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

DA Carlin‡⊥, RW Caster‡†, XK Wang †, CV Ryklansky†, SA Betzenderfer†, CX Chen†, VM Duong†, 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

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 recapitulating experimentally determined functional effects of mutations. This is primarily due to the lack of data sets 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 mutants 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 tested result in the intended functional effect.

****

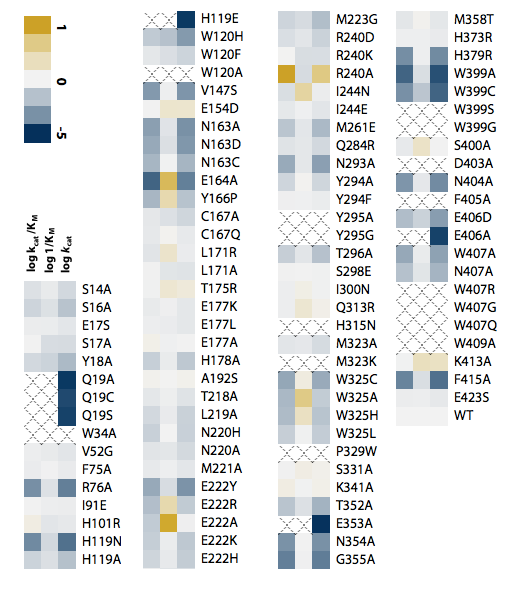
**Figure 1. Structure and catalyzed reaction of BglB** (A) Structure of BglB in complex with the modeled *p*-nitrophenyl-ß-D-glucoside utilized for the design efforts. Alpha carbons of residues mutated are shown as blue spheres (B) The BglB catalyzed reaction on p-nitrophenyl-ß-D-glucoside used to evaluate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

A primary factor that has limited development of improved enzyme redesign algorithms is the existence of large quantitative datasets correlating enzyme sequence and function. Large quantitative **data sets** exist for both protein-protein interfaces, as well as protein thermostability, and have played a critical role in evaluating and improving computational algorithms to accurately model and design novel functions. 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 >100 enzyme mutants, enabling both the assessment and potential improvement of modeling algorithms to evaluate enzyme structure-function relationships. This dataset is the first step towards developing an enzyme database equivalent to ProTherm, but relating enzyme sequence-function as opposed to protein sequence-thermostability. The enzyme we focused on developing as the first entry to this dataset is ß-glucosidase B from XXX (BglB), a family 1 glycoside hydrolase. Family 1 glycoside hydrolases have been the subject of numerous structural and kinetic studies due to their importance of being 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**.

Here, we report 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 4-nitrophenyl-ß-D-glucoside (pNPG). Greater than 90% of the designed mutations resulted in active and soluble protein. In the development of this dataset we discovered 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 potential path forward to develop and evaluate next generation enzyme reengineering algorithms.

****

**\*Change WT to BglB in Figure above\***

**Figure 2. Relative kinetic constants of 104 BlgB mutants**. The heatmap depicts the effect of the mutant 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 . 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 **\*\*\*Check to make sure true… were some soluble and just not active, if so we need to add a 4th column (maybe just +/-) or some way in indicating the protein was not able to be produced in a purified as 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 ± XXX M-1 min-1 for *k*cat/KM, *k*cat/KM, and *k*cat/KM respectively. A full table of kinetic constants and substrate vs. velocity curves for each is provided in the Supplemental Materials.

\*\*\*Double check WT numbers\*\*\*

# ▩ 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 [].

To establish a set of mutants to generate and kinetically characterize, two methods were used. 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 []. Each mutation was explicitly modeled and scored within 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 1 illustrates the positions throughout 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 site-specific mutants was generated via Kunkel mutagenesis and sequence verified using the Transcriptic cloud laboratory platform. Mutant plasmids were transformed into *E. coli* BLR(DE3) cells, protein expressed using IPTG based induction, and purified using immobilized metal affinity chromatography. After elution the 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 variance, 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. In Supplemental Figure [Yields] the distribution of yields for all 104 mutants are illustrated. 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 ± XXX M–1 min–1. Observed rates at 8 substrate concentrations were fit to the Michaelis-Menten equation using SciPy [Supp Fig XXX]. Kinetic constants for each mutant with the substrate pNPG are represented as a heatmap in **Figure 2**. In addition, a complete table with experimentally measured kinetic constants is reported in Supplemental Table 2 and the non-linear regression analysis fit to each measurement is reported in Supplemental Figure 2.

Based on the maximum concentration of enzyme used in our assays and colorimetric absorbance changes at the highest substrate concentration used we estimate our assays limit of detection for kcat/KM to be 10 M-1min-1. Of the 90 mutants purified as soluble protein, 6 were below the limit of detection. [WHICH WHERE THESE, ARE THEY A COMMON STRUCTURAL MOTIF… EVEN BOTH CATALYTIC RESIDUES HAD DETECTABLE ACTIVITY… LIKELY WORTH DISCUSSING BRIEFLY OUT HERE (and adding soluble protein in Figure will help)… MAYBE SIGNIFICATN STRUCTURAL CHANGES AND CRYSTAL NEEDED TO FIND OUT]. The highest catalytic efficiency observed is 5.6 x 105 M-1 min-1 for mutation R240A. In addition, while no substrate inhibition was observed on the wild type BglB, four mutants exhibited measurable substrate inhibition and are reported in Supplemental Table [SI].

## Sequence-structure-function relationships in BglB

In agreement with previous studies, our results clearly demonstrate the importance of the catalytic roles for E164, E353, and Y295. Mutating any of these residues to alanine results in more a more than 85,000-fold reduction in catalytic efficiency. However, the systematic alanine scan of every residue within 12 Å of the ligand 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 (F1GH), 1,554 non-redundant proteins, this glutamine is 95% conserved (Figure 4B). A structural analysis illustrates that both the nitrogen and oxygen of the amide are making interactions with the C-3 and C-4 sugar hydroxyls (**Figure 3**). While removing these interactions would be predicted to decrease catalytic efficiency, it was unexpected to observe a 57,000-fold reduction. This single mutation is equivalent to removing E353, which reduces activity 85,000-fold. However, unlike E353, the nucleophilic glutamate directly involved in the reaction chemistry, Q19 is only indirectly involved in the chemical reaction taking place. 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 XXX].

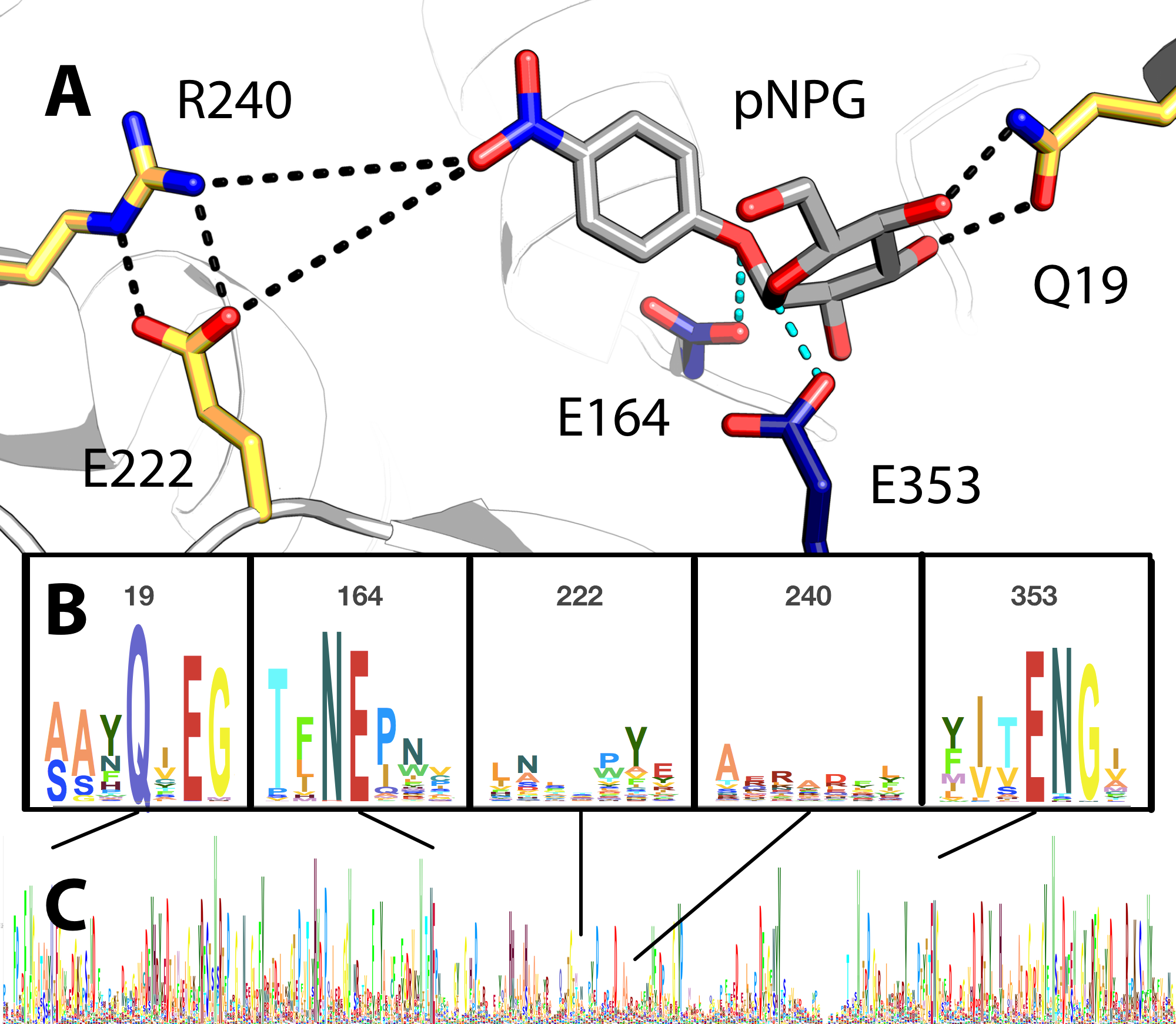


Figure 3. Active site model and conservation analysis of BglB (A) Docked model of pNPG in the active site of BglB showing established catalytic residues (navy) and a selection of residues mutated (gold). A multiple sequence alignment of the Pfam database’s collection of 1,554 family 1 glycoside hydrolases 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.

Make numbers in (B) bigger!

Another unexpected finding was a ten fold increase of *k*cat by a single point mutant, R240A. The BglB 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 substantially closer to the active site (Supp Figure XXX), resulting in a significant change of the electrostatic environment in the active site. In addition, the mutation E222A decreases *k*cat by ten fold. Both observations support the previously proposed hypothesis that the electrostatic environment of the enzyme active site is of primary importance to catalysis [CITE WARSHEL].

### 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 simulations protocol used approximates the numerous protocols previously used in successful enzyme reengineering efforts. Briefly, the algorithm used was a Monte Carlo protocol, with random perturbation of the ligand followed by functional constraint optimization through rigid body minization of the ligand and side chain, side chain repacking, and side chain and backbone minimization.. The command line and an example set of input files for wild type BglB are provided in Supplemental [Code].

For each mutant, we generated 100 models and selected for subsequent structural analysis the 10 models that had the lowest overall system energy. From this analysis, a set of 59 potentially informative features was identified and their value was calculated for each mutant. These included metrics of the predicted interface energy, number of hydrogen bonds to the ligand, and the 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 X. For *k*cat/KM and *k*cat the highest correlation observed is to the total number of non-local contacts (i.e. residues implicated in molecular interactions that are separated by more than 8 amino acid residues in sequence), with a Pearson Correlation Coefficient (PCC) of 0.56 (p-value < X. Wilcoxon test) and 0.43 (p-value < X. Wilcoxon test), respectively. For KM the highest PCC is 0.29 (p-value < X. Wilcoxon test) to the total number of hydrogen bonds within each BglB model. The Spearman Rank Correlation (SRC) follow similar trends to those for PCC for all three predicted constants (X,Y and Z for *k*cat/KM , *k*cat and KM respectively). The PCC and SRC values with respect to the kinetic constants for all features are available in supplemental file X.

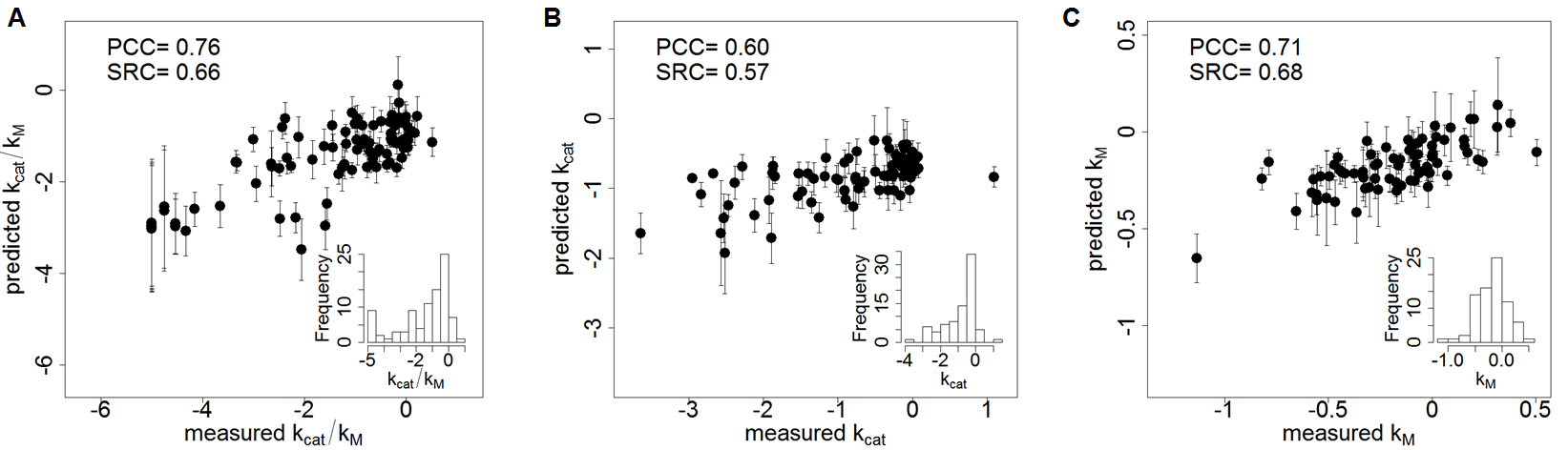


Figure 4. Correlation between machine learning predictions and experimentally-determined kinetic constants. The log value corresponding to the relative *k*cat/KM (A), *k*cat (B), and 1/KM (C) for each mutants experimentally-determined kinetic constants (equivalent to the values depicted in Figure 2) 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 by1,000 times cross validation for each datapoint. 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).

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. A set of 59 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 all features that have been assigned non-zero weights during training (9 for *k*cat/KM, 8 for *k*cat, 10 for KM). Since the value of each feature is normalized, its contribution to the kinetic constant prediction is indicative of its weight in determining the kinetic constant value (columns 1-3). Column 4 provides a description of each feature, and column 5 and 6 are the value ranges in the training dataset. The full feature table is available in Supplemental Table X. 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 | Packing of the system with pNPG | 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 total 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 around E353 without pNPG | 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 |

Next, 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 (ref), a constraint regression technique that uses both l1 and l2 regularization for feature selection (see Supplementary Materials). To increase robustness to sample size and remove bias, we used a bootstrapping aggregating (bagging) technique, where the predicted value was an average from 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 (PCC X vs Y±std for Ensemble and individual model, respectively; SRC X vs Y±std for Ensemble and individual model, respectively). As shown in Figure 4, the model predictions have a strong correlation with experimentally-measured values albeit with variability that differs for each constant but is consistent across all samples (X, Y, and Z the average std for all samples in the case of *k*cat/KM , *k*cat and KM respectively)

For each of the kinetic constant that is to be predicted, only 8-10 features were selected by the method to have non-zero weights (**Table 1** and Supp Table XXX). Interestingly, the automatically selected features by the learning method are also sound from a structural/chemical perspective. For example, the set of features identified that correlate to *k*cat and *k*cat/KM is the calculated hydrogen bonding energy between BglB and pNPG as well as the number of hydrogen bonds to pNPG. However, these metrics were found to oppositely correlate indicating that a few numbers of strong hydrogen bonds promotes catalytic efficiency. Overall, metrics correlated to *k*cat were found to be co-correlated with kcat/KM, and are mostly measures of hydrogen bonding. Conversely, structural features correlated to KM are predominantly packing statistics.

# ▩ 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 enzymes kinetic parameters. Here we construct the first large dataset of its kind for enzymes, enabling statistically significant evaluations of the ability to predict the functional effects of enzyme mutations.

The dataset generated here uncovered several new structure-function relationships for BglB, and for each amino acid in the active site provides its quantitative contribution towards catalysis. 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 over the entire Pfam family by >85% result in >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 eleven of the fourty-four residues are conserved at >85%. When mutating each residue spatially located near the active site to alanine only 19 resulted in a decrease in catalytic efficiency of >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 selected in machine learning as predictive of function. As the calculated interface energy is one of the most common metrics used to for evaluating designs in the Rosetta Molecular Modeling Suite this has significant implications on future design strategies. Specifically, 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.

While the data set generated here enabled the development of a predictive modeling algorithm, it is unclear of the same features utilized in the modeling algorithm will be consistent in other enzyme systems. More data sets of standardized kinetic constants is 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 hydrolases. In addition, further work is needed to directly integrate the utilization of these large data sets directly into enzyme redesign algorithms, enabling data driven design algorithms.

# ▩ 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 the structure-function relationship of BglB. In addition, it allowed a statistically significant assessment of the Rosetta Molecular Modeling Suite’s predictive ability for evaluating the effects of mutations on an enzyme’s kinetic properties. Finally, by implementing machine learning protocols on this large dataset we identified structural features closely correlated to kinetic properties. We believe this type of data set will be invaluable for the future 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.

<A little bit on modeling the ensemble of structures used for ML>

## Mutagenesis, expression, and purification

The BglB gene was codon optimized for *E. coli* and synthesized as a DNA String by Life Technologies. It wascloned 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 through the Transcriptic Cloud 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 XXX-YYY were aliquotted in triplicate in 25 µL volumes and 75 µL of 4-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 was measured every minute for 30-60 min at a wavelength of 420 nm 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 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. More information about the optimization and statistical procedure followed is available in Supplementary 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. XW is supported by X 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

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.