Using semi-automated curation workflows to collect, organize, and curate the data and models necessary to support the EPA CompTox chemical dashboard

Christopher M. Grulke[1], Indira Thillainadarajah[1], Risa Sayre[3], Kamel Mansouri[2], Antony J. Williams[1], Ann Richard[1]

[1] National Center for Computational Toxicology, ORD/U.S. EPA, RTP, NC, USA

[2] ORISE Fellow, NCCT, ORD/U.S. EPA, RTP, NC, USA

[3] ORAU, NCCT, ORD/U.S. EPA, RTP, NC, USA

With the increasing need to leverage data and models to perform cutting edge analyses within the environmental science community, collection and organization of that data into a readily accessible format for consumption is a pressing need. The EPA CompTox chemical dashboard is intended to provide a user-friendly interface to examine and export such data and models; however, such data needs first to be both curated and interconnected. The chemical dashboard is built on top of the long-running DSSTox database project which has been curating the chemicals and associated data of greatest interest to the environmental community for over a decade. In recent years, DSSTox has expanded to include chemicals with varying degrees of curation while maintaining strict documentation of the level of quality control (QC) that he been applied. This has enabled a marked increase of chemical space coverage through the use automated curation protocols, while prioritizing where manual review is applied to provide the highest confidence in the chemical-data associations. With the growth of computation modeling for estimating chemical parameters (e.g., LogP, Solubility, pKa), we have extended the scope of the DSSTox project to include the storage and management of several Quantitative Structure-Activity Relationship (QSAR) models, including, to the extent possible, providing formal documentation in accordance with regulatory guidelines. Additionally, we have recently embarked on a focused effort to collect pharmacokinetic parameters and models from literature. This presentation will clarify the areas of greatest concern when merging chemical data from various sources and the difficulties in dealing with complex relational data necessary to understand the relevance of the information flooding into the environmental science domain. *This abstract does not reflect U.S. EPA policy.*