ACS Spring 2016 National Meeting, San Diego, CA, March 13-17, 2016

Proposed abstract for Oral Presentation in ENVIR: Opportunities & Progress in Computational Prediction of Contaminant Toxicity, Fate & Transport Properties

(Submission Deadline: October 12, 2015)

Title: The influence of data curation on QSAR Modeling – examining issues of quality versus quantity of data

Authors: Kamel Mansouri[1], Christopher M Grulke[2], Ann M Richard[3], Antony J Williams[3]

[1] ORISE Post Doctoral Fellow, US EPA, Research Triangle Park, NC 27711
[2] Lockheed Martin – Contractor to the US EPA, Research Triangle Park, NC 27711
[3] National Center for Computational Toxicology, US EPA, Research Triangle Park, NC 27711

This presentation will examine the impact of data quality on the construction of QSAR models being developed within the EPA's National Center for Computational Toxicology. We have developed a public-facing platform to provide access to predictive models. As part of the work we have attempted to disentangle the influence of the quality versus quantity of data available to develop and validate QSAR models. We will present specific examples of data quality issues underlying the widely used EPISuite software that was initially developed over two decades ago. Relative to the era of EPISuite development, modern cheminformatics tools allow for more advanced capabilities in terms of chemical structure representation and storage, as well as enabling automated data validation and standardization approaches to examine data quality. This presentation reviews both our manual and automated approaches to examining key datasets related to the EPISuite training and test data, including; approaches to validate across chemical structure representations (e.g., mol file and SMILES) and identifiers (chemical names and registry numbers) and approaches to standardize data into QSAR-consumable formats for modeling. Our efforts to quantify and segregate data into quality categories has allowed us to investigate the resulting models that can be developed from these data slices and to examine to what extent efforts into the development of large high-quality datasets have the expected pay-off in terms of prediction performance. This abstract does not reflect U.S. EPA policy.