A machine learning-based QSAR approach to predict biological removal of organic micropollutants during wastewater treatment

J. A. Cordero[1], K. Fenner[1,2]*

[1]Department of Environmental Chemistry, Eawag: Swiss Federal Institute of Aquatic Science and Technology, [2]Department of Chemistry, University of Zürich

Accurate prediction of removal of micropollutants during wastewater treatment is urgently needed for improved chemical risk assessment and for the design of greener chemical substances and processes. One of the major challenges hindering development of accurate models is the lack of large homogenous databases. However, experimental data on removals of hundreds of micropollutants for several wastewater treatment plants (WWTPs) have been recently collected, thus providing new opportunities for modeling.

In this study, a first attempt to predict removals using molecular descriptors (i.e., Padel descriptors, MACCS fingerprints and enviPath biotransformation rules) and machine learning algorithms (i.e., Random Forests, Support Vector Machines and Neural Networks) is presented. Two independent datasets from two major sampling campaigns that collected removal data from WWTPs in Switzerland, Australia and Sweden were used. The first dataset (hereafter referred as AUS) includes data for 293 chemical compounds in 15 WWTPs and the second dataset (hereafter referred as AMAR) includes 384 compounds for 8 WWTPs. The AUS dataset is characterized by compounds with small breakthroughs ranging mostly between 0 and 0.2, that is, compounds that are largely removed. Differently, the AMAR dataset contains more compounds with breakthrough in the range of 0.2-0.4, but only few compounds with intermediate and large breakthroughs, challenging model training. Upon log transformation of breakthrough, it was possible to produce models that explain, at least partially, breakthrough in terms of the presence or absence of molecular substructures. The best performance (R^2_{test}= 0.1-0.4) was achieved using a random forest regressor and MACCS fingerprints as features. All models show large differences in performance with different random train-test splits, which is attributed to the large chemical space covered by the test set and the limited number of training examples. Therefore, we suggest the following strategies for further development: 1) applying transfer learning techniques using lab-generated data but relying on WWTP data to train the final model; 2) targeting more recalcitrant compounds through exhaustive suspect screening. Our models are available at: https://c4science.ch/source/pepper/repository/