Design and kinetic characterization of over 100 glycoside hydrolase mutants enabling the discovery of specific structural features correlated with kinetic constants

DA Carlin‡⊥, RW Caster‡†, SA Betzenderfer†, CX Chen†, VM Duong†, CV Ryklansky†, XK Wang \*†, A Alpekin†, N Beaumont†, H Kapoor†, N Kim†, H Mohabbot†, B Pang†, R Teel †, L Whithaus†, I Tagkopoulos^†, JB Siegel†∥▽

† Genome Center, University of California, Davis⊥ Biophysics Graduate Group∥ Department of Chemistry▽ Department of Biochemistry & Molecular Medicine, University of California, Davis \* Department of Biomedical Engineering, University of California, Davis ^ Department of Computer Science, University of California, Davis

ABSTRACT: The use of computational modeling algorithms to guide the design of novel enzyme catalysts is a rapidly growing field. However, many of the methods developed to date are optimized around indirect measures of function, such as active site sequence recovery, as opposed to direct correlation to experimentally–determined functional effects of mutations. This is due to the lack of datasets for which a large panel of enzymes has been produced, purified, and kinetic constants determined. Here we directly address this issue by constructing a dataset of over 100 mutant enzymes, each of which were produced, purified, and kinetic constants (i.e. *k*cat and KM) measured. We illustrate the importance of this type of data set for the potential future improvement of computational enzyme redesign algorithms by constructing molecular models for each mutant and using machine learning algorithms to elucidate which calculated structural features are correlated with the measured functional parameters. The dataset and analyses carried out in this study not only provide novel insight into how this enzyme functions, but provides a clear path forward for the improvement of computational enzyme redesign algorithms.

# ▩ INTRODUCTION

The ability to rationally reengineer enzyme function has the potential to allow the development of highly efficient and specific catalysts tailored for needs beyond what was selected for during natural evolution. A rapidly growing route for engineering enzyme catalysts is the use of computational tools to evaluate mutations *in silico* before experimentally characterizing the mutant proteins in the lab. Using the Rosetta Molecular Modeling Suite, reengineering of both specificity and chemistry has been accomplished. However, only a relatively small number of all designs have the intended functional effect.

****

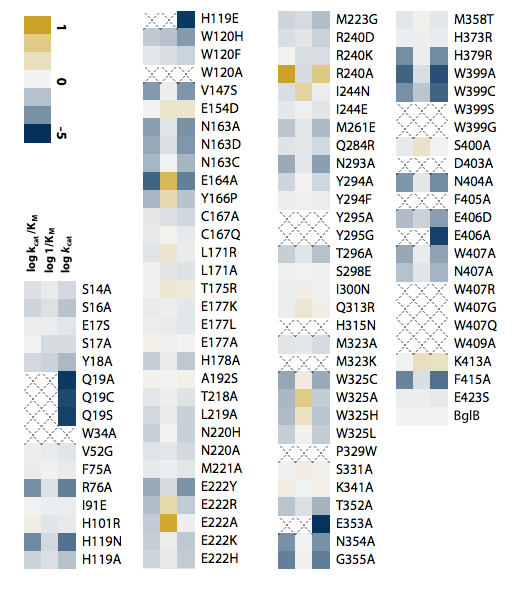
**Figure 1. Structure and catalyzed reaction of BglB** (A) Structure of BglB in complex with a modeled *p*-nitrophenyl-ß-D-glucoside used for design with alpha carbons of residues mutated shown as blue spheres (B) The BglB–catalyzed reaction on *p*-nitrophenyl-ß-D-glucoside used to evaluate kinetic constants of designed mutants

A major hinderance to improved enzyme redesign algorithms is the lack of large quantitative datasets correlating enzyme sequence and function. Large quantitative data sets exist for both protein–protein interfaces and protein thermostability, and have played a critical role in evaluating and improving computational algorithms to accurately model and design protein–protein interactions and thermostable proteins. [] However, there is no equivalent dataset of sequenced, purified, and kinetically characterized enzyme mutants. While many large mutant enzyme libraries have been produced and screened, often only a small subset of the libraries are produced, purified, and kinetically characterized to determine Michaelis-Menten constants.

Due to the lack of quantitative sequence-function datasets for enzymes, many efforts to evaluate and develop modeling algorithms have focused around sequence recovery as opposed to recapitulation of experimentally characterized effects. However, sequence recovery is a non-ideal metric as there are many mutations that are likely neutral or possibly beneficial to function. We aimed to address this by determining kinetic constants for a large number (>100) of enzyme mutants, enabling both the assessment and potential improvement of modeling algorithms to evaluate enzyme structure–function relationships. This dataset is the first step toward an enzyme database equivalent to ProTherm but relating enzyme sequence–function as opposed to protein sequence–thermostability.

The enzyme chosen as the first entry to this dataset is ß-glucosidase B from *Paenibacillus polymyxa* (BglB), a family 1 glycoside hydrolase. Family 1 glycoside hydrolases have been the subject of numerous structural and kinetic studies due to their importance as the penultimate step in cellular ligno-cellulose utilization. An X-ray crystal structure of BglB indicates that BglB follows a classical Koshland double-displacement mechanism in which E353 performs a nucleophilic attack on the anomeric carbon of the substrate’s glucose moiety. The leaving group is protonated by E164. A third active site residue, Y295, orients E353 for catalysis with a hydrogen bond. [Isorna] The protein structure and reaction scheme are provided in **Figure 1**.

Our results include a large data set of kinetic constants of 104 computationally designed variants of BglB, each of which was produced, purified, and kinetic constants (*k*cat, KM, Ki) measured using the reporter substrate *p*-nitrophenyl-ß-D-glucoside (pNPG).

****

**Figure 2. Relative kinetic constants of 104 BglB mutants**. The heatmap depicts the effect of the mutation on each kinetic constant relative to native BglB. As indicated in the color legend, gold is for higher value and blue for a lower value. If the kinetic constant was not measurable, an X is depicted in the box. Mutations with an X in all three columns did not result in soluble protein. Values are on a log scale and the ranges are as follows: 10–11,000 min-1 (*k*cat), 0.6–85 mM (KM), and 10–560,000 M-1min-1 (*k*cat/KM) with wild type constants of 880 ± 10 min–1, 5.0 ± 0.2 mM, and 171,000 ± 8000 M-1 min-1 for *k*cat/KM, *k*cat/KM, and *k*cat/KM respectively. A full table of kinetic constants and substrate versus velocity curves for each are provided in the Supplemental Materials.

The production of this dataset revealed several mutations to non-catalytic residues (i.e. those not directly involved in the proposed reaction chemistry) that are as important to the enzyme-catalyzed reaction as catalytic residues. In addition, we demonstrate the ability to predict effects on *k*cat, KM, and *k*cat/KM using molecular modeling. Finally, we illustrate how the use of machine learning can be used to identify calculated structural features from the molecular models that significantly improve the predictive accuracy of the molecular modeling. These analyses provide a unique insight into the factors important for catalysis in BglB as well as a path forward to for the development and evaluation of next-generation enzyme reengineering algorithms.

# ▩ RESULTS

## Computationally-directed engineering of BglB

The crystal structure of recombinant BglB with the substrate analog 2-deoxy-2-fluoro-alpha-D-glucopyranose bound was used to identify the substrate binding pocket and the catalytic residues. To generate a molecular model which approximates the first proposed transition state for the hydrolysis of pNPG, an SN2-like transition state was built and minimized in Spartan based on a 3D conformer of PubChem CID 92930. Functional constraints were used to define catalytic distances, angles, and dihedrals between pNPG, the acid-base E164, the nucleophile E353, and Y295, which is proposed to stabilize the nucleophilic glutamate. The angle between the attacking oxygen from E353, the anomeric carbon, and the phenolic oxygen was constrained to 180˚, in accordance with an SN2-like mechanism [].

Two methods were used to establish a set of mutants to generate and kinetically characterize. The first method was a systematic alanine scan of the BglB active site where each residue within 12 Å of the ligand in our model was individually mutated to alanine. In the second method, mutations predicted to be compatible with the modeled pNPG transition state in BglB structure were selected through the program Foldit, a graphical user interface to the Rosetta Molecular Modeling Suite []. Mutations were modeled and scored in Foldit and a selection of mutations that were either favorable or did not increase the energy of the overall system by greater than 5 Rosetta energy units were chosen to synthesize and experimentally characterize. Figure 1A illustrates the positions in the protein where mutations were introduced, and a full list of mutations selected is listed in Supplemental Table 1. Together, a total of 69 positions were covered over the 104 mutants made.

### Protein production and purification

Each of the 104 mutants was made via Kunkel mutagenesis [] using the Transcriptic cloud laboratory platform and sequence-verified mutant plasmids were transformed into *E. coli* BLR(DE3). After protein expression induced by 1 mM IPTG, proteins were purified via immobilized metal affinity chromatography. Absorbance at 280 nm was used to quantify protein yield and SDS-PAGE was used to evaluate purity. All proteins used in the study were greater than 80% pure.

A total of ten biological replicates of the native BglB were used to assess expression and purification. The average yield was found to be 1.2 ± 0.4 mg/mL. Of the 104 mutants synthesized, 90 were found to be expressed and purified as soluble protein. The distribution of yields for all 104 mutants is illustrated in the Supplemental Materials. Greater than 70% maintained the yields obtained for native BglB, and 14% were not expressed and purified as a soluble protein above our limit of detection (0.1 mg/mL) based on A280 and SDS-PAGE.

### Kinetic characterization of mutants

Michaelis-Menten kinetic constants for each of the 104 mutants were determined using the colorimetric assay of pNPG hydrolysis. Ten biological replicates of the wild type enzyme had an average *k*cat of 880 ± 10 min–1, an average KM of 5 ± 0.2 mM, and *k*cat/KM of 171,000 ± 8000 M–1 min–1. Observed rates at 8 substrate concentrations were fit to the Michaelis-Menten equation using SciPy. Kinetic constants for each mutant with the substrate pNPG are represented as a heatmap in **Figure 2**. In addition, a table of experimentally measured kinetic constants and nonlinear regression analysis is reported in the Supplemental Material.

Based on the maximum concentration of enzyme used in our assays and colorimetric absorbance changes at the highest substrate concentration used, we estimate our assay’s limit of detection for *k*cat/KM to be 10 M-1min-1. Of the 90 solubly purified mutants, 6 were below the limit of detection. The highest catalytic efficiency observed is 5.6 x 105 M-1min-1 for mutation R240A. In addition, while no substrate inhibition was observed on the wild type BglB, four mutants exhibited measurable substrate inhibition (the inhibition parameter KI for these mutants is reported in the Supplemental Material).

## Sequence–structure–function relationships in BglB

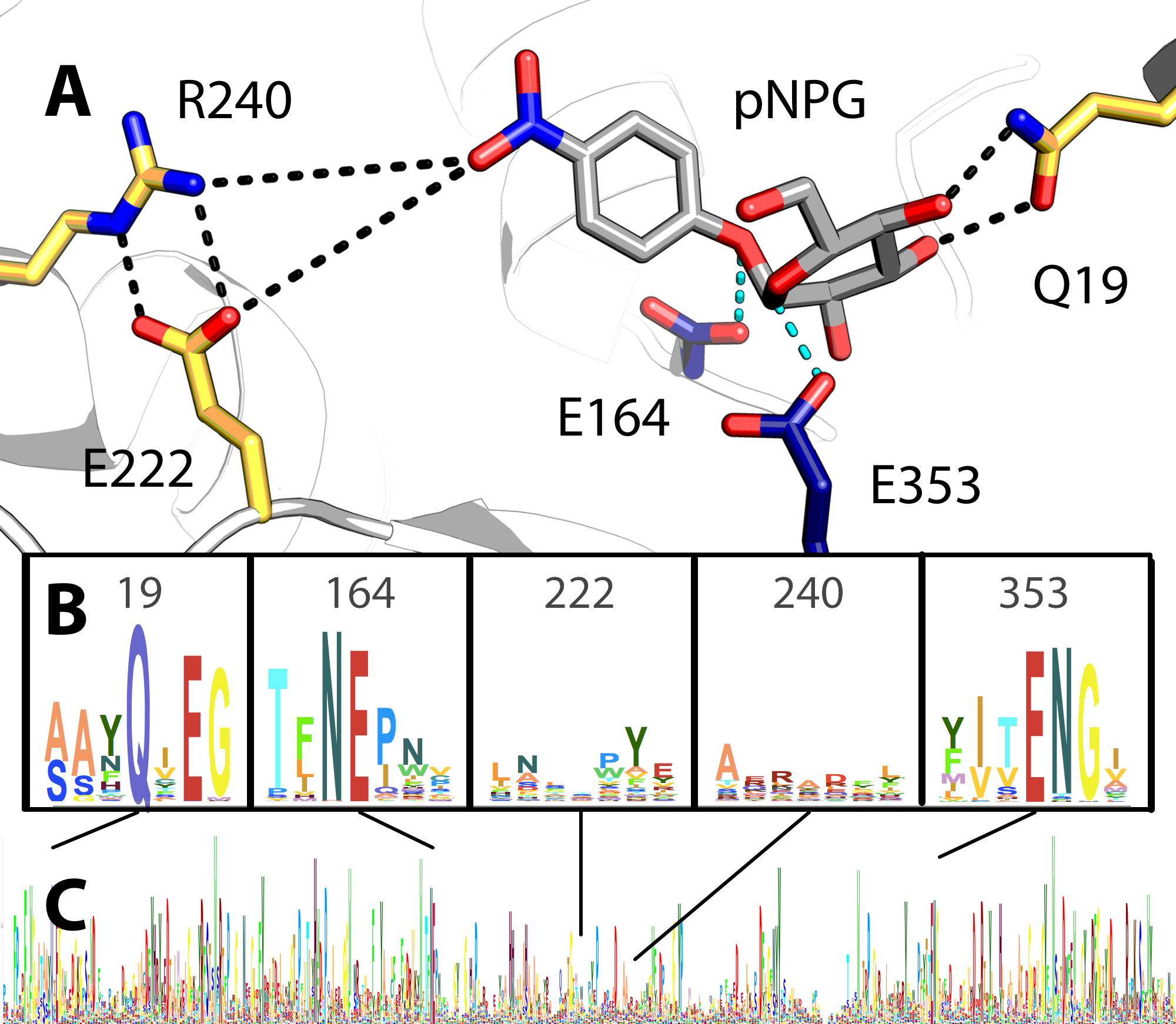


Figure 3. Active site model of BglB (A) Docked model of pNPG in the active site of BglB showing established catalytic residues (navy) and a selection of residues mutated (gold). A multiple sequence alignment of the Pfam database’s collection of 1,554 family 1 glycoside hydrolases was made and the sequence logo for (B) selected regions around specific residues discussed in the text and (C) over the entire BglB coding sequence is represented. The height for each amino acid indicates the sequence conservation at that position.

In agreement with previous studies, our results clearly demonstrate important catalytic roles for E164, E353, and Y295. Mutating any of these residues to alanine results in a greater than 85,000-fold reduction in catalytic efficiency. However, the systematic alanine scan of every residue within 12 Å of the ligand also revealed unexpected structure-function relationships in BglB.

One of the residues that had an unexpected effect on function was Q19. Based on a multiple sequence alignment of the Pfam database for the BglB enzyme family comprising 1,554 non-redundant proteins, Q19 is 95% conserved (Figure 4B). A structural analysis illustrates that both the nitrogen and oxygen of the amide interact with the substrate’s sugar hydroxyl groups **(Figure 3)**. While removing these interactions may be predicted to decrease catalytic efficiency, it was unexpected to observe a 57,000-fold reduction. This single mutation is equivalent to removing the established catalytic residue E353, which reduces activity 85,000-fold. However, unlike E353, the nucleophilic glutamate directly involved in the reaction chemistry, Q19 is not involved in hydrolysis of the substrate. A crystal structure in complex with the 2-deoxy-2-fluoro-alpha-D-glucopyranose inhibitor of the Q19A mutation may help elucidate the structural effect of this mutation. Based on molecular modeling, no major structural change for this mutant is predicted [Supp. Figure X].

Another unexpected finding was a ten fold increase of *k*cat by a single point mutant, R240A. This mutation is not predicted to make a direct molecular interaction with the substrate, so the mechanism by which this mutation increases the rate of catalysis is unclear. The crystal structure reveals that R240 forms two hydrogen bonds with E222 **(Figure 3)**. Molecular modeling of the R240A mutant predicts that E222 adopts an alternative conformation in which the acid functional group of the glutamate is 2 Å closer to the active site (Supp Figure XXX), resulting in a significant change of the electrostatic environment there. Further supporting the hypothesis that the electronegative environment in this region of the active site is important to catalysis is the mutation E222A, which decreases *k*cat by ten fold.

### Evaluation of computational modeling

Molecular models were generated for each of the 104 BglB mutants in order to evaluate the Rosetta Molecular Modeling Suite’s ability to evaluate the functional effects of mutations on BglB kinetic properties. For each mutant, the modeled pNPG previously described was docked into the active site. The docking and structural minimization protocol used approximates the numerous protocols previously used in successful enzyme reengineering efforts: a Monte Carlo simulation with random perturbation of the ligand followed by functional constraint optimization and side chain repacking. The command line and an example set of input files for wild type BglB are provided in the Supplemental Materials.

For each mutant, 100 models were generated the lowest 10 in overall system energy were selected for subsequent structural analysis. From this analysis, a set of 59 potentially informative features was identified and a value calculated for each mutant. Strutural features selected include predicted interface energy, number of hydrogen bonds to the ligand, and change in solvent accessible surface area upon ligand binding. A complete list of all features and their correlation to the experimentally measured constants is given in Supplemental Table 1. For *k*cat/KM and *k*cat, the highest correlation observed is to the total number of non-local contacts (i.e. residues separated by more than 8 sequence positions that interact), with a Pearson Correlation Coefficient (PCC) of 0.56 (p-value 0.009; Wilcoxon test) and 0.43 (p-value 0.004; Wilcoxon test), respectively. For KM the highest PCC is 0.29 (p-value 0.0005; Wilcoxon test) to the total number of hydrogen bonds within each BglB model. The Spearman Rank Correlation (SRC) follows similar trends to those for PCC for all three predicted constants (0.55, 0.42 and 0.38 for *k*cat/KM , *k*cat and KM respectively). The PCC and SRC values for all features are available in Supplemental Materials.

Table 1. Most informative structural features for kinetic constant prediction. For each mutant, ten out of hundred models were selected based on the lowest total system energy. Fifty-nine structural features were calculated for the selected models and the most informative features were selected based on a constrained regularization technique (elastic net with bagging; see Methods). The table contains the top features that have been assigned non-zero weights during training (9 for *k*cat/KM, 8 for *k*cat, 10 for KM). The relative contribution of each feature in determining the kinetic constant is given as a normalized weight (columns 1-3). Column 4 provides a description of each feature, and columns 5 and 6 show the range of observed values in the training dataset. The full feature table is available in Supplemental Table 2. *ns=feature not selected by the algorithm*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *k*cat/KM | *k*cat | KM | Description | Min. | Max. |
| -1.00 | ns | ns | Hydrogen bonding energy of pNPG | -4.53 | -1.8 |
| -0.63 | 1.00 | -0.03 | Total number of polar contacts in the system | 144 | 155 |
| -0.43 | ns | ns | Count of hydrogen bonds to pNPG | 4 | 9 |
| -0.03 | ns | ns | Hydrogen bonding energy of E164 | -0.93 | -0.21 |
| 0.29 | ns | -0.27 | Lennard-Jones repulsion of Y295 | 0.54 | 0.99 |
| 0.39 | 0.92 | ns | Change in solvent-accessible surface upon binding | 0.86 | 0.96 |
| 0.44 | 0.15 | 1.00 | Packing of the system without pNPG | 0.67 | 0.72 |
| 0.44 | 0.53 | 0.46 | Overall packing metric of the system | 0.67 | 0.73 |
| 0.98 | 0.09 | ns | Hydrogen bonding energy of Y295 | -1.28 | -0.5 |
| ns | -0.51 | ns | Packing with pNPG around E353 | 0.19 | 1 |
| ns | -0.10 | ns | Total system energy | -636.44 | -621.6 |
| ns | -0.01 | ns | Hydrogen bond energy of the whole system | -76.7 | -67.63 |
| ns | ns | 0.11 | Lennard-Jones repulsion around E353 | 0.67 | 1.41 |
| ns | ns | 0.27 | Average hydrophobic surface area without pNPG | 0.51 | 1.75 |
| ns | ns | 0.32 | Packing without pNPG around E353 | 0.37 | 0.99 |
| ns | ns | 0.34 | Packing around E164 without pNPG | 0.37 | 0.99 |
| ns | ns | 0.38 | Packing around Y295 without pNPG | 0.34 | 0.99 |
| ns | ns | 0.51 | Lennard-Jones repulsion of E164 | 0.83 | 1.53 |

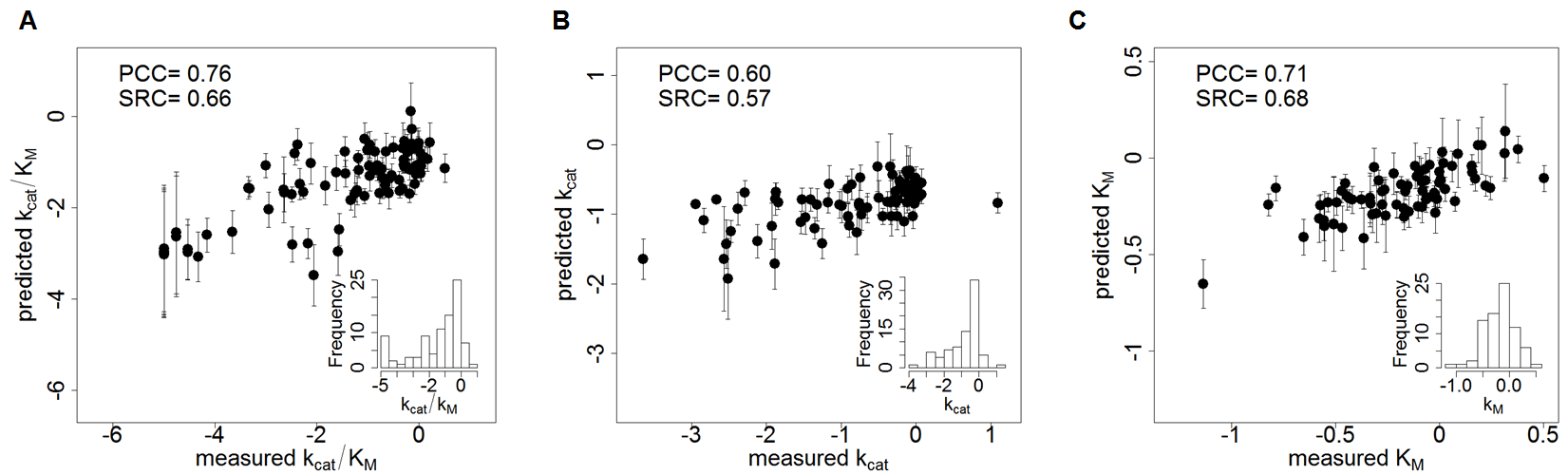


Figure 4. Correlation between machine learning predictions and experimentally-determined kinetic constants. For each of *k*cat/KM (A), *k*cat (B), and KM (C), experimentally-determined kinetic constants are shown on the *x* axis and machine learning predictions ± standard deviation are shown on the *y* axis. The standard deviation was calculated based on the prediction by 1,000 times stratified cross validation for each mutant. All values are normalized relative to wild type BglB and are in log scale. Inset: histograms display the distribution of experimentally-determined values in the data set (90, 80 and 80 samples for *k*cat/KM, *k*cat, and KM, respectively).

Since no single feature was correlated with the observed kinetic constants, we used all 59 features together with a regularized linear model to predict the kinetic constants based on the most informative feature subsets. For this task, we used elastic net regularization, a constraint regression technique that uses both l1 and l2 regularization for feature selection (see Supplemental Materials). [1] To increase robustness to sample size and remove bias, we used a bootstrapping aggregating (bagging) technique, where the predicted value was an average of 1,000 elastic net models, each trained on a different subset of the data (see Methods). The final prediction from this ensemble learning regression method outperformed all individual models. In the case of *k*cat/KM, the PCC is 0.77 versus 0.65 for ensemble and X.XX for individual models. For *k*cat and KM, PCC is 0.60 versus 0.48 and 0.71 versus 0.56, respectively. As shown in **Figure 4**, the machine learning predictions have a strong correlation with experimentally-measured values. Variability differs for each constant but is consistent across all samples, with the the absolute mean±average std for all samples 1.47 ± 0.38, 0.75 ± 0.19, and 0.16 ± 0.05 for *k*cat/KM , *k*cat and KM respectively.

Each kinetic constant is predicted by 8–10 structural features calculated by the Rosetta Molecular Modeling Suite (**Table 1**). The Michaelis-Menten constant KM, which is a measure of how tightly the enzyme-substrate complex binds, is predicted by Rosetta packing statistics, which quantify the area of empty space between atoms in a given region. The set of features identified that correlate to *k*cat and *k*cat/KM include calculated hydrogen bonding energy between BglB and pNPG as well as the number of hydrogen bonds to pNPG. Overall, metrics correlated to *k*cat were found to be co-correlated with kcat/KM, and are mostly measures of hydrogen bonding. Structural features correlated to KM are predominantly packing statistics. It is interesting to note that while no specific feature is strongly correlated to any kinetic constant, the set of features selected by machine learning that predict constant make sense from a physical perspective.

# ▩ DISCUSSION

The Rosetta Molecular Modeling Suite has been successfully used to guide the engineering wide range of enzyme functions. However, there has been a limited ability to benchmark its predictive ability for enzyme reengineering due to the lack of a large, kinetically quantitative, and uniformly collected dataset on the effects of mutations on kinetic parameters. Here, we construct the first large dataset of its kind for enzymes and report statistically significant evaluation of our ability to predict the functional effects of enzyme mutations.

The dataset generated here uncovered several new structure-function relationships in BglB, and provides the quantitative contribution towards catalysis of each amino acid in the active site. This systematic analysis revealed that several amino acids within the active site which are not directly involved in the reaction chemistry are as important to catalysis as the three residues which are directly involved in the chemistry. We observe that ten of the eleven positions in the active site that are conserved across the entire Pfam family by greater than 85% result in a greater than 100-fold decrease in activity when mutated to alanine (Supplemental Table X). This supports the widely held assumption that highly conserved residues within an enzyme active site are functionally important. However, within 12 Å of the active site only 11 of the 44 residues are greater than 80% conserved. When mutating each residue near the active site to alanine, only 19 mutations resulted in a decrease in catalytic efficiency of greater than 100-fold, and 10 mutations were not found to significantly affect catalytic efficiency. Based on this data there does not appear to be a strong correlation between the observed functional effect for mutations to residues conserved at less than 85% within the BglB active site. In addition, the mutation R240A, which is not observed in any natural variant in the glycosyl hydrolase 1 family, resulted in a 10-fold increase in *k*cat. This emphasizes the importance of not limiting design efforts to changes previously observed in nature when engineering function towards a non-natural substrate.

The large dataset of kinetic constants generated enabled the use of machine learning techniques which identified structural features correlated with function. It is interesting to note that the calculated interface energy was not found to be predictive of any kinetic parameter, and was not a feature identified in machine learning as predictive of function. This has significant implications for future design strategies since the calculated interface energy is one of the most common metrics used to for evaluating designs in the Rosetta Molecular Modeling Suite. It may be pertinent to develop training datasets, such as we have done here for BglB, in order to identify the appropriate metrics used for selecting designed mutants to functionally characterize in other enzyme systems.

While the dataset generated here enabled the development of a predictive modeling algorithm, it is unclear if the features selected by machine learning for BglB will be useful for prediction in other enzyme systems. More datasets of standardized kinetic constants are needed to determine if our results and the resultant machine learning based scoring function is applicable to every family 1 glycoside hydrolase, or even to other classes of hydrolase. Further work is needed to integrate the use of these large data sets into enzyme redesign algorithms to enable data driven design of novel enzymatic catalysts.

# ▩ CONCLUSION

In this work, over 100 computationally-designed mutants of a family 1 glucosidase were produced, purified, and kinetically characterized. This dataset revealed new insights into structure-function relationships in BglB. Machine learning protocols identified structural features closely correlated to kinetic properties. The development of this large data set allowed a statistically significant assessment of the Rosetta Molecular Modeling Suite’s ability to predict functional effects of mutations on this enzyme’s kinetic properties. We believe this type of data set will be invaluable for the development of computational enzyme engineering algorithms and providing insight into the physical basis of enzyme sequence-structure-function relationships.

# ▩ METHODS

## Molecular modeling for mutant selection

The crystal structure of recombinant BglB with the substrate analog 2-deoxy-2-fluoro-alpha-D-glucopyranose bound was used to identify the substrate binding pocket and the catalytic residues. Functional constraints were used to define catalytic distances, angles, and dihedrals among 4-nitrophenyl-ß-D-glucoside, E164, E353, and Y295. The structure was then loaded into Foldit, a graphical user interface to Rosetta. Point mutations to the protein were modeled and scored and those with reasonable energies (less than 5 Rosetta energy units higher than the native structure) were chosen.

Molecular models of each mutant enzyme were made using the Rosetta Molecular Modeling Suite by Monte Carlo optimization of total system energy.

## Mutagenesis, expression, and purification

The BglB gene was codon-optimized for *E. coli*, synthesized as a DNA String by Life Technologies, andcloned into a pET29b+ vector using Gibson assembly. Site-directed mutagenesis performed according to the method developed by Kunkel was used to generate mutations to BglB via the Transcriptic cloud laboratory platform. Variants were expressed and purified via immobilized metal ion affinity chromatography and assessed using 4-20% gradient SDS-PAGE Bolt Gels from Life Technologies.

## Kinetic characterization

The activity of the computationally designed enzyme variants was measured by monitoring the production of 4-nitrophenol. Mutant proteins ranging in concentration from 0.1 to 1.7 mg/mL were aliquotted in triplicate in 25 µL volumes and 75 µL of *p*-nitrophenyl-ß-D-glucoside (100 mM, 25 mM, 6.25 mM, 1.6 mM, 0.4 mM, 0.1 mM, or 0.02 mM) in enzyme storage buffer was added. Absorbance at 420 nm was measured every minute for 30-60 min and the rate of product production in M/min was calculated using a standard curve (see Supplemental Materials). A total of 2944 observed rates for 119 individual proteins (including biological replicates) were fit to the Michaelis-Menten equation using SciPy.

## Predictive modeling

Elastic net is a linear regression model that uses both l1 and l2 regularization for simultaneously enforcing sparsity and selecting the most informative features. To evaluate the prediction performance of the method, we used stratified 10-fold cross-validation together with bootstrap aggregating (bagging). Bagging was used to improve the stability and robustness of the predictor and entail in training 1,000 elastic net models with randomly drawn but stratified 10-fold cross-validation samples. The final three feature sets (one of each parameter to be estimated) were selected according to the averaged weight of each feature in all the 10,000 elastic net models (10 models per cross-validation, randomized 1,000 times). The weight of each selected feature in table 1 was normalized with respect to the weight with the largest absolute value. P-values were calculated based on the Wilcoxon signed-rank test after features and kinetic constants were normalized in the [0,1] interval. More information about the optimization and statistical procedure followed is available in Supplemental Materials.

ASSOCIATED CONTENT

**Supporting Information**. “This material is available free of charge via the Internet at http://pubs.acs.org.”

AUTHOR INFORMATION

Corresponding Author

\* jbsiegel@ucdavis.edu

Present Addresses

†If an author’s address is different than the one given in the affiliation line, this information may be included here.

Author Contributions

‡These authors contributed equally.

Funding Sources

IT is supported by ARO grant #201121557 and NSF grant #1254205. XKW is supported by CSC fellowship.

Notes  
Any additional relevant notes should be placed here.

ACKNOWLEDGMENT

(Word Style "TD\_Acknowledgments"). Generally the last paragraph of the paper is the place to acknowledge people (dedications), places, and financing (you may state grant numbers and sponsors here). Follow the journal’s guidelines on what to include in the Acknowledgement section.

ABBREVIATIONS

pNPG, *p*-nitrophenyl-ß-D-glucoside; IPTG, isopropyl β-D-1-thiogalactopyranoside

REFERENCES

(Word Style "TF\_References\_Section"). References are placed at the end of the manuscript. Authors are responsible for the accuracy and completeness of all references. Examples of the recommended formats for the various reference types can be found at <http://pubs.acs.org/page/4authors/index.html>. Detailed information on reference style can be found in The ACS Style Guide, available from Oxford Press.

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.