ToxStrategies

Identifying Structural Overlap Between EDSP Universe of Chemicals and ToxCast/Tox21 using Chemical Clustering and Uniform Manifold Approximation and Projection

B. Cook, M. Nelms, D. Harris, R.J. Bever, S.G. Lynn, D. Williams, S. Borghoff, S.W. Edwards, and K. Markey

Abstract:

Background and Purpose: The U.S. Environmental Protection Agency's Endocrine Disruptor Screening Program (EDSP) Universe of Chemicals (UoC) is a chemical list derived from the statutory scope of the EDSP-covered pesticidal active and inert chemicals as well substances of relevance to drinking water. Understanding the domain of applicability of EDSP test methods for the UoC is important for guiding data interpretation and identifying gaps where additional test method development, test validation, and/or testing may be required. Clustering chemicals based upon similarities in structure, physicochemical properties, and/or mode(s) of action has been used for decades to assist in predicting the properties and potential hazards of chemicals, especially when combined with toxicity and/or bioactivity data. Integrating chemical cluster data with knowledge of toxicity pathways (e.g., adverse outcome pathways) can further enhance hazard characterization by linking molecular-level perturbations to adverse effects. Additionally, methods such as read-across and quantitative structure-activity relationships ((Q)SARs) can use data from tested chemicals to predict the properties of untested ones to fill data gaps; thereby, helping guide and prioritize further testing. The goal of this study was to utilize chemical clustering to investigate the coverage of the EDSP UoC tested in the ToxCast/Tox21 program and identify areas of the EDSP UoC chemical space not covered by ToxCast/Tox21 chemicals. Fulfilling this goal will allow the identification of any gaps that may reduce the predictive capacity.

Methods: A total of 6,947 chemicals from the EDSP UoC with structure information in the EPA's CompTox Chemicals Dashboard were grouped into 826 clusters based upon their SMILES-based ToxPrint fingerprints. An additional 3,325 chemicals lacked defined structural information and were not used in clustering. Structure information was then collected for 8,630 of the ToxCast/Tox21 chemicals. The overlap in chemical space was evaluated in two ways: by mapping the ToxCast/Tox21 chemicals to the EDSP UoC clusters using structural similarity and using Uniform Manifold Approximation and Projection (UMAP) to perform dimensionality reduction and visualize the overlaps between the datasets.

Results: There is a large overlap between the chemicals in the EDSP UoC and those in ToxCast/Tox21, with approximately 3,650 chemicals (30.5%) present in both datasets. After binning the UMAP projection data into 1x1 grids we calculated a Jaccard index overlap of ~54%, further suggesting that there is a large overlap in the structural space between the chemicals in the two datasets. However, there are some areas of the structural space (14 of the 489 UMAP grids, covering 18 chemicals in EDSP UoC) where the EDSP UoC are not covered by comparable chemicals in ToxCast/Tox21. Similar observations were made when investigating the UMAP results using physicochemical properties, where 6 of 346 UMAP grids contain only EDSP UoC chemicals (42 chemicals) without any corresponding ToxCast/Tox21 chemicals.

Conclusions: This study illustrates that whilst there is a substantial overlap between chemicals in the EDSP

UoC and ToxCast/Tox21, some gaps do remain for some number of well-defined EDSP UoC chemicals. These gaps highlight areas where further testing is needed to enhance the predictive capacity and extendibility of the ToxCast/Tox21 data to the EDSP UoC. Together, these findings underscore the utility of using clustering approaches to investigate the domain of applicability for chemicals with bioactivity data to a set of chemicals of interest and illustrate how they may be used to assist with guiding additional testing. This abstract neither constitutes nor necessarily reflects U.S. EPA policy.