Assessing the domain of applicability of polyparameter linear free energy relationships (PP-LFERs)

SATOSHI ENDO

Osaka City University, Urban Research Plaza & Graduate School of Engineering, Sugimoto 3-3-138, Sumiyoshi-ku, 558-8585 Osaka, Japan, ensato.de@googlemail.com

Polyparameter linear free energy relationships (PP-LFERs) are increasingly used to predict partition coefficients that are relevant for environmental assessments of contaminants. PP-LFERs appear, for example,

$$\log K = c + sS + aA + bB + vV + lL$$

where K is the partition coefficient and S, A, B, V, and L are the solute descriptors that describe the properties of the solute. The lowercase latters are fitting coefficients and determined via an empirical calibration, typically with several tens of experimental data. PP-LFERs provide accurate predictions for partition coefficients of a number of neutral organic chemicals because of the well-selected descriptors which capture all relevant molecular interactions occuring in solvents and sorbing media.

Because PP-LFERs are empirical fitting models, their domain of applicability is defined by the training chemicals used to calibrate the fitting coefficients. It is natural to expect that PP-LFER predictions are less accurate for chemicals that are outside the domain of applicability defined by the calibration chemicals. The domain of applicability for a PP-LFER, however, has not been considered or only vaguely defined in previous publications. The model applicability domain is important, particularly because the chemical domain of interest for environmental science is extremely large, as practically all industrial chemicals need to be screened for their potential environmental hazards and risks.

In this work, leverage calculations were applied to PP-LFERs to evaluate their domain of applicability. The leverages (h) are diagonal elements of the hat matrix and often used to identify outliers and extrapolations for multiple linear regression models. As a first attempt, 248 data for the oil/water partition coefficient were used to evaluate the relevance of leverage calculations. Twenty-five data were randomly selected and used to train the PP-LFER model. The rest chemicals were used to evaluate prediction accuracy. This calculation was repeated 500 times. The root mean square error (RMSE) was 0.24±0.05 for training chemicals, 0.31±0.02 for test chemicals with $h<2p/n$ (i.e., interpolation), and 0.43±0.12 for test chemicals with $h>2p/n$ (i.e., extrapolation) ($2p/n$ is two times the number of fitting parameters divided by the number of observations, often used as a cut-off for interpolation). These results suggest that h can help identify the chemicals that are tendencially less well predicted by the calibrated PP-LFER. Similar results were obtained for other partition coefficients.

Successively, published PP-LFERs for various partition coefficients such as octanol-water, organic matter-water, air-water, and phospholipid membrane-water partition coefficients were retrospectively evaluated for their domain of applicability. For this purpose, 20 environmentally relevant chemicals that are distinctly different from each other in terms of their PP-LFER descriptors were used as probes, and h values were calculated for these chemicals. The results indicate that none of the evaluated PP-LFER equations encompasses all 20 probe chemicals within their domain of applicability. Most often, polyfluorinated and organosilicon compounds (as represented by D5 and 8:2 fluorotelomer alcohol) are out of the model applicability domain. Also, relatively large, polar compounds such estradiol, metolachlor, and bisphenol A are often not covered by the current PP-LFERs. In contrast, relatively small chemicals with or without a polar functional group are usually within the calibration set. This study demonstrates that the leverage calculation is a simple approach to quantitatively evaluate the domain of applicability of PP-LFERs and is useful for both users and developers of PP-LFERs.