The Role of Feature Selection and Statistical Weighting in Predicting *In Vivo* Toxicity Using *In Vitro* Assay and QSAR Data

Wignall, J.[1], Martin, M.[2], Varghese, A.[1], Trgovcich, J.[1]

1. ICF International, Fairfax, VA
2. US EPA, RTP NC

Our study assesses the value of both *in vitro* assay and quantitative structure activity relationship (QSAR) data in predicting *in vivo* toxicity using numerous statistical models. The models were built on datasets of (i) 586 chemicals for which both *in vitro* and *in vivo* data are currently available in EPA's Toxcast and ToxRefDB databases, and (ii) 769 chemicals for which both QSAR data and *in vivo* data exist. Similar to a previous study (based on just 309 chemicals, Thomas et al. 2012), after converting the continuous values from each dataset to binary values, the majority of more than 1,000 *in vivo* endpoints are poorly predicted. Even for the endpoints that were well predicted (about 40 with an F1 score of >0.75), class imbalances in *in vivo* endpoint data or cytotoxicity across *in vitro* assays may be skewing results. We investigated whether use of best practices for data preprocessing and model fitting in real-world contexts would improve model predictions. This included options for dealing with missing data, including omitting observations, aggregating variables, and imputing values. We also examined the impacts of feature selection (from both a statistical and biological perspective) on performance and efficiency, and we weighted outcome data to reduce endpoint imbalances to account for potential chemical selection bias and assessed revised performance. For example, initial weighting strategies decrease the number of models with an F1 score >0.75 drastically (to 6), but these models are more able to predict nontoxic chemicals in certain contexts. The results of these analyses can be used to inform screening or other decisions, especially in the context of future data enhancements, such as more biologically relevant *in vitro* assays, additional *in vivo* endpoint data, and extension of chemical space.

Disclaimer: The views expressed herein are those of the authors and do not necessarily represent the views of the U.S. EPA.

Category: Risk Assessment