Developing tools for high resolution mass spectrometry-based screening via the EPA's CompTox Dashboard

Andrew D. McEachran[1,2]*, Kamel Mansouri[3], Hussein Al-Ghoul[4], Alex Chao[4], Chris Grulke[2], Jon R. Sobus[5], and Antony J. Williams[2]

[1]Oak Ridge Institute of Science and Education (ORISE) Research Participant, Research Triangle Park, NC
[2]U.S. Environmental Protection Agency, Office of Research and Development, National Center for Computational Toxicology, Research Triangle Park, NC
[3]Integrated Laboratory Systems, Inc., Morrisville, NC | [4]Oak Ridge Associated Universities, Research Triangle Park, NC
[5]U.S. Environmental Protection Agency, Office of Research and Development, National Exposure Research Laboratory, Research Triangle Park, NC

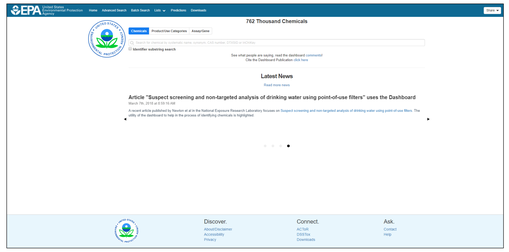
* mceachran.andrew@epa.gov | ORCiD: 0000-0003-1423-330X

Problem Definition and Goals

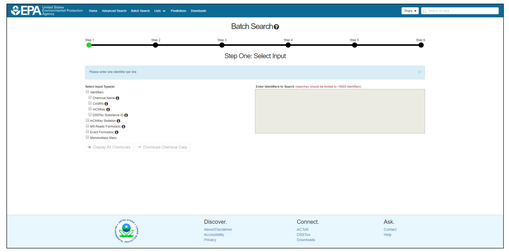
Problem: Non-targeted and suspect screening studies using high resolution mass spectrometry (HRMS) have revolutionized the detection of chemicals in complex matrices. However, data processing remains challenging due to the vast number of chemicals detected in samples, software and computational requirements of data processing, and inherent uncertainty in confidently identifying chemicals from candidate lists.

Goals: Develop tools, data, and visualization approaches within an open chemistry resource to provide a freely available software tool to support structure identification and non-targeted analysis.

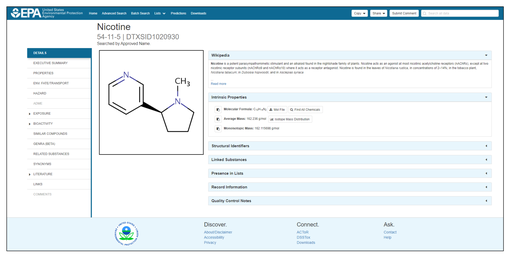
The CompTox Dashboard



Home page. The CompTox Dashboard (https://comptox.epa.gov) is a comprehensive chemistry resource containing chemistry data on more than 760,000 chemical substances.

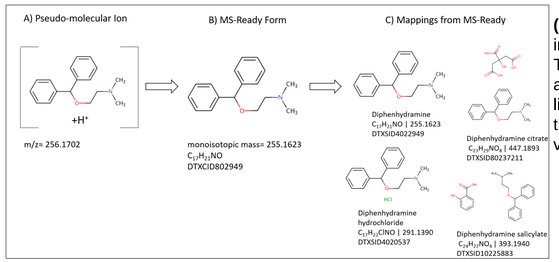


Batch Search Page. The Batch Search page enables users to query the underlying database using data collected from HRMS studies as molecular formulae or monoisotopic masses (i.e. 10s to 100s at a time). Metadata, structural information, presence in chemical lists, and many more pieces of data can be included in the download.

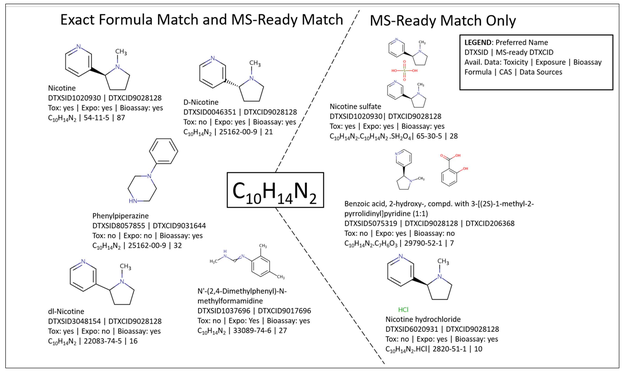


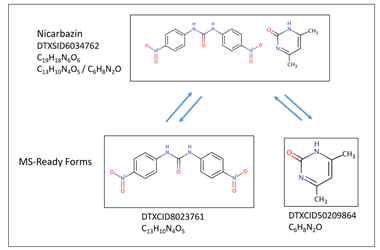
Single Chemical Page. A single chemical page includes chemical structures, experimental and predicted physicochemical and toxicity data, exposure data, and much more.

MS-Ready Structures for Database Searching



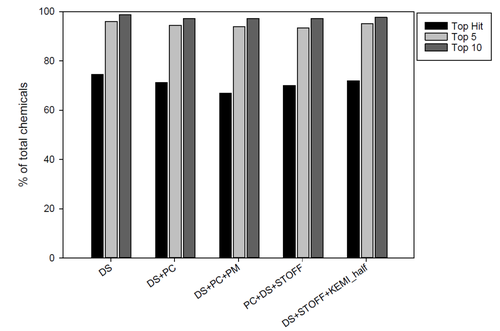
(Left and Below) To facilitate database searching, structures in DSSTox were processed into their "MS-Ready" form [4]. This processing workflow removes salts and stereochemistry and separates components of mixtures while retaining linkages to the original structure and substance. This enables the form of a structure observed via HRMS to be related to all variants of a structure in the database.





(Left) The results of an MS-Ready query include all those substances where one component of a substance matches the input query terms [4]. In this example, $C_{10}H_{14}N_2$ returns single component chemicals, mixtures, and salts. Returning all linked DTXSIDs enables access to varied metadata.

Chemical metadata for ranking

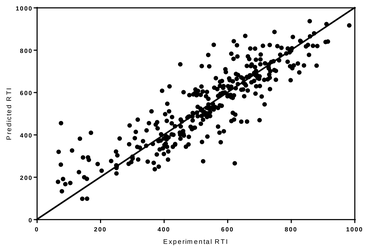


Percentage of chemicals identified using data source ranking combining multiple metadata sources. Total n=1748 unique chemicals from the ENTACT trial and CASMI 2016 contest (training and test sets). DS=DSSTox Data Sources, PC= PubChem Data Sources, PM=PubMed Reference Counts, STOFF= Presence in STOFF-IDENT, KEMI_half= Swedish Chemicals Agency (KEMI) Market List, weighted by 0.5.

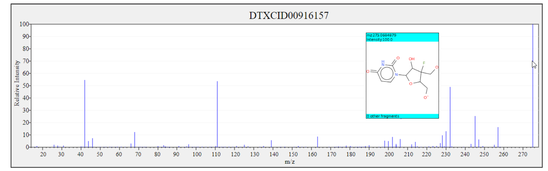
Retention Time Prediction

	logP	ChromGenius	OPERA-RT
Combined Test and Training Set (n=97)			
R^2	0.66	0.83	0.86
RMSE (min)	5.50	3.93	3.60
Absolute Mean Error (min)	4.65	3.03	2.93

(Top) Results from [2] indicated that an in-house QSRR model termed OPERA-RT performed on par with the commercial software ACD/ChromGenius.
(Bottom) Evaluation of retention time index (RTI) modeling developed by Aalizadeh *et al* (submitted manuscript). Experimental vs. predicted RTI value from ENTACT mixtures.



MS/MS Data in NTA/SSA



Using CFM-ID command line tools [5], MS/MS spectra were predicted for >700,000 structures in ESI+, ESI-, and EI mode (manuscript in prep). The ESI+/- data were used to analyze the CASMI 2016 datasets. Data below are structure identification ranks for the 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%

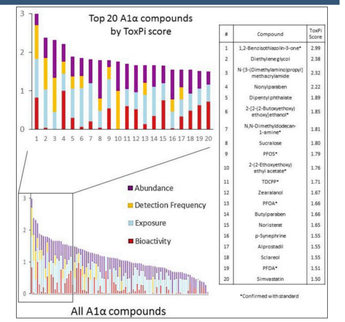
Applications

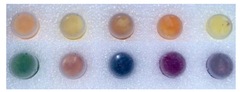
Suspect screening of drinking water using point of use filters [6].

- Dashboard tools used for identification (data source ranking and batch searching)
- Dashboard tools used for prioritization of identified compounds (exposure, bioactivity)

EPA's Non-targeted Analysis Collaborative Trial (ENTACT):

- MS-Ready structures underpinning analyses and provided to participants, improving accessibility to data
- In-house analysis utilizing data source ranking, metadata in Dashboard





References

1. Williams, AJ, *et al.* 2017. The CompTox Chemistry Dashboard: a community data resource for environmental chemistry. *J Cheminf*. https://doi.org/10.1186/s13321-017-0247-6
2. McEachran, AD, *et al.* 2017. Identifying known unknowns using the US EPA's CompTox Chemistry Dashboard. *Anal. Bioanal. Chem*. 409(7): 1729-1735. doi:10.1007/s00216-016-0139-z
3. McEachran, AD, *et al.* 2017. A comparison of three chromatographic retention time prediction models. *Talanta*. https://doi.org/10.1016/j.talanta.2018.01.022
4. McEachran, AD, *et al.* 2018. "MS-Ready" structures for non-targeted high-resolution mass spectrometry screening studies. *J Cheminf*. Accepted for publication.
5. Allen, *et al.* 2015. Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification. *Metabolomics*. https://doi.org/10.1007/s11306-014-0676-4
6. Newton, SR, *et al.* 2018. Suspect screening and non-targeted analysis of drinking water using point-of-use filters. *Environ Poll*. https://doi.org/10.1016/j.envpol.2017.11.033

Acknowledgements

The authors would like to thank Emma Schymanski and Christoph Ruttkies for their valuable contributions on the MS-Ready Structures, associated publication, and integration to MetFrag. The authors would also like to thank Reza Aalizadeh and Nikolaos Thomaidis for use of the UOA-RTI software for retention time index prediction (manuscript submitted for publication). The authors would like to acknowledge the Dashboard development team for their efforts in surfacing all functionality. This work was supported in part by an appointment to the ORISE Research Participation Program at the Office of Research and Development, U.S. EPA, through an interagency agreement between the US EPA and DOE.