US-EPA CompTox Chemicals Dashboard as a web-based data resource to help identify contaminants in water

Antony Williams, Jon Sobus and Elin Ulrich

U.S. Environmental Protection Agency, Office of Research and Development, Center for Computational Toxicology and Exposure, Research Triangle Park, NC

ORCID: 0000-0002-2668-4821

Antony Williams I williams.antony@epa.gov I 919-541-1033

SETAC 2019
Toronto, CAN
November 3-7, 2019

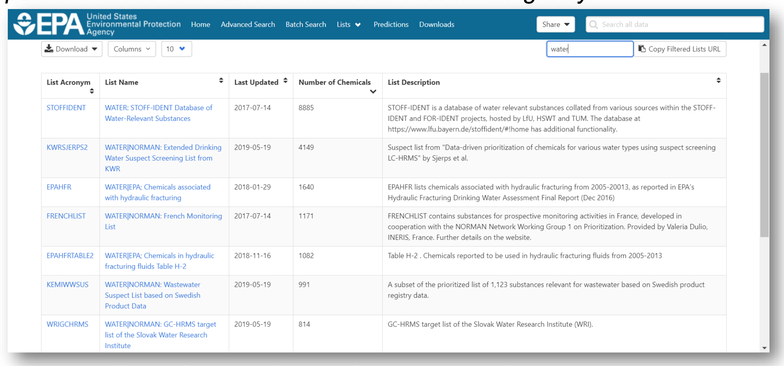
Problem Definition and Goals

Problem: There are few sources of curated data, and integrated workflows, available online to support structure identification using mass spectrometry approaches.

Goals: Deliver online access to hundreds of thousands of chemicals of interest to environmental science and computational toxicology. Provide lists of suspect screening chemicals that have been, or could be detected in water via a simple to use web-based interface. Deliver application to support diverse types of data including experimental and predicted physicochemical properties, *in vivo* hazard data and *in vitro* toxicity and toxicokinetic data. Make the data available as downloadable data for reuse and repurposing in other databases.

Abstract

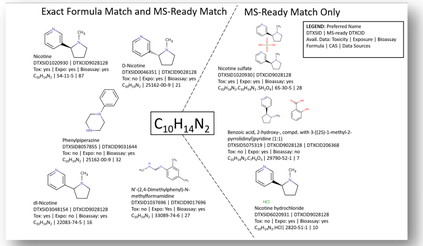
Non-targeted, targeted and suspect screening, as well as “Known Unknowns” and “Unknown Unknowns” are now common terms in the field of water analysis. While data processing can be highly automated, the identification of chemicals from extracted masses, formulae or fragmentation utilizes reference spectral libraries or identification and ranking of tentative candidate lists from large structure libraries. The US EPA CompTox Chemicals Dashboard (https://comptox.epa.gov/dashboard) provides access to data for ~875,000 substances, searchable by mass and formula and then ranked using associated meta-data. Cheminformatics approaches are also utilized to provide mapped relationships between individual substances and their “MS-Ready” (desalted, non-stereospecific) forms. *This abstract does not necessarily represent the views or policies of the U.S. Environmental Protection Agency.*



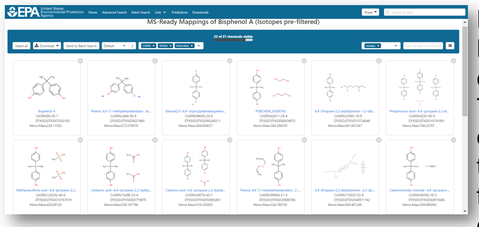
A set of lists of chemicals identified in water included in the data sets

MS-Ready Structures for Database Searching

To facilitate searching, structures are processed into “MS-Ready” forms [4]. This removes salts and stereochemistry and separates mixture components while retaining linkages to the original structures. This enables the form of a structure observed via MS to be related to all variants of a structure.



MS-Ready mappings based on formula

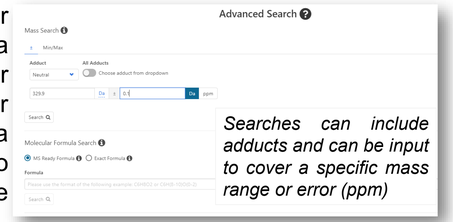


MS-Ready mappings for BPA, includes all salts and multicomponent chemicals

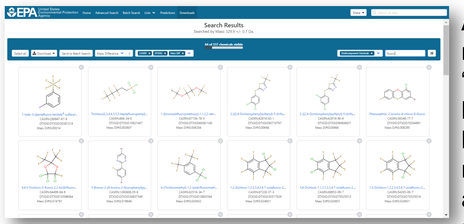
From the Chemical Details page all MS-Ready forms for a chemical can be displayed with a single click. This list of chemicals can be downloaded into an Excel file together with metadata such as formula, mass, predicted property data and experimental toxicity data.

Advanced Searching for Chemical Identification Using MS Data

Advanced searching includes mass or formula. The user can select from a set of potential adducts and for formula searching can perform either MS-ready formula or Exact formula searches. It is also possible to generate matching formulae in the database from mass.



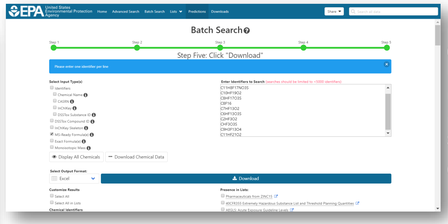
Advanced Search – by Mass or Formula



Search for mass 329.9+/-0.1 Da. Filter: fluoro

A search on mass 329.9+/-0.1 Da returns 337 hits. Filtering based on “fluoro” reduces the hit list to 67 hits. Candidate hits can be ranked based on metadata [2] such as number of associated data sources, associated PubMed articles and presence in commercial products

Batch Searching of Thousands of Masses and Formulae



Batch searching based on list of masses or formulae provides a hit list of chemicals based on an MS-Ready search. The resulting hit list can be exported, with metadata, into an Excel file.

An Excel file, see below, includes the mappings between chemicals returned as hits based on a formula search and all related mappings. A neutral chemical will map to all related chemicals based on MS-Ready mappings to include salts and multicomponent chemicals

	A	B	C	D	E	F	G
	INPUT	FOUND_BY	DTXCID_INDIVIDUAL	FORMULA_INDIVI	DTXSID	PREFERRED_NAME	CASRN
	C8F18O2S	MS Ready Formula	DTXCID107140	C8F18O2S	DTXSID5027140	Perfluorooctanesulfonyl fluoride	307-35-7
	C11H8F17NO3S	MS Ready Formula	DTXCID607831	C11H8F17NO3S	DTXSID7027831	N-Methyl-N-(2-hydroxyethyl)perfluorooctanesulfonamide	24448-09-7
	C11H8F17NO3S	MS Ready Formula	DTXCID607831	C11H8F17NO3S	DTXSID40893635	N-Alkyl perfluoroalkyl sulfonamidoethanols	NOCAS_893635
	C10HF19O2	MS Ready Formula	DTXCID1011860	C10HF19O2	DTXSID3031860	Perfluorodecanoic acid	335-76-2
	C10HF19O2	MS Ready Formula	DTXCID1011860	C10HF19O2	DTXSID60880027	Ammonium perfluorodecanoate	3108-42-7
	C10HF19O2	MS Ready Formula	DTXCID40896727	C10HF19O2	DTXSID40379808	Perfluoro-3,7-dimethyloctanoic acid	172155-07-6
	C10HF19O2	MS Ready Formula	DTXCID1011860	C10HF19O2	DTXSID20880028	Sodium perfluorodecanoate	3830-45-3
	C10HF19O2	MS Ready Formula	DTXCID1011860	C10HF19O2	DTXSID40892481	Perfluorodecanoate	73829-36-4
	C10HF19O2	MS Ready Formula	DTXCID1011860	C10HF19O2	DTXSID30892718	Nonadecafluorodecan(~2~H)oic acid	NOCAS_892718
	C10HF19O2	MS Ready Formula	DTXCID1011860	C10HF19O2	DTXSID00893893	Perfluoroalkyl (linear) carboxylic acids	NOCAS_893893
	C10HF19O2	MS Ready Formula	DTXCID1011860	C10HF19O2	DTXSID80892980	Perfluoroalkyl carboxylates	NOCAS_892980
	C10HF19O2	MS Ready Formula	DTXCID1011860	C10HF19O2	DTXSID20894100	nonadecafluoro(1,2-13C2)decanoic acid	NOCAS_894100
	C8HF17O3S	MS Ready Formula	DTXCID1011864	C8HF17O3S	DTXSID3031864	Perfluorooctanesulfonic acid	1763-23-1
	C8HF17O3S	MS Ready Formula	DTXCID1011864	C8HF17O3S	DTXSID8037706	Potassium perfluorooctanesulfonate	2795-39-3
	C8HF17O3S	MS Ready Formula	DTXCID1011864	C8HF17O3S	DTXSID2032421	Lithium perfluorooctanesulfonate	29457-72-5

Future Work

- The prediction of mass spectral fragmentation data (LC-MS positive and negative ion mode (10/20/40eV) and GS-MS data allowing for searching of experimental vs predicted data is in testing [4]
- Searching based on structure, substructure and similarity, including filtering based on mass. formula and presence/absence of specific elements is already available as an internal prototype.

References

1. Williams *et al.* The CompTox Chemistry Dashboard: a community data resource for environmental chemistry, J. Cheminformatics, 9, Article number: **61** (2017), DOI: 10.1186/s13321-017-0247-6
2. McEachran, AD, *et al*. Identifying known unknowns using the US EPA's CompTox Chemistry Dashboard. Anal. Bioanal. Chem. **409**(7): 1729-1735 (2017). DOI:10.1007/s00216-016-0139-z
3. McEachran, AD, *et al*. “MS-Ready” structures for non-targeted high-resolution mass spectrometry screening studies. J Cheminformatics, **10**, 45 (2018). DOI: 10.1186/s13321-018-0299-2
4. McEachran, AD, *et al*. Linking in silico MS/MS spectra with chemistry data to improve identification of unknowns, Scientific Data volume 6, **141** (2019), DOI:10.1038/s41597-019-0145-z

Acknowledgements

The authors thank the chemical curation team for their rigorous work and the software development team for the development of the dashboard.