The EPA CompTox Chemistry Dashboard: A Web-Based Data Integration Hub for Toxicology Data

[1]Antony Williams*, [1]Chris Grulke, [2]Kamel Mansouri, [1]Jennifer Smith, [1]Jeremy Fitzpatrick, [1]Grace Patlewicz, [1]Imran Shah, [1]Ann Richard, and [1]Jeff Edwards

[1]U.S. Environmental Protection Agency, Office of Research and Development, National Center for Computational Toxicology (NCCT), Research Triangle Park, NC

[2]Oak Ridge Institute for Science and Education (ORISE) Participant, Research Triangle Park, NC

1387/P125

Society of Toxicology Annual Meeting

Baltimore, MD

March 12-16, 2017

ORCID: 0000-0002-2668-4821

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

Problem Definition and Goals

Problem: Data of value to toxicologists is hard to locate as it is distributed across many databases and resources.
Goals: To provide a web-based ***integration hub*** (at http://comptox.epa.gov) to improve accessibility to data, algorithms, searches to support computational toxicology, and Open Data for redistribution and reuse.

Abstract

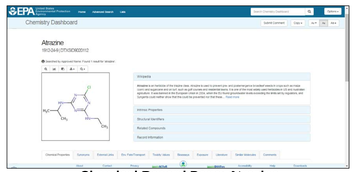
The U.S. Environmental Protection Agency Computational Toxicology Program integrates advances in biology, chemistry, and computer science to help prioritize chemicals for further research based on potential human health risks. This work involves computational and data driven approaches that integrate chemistry, exposure and biological data. As an outcome of these efforts the National Center for Computational Toxicology (NCCT) has measured, assembled and delivered an enormous quantity and diversity of data for the environmental sciences including high-throughput *in vitro* screening data, *in vivo* and functional use data, exposure models and chemical databases with associated properties. These data were aggregated from both agency resources and public databases. A series of software applications and databases have been produced over the past decade but our new software architecture assembles the resources into a *single* platform. This new web application, the CompTox Chemistry Dashboard (at http://comptox.epa.gov) provides access to data associated with ~740,000 chemical substances. These data include experimental and predicted physicochemical property data, bioassay screening data associated with the ToxCast program, product and functional use information and a myriad of related data of value to environmental scientists. The dashboard provides chemical-based searching based on chemical names, synonyms and CAS Registry Numbers. Flexible search capabilities allow for chemical identification based on non-targeted analysis studies using mass spectrometry. Chemical identification using both mass and formula-based searching utilizes rank-ordering of results via functional use statistics, thereby providing a solution to help prioritize chemicals for further review when detected in environmental media.

~740,000 Chemicals and ~10 million properties

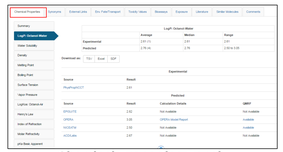
Dashboard Entry Page

The entry to the dashboard is a simple text entry box allowing a type-ahead search for systematic, trade and trivial names, CAS Registry Numbers and InChI identifiers. Searches can be filtered, for example, to return single component chemicals (not mixtures). An advanced search allows for searching based on molecular mass or molecular formula, specifically to support non-targeted analysis or searching for “known-unknowns”

For those records with associated chemical structure representations various inherent properties (molecular formula, mass, systematic name) and predicted physicochemical properties (logP, water solubility etc.) are provided. Where possible, links are provided to related Wikipedia articles. An associated molfile is available for download to the desktop, and a summary report containing record data can be generated as a PDF file.

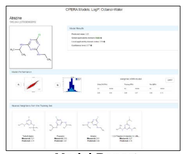


Chemical Record Page: Atrazine



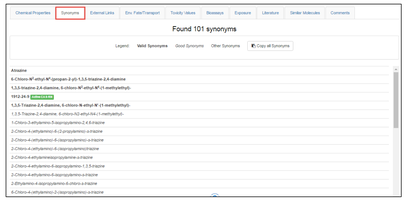
Chemical Properties Panel

Chemical property predictions are based on the **Open** structure activity **R**elationships **A**pplication (OPERA) models developed from curated datasets. QSAR Model Reporting Formats (QMRFs) are available for all OPERA models which include descriptions of global and local applicability domains and nearest neighbors in the training set.

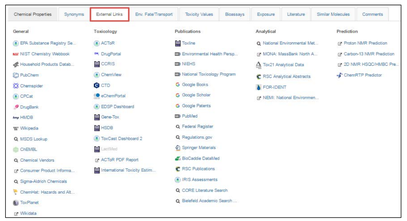


Model Report

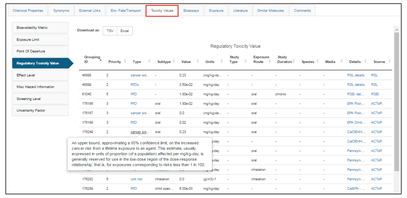
The Chemistry Dashboard – Data Tabs and Sources



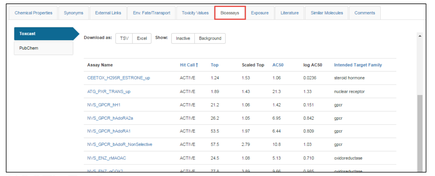
100s of thousands of synonyms and registry numbers underpin the *type-ahead* for efficient chemical searching



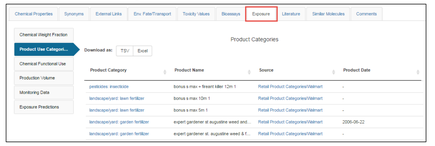
Dozens of external links to connect agency and public resources. The Dashboard acts as a data hub.



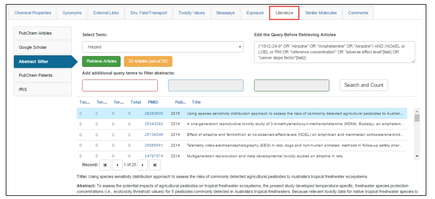
Toxicity Value data are sourced from agency and public resources and integrated into multiple tables.



NCCT’s ToxCast data is available for review and download. Rank order based on AC50 Values, include/exclude both inactive and background result data.



Exposure data includes access to NHANES data, ExpoCast predictions and CPDat data regarding consumer products and functional use of chemicals.

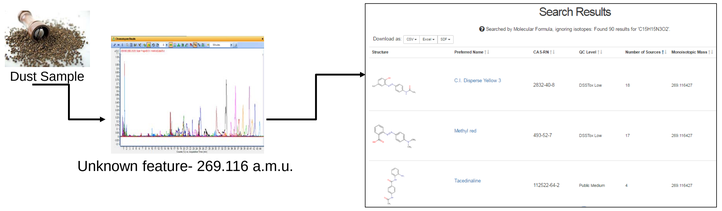


The Literature Tab integrates EPA documents, e.g. IRIS reports, to the PubChem literature and patents widgets, to Google Scholar and allows “Pubmed Abstract Sifting”.

Chemical Structures, Families & UVCB Substances

~720,000 of the chemical substances on the dashboard have defined chemical structures. Many of the chemicals of interest to EPA are classed as “UVCBs” - Unknown or Variable Composition, Complex Reaction Products and Biological Materials. The data model includes hosting chemical families, for example, the polychlorinated biphenyls record includes all 209 members of the family. For polymers, a chemical record can include associated monomers, and mixtures can include all components (as an example see the various possible components of the xylenes family). The record information accordion contains information regarding our confidence in the chemical structure-identifier data quality for a record (e.g. Level 2, Expert curated, unique chemical identifiers confirmed using multiple public sources).

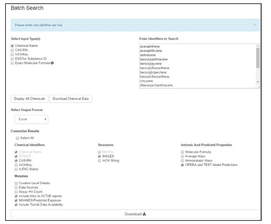
Non-Targeted Mass Spectrometry Analysis Support



- Monoisotopic masses or chemical formulae associated with spectral features are searched within the Dashboard and the results are rank-ordered.
- Rank-ordering can be performed using various criteria including number of associated data sources.

Batch Searching

- Searches of thousands of chemical names, CAS Numbers or InChIKeys can be performed to source:
  - Chemical structure SDF files
  - Inherent structure data (mass, formula) and predicted property data
  - Availability of *in vitro* bioassay, exposure and toxicity data
  - Metadata regarding data curation levels
- Data can be downloaded as tab-separated files, Excel files or SDF files for use in cheminformatics software packages.



Future Work Will Include

- Structure, substructure and similarity searching for chemicals.
- Interactive online prediction tools for physchem properties and toxicity will be made available.
- Enhanced visualization and searching of *in vitro* bioassay data (bioassay analysis, Hill curves etc.).
- Support for Generalized Read-Across (GenRA) approaches.
- Delivery of web-services and application programming interface based access to data.

References

1. Mansouri K, Grulke CM, Richard AM, Judson RS, Williams AJ. 2016. An automated curation procedure for addressing chemical errors and inconsistencies in public datasets used in QSAR modelling. SAR QSAR Environ. Res. 27(11): 939-965. doi:10.1080/1062936X.2016.1253611
2. McEachran AD, Sobus JR, Williams AJ. 2017. Identifying known unknowns using the US EPA's CompTox Chemistry Dashboard. Anal. Bioanal. Chem. 409(7): 1729-1735. doi:10.1007/s00216-016-0139-z

Acknowledgements

We acknowledge our colleagues for valuable support and input: Jon Sobus and Andrew McEachran (Non-targeted Analysis), John Wambaugh (ExpoCast), Kathie Dionisio and Katherine Phillips (CPDat), Richard Judson (Toxicity Value data)