Supplemental material

$$\hat{\omega_0}, \hat{\omega_i}^{OLS} = argmin_{\omega_0,\omega_i} \sum_{i=1}^{N} (y_i - \omega_0 - \sum_{i=1}^{p} \omega_i x_i)^2$$

The OLS estimation tends to perform not well since it does not consider the trade-off between variance and bias resulting from the fact that all the features are utilized to construct the regression model. One direct consequence is that the prediction power of an informative feature subset is diluted due to the incorporation of noisy features to the model.

1 Stratified 10-fold cross validation

Cross validation is a common method to validate a machine learning model and a guide to determine its tuning parameters. The idea is to set aside part of the dataset as a testing set and train the model with the rest of the datapoints. One common approach is 10-fold cross validation, where the dataset is partitioned into ten folds followed by ten rounds of training, where a different fold is the testing set each time. In order to make sure each fold is a good representative of the whole dataset, we first stratified the folds, by using an even distribution of experimentally-measured values in the different folds (ref [2]). The number of possibilities for choosing 10 samples from a dataset increases with the dataset's size in a factorial manner, which quickly render it impractical to exhaustively assess all possible fold combinations. To reduce bias, we run the cross validation for 1,000 times and set the final prediction parameters as an average of all runs. The contribution of each feature to the kinetic constants was evaluated by its weight. In the 1,000 times' running, ten regularized linear models were generated each time. The number of features with non-zero weights varies in each round of training. The final selected features were the ones that were selected by most of the models and had a weight that was its average value in all models (in which it was selected) and scaled with respect to the largest absolute weight in the set.