The US-EPA CompTox Chemicals Dashboard – an online data integration hub supporting non-targeted analysis

Antony J. Williams* and Charles Lowe

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

ORCID: 0000-0002-2668-4821
Antony Williams I williams.antony@epa.gov I 919-541-1033

SETAC NTA Meeting
Durham, NC
May 22-25, 2022

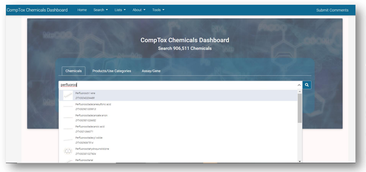
Problem Definition and Goals

Problem: Structure identification workflows in non-targeted analyses (NTA) have historically identified less than 10% of observed chemical features in environmental samples. Improvements in workflows require the incorporation of high quality data from a variety of resources to confidently identify structures. **Goals:** To develop NTA identification tools and provide functionality within the US EPA's CompTox Chemicals Dashboard.

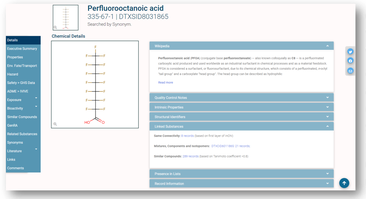
Abstract

The CompTox Chemicals Dashboard is a publicly accessible database provided by the Center for Computational Toxicology and Exposure at the US-EPA. The Dashboard provides access to a database containing ~900,000 chemicals and integrates a number of our public-facing projects (e.g., ToxCast and ExpoCast). The available data provide a valuable foundation to mass-spectrometry based structure identification, especially of known-unknowns (1,2) and the Dashboard is used to assist in identifying chemicals present in environmental media including house dust and water. This poster will review the data and functionality available in the Dashboard to support structure identification using mass spectrometry data. Specifically, we have developed approaches to rank-order hit lists of chemicals based on mass and formula-based searching using candidate metadata ranking, have developed MS-Ready structure mappings as an underpinning of our approach, and used targeted lists to assist in the ranking process. *This abstract does not necessarily represent the views or policies of the U.S. Environmental Protection Agency.*

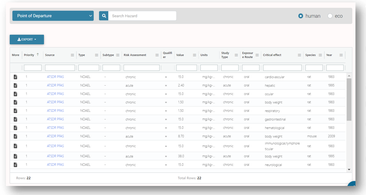
CompTox Chemicals Dashboard



Dashboard Homepage.
The home page to the Dashboard is at http://comptox.epa.gov/dashboard. The landing page provides a simple text entry box with a type-ahead search for systematic and trivial names, CASRNs and InChI identifiers. Product/Category and Assay/Gene searches are available.



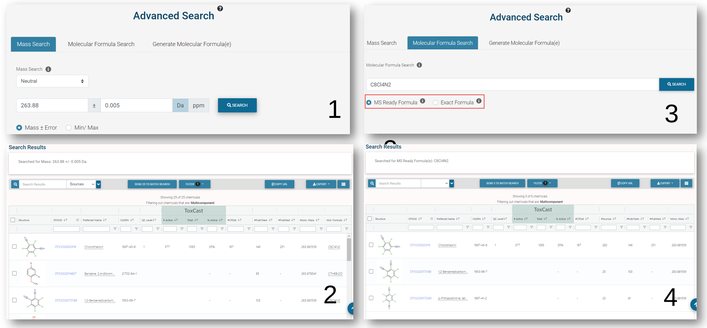
Chemical Record Page: PFOA
Where possible, the lede for the Wikipedia articles is displayed, as well as a link out. Structure file formats are available for download to the desktop (SMILES and molfile) and an executive summary report regarding chemical toxicity is provided. Structures can be downloaded as Molfiles and searches using InChIKeys link out to the web.



Toxicity Data Values Panel
The Hazard tab provides access to data assembled from a series of public resources including EPA data (i.e., IRIS and PPRTV reports, ToxRef DB), public domain databases (e.g., ECHA) and associated with tens of thousands of peer-reviewed literature articles. Data can be downloaded as TSV and Excel files.

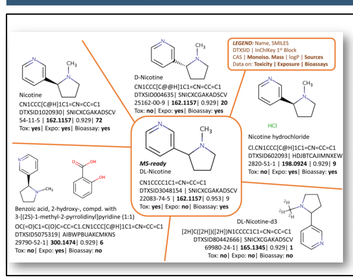
Toxcast Bioactivity Summary Page.
ToxCast/Tox21 Bioactivity data have been measured over the past decade and are displayed under the Bioactivity Tab. Data can be downloaded as CSV and Excel files. New data are generated each year including new chemicals and often times new assays. New data to be released in the future include high-throughput phenotypic profiling and transcriptomics data.

Advanced Searching for Identification of Unknowns

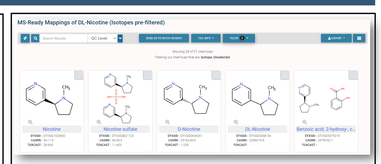


Advanced Searches allow both mass and formula-bases searches. In the results sorting candidate structures by metadata, including the number of associated data sources within the database (1), is a valuable assist to identification. Tiles 1 & 2 show searching by monoisotopic mass (+/- error) observed via NTA and rank-ordering the results to bring the most likely candidate structures to the top of the search results. Tiles 3 & 4 demonstrate the process when a user has already generated a molecular formula. Users can take advantage of pre-generated MS-Ready forms of the chemicals in their searches of the Dashboard. This includes in searches of both single chemicals by mass and formula or in batch searches as described below.

MS-Ready Structural Forms Data

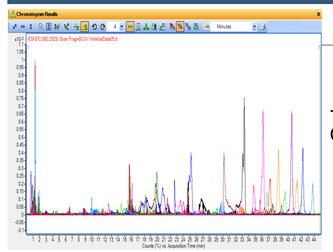


MS-ready structures (3,4) relate compounds via their chemical structures to their salts, their presence in other multi-component chemical substances, their stereoisomeric forms and isotopomers. As an example "MS-ready" structures related to Nicotine (Schymanski and Williams), together with available selected data from the Dashboard, are displayed. MS will detect, e.g., [M + H]+ 163.1235 (structures top left, top middle, center), not salts or mixtures. Various toxicity, exposure, bioactivity, and reference data exist for all forms (bold values). MS-Ready structures are shown as "Linked Substances"

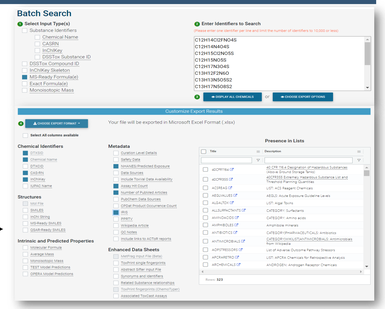


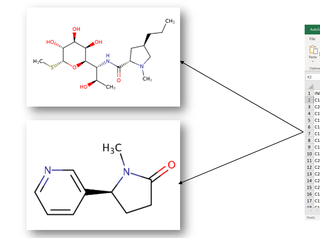
in the Dashboard and can be viewed in both a tile format (shown above) and in a tabular format containing related metadata for sorting and candidate ranking. The data can be downloaded as Excel or SDF files or pasted over to the Batch search to source additional data.

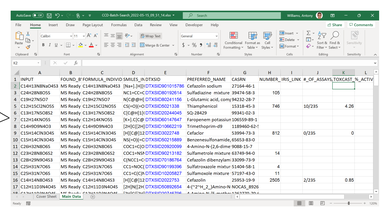
Batch Searching of Unknowns



-Database Matching for formula(e) C18H34N2O6S, C10H12N2O, etc.



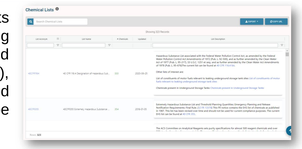


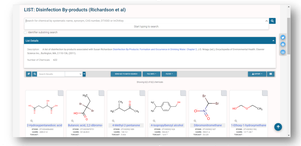


Excel export of batch search of molecular formulae in the Dashboard. Data included in the download consists of CASRN, formula, SMILES, InChIKey, mass, bioactivity, exposure potential, etc. It is possible to select multiple other forms of meta data to include in the download via the Batch Search screen (shown above) and include in the file.

Dashboard Chemical Lists for Identification

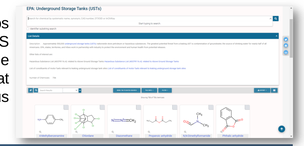
The Dashboard presently hosts over 320 chemical lists that are segregated into different groups including regulatory lists, lists directly related to non-targeted analysis (for example, chemicals of emerging concern), and recent efforts have produced a number of lists related to PFAS chemicals, almost 40 in total. These lists can be used as meta data flags in the batch search.





Lists can be specifically oriented to research projects or to focused datasets. For example, to chemicals of emerging concern (CEC) or, as in the case of the example shown, to disinfectant by-products. The number of lists continues to grow as well as being versioned and maintained with each release.

Lists can be cross-referenced in terms of relationships including multiple versions of a list (e.g., the lists of PFAS chemicals with structures), the multiple versions of the TSCA (Toxic Substances Control Act inventory) updated at least annually, and, in the example shown, the various types of lists associated with storage tanks.

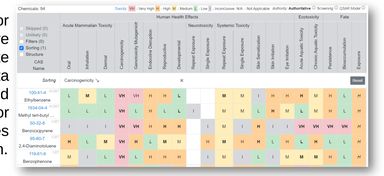


Future Work

There are many additional efforts presently underway to develop computational support for NTA.

- Software tools that handle experimental instrument data directly to extract signals and search against *in silico* predicted spectra and perform candidate ranking using DSSTox data
- Expand the existing collection of *in silico* spectra (5,6) to include over 1 million substances
- Assemble and homogenize a collection of public domain spectral data to allow for searching of experimental spectra against experimental spectra
- Include an integration to allow candidates to be profiled according to chemical hazard (using proof-of-concept cheminformatics modules).

The Hazard Comparison module allows for users to generate a profile of one or more chemicals including various types of acute and chronic toxicities using sources of data including authoritative, screening and predicted toxicity. This approach allows for the ranking of potential candidates identified by NTA for follow-up identification.



References

1) Little JL, Williams AJ, Pshenichnov A, Tkachenko A. (2012). Identification of "known unknowns" utilizing accurate mass data and ChemSpider. *J Amer Soc Mass Spectrom*. 23(1): 179.
2) McEachran AD, Sobus JR, Williams AJ. (2017). Identifying known unknowns using the US EPA's CompTox Chemistry Dashboard. *Anal. Bioanal. Chem*. 409(7): 1729.
3) McEachran et al. (2018) "MS-Ready" structures for non-targeted high-resolution mass spectrometry screening studies. *J Cheminform* 10:45.
4) Schymanski and Williams (2017) Open Science for Identifying "Known Unknown" Chemicals. *Environ. Sci. Technol.* 51, 5357
5) McEachran et al. (2020) Revisiting Five Years of CASMI Contests with EPA Identification Tools. *Metabolites*. 10(6): 260.
6) McEachran *et al.* (2019) Linking in silico MS/MS spectra with chemistry data to improve identification of unknowns, Scientific Data 6(141)

Acknowledgements

The authors would like to acknowledge the curation team for their efforts to curate chemical data into the underlying DSSTox database, the development team building the CompTox Chemicals Dashboard, and our colleagues in the mass spectrometry team who have used the application and given significant feedback to guide its development to support NTA research in the center. We specifically appreciate the support of Jon Sobus and Elin Ulrich.