A library of over 100 mutant enzymes to discover structural features correlated with kinetic constants

DA Carlin‡⊥, RW Caster‡†, SA Betzenderfer†, CX Chen†, VM Duong†, CV Ryklansky†, X Wang †, A Alpekin†, N Beaumont†, RJ Boyd†, H Kapoor†, N Kim†, JE Lucas†, 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: Computational enzyme design has had enormous success in the development of novel catalysts that perform desired chemical reactions which have no known biological catalyst []. At the forefront of computational design methodologies has been the Rosetta Molecular Modeling Suite [Mak]. However, enzyme design methods developed in this suite are currently developed and benchmarked around 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. kcat 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, current design efforts often require screening hundreds of designs, only a small number of which introduce the intended functional effect. This has led to significant efforts to improve the design and modeling protocols to improve predictive capabilities.

****

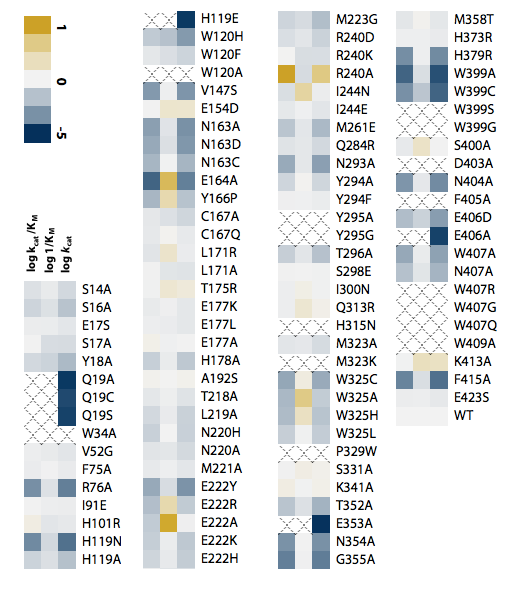
**Figure 1. Structure and catalyzed reaction of BglB** (A) Structure of BglB in complex with a modeled *p*-nitrophenyl-ß-D-glucoside. Alpha carbons of residues mutated shown as blue spheres (B) Reaction of BglB on p-nitrophenyl-ß-D-glucoside

A primary factor that has limited development of enzyme redesign algorithms is the existence of large quantitative datasets correlating 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. Mutant enzyme libraries are commonly screened for functional effects with measuring protein concentration or the effects of the mutations on the intrinsic kinetic constants of the enzyme. From large library screens often only one to ten mutants are characterized at the level of Michaelis-Menten constants.

Due to the lack of quantitative sequence-function datasets for enzymes, most 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 which 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. In essence, 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, 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 (in natural systems, another sugar) is protonated by E164. 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 (kcat, 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 are able to report the first analysis of the ability to predict effects on kcat, KM, and kcat/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 enzyme catalysis as well as a potential path forward to develop and evaluate next generation enzyme reengineering algorithms.

****

**Figure 2. Kinetic constants for 104 mutants of BglB relative to native enzyme**. The heatmap is colored gold for a higher value compared to wild type kinetic constants and blue for a lower value. Values are on a log scale and the ranges are as follows: 10–11,000 min-1 (kcat), 0.6–85 mM (KM), and 10–560,000 M-1min-1 (kcat/KM) with wild type constants 880 ± 10 min–1 (kcat), KM of 5 ± 0.2 mM, and kcat/KM of 171,000 ± 0.05 1/M\*min.

# ▩ 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 stabilizes the attacking species. 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 A of the ligand in our model was individually mutated to alanine. In the second method, mutations predicted to be energetically favorable by the program Foldit were selected. Each mutation was explicitly modeled and scored within Foldit and a selection of mutations that did not increase the energy of the 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.

Of the 104 mutants synthesized, 90 were found to be expressed and purified as soluble protein. 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 0.5 ± 0.3 mg/mL. 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 (0.2 mg/mL).

### 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 kcat of 880 ± 10 min–1, an average KM of 5 ± 0.2 mM, and kcat/KM of 171,000 ± 0.05 1/M\*min. Kinetic constants for each mutant with the substrate pNPG are represented as a heatmap in Figure 2. Observed rates at 8 substrate concentrations were fit to the Michaelis-Menten equation using SciPy.

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 solubly purified mutants, 6 are below this limit of detection. 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 important 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 A 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-5 and C-5 sugar hydroxyls (Figure 4A). 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 his mutant is predicted. Drawings of the lowest-energy models for this mutant can be found in Supplemental X.

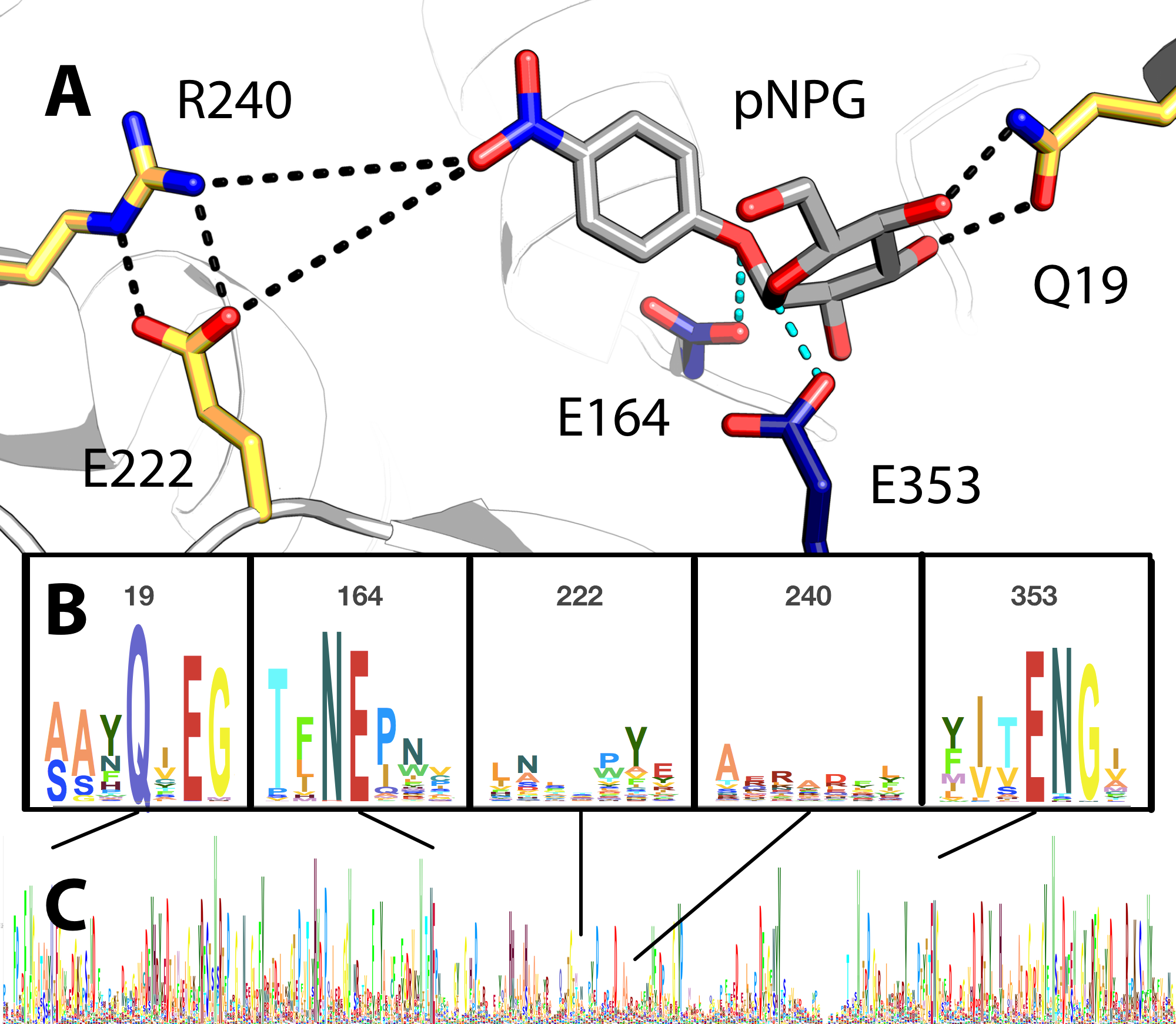


Figure 3. Active site model of BglB (A) Docked model of pNPG in the active site of BglB showing established catalytic residues (navy) and a selection of residues mutated (gold). Multiple sequence alignment of the Pfam database’s collection of 1,554 family 1 glycoside hydrolases with (B) selected regions enlarged and (C) over the whole BglB coding sequence.

Another unexpected finding was an increase in the rate of catalysis by ten fold by a single point mutant, R240A. R240 is not predicted to make a direct molecular interaction with the substrate, so it is unclear by what mechanism this mutation increased the rate of catalysis by ten fold while increasing the pseudo–dissociation constant *KM* about 4 fold. The crystal structure reveals that the sidechain of R240 forms two hydrogen bonds with the sidechain E222. Mutation of R240 to alanine, removing these interactions, is predicted by molecular modeling to stabilize an alternate rotamer of E222, with the acid functional group of the glutamate moving closer to the active site (Figure XXXB). This will result in a significant change in the electrostatic environment of the active site. Further supporting the hypothesis that the electronegative environment in this region of the active site is important to catalysis is that the mutation E222A decreases kcat 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 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 and side chain repacking. The command line and an example set of input files for wild type BglB are provided in Supplemental [Code].

The lowest 10% in total energy models were selected to represent each mutant, and 59 structural metrics for each mutant were calculated (for packing, for hydrogen-bonding, polar contacts, entire system as well as specific to ligand and catalytic residues). 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 is given in Supplemental X The value for each calculated metric was averaged for the 10 structures and evaluated for its correlation to each experimentally measured kinetic constant (Supplemental Table XXX—PCC of each individual Feature vs. Constant).

The highest correlation observed is a PCC of 0.56 and 43 for kcat/Km and kcat, respectively, with the total number of non-local contacts (i.e. residues making molecular interactions that are separated by more than 8 amino acid residues in sequence). For KM the most significant PCC is 0.29 to the total number of hydrogen bonds within each BglB model. Given the relatively low PCC correlation coefficients to between any single calculated feature and measured kinetic constant we explored the use of machine learning to determine if a combination of features could recapitulate the experimental data.

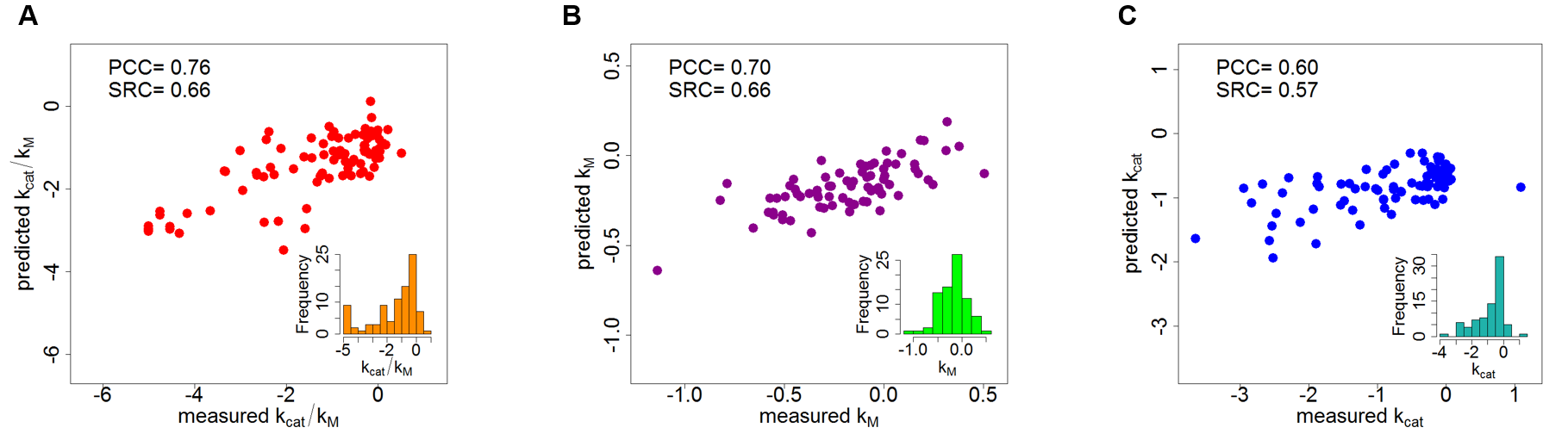
[TALK ABOUT MACHINE LEARNING RESULTS] The correlations between calculated structural features and experimentally-determined values of kcat, KM, and kcat/KM for each mutant are given in Figure 4.

For each of the kinetic constant 6-9 features were found that significantly improved functional predictions from the calculated features. An example feature identified that correlated to kcat and kcat/KM by machine learning corresponds to computational estimates of the hydrogen bonding energy between BglB and pNPG. Metrics correlated to kcat were found to be co-correlated with kcat/KM, and include a count of the number of hydrogen bonds, which is negatively correlated, and an estimation of the hydrogen bond energy, which is positively correlated, indicating that catalysis is predicted by the strength of a small number of strong hydrogen bonds rather than by their quantity. Conversely, structural features correlated to KM are predominantly packing statistics, with the addition of Lennard-Jones repulsion for Y295. Of the 59 metrics assessed, 9 were found to correlate significantly with kinetic constants. Complete weights for each feature are given in Supplemental Table X.

Table 1. Descriptions of the structural features identified by machine learning and weights with which they contribute to predictive ability. One hundred models of each mutant protein were generated and scored on a set of 59 common structural features currently used for enzyme design. An elastic net algorithm identified features with significant correlation to the kinetic constants kcat/KM, KM, and kcat and assigned a weight to each, columns 1–3. A description of each feature is included in column 4, and ranges observed are in included in columns 5 and 6. A full table of features is available in Supplemental Table XXX.

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

*nc = no correlation.*



\*\*\*PLEASE CHANGE TO BLACK AND WHITE… JUST DID RECOLOR HERE SO NOT AS PRETTY AS COULD BE\*\*\* 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 are shown on the *y* axis. All values are normalized relative to wild type BglB and are in log scale. Inset Pearson and Spearman correlation coefficients demonstrate the strength of the predictions, and histograms display the frequency of each experimentally-determined value in the data set used for machine learning.

# ▩ DISCUSSION

The Rosetta Molecular Modeling Suite has been successfully used to direct the engineering of over 30 enzymes. 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 observed that of the 11 positions in the active site that are conserved over the entire Pfam family by >85%, 10 of them resulted 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 11 of the 44 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. The metric with the best correlation to overall efficiency was hydrogen bonding energy of the substrate. While the metrics identified by machine learning are consistent with chemical principles, more data sets of standardized kinetic constants on enzyme classes beyond the glycoside hydrolase family will be needed to determine if these features are system-specific or generally correlated with function among many enzyme classes.

Currently, due to lack of large data sets of kinetic constants, it is unclear whether 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 develop large data sets of kinetic constants to further evaluate enzyme redesign algorithms. It will be desirable in future to kinetically characterize large numbers of enzyme variants to enable data driven design algorithms.

# ▩ CONCLUSION

In this work, over 100 computationally-designed mutants of a family 1 glucosidase were produced, purified, and kinetically characterized. To the best of our knowledge, this is the largest data set of mutant enzymes produced and kinetically characterized in a uniform manner to date. 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.

## Mutagenesis, expression, and purification

A sequence coding for BglB was synthesized by Life Technologies as a DNA String codon optimized for E. coli and cloned 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, and variants were expressed and purified via immobilized metal ion affinity chromatography.

## Kinetic characterization

The activity of the computationally designed enzyme variants was measured by the appearance of the highly colored product 4-nitrophenol (see Scheme 1). Mutant proteins 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 rate termed kobs (1/min) was calculated by dividing the rate observed (1/min) by the enzyme concentration (M).

### Statistical analysis

Data including 2944 observed rates for 119 individual proteins (including replicates) were fit to the Michaelis-Menten equation using SciPy.

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

Any funds used to support the research of the manuscript should be placed here (per journal style).

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.