# Predictive Quantitative Structure-Activity Relationship (QSAR) Modeling of Fathead Minnow Aquatic Toxicity

Joanna Kopko, Xuelian Jia, Hao Zhu
Rutgers, The State University of New Jersey

Chemical Toxicity has been a major threat to people and animals with more new compounds available on the market. Traditional experimental testing using animal models are expensive and time-consuming. Artificial intelligence techniques, such as those based on machine learning, are promising to advance toxicology by providing cheap and instant chemical toxicity evaluation tools. Quantitative structure-activity relationship (QSAR) modeling is a computational modeling strategy which builds the relationship between chemical structures and biological activity and then uses resulted models to predict the activities of new chemicals based on their structures. Based on QSAR strategy, this study aimed to develop predictive models to evaluate compounds' aquatic toxicity shown as LC50 (lethal concentration by 50%) of fathead minnow testing. The dataset contained 754 unique compounds with their experimental LC50 values. The molecular structures of these compounds were annotated using simplified molecular-input line-entry system (SMILES), and chemical descriptors were calculated based on the SMILES outputs. A chemical space, which was based on the principal component analysis (PCA), was created to visualize the distribution of toxic and non-toxic compounds in this dataset. Then, Multiple Linear Regression (MLR), k-Nearest Neighbor (kNN), and Random Forest (RF) were used to develop QSAR models using the same chemical descriptors. Accuracy, sensitivity, specificity, and the correct classification rate (CCR) were used to evaluate the performance of the models. For all three models, the accuracy, sensitivity, specificity and CCR ranged 79-95%, 71-96%, 81-96% and 78- 96% respectively. The RF model has the best performance (e.g. CCR = 96%) and MLR has the poorest performance (e.g. CCR= 78%). All models have satisfactory performance (CCR higher than 70%), indicating that the models are ready to predict the toxicity of new compounds purely based on their structures. Therefore, machine learning has significantly advanced Computational Toxicology and proven itself to be a promising alternative to traditional experimental procedures. Funded by R25ES020721.