Cheminformatics approaches supporting UVCB chemical representations in the US-EPA CompTox Chemicals Dashboard

Antony Williams and Charles Lowe

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

HESI
UVCB Workshop
September 18-19, 2023

United States Environmental Protection Agency

Description

The Center for Computational Toxicology and Exposure (CCTE) at the US Environmental Protection Agency (EPA) has assembled data for ~1.2 million chemical substances and made these data available to the community via the CompTox Chemicals Dashboard. This includes high-throughput *in vitro* screening data, legacy *in vivo* animal data, and functional use data, exposure models and chemical databases with associated properties. The data include tens of thousands of substances deemed to be from the class of chemicals known as "**UVCB**s": Unknown or Variable Composition, Complex reaction products and Biological materials. We curate chemicals from various UVCB classes and utilize Markush representations and enumeration approaches to provide mappings between UVCBs and their related structural representations. Structural representation of UVCB classes can help in profiling these complex chemicals, for example in physicochemical parameter ranges, and these data can potentially be of value in supporting hazard and risk assessment.

DSSTox and the CompTox Chemicals Dashboard

The DSSTox database (1) is the chemistry database underlying all software applications from the US-EPA's Center for Computational Toxicology and Exposure, including the publicly accessible CompTox Chemicals Dashboard (https://comptox.epa.gov/dashboard). The database has a complex data model allowing for mappings between substances including "successor substances", for example monomer to polymer mappings. The DSSTox database expands using the expert input and guidance of a team of highly skilled chemical curators. Chemical substance registrations include, as appropriate, the addition of chemical structures mapped to relevant identifiers such as CAS Registry Numbers, chemical names and synonyms, chemical structure representations, and other numeric identifiers such as EINECS and EC numbers. Chemical curation includes harvesting of data from public domain databases including PubChem, ECHA's substance database and other public domain resources. Thorough processes are applied to curate and expand chemical substance registrations to release in the Dashboard.

Acknowledgements

The authors thank the data curation team for their rigorous work in annotating and identifying information in the records. Chemical data extraction, curation and annotation is an essential part of this work.

Cedar Oil (CAS RN: 8000-27-9)

UVCB chemicals can be mapped to "related substances". These related substances can be components of a UVCB complex mixture, homologous series members, transformation products or metabolites. The association of certain related substances was based on searching for associated documentation online including NIH chemical selection working group documents. Inclusion as a related substance is a subjective decision.

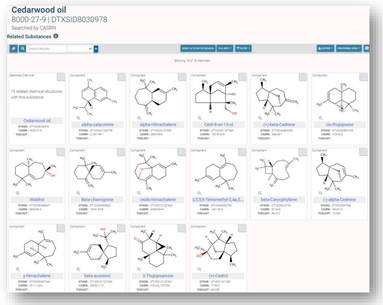

Fig. 1: Related substances mapped to Cedarwood Oil

Coal Tar Pitch (CAS RN: 65996-93-2)

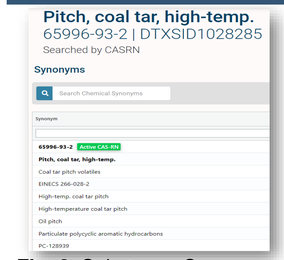

Fig. 2: Substance Synonyms

Coal Tar pitch with the CAS RN 65996-93-2 is an example of the variability of the different synonyms that can be associated with a single substance in DSSTox. For this substance the list includes a Pesticide Code (PC-128939), an EINECS number and multiple names. A search of CAS' SciFinder for the CAS RN shows that the preferred name "Pitch, coal tar, high-temp." is consistent.

The Dashboard provides access to different data types including data from the "ToxVal" database. The data are segregated into general Hazard data and including Cancer, Genotoxicity and Skin/Eye hazard data. Other available data include exposure data with chemical weight fraction data for >50 products containing coal tar pitch.

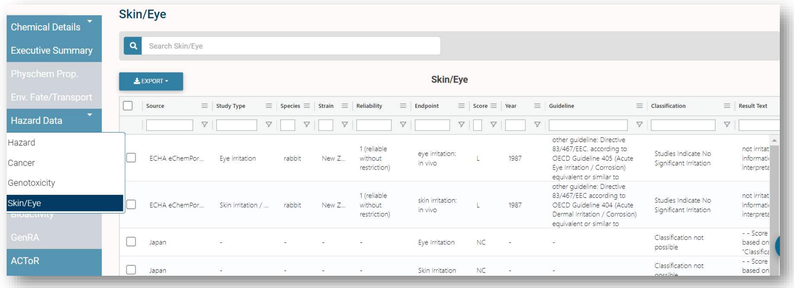

Fig. 3: Skin/eye data under the Hazard Tab Dashboard shows data from ECHA & Japan.

Linear Alkylbenzenes (CAS RN: 67774-74-7) and their sulfonates

Some chemicals can be represented as "Markush structures", chemicals which are generic representations which, when enumerated, represent a class of chemicals. The Markush parent maps onto its "children" as related members of the family.

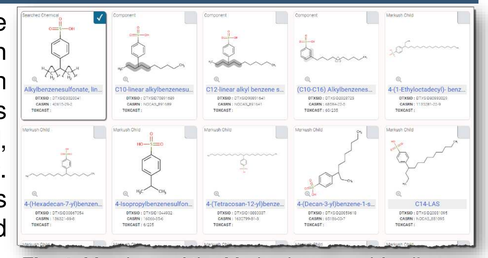

Fig. 4: Members of the Markush mapped family

For those substances that can be represented in Markush structure forms this approach offers a beneficial way to classify chemicals under overarching structural forms. For example, the PFASMARKUSH list (URL: https://comptox.epa.gov/dashboard/chemical-lists/PFASMARKUSH) at present contains 326 Markush structures in total, thereby assembling linear PFAS-sulfonic acids, such as linear-PFOS, into a set of related structures. As should be obvious from Fig. 5 there are many such classes.

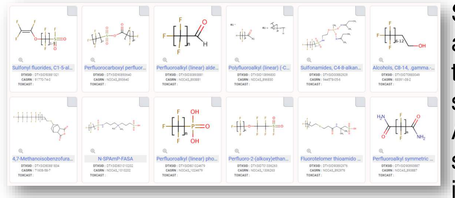

Fig. 5: The PFASMARKUSH list has 326 members

Such Markush representations are presently being applied to the development of a set of surfactants from the Quaternary Ammonium Compounds (QACs) set. This set of chemicals is of increasing interest at present to a number of regulatory agencies.

The DSSTox data model, and the manner in which the data can be made accessible via the CompTox Chemicals Dashboard, is an enabling technology to handle UVCB data management from the US EPA.

References

1. EPA's DSSTox database: History of development of a curated chemistry resource supporting computational toxicology research: https://doi.org/10.1016/j.comtox.2019.100096
2. The CompTox Chemistry Dashboard: a community data resource for environmental chemistry: https://doi.org/10.1186/s13321-017-0247-6
3. Cedarwood Oil Chemical Selection Working Group documentation: https://ntp.niehs.nih.gov/sites/default/files/ntp/htdocs/chem_background/exsumpdf/cedarwood_oil_508.pdf
4. Linear alkylbenzene sulfonate related substances including Markush parent: https://comptox.epa.gov/dashboard/chemical/related-substances/DTXSID3020041

Disclaimers

The views expressed in this poster are those of the authors and do not necessarily reflect the views or policies of the US EPA.