Automated workflows for data curation and standardization of chemical structures for QSAR modeling

Kamel Mansouri[1,2,3], Andrew D. McEachran[1,2], Chris M. Grulke[1], Ann M. Richard[1], Richard S. Judson[1] and Antony J. Williams[1]

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

[2] Oak Ridge Institute for Science and Education, Oak Ridge, TN, USA

[3] ScitoVation LLC, Research Triangle Park, NC, USA

Large collections of chemical structures and associated experimental data are publicly available, and can be used to build robust QSAR models for applications in different fields. One common concern is the quality of both the chemical structure information and associated experimental data. Here we describe the development of automated KNIME workflows to both assist in the curation of data and to standardize the chemical structures according to a set of standard rules. The publicly available PHYSPROP physicochemical properties and environmental fate datasets were used as case studies to reveal commonly encountered errors and develop a set of rules to correct them. The workflow first assembles structure–identity pairs using up to four provided chemical identifiers, including chemical names, CASRNs, SMILES, and MolBlocks. Problems detected included errors and mismatches in chemical structure formats, identifiers and various structure validation issues, including hypervalency and stereochemistry descriptions. Subsequently, a structure standardization KNIME workflow was used to generate "QSAR-ready" forms prior to calculating molecular descriptors. This workflow performs a series of operations on the 2D structures including desalting, stripping stereochemistry, standardizing tautomers and nitro groups, correcting valence, neutralizing when possible and removing duplicates. A machine learning procedure was applied to evaluate the impact of this curation process. The models based on the curated data and standardized structures showed statistically improved predictive performance. These workflows were used to curate and standardize the full list of PHYSPROP datasets that were used to develop OPERA models available on the EPA's CompTox Chemistry Dashboard (https://comptox.epa.gov). They were also applied on thousands of other datasets that were used in international consortiums such as CERAPP and CoMPARA. The QSAR-ready workflow was

modified to generate “MS-ready structures” to support mass spectrometry non-targeted analysis. All workflows, data and models are open-source and freely available on GitHub (https://github.com/kmansouri) for further usage and integration by the scientific community. *This work does not reflect U.S. EPA policy.*