Abstract 131 — Poster Presentation: Session II-3 "Computational Modelling and Chem-Informatics"

Prediction of Physicochemical Properties of Environmental Chemicals Using Molecular Fingerprints and Machine Learning Methods

Zang Q[1], Mansouri K[2], Allen D[1], Casey W*[3], Judson R[2]

[1]ILS/NICEATM, RTP, NC, USA; [2] EPA/ORD/NCCT, RTP, NC, USA; [3]NIH/NIEHS/DNTP/NICEATM, RTP, NC, USA

**Presenting author*

Abstract

Novel methods are presented for the estimation of six physicochemical properties (octanol/water partition coefficient, water solubility, boiling point, melting point, vapor pressure and bioconcentration factor) of environmental chemicals using simple binary molecular fingerprints. Quantitative structure-property relationship (QSPR) models were developed using four approaches with differing complexity: multiple linear regression (MLR), random forest (RF) regression, partial least squares regression (PLSR), and support vector regression (SVR). Genetic algorithms (GA) and RF methods were employed to select the most information-rich subset of descriptors for obtaining reliable and robust regression models. MLR, PLSR, and SVM exhibited satisfactory predictive results with low prediction errors and all substantially outperformed RF. The approach in which MLR was coupled with GA for descriptor selection was superior to all other approaches and achieved high correlation coefficients between the calculated and experimental data (R2 >0.90). This study demonstrates that (1) molecular fingerprints are useful descriptors, (2) GA is an efficient feature selection tool from which selected descriptors can effectively model these properties, and (3) simple methods such as MLR give better results than more complicated methods. This project was funded in whole or in part with Federal funds from the NIEHS, NIH under Contract No. HHSN27320140003C.