OPERA: A free and open source QSAR tool for predicting physicochemical properties and environmental fate endpoints.

Kamel Mansouri[1,2,3], Chris M. Grulke[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

Collecting the chemical structures and data for necessary QSAR modeling is facilitated by available public databases and open data. However, QSAR model performance is dependent on the quality of data and modeling methodology used. This study developed robust QSAR models for physicochemical properties and environmental fate endpoints that can be used for regulatory purposes. Publicly available data were collected from the PHYSPROP database among other sources. These data sets have undergone extensive curation using an in-house automated workflow to enhance the quality of the data. The chemical structures were standardized to "QSAR-ready form" prior to calculation of the molecular descriptors. The modeling procedure was based on the five OECD principles for QSAR models to produce reliable yet simple models. Genetic algorithms were used to select the most pertinent and mechanistically interpretable descriptors (from 2 to 15 with an average of 11 descriptors). The sizes of the modeled datasets

varied from 150 chemicals for biodegradability half-life to 14,050 chemicals for logP, with an average of 3222 chemicals across all endpoints. The optimal models were built on randomly selected training sets (75%) and validated using 5-fold cross-validation (CV) and test sets (25%). The CV Q^2 of the models varied from 0.72 to 0.95 with an average of 0.86 and an R^2 test from 0.71 to 0.96 with an average of 0.82. Modeling and performance details were described in QSAR model reporting format (QMRFs) and validated by the European Commission's Joint Research Center (JRC) for OECD compliance. All models are delivered as a free, open source/open data application called OPERA (OPEn structure-activity Relationship App) used to predict properties for ~750,000 chemicals. The predicted data are freely available on the EPA's CompTox Chemistry Dashboard (https://comptox.epa.gov). *This work does not reflect U.S. EPA policy*