Universal LD_{50} predictions using deep learning

Risa R. Sayre[1,2] & Christopher M. Grulke[1]

U.S. Environmental Protection Agency, Office of Research and Development, National Center for Computational Toxicology 2. ORISE Research Participant

Our approach uses an ensemble of multilayer perceptron regressions to predict rat acute oral LD50 values from chemical features. Features were generated from QSAR-ready SMILES with MOE 2016.08 (all 2D fingerprints), CORINA Symphony 14698 (Toxprint chemotypes), and RDKit 2016.09.4 (more than 70 numeric or vector fingerprints). LD50 values were log transformed, then scaled from 0 to 1, and feature values were standardized to zero mean with unit variance. All numeric fingerprints from MOE and RDKit were aggregated into single vectors. We used the modeling environment of Keras 2.1.1 in Python 3.5 with a Tensorflow 1.4.0 backend. Each one-hidden-layer Sequential model built per descriptor was optimized with a grid search of the following hyperparameters: (batch size, optimization function, loss function, learning rate, number of hidden dimensions, hidden layer activation function, and output layer activation function). The sum of the distances from the mean of each feature's training set defines the applicability domain index. The instantiation with the highest validation R2 for a given feature set was tested against a predictivity threshold of training R2 > 0.5, validation R2 > 0.45, RMSE < RMSE using the mean LD50 as a prediction, and a significant Spearman's rho over 0.6. They were further validated by comparing y randomized results. The mean of predictions for each model above the threshold created the ensemble prediction value, which had a validation set R2 of 0.57***. This approach does not necessarily reflect US EPA policy.***