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‡†, XK Wang *†, SA Betzenderfer†, CX Chen†, VM Duong†, CV Ryklansky†, 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 address this by constructing a dataset of over 100 mutant enzymes, each of which were produced, purified, and kinetic constants (i.e. k_{cat} and K_M) measured. We illustrate the importance of this type of data for the 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 tested have the intended functional effect.

A major hindrance to improved enzyme redesign algorithms is the lack of large quantitative datasets correlating enzyme sequence and function. Large quantitative datasets exist for both protein–protein interfaces and protein thermostability, and have played a key 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 for each mutant. Due to the lack of quantitative sequence-function datasets for enzymes, many efforts to develop and evaluate 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 neutral or beneficial to function being assessed.

We aimed to address this by determining kinetic constants for >100 enzyme mutants, which provides a large enough dataset to

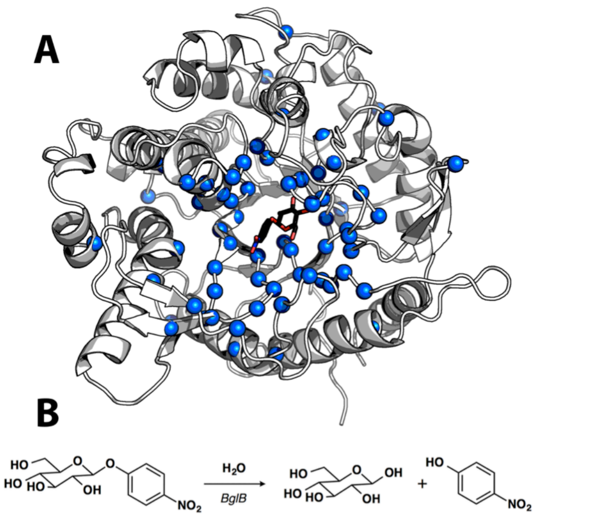

Figure 1. Structure and catalyzed reaction of BglB (A) Structure of BglB in complex with the modeled *p*-nitrophenyl-ß-D-glucoside used for design. Alpha carbons of residues mutated shown as blue spheres (B) The BglB–catalyzed reaction on *p*-nitrophenyl-ß-D-glucoside used to evaluate kinetic constants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

use machine learning approaches to select calculated structural features correlated to function. 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**.

Here, we report a large dataset of kinetic constants of 104 computationally designed variants of BglB, each of which was produced, purified, and kinetic constants (k_{cat}, K_M, K_i) measured using the reporter substrate *p*-nitrophenyl-ß-D-glucoside (pNPG). 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}, K_M, and k_{cat}/K_M using molecular modeling. Finally, we illustrate how machine learning can be used to identify structural features from the molecular models that significantly improve the predictive accuracy of the molecular modeling. These analyses provide insight into the factors important for catalysis in BglB as well as a path forward 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 S_N2-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 S_N2-like mechanism.

Two approaches were used to establish a set of mutants to generate and kinetically characterize. The first approach 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 approach, 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. 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 on the Transcriptic cloud laboratory platform and sequence-verified. Mutant plasmids were transformed into *E. coli* BLR(DE3), and 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 this 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 (**Figure 2**). The distribution of yields for all 104 mutants is illustrated in Supplemental Figure X. Greater than 35% maintained the yields obtained for native BglB, and 15% were not expressed and purified as a soluble protein above our limit of detection (0.1 mg/mL) based on A280 and SDS-PAGE.

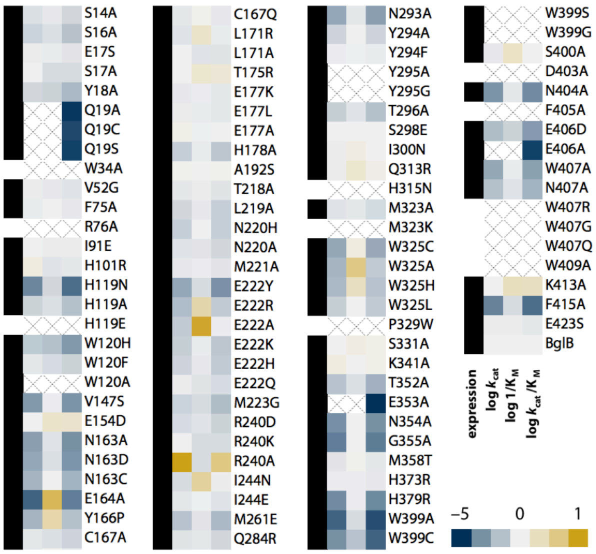

Figure 2. Log scale relative kinetic constants of 104 BglB mutants. The heatmap depicts the effect of the mutation on each kinetic constant relative to native BglB. As indicated in the color legend, gold is for higher value and blue for a lower value. If the kinetic constant was not measurable, an X is depicted in the box. Proteins that were expressed as soluble protein with a purification yield of >0.1 mg/mL, and validated by SDS-PAGE are labeled with a black box in the first column. Those below our limit of detection of 0.1 mg/mL are labeled with an empty box. Values are on a log scale and the ranges are as follows: 10–11,000 min^{-1} (k_{cat}), 0.6–85 mM (K_M), and 10–560,000 $M^{-1}\text{min}^{-1}$ (k_{cat}/K_M) 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}/K_M, k_{cat}/K_M, and k_{cat}/K_M respectively. A full table of kinetic constants and substrate versus velocity curves for each are provided in the Supplemental Materials.

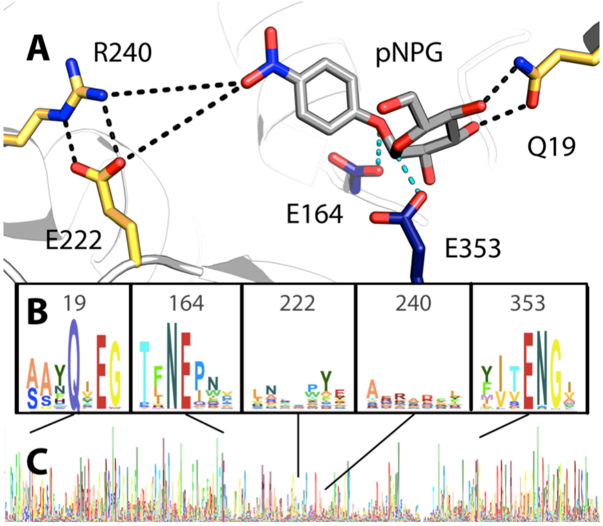

Figure 3. Active site model and conservation analysis of BglB (A) Docked model of pNPG in the active site of BglB showing established catalytic residues (navy) and a selection of residues mutated (gold). A multiple sequence alignment of the Pfam database's collection of 1,554 family 1 glycoside hydrolases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

Kinetic characterization of mutants

Michaelis-Menten kinetic constants for each of the 104 mutants were determined using the colorimetric assay of pNPG hydrolysis and are represented as a heatmap in **Figure 2**. Ten biological replicates of the wild type enzyme have an average k_{cat} of 880 ± 10 min^{-1}, K_M of 5 ± 0.2 mM, and k_{cat}/K_M of 171,000 ± 8000 M^{-1} min^{-1}. To determine kinetic constants, observed rates at 8 substrate concentrations were fit to the Michaelis-Menten equation. Experimentally measured kinetic constants for each mutant and non-linear regression analyses can be found in Supplemental Table X.

Based on the maximum concentration of enzyme used in our assays and colorimetric absorbance changes at the highest substrate concentration used, we estimate our limit of detection for k_{cat}/K_M to be 10 $M^{-1}min^{-1}$. Of the 90 solubly purified mutants, 6 were below the limit of detection. The highest catalytic efficiency observed is 5.6 x 10^5 $M^{-1}min^{-1}$ for mutation R240A. In addition, while no substrate inhibition was observed on the wild type BglB, four mutants exhibited measurable substrate inhibition (the inhibition parameter K_i for these mutants is reported in Supplemental Table X).

Observed sequence–structure–function relationships in BglB

In agreement with previous studies, our results demonstrate the importance of E164, E353, and Y295 for catalysis. Mutating any of these residues to alanine results in a >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. A structural analysis illustrates that both the nitrogen and oxygen of the amide interact with the substrate's sugar hydroxyl groups (**Figure 3A**). 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 3B**). While removing these interactions might be predicted to decrease catalytic efficiency, it was unexpected to observe a 57,000-fold reduction. The mutation Q19A is almost 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 chemistry of the reaction. 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 (Supplemental Figure X).

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 3A**). 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 (Supplemental Figure X), resulting in a significant change of the electrostatic environment there. 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 [Warshel].

Computational modeling and evalutation of predictive ability

In order to evaluate the Rosetta Molecular Modeling Suite's ability to evaluate the functional effects of mutations on BglB kinetic properties, molecular models were generated for each of the 104 BglB mutants. 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 through rigid body minimization of the ligand and sidechains, sidechain repacking, and sidechain and backbone minimization. [] An example set of input files for wild type BglB are provided in the Supplemental Materials.

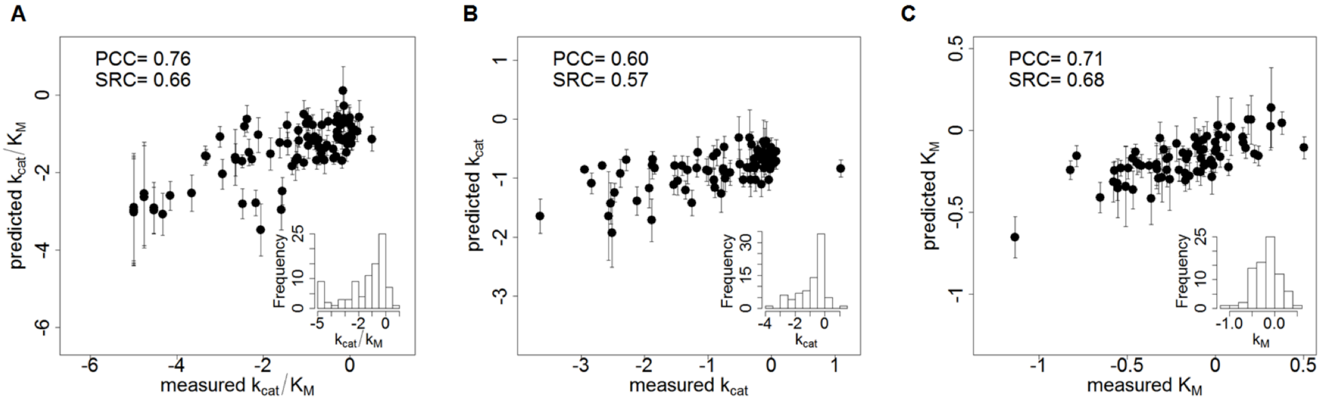

Figure 4. Correlation between machine learning predictions and experimentally-determined kinetic constants. The log value corresponding to the relative k_{cat}/K_M (A), k_{cat} (B), and $1/K_M$ (C) for each mutant's experimentally-determined kinetic constants (equivalent to the values depicted in Figure 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 by 1000-fold cross validation for each point. 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}/K_M, k_{cat}, and K_M, respectively).

For each mutant, 100 models were generated as described above and the lowest 10 in overall system energy were selected for subsequent structural analysis. A value for each of 59 potentially informative features was calculated for each model. Structural features included predicted interface energy, number of hydrogen bonds between protein and ligand, and change in solvent accessible surface area upon ligand binding. Correlation of the calculated structural features to each kinetic constant was assessed using the Pearson Correlation Coefficient (PCC) and Spearman Rank Correlation (SRC). For both k_{cat} and k_{cat}/K_M, the strongest correlation observed is to the total number of non-local contacts (count of residues separated by more than 8 sequence positions that interact with each other), with a PCC of 0.56 (p-value 0.009; Wilcoxon test) and 0.43 (p-value 0.004; Wilcoxon test), respectively. For K_M, the highest PCC is 0.29 (p-value 0.0005; Wilcoxon test) to the total number of hydrogen bonds in each BglB model. The SRC follows similar trends to those for PCC for all three predicted constants (largest SRC of 0.55, 0.42

Table 1. Most informative structural features for kinetic constant prediction. For each mutant, 10 out of 100 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 features that have been assigned non-zero weights during training (9 for k_{cat}/K_M, 8 for k_{cat}, 10 for K_M). 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 X. *ns=feature not selected by the algorithm*

k_{cat}/K_M	k_{cat}	$1/K_M$	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	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 pNPG 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

and 0.38 for k_{cat}/K_M, k_{cat} and K_M respectively). The PCC and SRC values for all features are available in Supplemental Table X.

Machine learning prediction of kinetic constants

Since no single structural feature predicts k_{cat}, K_M, or k_{cat}/K_M with PCC > 0.7, machine learning was used to identify a subset of calculated features correlated to observed kinetic constants. Elastic net regularization, a constraint regression technique that uses both l_1 and l_2 regularization for feature selection was used to identify 8–10 structural features and a regularized linear model that predicted each kinetic constant. To increase robustness to sample size and remove bias, we used a bootstrapping aggregating (bagging) technique, where the predicted value was an average of 1000 elastic net models, each trained on a different subset of the data (see Methods). The final prediction from this ensemble learning regression method outperformed single feature selection for each kinetic constant. For k_{cat}/K_M, the PCC increased to 0.76 from 0.56, in the case of k_{cat} to 0.60 from 0.56, and for K_M to 0.71 from 0.29. **Figure 4** illustrates the correlations between machine learning predictions and experimentally-measured values.

The primary features found to correlate to $1/K_M$ are metrics of protein packing without the ligand present. Interestingly, all of these packing features are positively correlated to $1/K_M$. While K_M is a complex metric, the substrate binding affinity is expected to have a significant effect on its value. Therefore, this suggests that BglB requires room around each catalytic residue and the entire protein in order to optimally accommodate K_M. This would be indicative that BglB operates through an induced fit mechanism. However the induced structural changes must be small since the RMSD between the apo and transition state analogue bound forms of the proteins is <0.2 A.

The features selected by the algorithm as predictive of k_{cat} include a count of polar contacts, consistent with our understanding that BglB stabilizes the accumulated positive charge on the oxocarbenium ion in the proposed transition state [Withers]. Further supporting this hypothesis is the selection of a ligand burial term (change in solvent accessible surface area on binding). Taken together with the finding that residues such as E222 which do not make direct molecular interactions with the substrate nonetheless play a key role in catalysis, the selection of these features by the machine learning algorithm is consistent with the finely-tuned electrostatic environment of the BglB active site being of primary importance for catalysis.

In BglB, the most informative feature predicting k_{cat}/K_M is the calculated hydrogen bonding energy of the substrate. The identification of this feature indicates the importance of protein–ligand hydrogen bond interactions in positioning the substrate and the protein sidechains for catalysis ("orbital steering") []. This is further supported by the finding that the mutation Q19A, which removes two hydrogen bond interactions between Q19 and pNPG is equivalent to the catalytic knockout E353A in reducing k_{cat}/K_M.

It is interesting to note that the only significant features correlated to all three kinetic constants are total system metrics, while features correlated to only one kinetic constant were metrics capturing a particular aspect (e.g. packing or hydrogen bond energy) of the BglB structure. It is also interesting that the most common metric used for successful enzyme designs, interface energy [], was not selected by the algorithm to be predictive of any kinetic constant.

▩ 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 of the effects of mutations on kinetic parameters. Here, we construct the first such dataset 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 two residues which are directly involved in the chemistry.

Of the 44 positions in the active site systematically mutated to alanine, 11 are conserved by >85% in amino acid identity. When any one of these amino acids is mutated to alanine, catalytic efficiency decreases >100-fold (Supplemental Table X). This supports the widely held assumption that highly conserved residues within an enzyme active site are functionally important. However, only 11 of the 44 residues within 12 Å of the active site are >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 these findings, there does not appear to be a strong correlation between conservation and function if conservation is <85%. Finally, 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 by machine learning as predictive of function. This has significant implications for future design strategies since the interface energy is one of the most common metrics used to evaluate enzyme designs. It may be pertinent to develop training datasets, such as we have done here for BglB, in order to identify the appropriate metrics to be used for selecting designed mutants to functionally characterize in other enzyme systems.

While the dataset generated here enabled the development of a machine learning–based scoring function, 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 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. This 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.

Mutagenesis, expression, and purification

The BglB gene was codon-optimized for *E. coli*, synthesized as a DNA String by Life Technologies, 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 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

One hundred molecular models of each mutant enzyme were made using the Rosetta Molecular Modeling Suite by Monte Carlo optimization of total system energy and the lowest 10 selected for feature generation. Elastic net regularization was used to select the most informative features. To evaluate the prediction performance of the method, stratified 10-fold cross-validation together with bootstrap aggregating (bagging) was used. 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 X.

ASSOCIATED CONTENT

Supporting Information. A full list of mutations selected, the distribution of yields for all 104 mutants, experimentally measured kinetic constants for each mutant, nonlinear regression analyses, the inhibition parameter K_i for mutants exhibiting substrate inhibition, models of Q19A and R240A, an example set of Rosetta input files for wild type BglB, and PCC and SRC values for all features are included as supporting information. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

jbsiegel@ucdavis.edu

Author Contributions

‡These authors contributed equally.

ACKNOWLEDGMENT

This work was supported by ARO #201121557 and NSF #1254205 (IT) and X and Y (JBS).

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.