Integrating Structure-based Chemical Taxonomies to Focus Queries of NAMs Data

V. Hull[1], T. LaPratt[1], E. Reinke[1], N. Kleinstreuer[2]

[1]Inotiv, United States; [2]NIH/NIEHS/DTT/NICEATM, United States

Characterizing and navigating chemical space is a complex challenge, especially when dealing with large and diverse data sets that encompass a variety of chemical structures. This is particularly true for data generated using new approach methodologies (NAMs), as in silico and in vitro approaches typically yield larger volumes of information than traditional animal methods. One way to focus chemical space exploration is through tools such as ClassyFire (https://doi.org/10.1186/s13321-016-0174-y), an automated, structure-based chemical taxonomy tool. Here, we describe efforts to incroporate ClassyFire chemical taxonomies into the Integrated Chemical Environment (ICE; https://ice.ntp.niehs.nih.gov/), an open-access web-based resource for toxicologically relevant data and computational tools. We will also present case studies that illustrate how ClassyFire chemical taxonomies can facilitate chemical space exploration within ICE.

We obtained the full list of chemicals in the U.S. Environmental Protection Agency's (EPA's) Distributed Structure-Searchable Toxicity (DSSTox) chemicals database from EPA's Comptox Chemicals Dashboard in August 2024. The 1,125,991 unique InChIKeys in DSSTox were run through the ClassyFire API, resulting in classifications for 1,125,328 structures. These classifications fell into 3,455 unique groupings, with 10 of the 11 levels of the ClassyFire hierarchy represented. "Organics" was the most abundant kingdom (the highest level of the taxonomy), "Benzenoids" was the most abundant superclass (the second level), and "Benzene and Substituted Derivatives" was the most abundant class (the third level). We then integrated these classifications into ICE via a downloadable file.

To demonstrate the utility of this new feature, we used classifications for chemicals tested in EPA's Tox21 program to focus queries of bioactivity and chemical use data in ICE. There were 8,583 Tox21 chemicals with ClassyFire classifications spanning eight levels of the hierarchy. "Benzenoids" and "Organoheterocyclic" were the most abundant superclasses, with 2,744 and 1,959 chemicals, respectively. A total of 3,215 chemicals were bioactive for at least one assay mapped to "Cellular Stress Response." The most abundant superclass for these active chemicals was "Benzenoids" (1,232 chemicals), and the most active class within Benzenoids was "Benzene and Substituted Derivatives" (918 chemicals). We also ran chemicals classified as "Benzenoids" and "Organoheterocyclic Compounds" through the ICE Chemical Characterization tool to examine potential links between chemical class and product use category. Use data were available for 360 benzenoids and 193 organoheterocyclic compounds. For both chemical groups, "Consumer Goods" was the most abundant use category, followed by "Construction/Maintenance" and "Cleaning."

By linking taxonomy to bioactivity and use, users can explore chemical structures and use cases that potentially contribute or relate to toxicological endpoints. This can help focus chemical selection, aid in selection of alternative chemicals, and identify chemical substructures that are correlated with chemical activity. This project was funded in whole or in part with federal funds from the NIEHS, NIH under Contract No. HHSN273201500010C. The views expressed above do not necessarily represent the official positions of any federal agency.