Supplemental Material Machine learning part

**1. Prediction and feature selection via Elastic net**

A regularized linear regression model, Elastic Net (EN), was chosen to fit the dataset of the kinetic constants, each constant fitted independently. Comparing to ordinary least square regression, an EN model is able to make a prediction and select the most informative feature set simultaneously as  and  penalties are applied to the regression weights (ref [1]). The weight of each structural feature generated by the BglB model in an EN regression model are estimated as

------------------------------------------(1)

Where:

: the intercept;

: the weight of structural feature i in the regression model;

p: the number of structural features generated by the BglB model;

: the kinetic constant (the dependent variable to be predicted);

: structural features generated by the BglB model (the independent variables);

,: parameters tuning the constraints on the weights.

Since the structural feature were measured in different range and unit, we first normalized all the features to be zero-centered with variance being one by subtracting the mean and dividing by the variance of the feature value. All the features are on the same scale to compare their contribution to the kinetic constants after the normalization. The tuning parameters,  are determined one by one via stratified 10-fold cross validation by searching a grid of and. Each round of cross validation generated a linear regression model. In order to build a more generalized model, cross validation were run 1,000 times, a different partitioning of the dataset each time. The final prediction of a mutant’s kinetic constant was an average of all the predictions during the 1,000 rounds of training. The average number of non-zero weights when predicting *k*cat/KM, *k*cat and KM were 9, 8 and 10 respectively. The top features were chosen and listed in table 1 with their averaged weights among all the models (9 for *k*cat/KM, 8 for *k*cat, 10 for KM)

**2. Stratified 10-fold cross validation**

Stratified 10-fold cross validation was implemented to validate the EN model (ref [2]). Specifically, all the mutants were first ranked according to the experimentally-measured value of the kinetic constant to be predicted and every 10 adjacent datapoints were randomly marked with an index using integers from 1 to 10 without duplication. Finally, all the datapoints with the same index were grouped together, resulting in ten folds. Since the datapoints in each folds comes from different level of the dataset, this guarantees every fold is a good representative of the dataset. Take the KM kinetic constant prediction as an example. Ten folds, each of which has 9 datapoints with KM ranging from high to low, were generated in this way. In order to build a robust prediction model, the cross validation was run 1,000 times, the dataset split into training set and testing set differently each time.

3. **Reference**

1.Zou H, Hastie T. Regularization and variable selection via the elastic net[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2005, 67(2): 301-320.

2.Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection[C]//Ijcai. 1995, 14(2): 1137-1145.