Supporting information for

**Kinetic characterization of over 100 glycoside hydrolase mutants enables the discovery of structural features correlated with kinetic constants**

Dylan Alexander Carlin1\*, Ryan W. Caster2\*, Xiaokang Wang3, Stephanie A. Betzenderfer2, Claire X. Chen2, Veasna M. Duong2, Carolina V. Ryklansky2, Alp Alpekin2, Nathan Beaumont2, Harhul Kapoor2, Nicole Kim2, Hosna Mohabbot2, Boyu Pang2, Rachel Teel2, Lillian Whithaus2, Ilias Tagkopoulos2,6, Justin B. Siegel2,3,4

Author affiliations:

1. Biophysics Graduate Group, University of California, Davis

2. Genome Center, University of California, Davis

3. Department of Chemistry, University of California, Davis

4. Department of Biochemistry & Molecular Medicine, University of California, Davis

5. Department of Biomedical Engineering, University of California, Davis

6. Department of Computer Science, University of California, Davis

**Table of contents**

Supplemental Table 1 (list of mutants) (**Alex paste in**)

Supplemental Figure 1 (Gel images) **DONE**!

Supplemental Figure 2 (Q19A and R240A) **DONE**!

Supplemental Table 2 (PCC and SRC values) **DONE**!

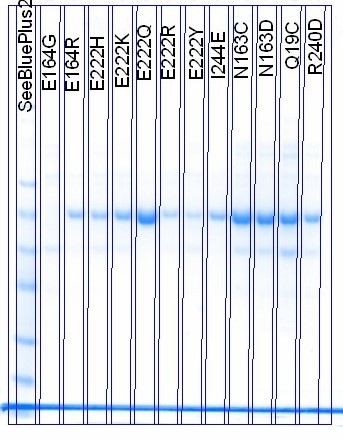
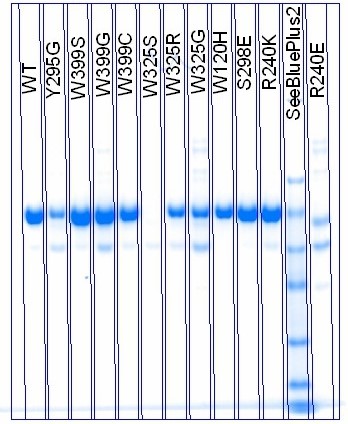
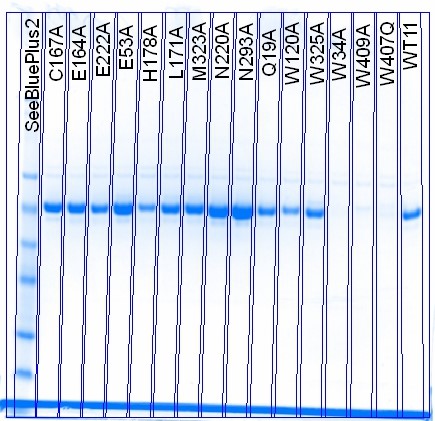
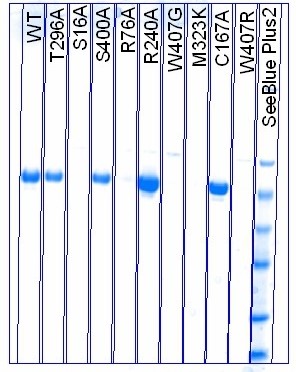
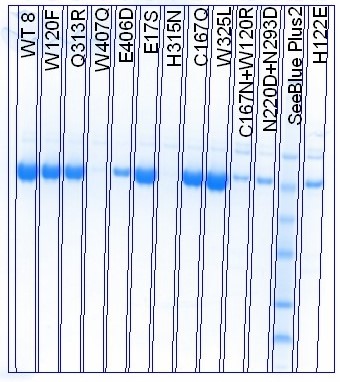
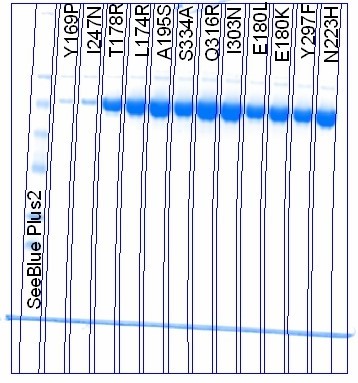
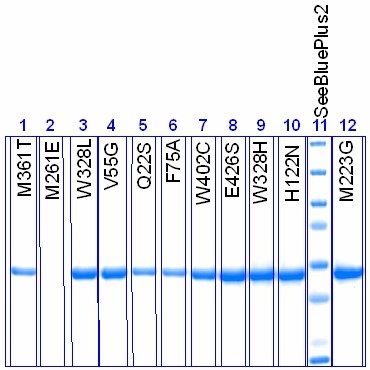
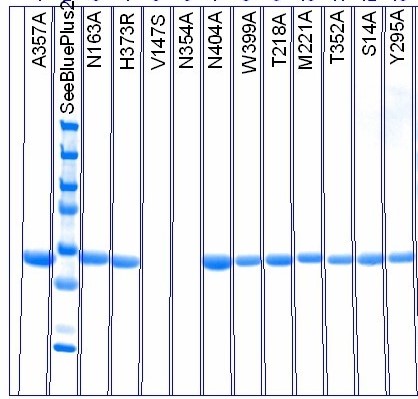
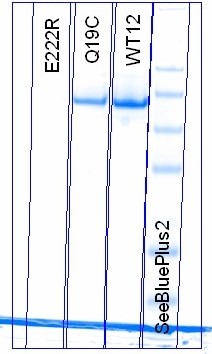
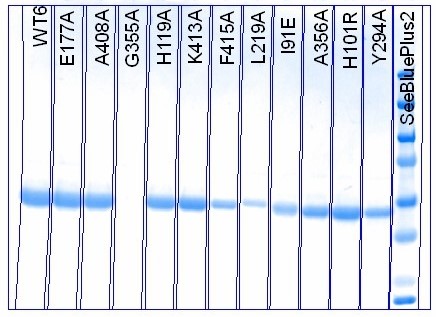
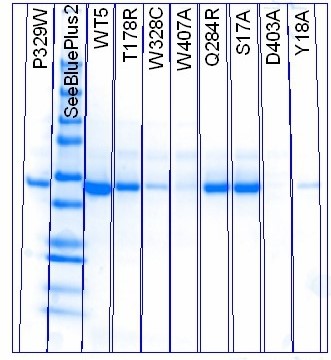
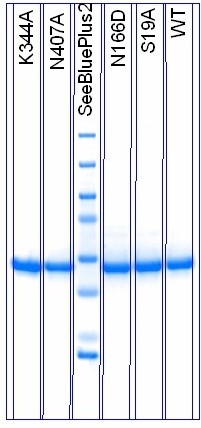
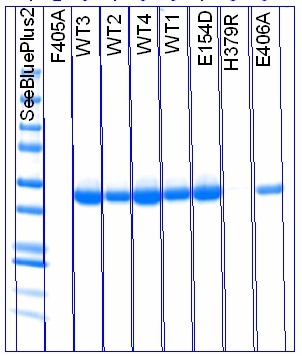
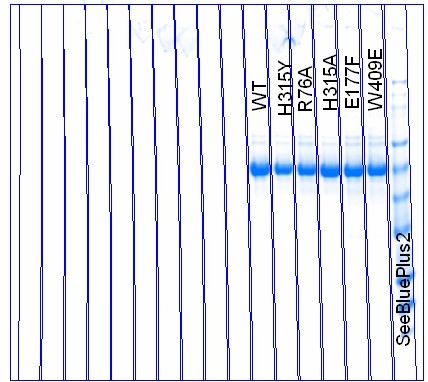
Supplemental Figure 3 (diagnostic plots for all mutants) (**Alex paste in**)

Supplemental materials and methods **DONE**!

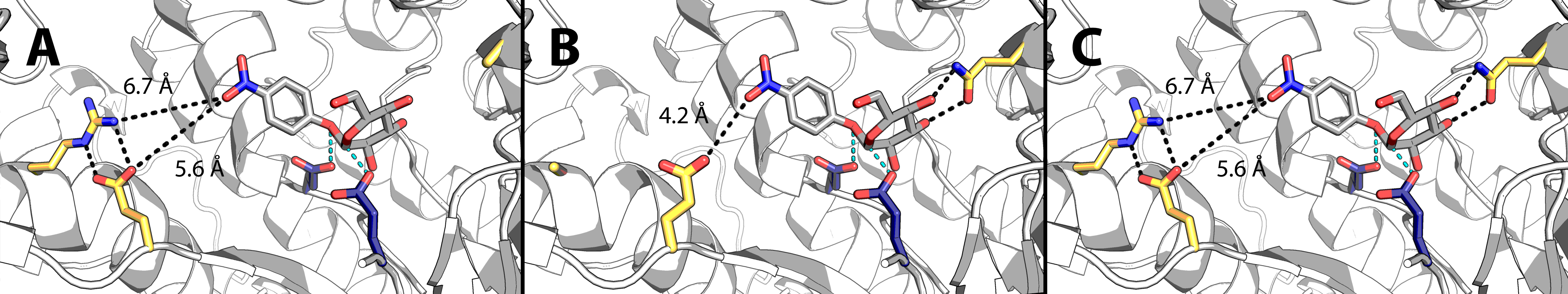
**Supplementary Table 1: List of 104 computationally designed mutants with their experimentally derived kinetic constants, nonlinear regression analysis, Ki for mutants that displayed substrate inhibition, and total yields**. The comprehensive list of 104 mutants experimentally designed, built, and assayed using the methods below.

**Supplemental Table 2. PCC and SRC values for each individual structural feature for each mutant**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Feature** | **PCC (Kcat/KM）** | **SRC (Kcat/KM）** | **PCC (1/KM)** | **SRC (1/KM)** | **PCC (Kcat)** | **SRC (Kcat)** |
| all\_cst | 0.182 | 0.116 | 0.039 | 0.071 | 0.140 | 0.120 |
| fa\_rep | 0.289 | 0.268 | 0.253 | 0.155 | 0.064 | 0.084 |
| hbond\_sc | -0.352 | -0.352 | -0.248 | -0.297 | -0.309 | -0.266 |
| SR\_1\_all\_cst | -0.061 | -0.039 | -0.148 | -0.114 | -0.163 | -0.119 |
| SR\_1\_burunsat\_pm | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| SR\_1\_fa\_rep | 0.195 | 0.180 | 0.193 | 0.309 | 0.155 | 0.042 |
| SR\_1\_hbond\_pm | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| SR\_1\_hbond\_sc | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| SR\_1\_nlpstat\_pm | 0.142 | 0.087 | 0.263 | 0.317 | 0.007 | 0.055 |
| SR\_1\_pstat\_pm | 0.164 | 0.096 | 0.050 | 0.064 | -0.131 | -0.069 |
| SR\_1\_total\_score | -0.078 | 0.093 | -0.092 | -0.050 | -0.081 | 0.024 |
| SR\_2\_all\_cst | 0.039 | -0.064 | -0.089 | -0.059 | -0.005 | -0.096 |
| SR\_2\_burunsat\_pm | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| SR\_2\_fa\_rep | 0.074 | 0.168 | 0.258 | 0.168 | 0.011 | 0.014 |
| SR\_2\_hbond\_pm | 0.087 | 0.092 | 0.162 | 0.191 | 0.149 | 0.116 |
| SR\_2\_hbond\_sc | -0.149 | -0.235 | -0.142 | -0.229 | -0.126 | -0.138 |
| SR\_2\_nlpstat\_pm | 0.168 | 0.142 | 0.031 | 0.038 | 0.080 | 0.092 |
| SR\_2\_pstat\_pm | 0.102 | 0.070 | -0.042 | -0.017 | -0.023 | 0.013 |
| SR\_2\_total\_score | 0.071 | 0.002 | -0.078 | -0.079 | -0.055 | -0.134 |
| SR\_3\_all\_cst | -0.061 | -0.039 | -0.148 | -0.114 | -0.163 | -0.119 |
| SR\_3\_burunsat\_pm | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| SR\_3\_fa\_rep | 0.195 | 0.180 | 0.193 | 0.309 | 0.155 | 0.042 |
| SR\_3\_hbond\_pm | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| SR\_3\_hbond\_sc | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| SR\_3\_nlpstat\_pm | 0.142 | 0.087 | 0.263 | 0.317 | 0.007 | 0.055 |
| SR\_3\_pstat\_pm | 0.164 | 0.096 | 0.050 | 0.064 | -0.131 | -0.069 |
| SR\_3\_total\_score | -0.078 | 0.093 | -0.092 | -0.050 | -0.081 | 0.024 |
| SR\_4\_all\_cst | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| SR\_4\_burunsat\_pm | 0.123 | 0.135 | 0.089 | 0.078 | 0.067 | 0.138 |
| SR\_4\_fa\_rep | -0.118 | -0.024 | -0.235 | -0.212 | -0.153 | -0.103 |
| SR\_4\_hbond\_pm | -0.197 | -0.206 | 0.046 | 0.069 | -0.211 | -0.182 |
| SR\_4\_hbond\_sc | 0.299 | 0.092 | -0.009 | -0.063 | 0.058 | 0.061 |
| SR\_4\_nlpstat\_pm | 0.067 | 0.174 | 0.160 | 0.244 | -0.007 | -0.015 |
| SR\_4\_pstat\_pm | 0.040 | 0.098 | 0.105 | 0.152 | -0.025 | -0.072 |
| SR\_4\_total\_score | 0.096 | 0.045 | 0.193 | -0.121 | -0.181 | -0.076 |
| SR\_5\_all\_cst | 0.128 | 0.074 | -0.091 | -0.048 | -0.062 | -0.082 |
| SR\_5\_burunsat\_pm | 0.055 | 0.052 | 0.061 | -0.031 | 0.136 | 0.099 |
| SR\_5\_dsasa\_1\_2 | 0.235 | 0.189 | 0.037 | 0.147 | 0.215 | 0.201 |
| SR\_5\_fa\_rep | -0.012 | -0.064 | 0.165 | -0.019 | 0.058 | -0.030 |
| SR\_5\_hbond\_pm | 0.514 | 0.410 | 0.267 | 0.334 | 0.267 | 0.146 |
| SR\_5\_hbond\_sc | -0.524 | -0.447 | -0.266 | -0.341 | -0.273 | -0.194 |
| SR\_5\_interf\_E\_1\_2 | -0.461 | -0.469 | -0.237 | -0.231 | -0.266 | -0.278 |
| SR\_5\_total\_score | -0.462 | -0.471 | -0.237 | -0.230 | -0.267 | -0.279 |
| tot\_burunsat\_pm | -0.078 | -0.101 | -0.036 | -0.066 | 0.154 | 0.056 |
| tot\_hbond\_pm | 0.419 | 0.396 | 0.291 | 0.380 | 0.315 | 0.323 |
| tot\_NLconts\_pm | 0.570 | 0.548 | 0.246 | 0.238 | 0.432 | 0.421 |
| tot\_nlpstat\_pm | 0.334 | 0.313 | 0.223 | 0.164 | 0.277 | 0.277 |
| tot\_nlsurfaceE\_pm | -0.285 | -0.284 | -0.193 | -0.276 | -0.176 | -0.111 |
| tot\_pstat\_pm | 0.306 | 0.209 | 0.118 | -0.009 | 0.221 | 0.161 |
| tot\_seq\_recovery | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| tot\_total\_charge | 0.051 | 0.053 | 0.199 | 0.153 | 0.103 | 0.123 |
| tot\_total\_neg\_charges | -0.068 | -0.063 | -0.148 | -0.090 | -0.044 | -0.034 |
| tot\_total\_pos\_charges | 0.010 | 0.014 | 0.177 | 0.209 | 0.133 | 0.129 |
| total\_score | -0.340 | -0.396 | -0.059 | -0.014 | -0.320 | -0.362 |



**Supplemental Figure 1: SDS-PAGE gel images for all mutants.** SDS-PAGE gels showing greater than 90% purity for all samples. Protein ladder used was SeeBlue® Plus2 Pre-stained Protein Standard (Life Technologies). Yields were also confirmed by comparing with visual gel yields.

****

**Supplementary Figure 2: Models of the active sites of mutants Q19A, R240A, and wild type BglB.** The estimated distances and predicted hydrogen bonding between the substrate, *p*-nitrophenyl-ß-D-glucoside, and the BglB molecule are indicated by either black or blue dotted lines, the former being non-catalytic and the latter being catalytic bonds. Mutation of the arginine at position 240 to an alanine is predicted to allow for new conformers of the glutamate at position 222 that interact with the substrate. Mutation of the glutamine at position 19 to an alanine removes two hydrogen bonds to the substrate drastically affecting catalytic efficiency.

**Supplemental materials and methods**

Mutants were designed using the computer software Foldit, a protein design game based on Rosetta. The program predicts the stability of the mutant enzyme and offers a visual of the new chemical environment, including molecular interactions amongst varying single or double point mutations. Mutants were chosen based on proximity to the active site as well as Foldit’s predicted energy. When designing mutants, if the change increased the energy by less than 5 energy units it was selected for testing.

A sequence coding for BglB (Uniprot P22505) was codon-optimized for *Escherichia coli* and manufactured as a DNA String by Life Technologies.  The gene was inserted between NdeI and XhoI to encode a C-terminal His tag onto the protein sequence.  Kunkel mutagenesis was used to create site-specific mutations, and all plasmids were sequence-verified on the Transcriptic cloud platform.

20 μL of chemically-competent *Escherichia coli* BL21(DE3) (Novagen) were transformed on ice with 1 μL of plasmid in buffer at a concentration of 130 to 90 ng/μL. The competent cell-plasmid mixture was temperature shocked to induce plasmid uptake by heating at 42°C for one minute and then chilling on ice for one minute. Cells were recovered in 200 μL Terrific Broth (TB) media at 37 °C for one hour. They were then plated onto an LB agarose plate containing 50 mg/mL kanamycin, and incubated for 24 hours at 37 °C.

For each mutant, a 50 mL Falcon tube containing 5 mL TB with 50 mg/mL kanamycin was inoculated with one colony from the plate. Tubes were covered with breathable seals and incubated with shaking for 24 hours at 37 °C.

Growth cultures were pelleted by centrifugation at 4700 RPM for 10 minutes and the supernatant was discarded. The cell pellet was resuspended in 5 mL of induction medium (TB with 1 mM isopropyl-β-D-thiogalactopyranoside and 50 mg/mL kanamycin). The tubes were covered again with breathable seals and incubated with shaking at 18 °C for 24 hours.

The 5 mL expression culture was pelleted by centrifugation at 4700 for 10 minutes and the supernatant was discarded. The resulting pellet was suspended in 500 μL wash buffer (50 mM HEPES, 150 mM sodium chloride, 15 mM imidazole, pH 7.50) and lysed with a mixture of BugBuster protein extraction reagent (Millipore), 1 mg lysozyme, 0.1 mg DNase, and 0.1 mg phenylmethylsulfonyl fluoride. Cell were lysed for 20 minutes and then centrifuged at 14,700 RPM for ten minutes.

Protein micro-columns [Bio-Spin Chromatography Columns #732-6008] were prepared over a waste collector with 100 μL of 50% nickel resin slurry in each column. The nickel resin was washed with 500 μL wash buffer, then two 500 μL aliquots of supernatant were added to the columns.  Six rounds of 500 μL wash buffer were then allowed to drip through the columns. Resulting protein micro-columns were then transferred to 2 mL tubes for elution. Protein was eluted in 2x100 μL elution buffer (50 mM HEPES, 150 mM  sodium chloride, and 25 mM EDTA, pH 7.50). A brief centrifugation at 4000 RPM ensured all protein was collected. Protein yield was then determined via absorbance at 260 and 280 nm and SDS-PAGE.

Proteins were assayed at 8 substrate concentrations via a four-fold serial dilution of a 100 mM substrate stock solution. Protein activity was tested by mixing 90 μL substrate stock and 10 μL purified protein. Protein solutions that turned yellow within five minutes displayed sufficient catalytic activity to be diluted to 1/100 concentrations. This was performed by combining 990 μL elution buffer and 10 μL purified protein. Protein solutions that did not turn yellow were converted into 1/10 concentrations by combining 100 μL purified protein with 900 μL elution buffer.

Diluted protein solution was dispensed in 25 μL aliquots into non-binding 96-well white [Corning Costar #07-200-845]. Mutants were assayed in triplets, allowing for four mutants per plate. 150 μL substrate stock (100 mM 4-nitrophenyl-beta-D-glucopyranoside in elution buffer) were dispensed into the first row of 96-well PCR plate, with 112.5 uL protein buffer dispensed into the rest of the rows. 37.5 μL of the substrate stock were pipetted from the first row into the second row, then mixed. Serial dilution was repeated for all rows, excluding the last row for a control. The assay was initiated by multi-channel pipetting 75 μL substrate from each row of the PCR plate into the corresponding rows of the non-binding white plate. The rate was determined by monitoring absorbance of light at 420 nm every 1 min for 60 min.

Unless otherwise noted, all supplies were purchased from Sigma-Aldrich.

**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.1 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. The weight of each structural feature generated by the BglB model in an EN regression model are estimated as



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)

**Stratified 10-fold cross validation**

Stratified 10-fold cross validation was implemented to validate the EN model.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.

**Supplemental references**

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

2. Kohavi, R. In *A study of cross-validation and bootstrap for accuracy estimation and model selection*, Ijcai, 1995; pp 1137-1145.