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

EPA
United States
Environmental Protection
Agency

Leveraging chemistry data to improve exposure analyses using the EPA's CompTox Dashboard

Andrew D. McEachran
Katherine Phillips, Kristin Isaacs, Seth Newton, Chris Grulke, Jon Sobus, and Antony J. Williams

AGRO 29

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 19, 2018

Exposure Science in the 21st Century: What are Key Drivers?

1) Understanding causes of disease

"…70-90% of disease risks are probably due to differences in environments"

EPIDEMIOLOGY

Environment and Disease Risks

A new paradigm is needed to assess how a lifetime of exposure to environmental factors affects the risk of developing chronic diseases.

Stephen M. Rappaport and Martyn T. Smith

Although the risks of developing chronic diseases are attributed to both genetic and environmental factors, 70 to 90% of disease risks are probably due to differences in environments (*1*–*3*). Yet, epidemiologists increasingly use genome-wide association studies (GWAS) to investigate diseases, while relying on questionnaires to characterize "environmental exposures." This is because GWAS represent the only approach for exploring the totality of any risk factor (genes, in this case) associated with disease prevalence. Moreover, the value of costly genetic information is diminished when inaccurate and imprecise environmental data lead to biased inferences regarding gene-environment interactions (*4*). A more comprehensive and quantitative view of environmental exposure is needed if epidemiologists are to discover the major causes of chronic diseases.

An obstacle to identifying the most important environmental exposures is the fragmentation of epidemiological research along lines defined by different factors. When epidemiologists investigate environmental risks, they tend to concentrate on a particular category of exposures involving air and water pollution, occupation, diet and obesity, stress and behavior, or types of infection. This slicing of the disease pie along parochial lines leads to scientific separation and confuses the definition of "environmental exposures." In fact, all of these exposure categories can contribute to chronic diseases and should be investigated collectively rather than separately.

To develop a more cohesive view of environmental exposure, it is important to recognize that toxic effects are mediated through chemicals that alter critical molecules, cells, and physiological processes inside the body. Thus, it would be reasonable to consider the "environment" as the body's internal chemical environment and "exposures" as the amounts of biologically active chemicals in this internal environment. Under this view, exposures are not restricted to chemicals (toxicants) entering the body from air, water, or food, for example, but also include chemicals produced by inflammation, oxidative stress, lipid peroxidation, infections, gut flora, and other natural processes (*5*, *6*) (see the figure). This internal chemical environment continually fluctuates during life due to changes in external and internal sources, aging, infections, life-style, stress, psychosocial factors, and preexisting diseases.

The term "exposome" refers to the totality of environmental exposures from conception onwards, and has been proposed to be a

School of Public Health, University of California, Berkeley, CA 94720–7356, USA. E-mail: srappaport@berkeley.edu

Published by AAAS

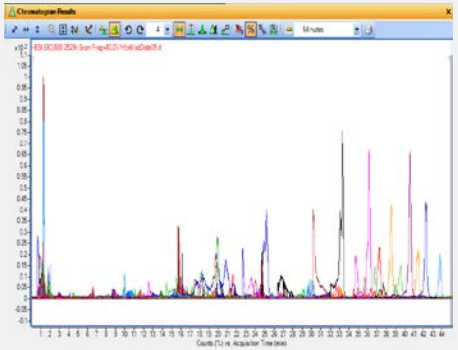
2) Ensuring chemical safety and human/eco health

Slide from Sobus

EPA United States Environmental Protection Agency

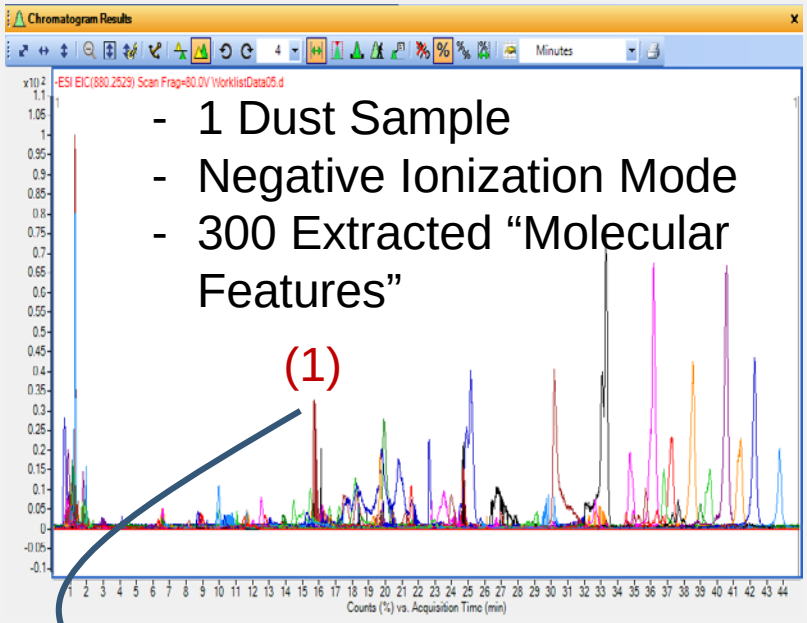
Comparing Analysis Approaches

- Targeted Analysis:
  - We know exactly what we're looking for
  - 10s – 100s of chemicals
  - <<1% of the exposome
- Suspect Screening Analysis (SSA):
  - We have chemicals of interest
  - 100s – 1,000s of chemicals
  - ~5-10% of the exposome
- Non-Targeted Analysis (NTA):
  - We have no preconceived lists
  - 1,000s – 10,000s of chemicals
  - 90-95% of the exposome
  - In dust, soil, food, air, water, products- potential exposure source for plants, animals, and humans



Slide from Sobus, Williams

General Goals of SSA/NTA



1) Prioritize "Molecular Features"

2) Correctly assign formulas

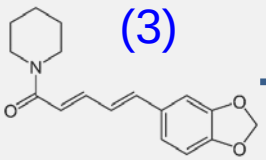
3) Correctly assign structures

4) Identify chemical sources

5) Predict chemical concentrations

EXPOSURE

(2) C17H19NO3



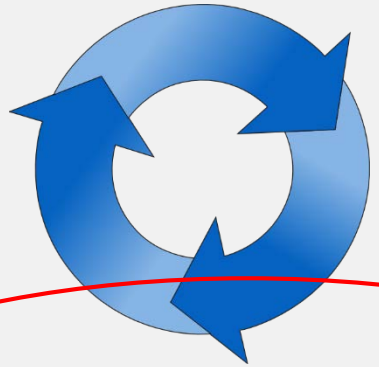
(4)

(5) 12 µg/g

EPA United States Environmental Protection Agency

The General Approach

Analytical Instruments



Comp. Tools & Workflows

FOR IDENT

Xcms Online

MetFrag

Databases

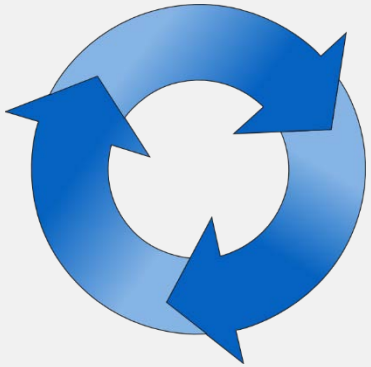
ChemSpider
Search and share chemistry

PubChem

MassBank

The General Approach

Analytical Instruments



Comp. Tools & Workflows

FOR IDENT

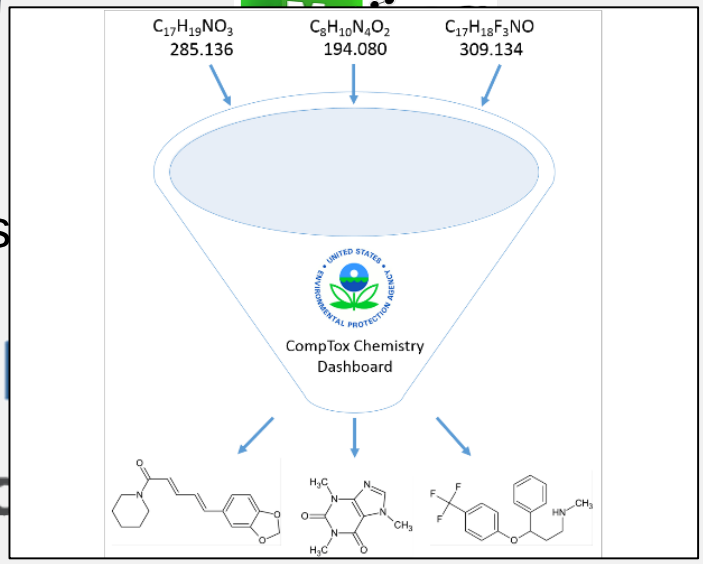
Xcms Online

Databases

ChemSpider
Search and share chemistry

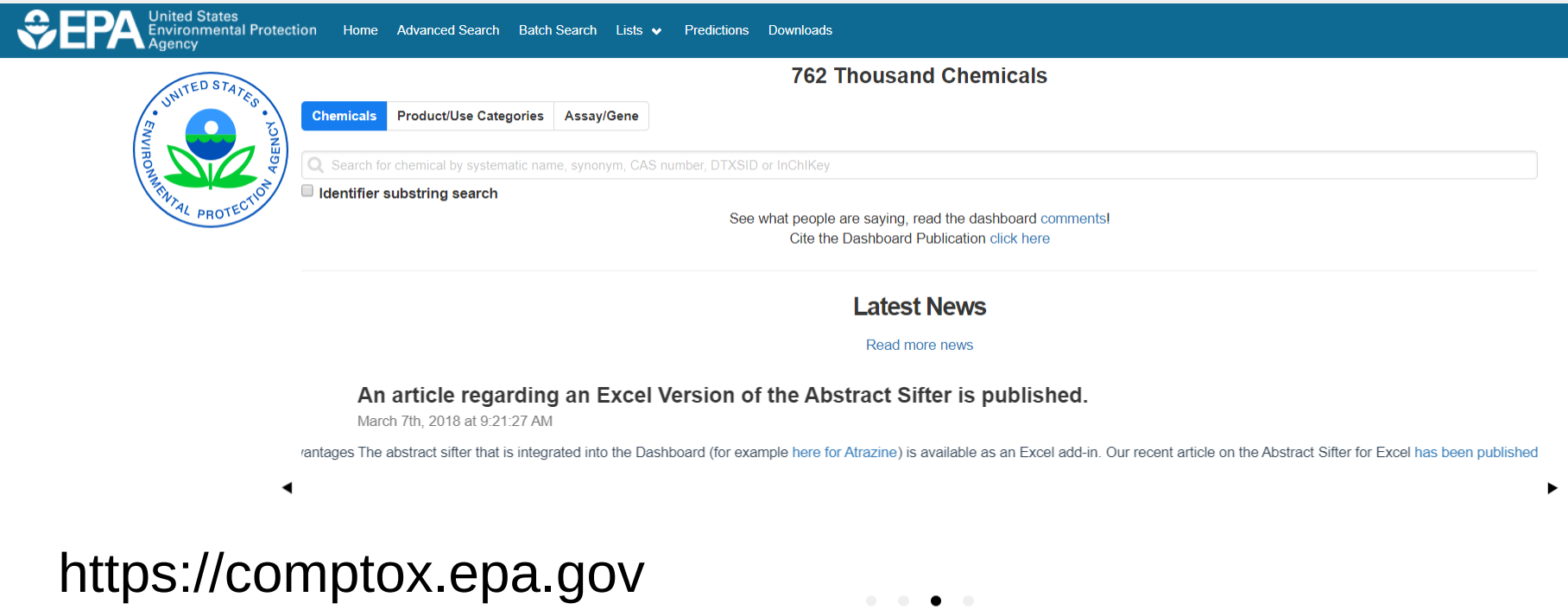
Pu

MassBa



EPA United States Environmental Protection Agency

CompTox Dashboard



https://comptox.epa.gov

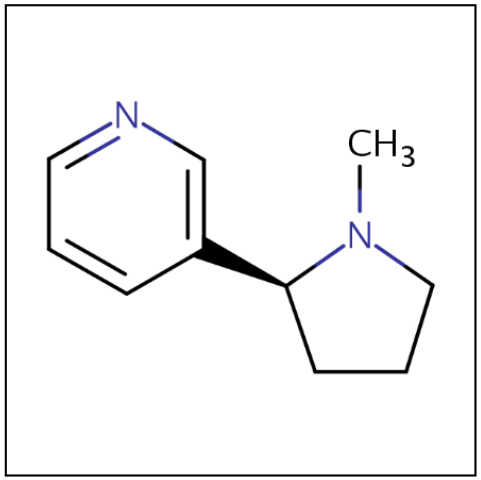
EPA United States Environmental Protection Agency

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

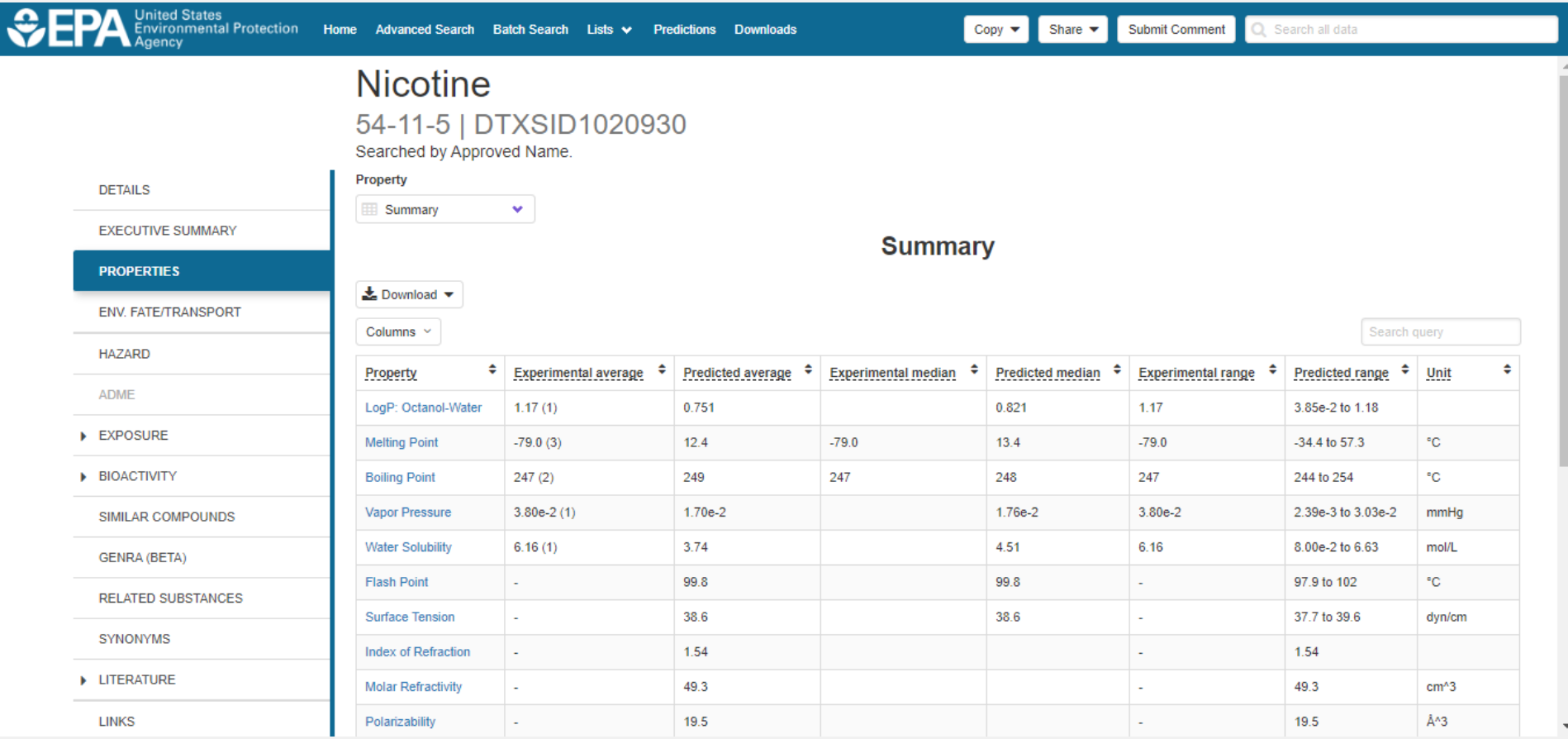
Linked Substances

Presence in Lists

Record Information

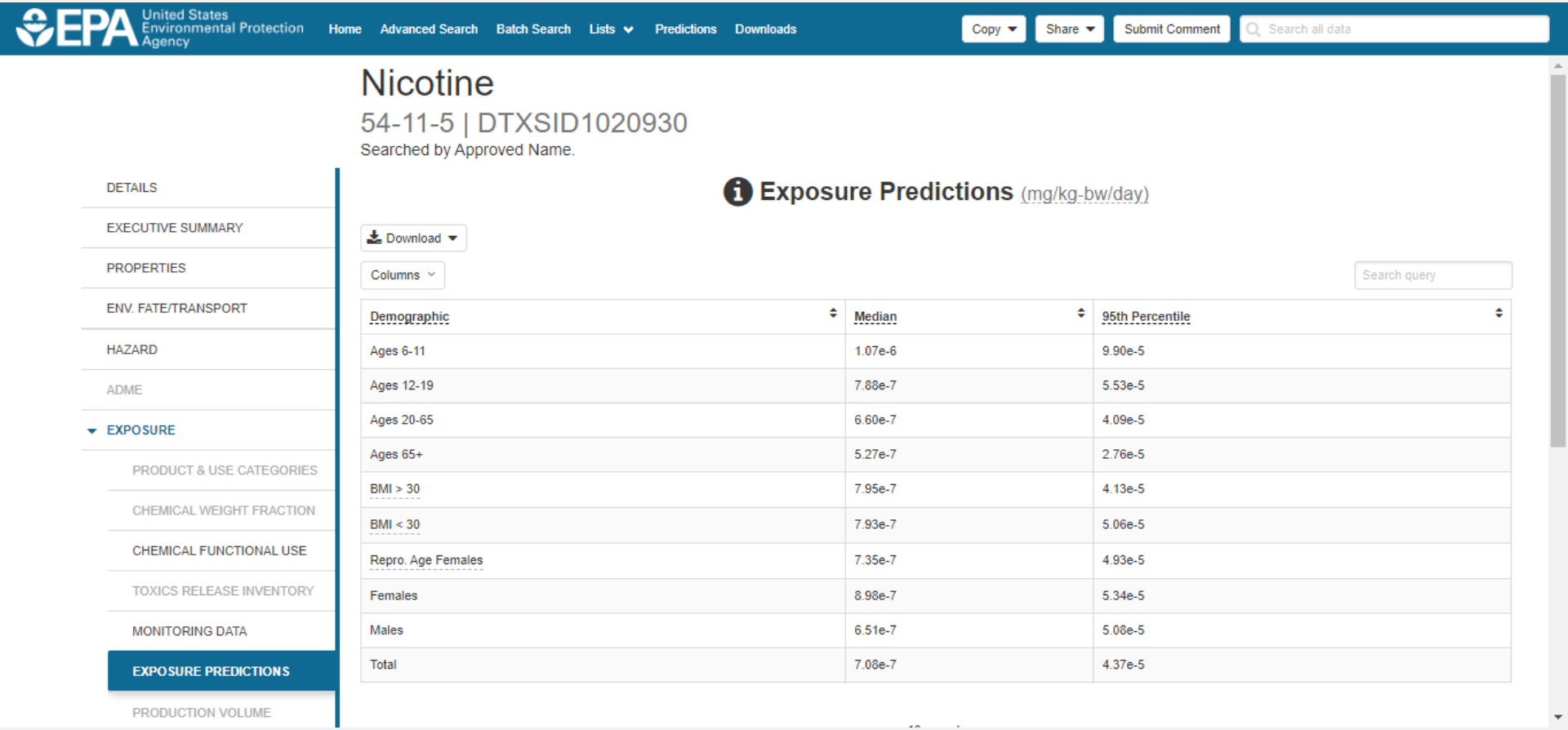
Quality Control Notes

EPA United States Environmental Protection Agency

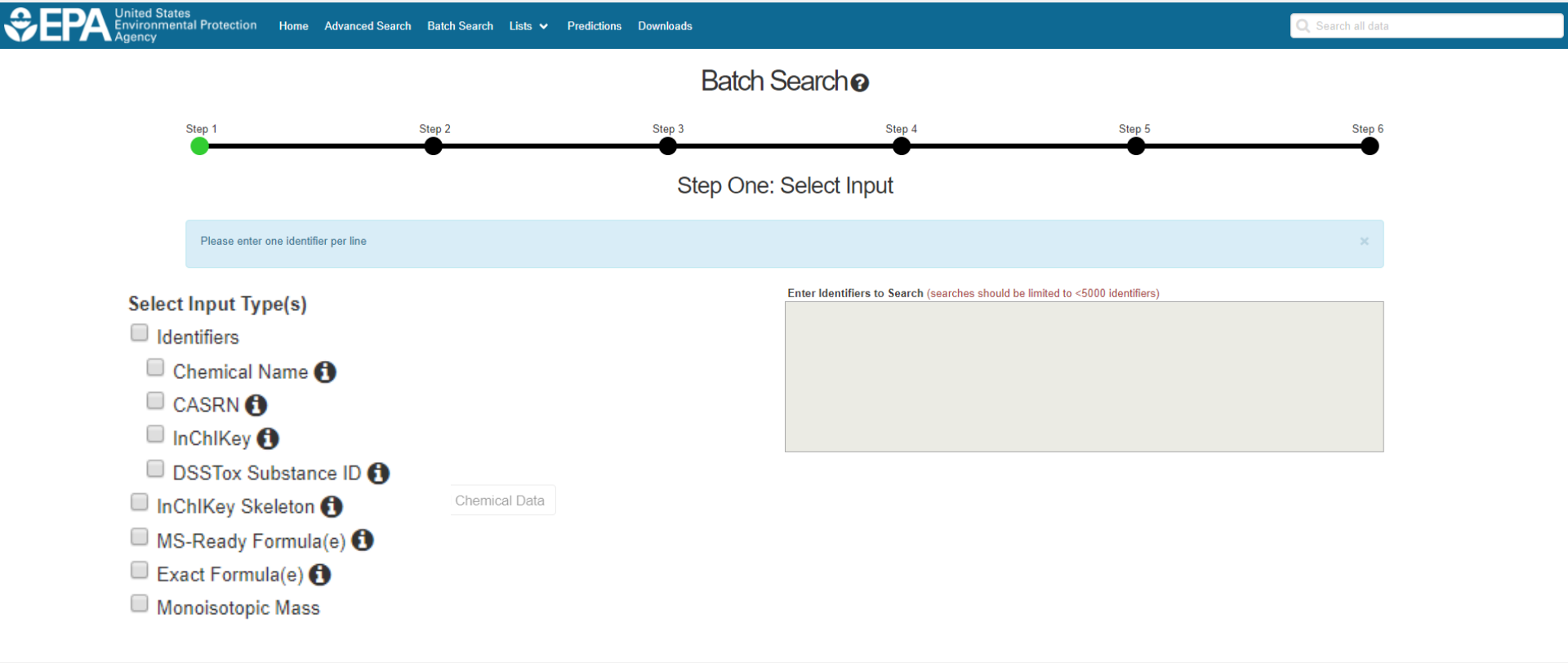


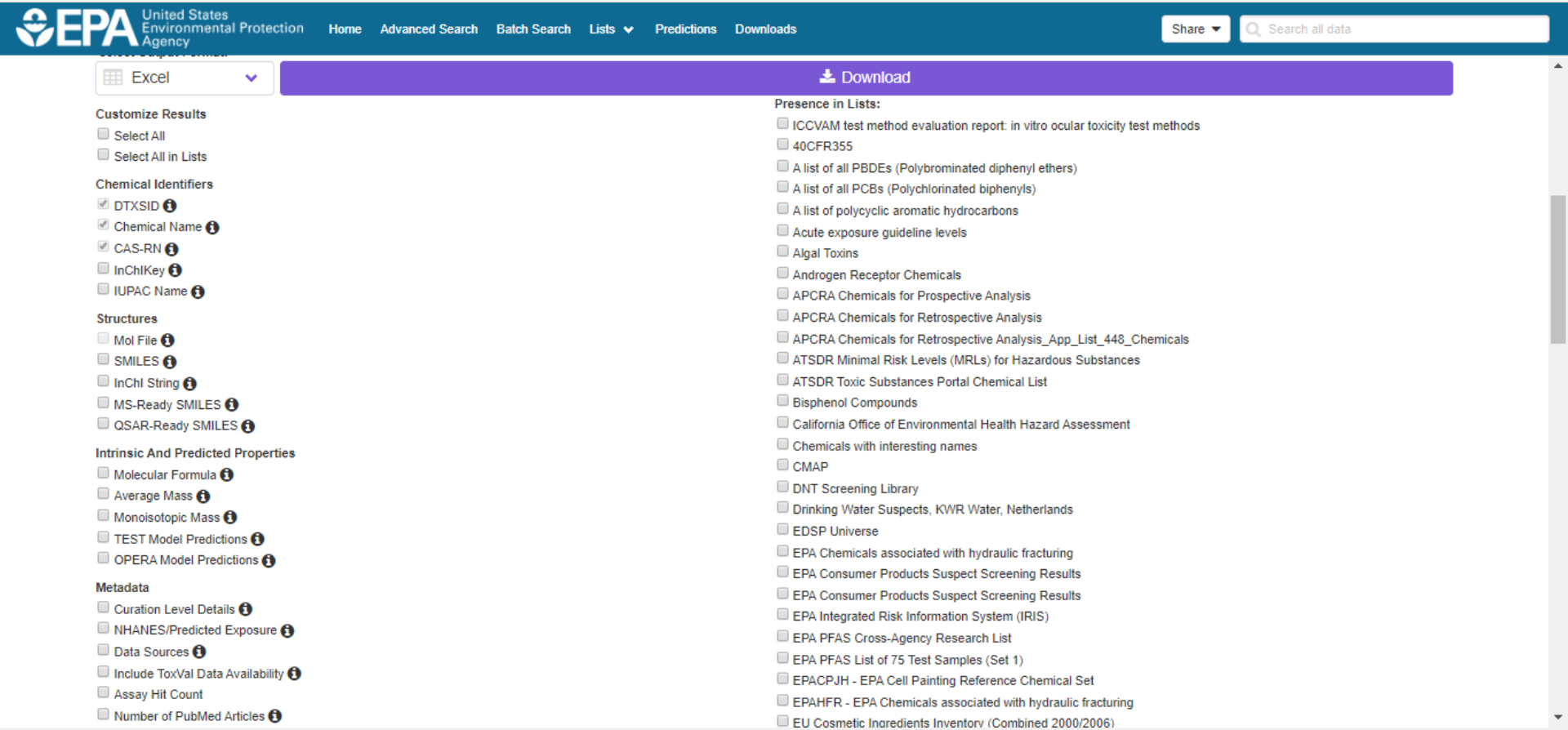
Property	Experimental average	Predicted average	Experimental median	Predicted median	Experimental range	Predicted range	Unit
LogP: Octanol-Water	1.17 (1)	0.751		0.821	1.17	3.85e-2 to 1.18	
Melting Point	-79.0 (3)	12.4	-79.0	13.4	-79.0	-34.4 to 57.3	°C
Boiling Point	247 (2)	249	247	248	247	244 to 254	°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

EPA United States Environmental Protection Agency



EPA United States Environmental Protection Agency




EPA United States Environmental Protection Agency
Home
Advanced Search
Batch Search
Lists
Predictions
Downloads
Share
Search all data
Excel
Download
Customize Results
Select All
Select All in Lists
Chemical Identifiers
DTXSID
Chemical Name
CAS-RN
InChIKey
IUPAC Name
Structures
Mol File
SMILES
InChI String
MS-Ready SMILES
QSAR-Ready SMILES
Intrinsic And Predicted Properties
Molecular Formula
Average Mass
Monoisotopic Mass
TEST Model Predictions
OPERA Model Predictions
Metadata
Curation Level Details
NHANES/Predicted Exposure
Data Sources
Include ToxVal Data Availability
Assay Hit Count
Number of PubMed Articles
Presence in Lists:
ICCVAM test method evaluation report: in vitro ocular toxicity test methods
40CFR355
A list of all PBDEs (Polybrominated diphenyl ethers)
A list of all PCBs (Polychlorinated biphenyls)
A list of polycyclic aromatic hydrocarbons
Acute exposure guideline levels
Algal Toxins
Androgen Receptor Chemicals
APCRA Chemicals for Prospective Analysis
APCRA Chemicals for Retrospective Analysis
APCRA Chemicals for Retrospective Analysis_App_List_448_Chemicals
ATSDR Minimal Risk Levels (MRLs) for Hazardous Substances
ATSDR Toxic Substances Portal Chemical List
Bisphenol Compounds
California Office of Environmental Health Hazard Assessment
Chemicals with interesting names
CMAP
DNT Screening Library
Drinking Water Suspects, KWR Water, Netherlands
EDSP Universe
EPA Chemicals associated with hydraulic fracturing
EPA Consumer Products Suspect Screening Results
EPA Consumer Products Suspect Screening Results
EPA Integrated Risk Information System (IRIS)
EPA PFAS Cross-Agency Research List
EPA PFAS List of 75 Test Samples (Set 1)
EPACPJH - EPA Cell Painting Reference Chemical Set
EPAHFR - EPA Chemicals associated with hydraulic fracturing
EU Cosmetic Ingredients Inventory (Combined 2000/2006)

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

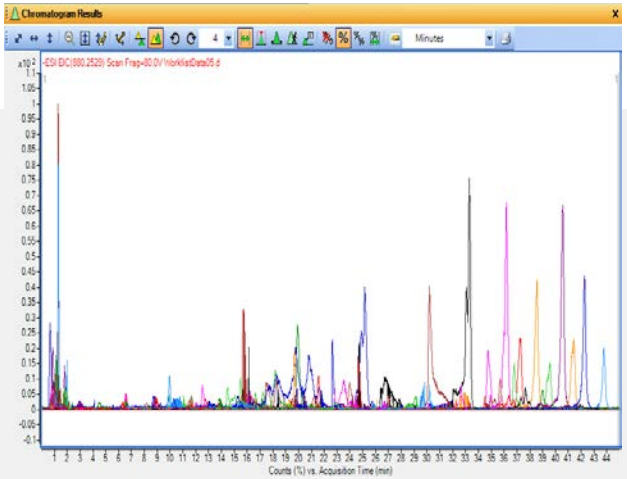
Select Input Type(s)

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

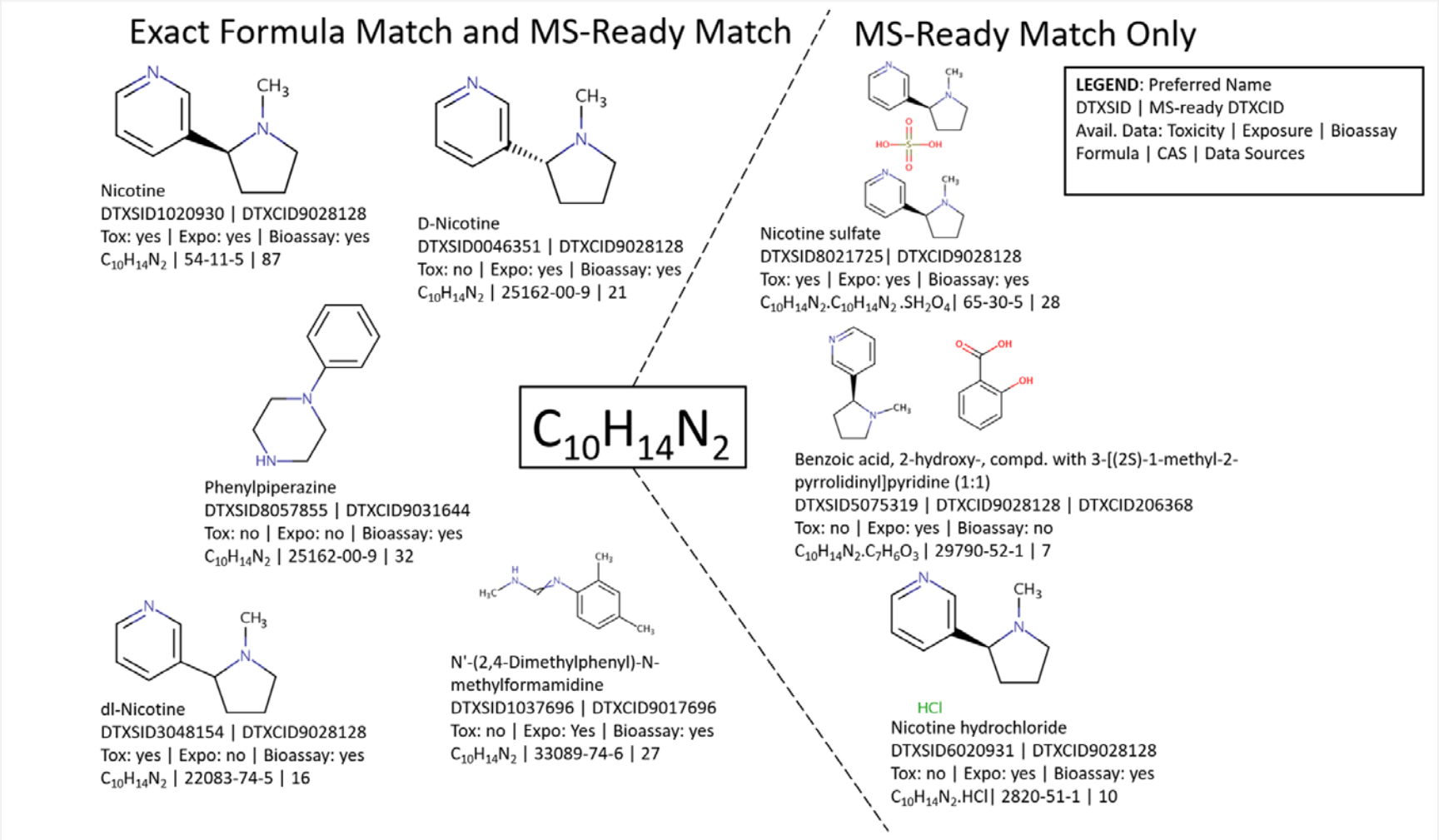
Enter Identifiers to Search (searches should be limited to <5000 identifiers)

Chemical Data

C10H14N2



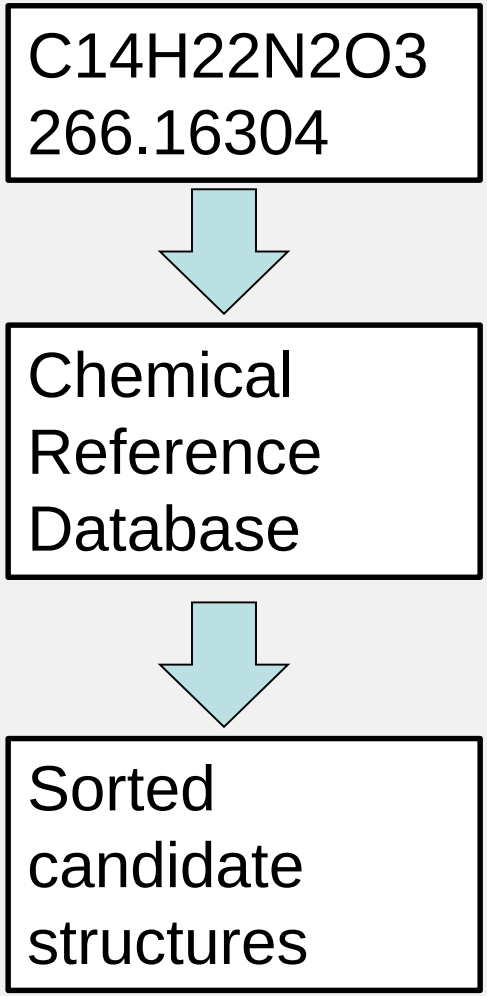
MS-Ready Structures improve database searching



McEachran, *et al*, accepted manuscript

Data Source Ranking for Identification in SSA/NTA

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

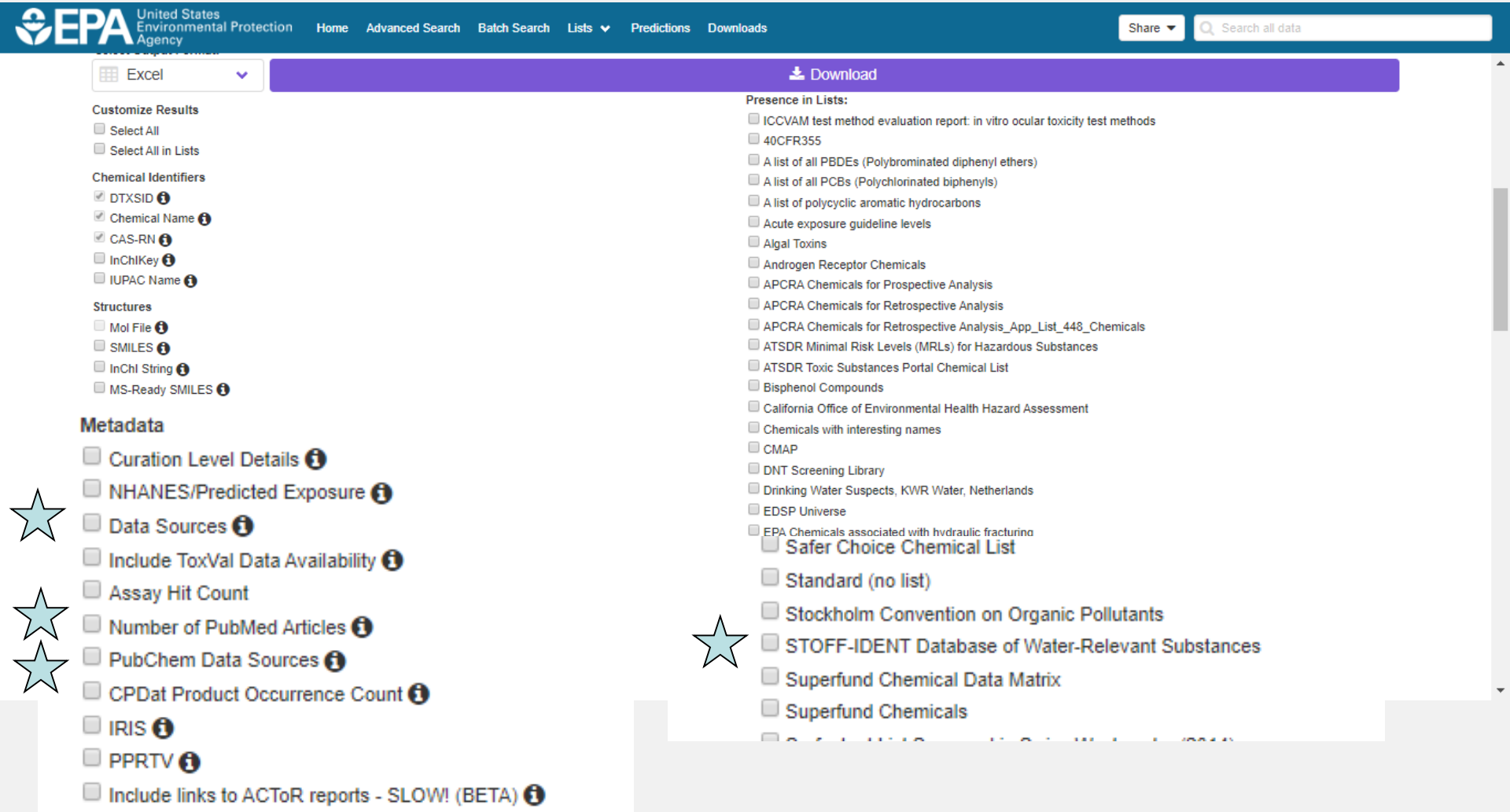
EPA United States Environmental Protection Agency

Increasing Data Streams to Improve Identifications

Chemistry Dashboard

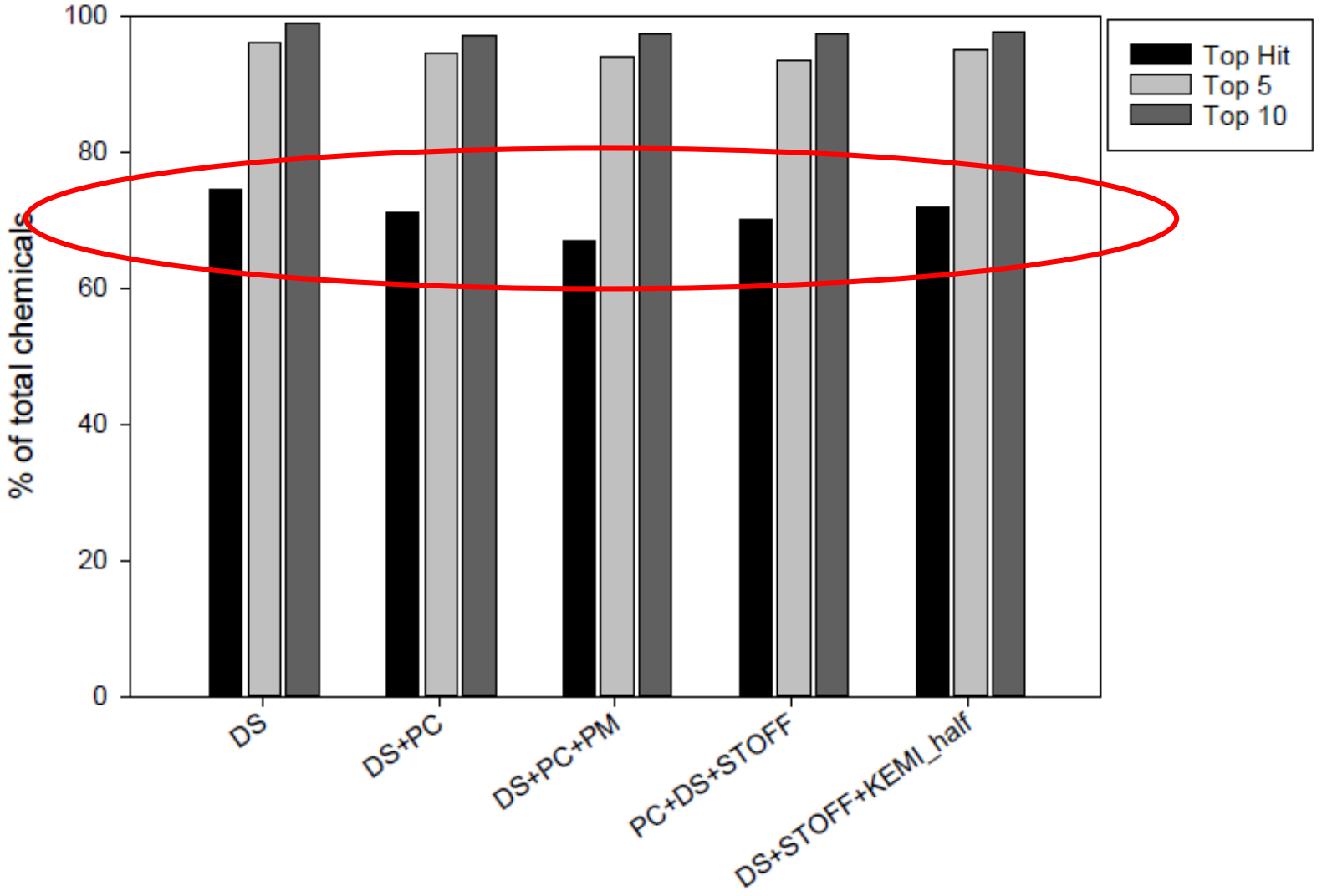
- US EPA CompTox Dashboard Data Sources (DS)
- PubChem Data Source Count
- PubMed Reference Count
- Presence in STOFF-IDENT Database
- Predicted Environmental Media Occurrence
- CPDat Product Occurrence Count
- OPERA PhysChem Properties
- NORMAN Network Priority List

All available via Batch Search:

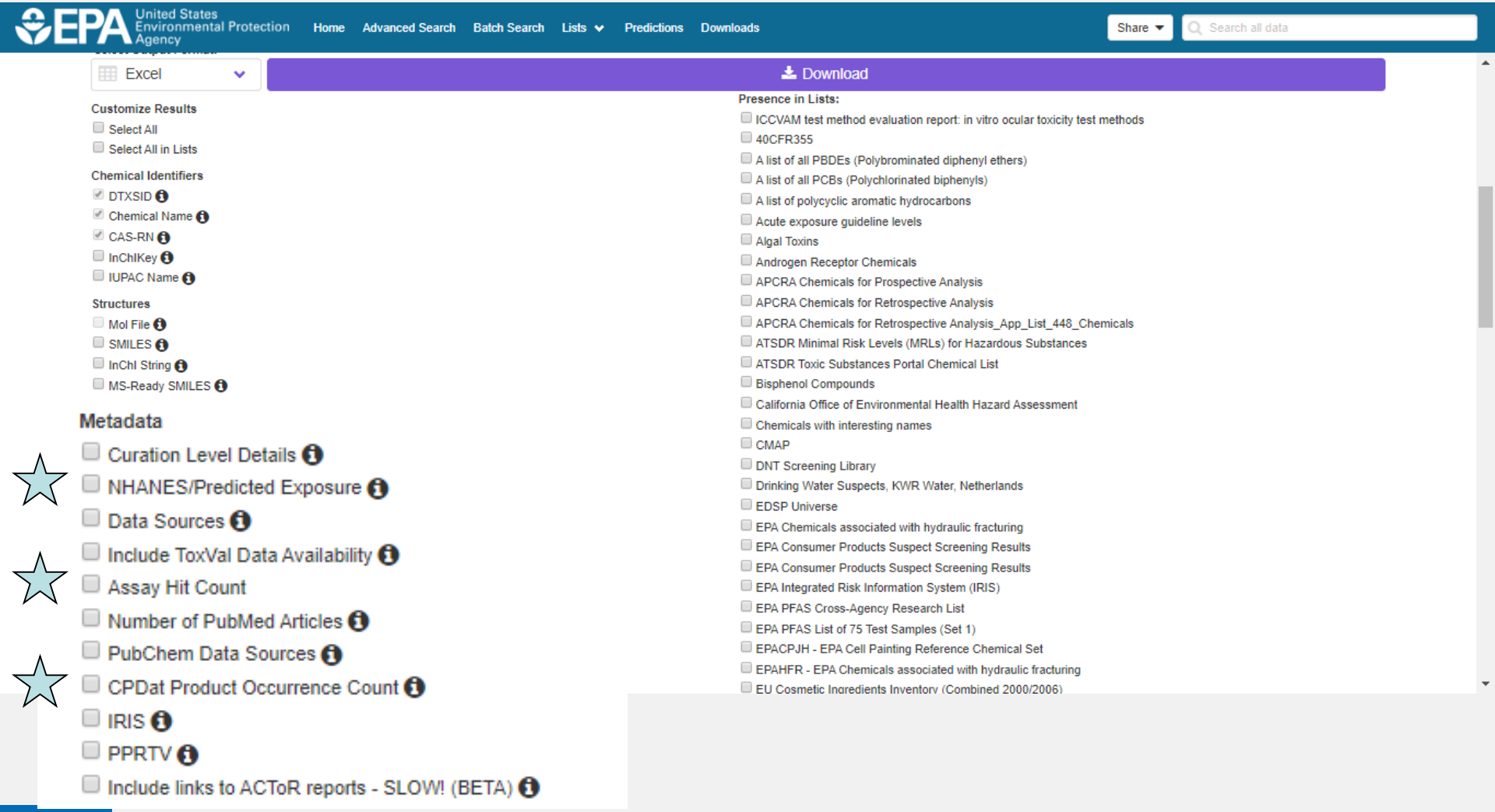


Identification ranks for 1783 chemicals using multiple data streams

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



Linking exposure potential and bioactivity for prioritization



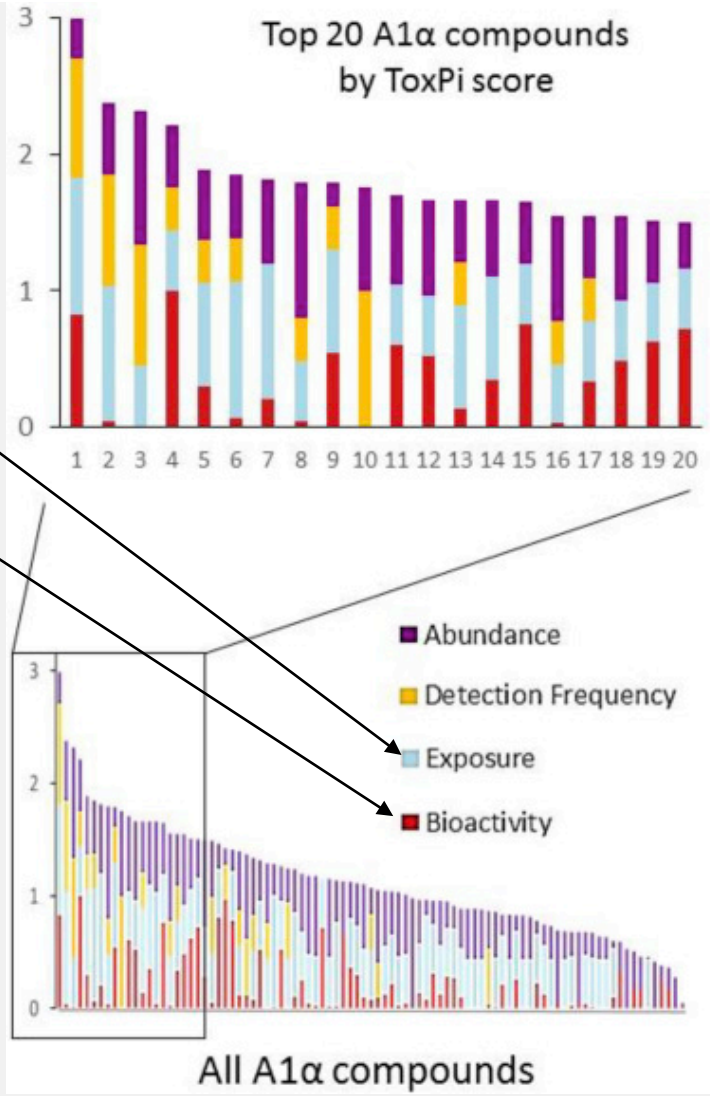
Environmental Pollution
Volume 234, March 2018, Pages 297-306

Suspect screening and non-targeted analysis of drinking water using point-of-use filters ☆

Seth R. Newton [a], Rebecca L. McMahen [a, b], Jon R. Sobus [a], Kamel Mansouri [b, c, 1], Antony J. Williams [c], Andrew D. McEachran [b, c], Mark J. Strynar [a]

Metadata

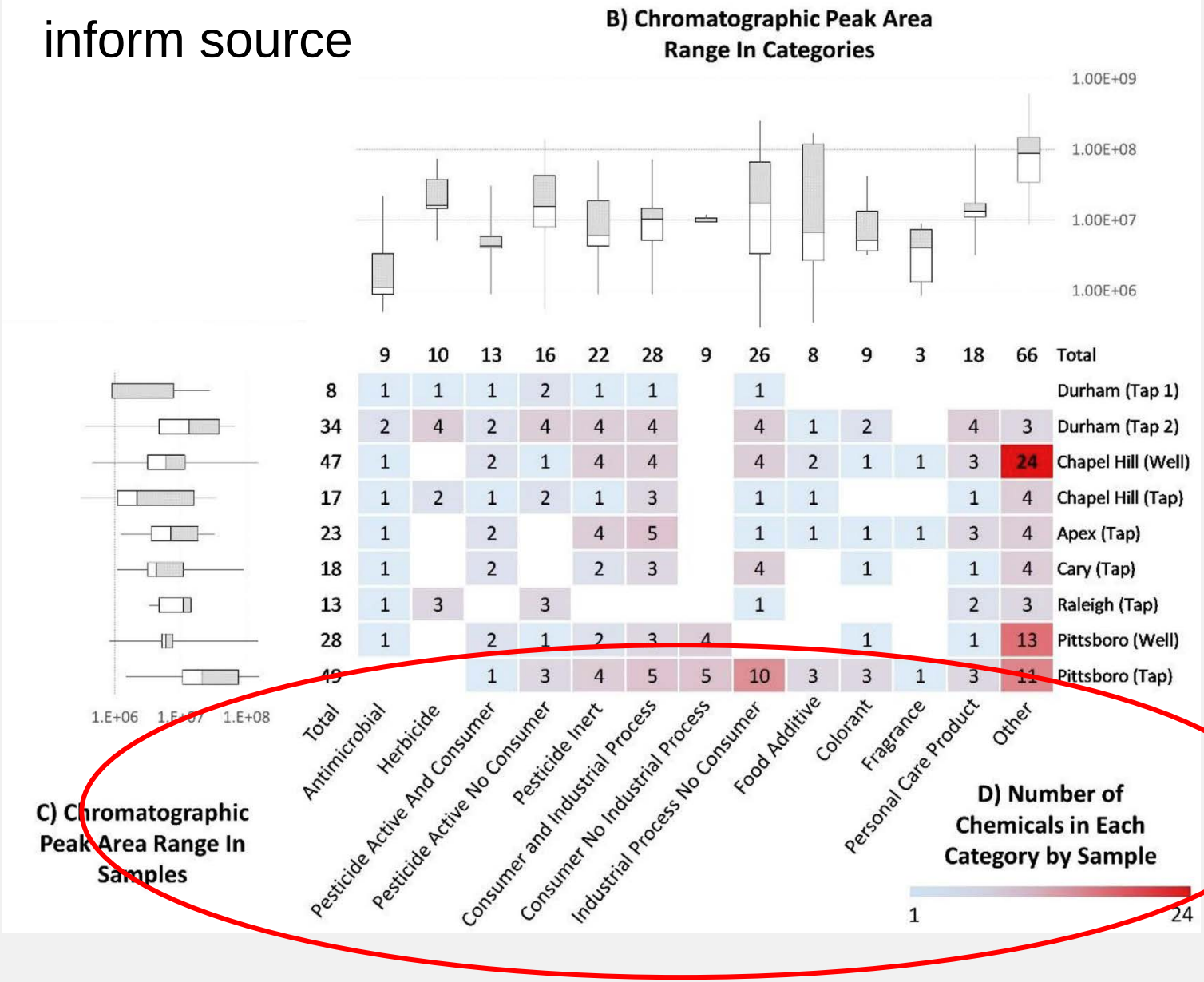
- Curation Level Details
- NHANES/Predicted Exposure
- Data Sources
- Include ToxVal Data Availability
- Assay Hit Count
- Number of PubMed Articles
- PubChem Data Sources
- CPDat Product Occurrence Count
- IRIS
- PPRTV
- Include links to ACToR reports - SLOW! (BETA)



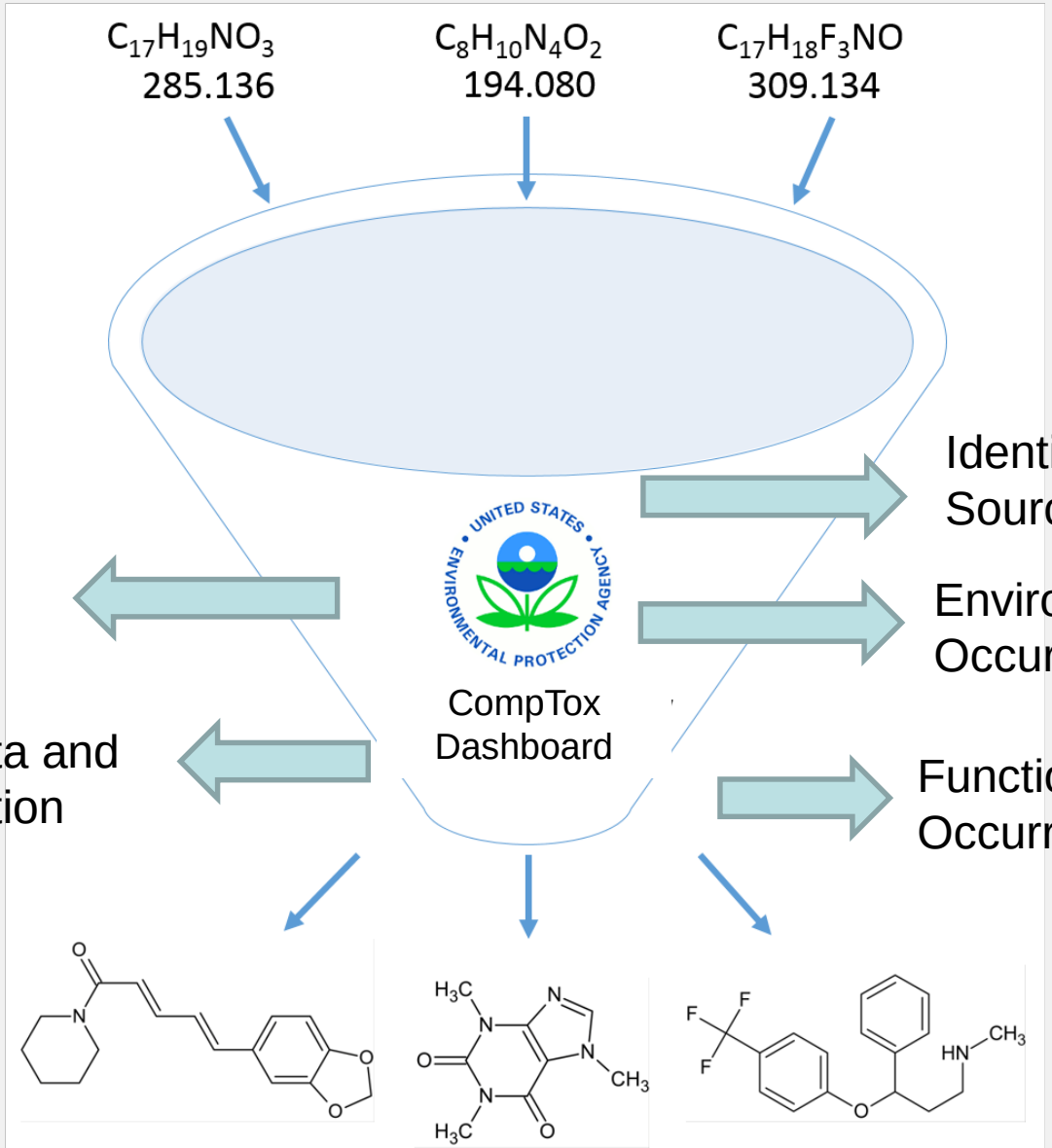
#	Compound	ToxPi Score
1	1,2-Benzisothiazolin-3-one*	2.99
2	Diethyleneglycol	2.38
3	N-[3-(Dimethylamino)propyl] methacrylamide	2.32
4	Nonylparaben	2.22
5	Dipentylphthalate	1.89
6	2-[2-(2-Butoxyethoxy) ethoxy]ethanol*	1.85
7	N,N-Dimethyldodecan-1-amine*	1.81
8	Sucralose	1.80
9	PFOS*	1.79
10	2-(2-Ethoxyethoxy) ethyl acetate*	1.76
11	TDCPP*	1.71
12	Zearalanol	1.67
13	PFOA*	1.66
14	Butylparaben	1.66
15	Noristerat	1.65
16	p-Synephrine	1.55
17	Alprostadil	1.55
18	Sclareol	1.55
19	PFDA*	1.51
20	Simvastatin	1.50

*Confirmed with standard

CPCat terms can inform source



Dashboard in SSA/NTA Workflows



Retention Time Prediction

MS/MS Data and Fragmentation Prediction

Identification by Data Source Ranking

Environmental Media Occurrence

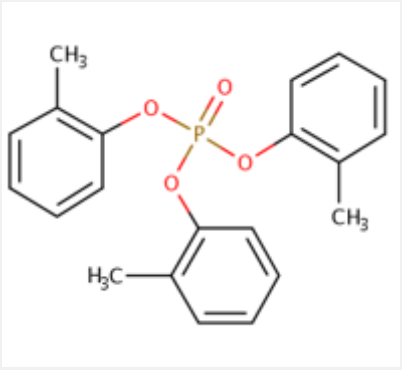
Functional Use/Product Occurrence

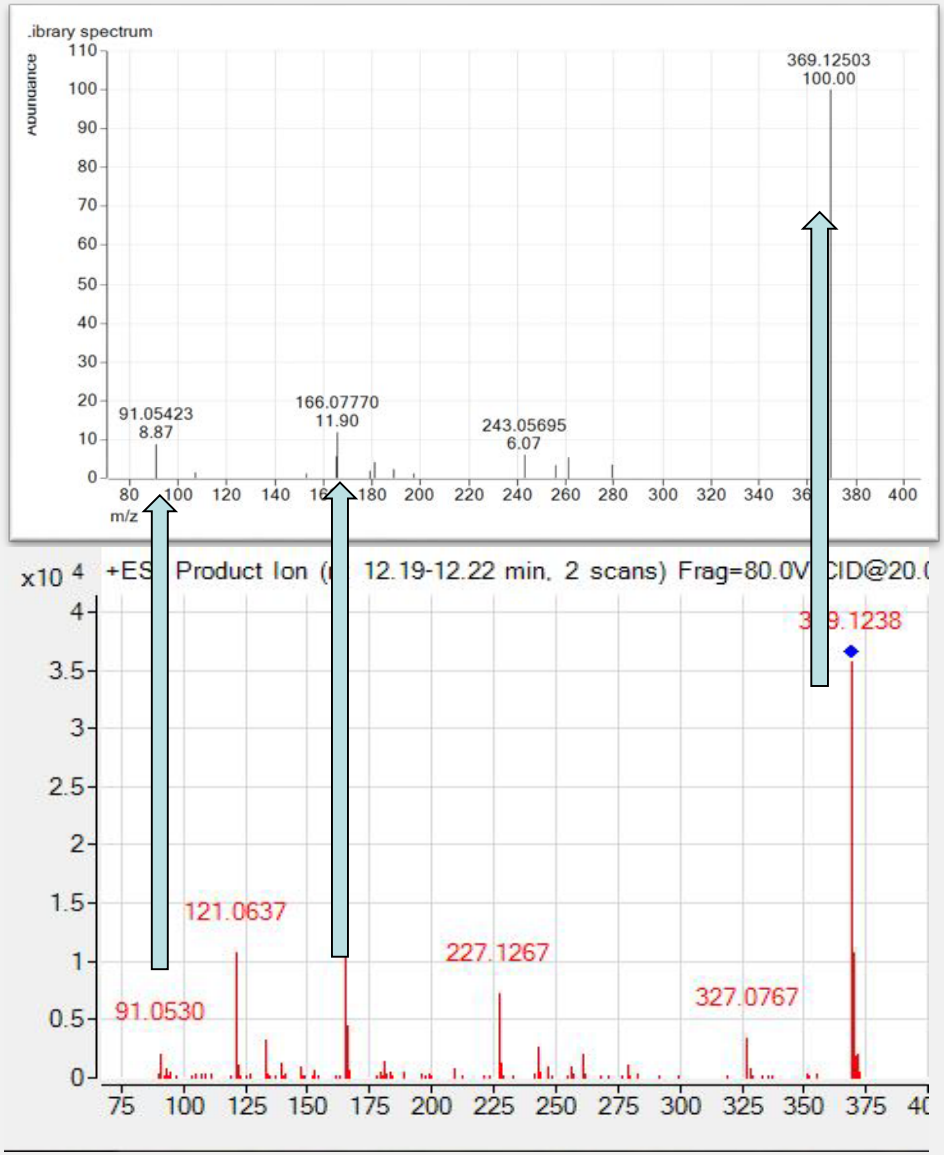
Future Directions

- Media occurrence prediction models
- Combined data visualization
- MS/MS match score calculations via the Dashboard
- Retention time index (RTI) predictions
- Ongoing expansion of the database
- Integration to public MS databases

MS/MS Spectral Matching for Identification

Library Fragmentation Spectra (20eV)

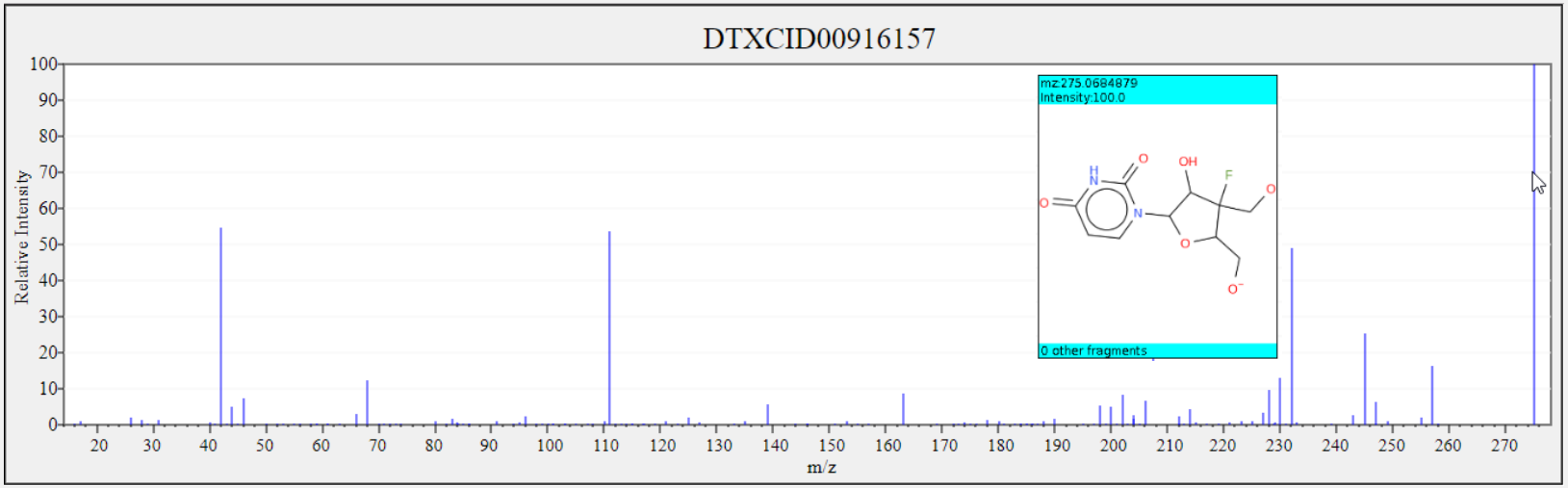




Observed Fragmentation Spectra (20eV)

Match Score

Predicted MS/MS Spectral Matching for Identification



CASMI 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%

EPA United States Environmental Protection Agency

Conclusions

- CompTox Dashboard provides access to a wealth of chemistry data that can be leveraged for exposure analyses
- NTA/SSA using the Dashboard results in high performance of identifications
- New data streams and predictions linked within the Dashboard further enhance NTA/SSA capabilities

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
Hussein Al-Ghoul*

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

*ORAU Participant

Questions?

- mceachran.andrew@epa.gov
- http://orcid.org/0000-0003-1423-330X
- Associated presentations:
  - 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
  - AGRO 107: Consensus ranking and fragmentation prediction for identification of unknowns in high resolution mass spectrometry