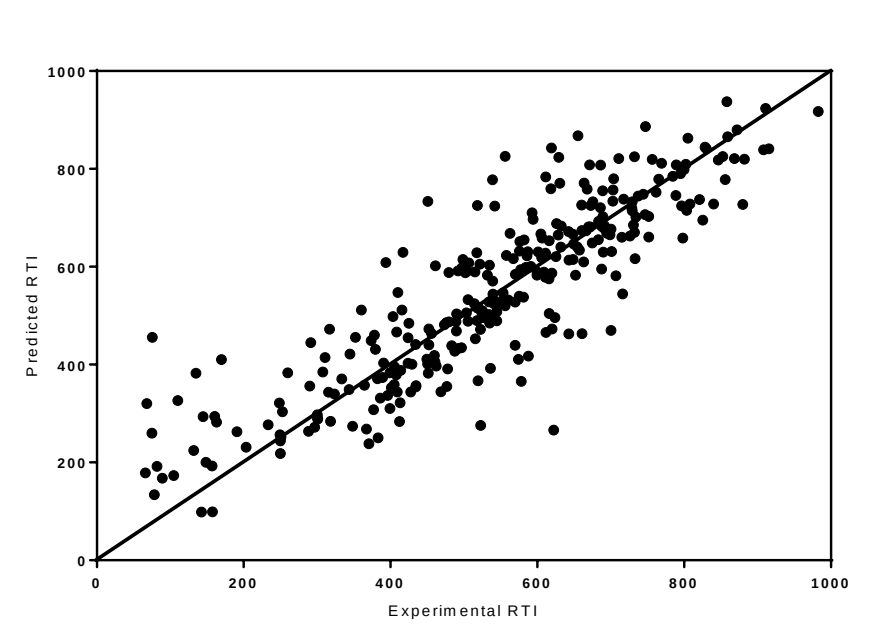
Functional Use/Product Occurrence

Retention Time Prediction

MS/MS Data and Fragmentation Prediction

Future Directions

- Combined data visualization
- Retention time index (RTI) predictions
- Ongoing expansion of the database
- Integration to public MS databases



Conclusions

- Databases are effective resources in SSA/NTA
- CompTox Dashboard provides access to chemistry data for >760,000 chemical substances
- Predicted MS/MS spectra linked within the CompTox Dashboard further enhances effectiveness and increases confidence in identifications

Acknowledgements

EPA NCCT
Tony Williams
Chris Grulke
Jeff Edwards

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

External Collaborators
Emma Schymanski- Univ. Luxembourg
Christoph Ruttkies- IPB, Halle
Kamel Mansouri- ILS, Inc

*ORISE Research Participant

Questions?

- mceachran.andrew@epa.gov
- http://orcid.org/0000-0003-1423-330X
- Associated presentations:
  - AGRO 29: Leveraging chemistry data to improve exposure analyses using the EPA's CompTox Chemistry Dashboard
  - ANYL 100: Developing tools for high resolution mass spectrometry-based screening via the EPA's CompTox Chemistry Dashboard
  - ENVR 152: EPA Comptox Chemistry Dashboard as a data integration hub for environmental chemistry data