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

Dylan Alexander Carlin[1]*, Ryan W. Caster[2]*, Xiaokang Wang[3], Stephanie A. Betzenderfer[2], Claire X. Chen[2], Veasna M. Duong[2], Carolina V. Ryklansky[2], Alp Alpekin[2], Nathan Beaumont[2], Harhul Kapoor[2], Nicole Kim[2], Hosna Mohabbot[2], Boyu Pang[2], Rachel Teel[2], Lillian Whithaus[2], Ilias Tagkopoulos[2,6], Justin B. Siegel[2,3,4]

Author affiliations:

1. Biophysics Graduate Group, University of California, Davis

2. Genome Center, University of California, Davis

3. Department of Chemistry, University of California, Davis

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

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

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

ABSTRACT

The use of computational modeling algorithms to guide the design of novel enzyme catalysts is a rapidly growing field. Force-field based methods have now been used to engineer both enzyme specificity and activity, however successful hit rates are often under ten percent. One potential reason for this is that current force-field based approaches are frequently trained using indirect measures of function rather than direct correlation to experimentally-determined functional effects . This is due

to the lack of datasets for which a large panel of enzyme variants has been produced, purified, and kinetically characterized. Here we report the k_{cat} and K_M values of over 100 purified mutants of a glycoside hydrolase enzyme. We demonstrate the utility of this data set by using machine learning to train a new algorithm that enables prediction of each kinetic parameter based on readily modeled and calculated structural features. 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 using computational approaches has the potential to enable rapid development of highly efficient and specific catalysts tailored for needs beyond those selected for during natural evolution.[1] A growing route for engineering enzyme catalysts is the use of computational tools to evaluate potential mutations *in silico* prior to experimental characterization. Using the Rosetta Molecular Modeling Suite, reengineering of both specificity and chemistry has been accomplished.[2,3,4,5,6] However, using this force-field based approach only a relatively small number of all designs tested have the intended functional effect. Furthermore, there have been no reports evaluating the predictive power of the Rosetta Molecular Modeling Suite on the functional effects of enzyme mutations. Therefore efforts to both evaluate and improve the predictive

power of this computationally inexpensive and widely accessible algorithm is of the utmost importance.

One likely reason for the low success rate of designs is that current algorithms are not directly trained on experimentally measured effects, but are trained on indirect measures such as sequence recovery (*i.e.* the ability to recapitulate a known active site after running a design simulation). These indirect measures are used because there are no reported datasets of kinetically characterized enzyme mutants that encompass a wide dynamic range of activity and have enough independent data points to enable training and cross validation of algorithms. Yet, there is significant precedence in the use of large datasets to train and evaluate force-field based algorithms for protein function. A closely related example is the ProTherm database which has over twenty thousands of measured effects of mutations on thermostability.[7] This database is the gold standard used for the development of numerous algorithms for predicting effects of mutations on thermostability.[8,9]

Here, we take the first key step towards developing a data set of enzyme mutants with measured effects on kinetic constants (*i.e.* k_{cat} and K_M) that is both large enough and has a wide enough dynamic range to enable training of computational protein design algorithms. The initial enzyme of focus is a family 1 glycoside hydrolase: ß-glucosidase B (BglB) from *Paenibacillus polymyxa*. The 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.[10] An X-ray crystal structure of BglB indicates that it 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.[10] The protein structure and reaction scheme are provided in Figure 1.

In this study we report the largest data set of its kind, in which 104 mutants of BglB are produced, purified, and kinetically characterized (*i.e.,* 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 use this dataset to train computational algorithms for the prediction of k_{cat}, K_M, and k_{cat}/K_M using metrics derived from force-field based 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 (PDB 2JIE) 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 approximating 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 (See supplemental information for details).

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.[4,11] 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[12] using the Transcriptic cloud laboratory platform and sequence-verified. Mutant plasmids were transformed into *Escherichia coli* BL21(DE3), and after expression 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, and fresh resin was used for each mutant to prevent wild type contamination.

A total of ten biological replicates of the native BglB were used to assess expression and purification. The average yield is found to be 1.2 ± 0.4 mg/mL. Of the 104 mutants synthesized, 90 are found to express and purify as soluble protein (Figure 2). The yields for all 104 mutants are included in Supplemental Table 1. Greater than 35% maintained the yields obtained for native BglB, and 15% are 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 and are represented as a heatmap in Figure 2. Ten biological replicates of the wild type enzyme has 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 and nonlinear regression analysis for each mutant can be found in Supplemental Table 1.

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 are below the limit of detection. The highest catalytic efficiency observed is 560,000 $M^{-1}min^{-1}$ for mutation R240A. In addition, while no substrate inhibition is observed for the wild type BglB, four mutants exhibit measurable substrate inhibition (the inhibition parameter K_i for these mutants is reported in Supplemental Table 1).

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 (k_{cat}/K_M). However, beyond the catalytic residues, the systematic alanine scan of every residue within 12 Å of the ligand revealed mutations which have an equivalent functional effect to mutating the established catalytic residues to alanine.

One mutation for which dramatic effects on function is observed is Q19A, which decreases catalytic efficiency 57,000-fold. An analysis of the crystal structure of BglB illustrates that both the nitrogen and oxygen of the amide sidechain interact with hydroxyl groups on the substrate (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 that this mutation has an almost equivalent effect to removing the established catalytic residue E353. 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-α-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 2A).

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 2 Å closer to the active site (Supplemental Figure 2B). This would result in a significant change of the electrostatic environment around the active site, suggesting that the electronegative environment enhances catalysis of pNPG hydrolysis. Consistent with this hypothesis is the observation that the mutation E222A decreases k_{cat} by ten fold. Both observations support the previously proposed hypothesis that the electrostatic environment of an enzyme active site is of primary importance to catalysis.[13]

Conservation analysis of the BglB active site

Of the 44 positions in the active site systematically mutated to alanine, 11 are conserved by >85% in amino acid identity with respect to 1,554 homologues in the Pfam database. When any one of these amino acids is mutated to alanine, catalytic efficiency decreases >100-fold (Supplemental Table 1). 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 >85% conserved. Of the 33 residues within 12 Å of the active site that are <85% conserved, only 8 alanine mutations resulted in a decrease in catalytic efficiency of greater than 100-fold, and 10 of these 33 mutations were not found to significantly affect catalytic efficiency.

Based on these findings, there does not appear to be a strong correlation between residue identity and function if a particular residue is <85% conserved. 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.

Computational modeling and evaluation 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. A Monte Carlo simulation with random perturbation of the ligand followed by functional constraint optimization

through rigid body minimization of the ligand, sidechain and ligand conformational sampling, and finally ligand, sidechain, and backbone minimization was used to approximate protocols used in successful enzyme reengineering efforts.[2] An example set of input files for wild type BglB are provided in the Supplemental Materials.

For each mutant, 100 models were generated as described above and the lowest 10 in overall system energy for each mutant were selected for subsequent structural analysis. A value for each of 59 potentially informative features, such as predicted interface energy, number of hydrogen bonds between protein and ligand, and change in solvent accessible surface area upon ligand binding, was calculated for each model. Correlation of the average calculated structural features to each kinetic constant was assessed using 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 $1/K_M$, 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 PCC for all three predicted constants (SRC of 0.55, 0.42 and 0.38 for k_{cat}/K_M, k_{cat} and $1/K_M$ respectively). The PCC and SRC values for all features are available in Supplemental Table 2.

Machine learning prediction of kinetic constants

Because no single structural feature predicts k_{cat}, $1/K_M$, or k_{cat}/K_M with high accuracy, machine learning techniques were 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 structural features and predict 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.

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 $1/K_M$ 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 $1/K_M$, meaning that in BglB an increase in structural packing around the catalytic residues and protein results in a higher K_M. Since substrate binding has a direct effect on K_M, this suggests physical space is needed around each catalytic residue and the entire protein in order to optimally accommodate the substrate. These features and interpretation of their selection is consistent with an induced fit mechanism. However, if BglB employs an induced fit mechanism, the structural changes would likely be relatively small since the RMSD between the apo and transition state analogue bound forms of BglB is < 0.2 Å.

The features selected by the algorithm as predictive of k_{cat} include a count of polar contacts, consistent with mechanistic studies that indicate BglB stabilizes the positive charge on the oxocarbenium ion in the proposed transition state.[14] Further supporting the importance of stabilizing the transition state is the selection of a ligand burial term (change in solvent accessible surface area on binding) by the elastic net algorithm, which indicates that tight packing and shape complementary are critical to catalysis. The selection of these features by the machine learning algorithm is consistent with hypothesis that the finely-tuned electrostatic environment of the BglB active site is of primary importance for catalysis.[13]

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 for optimally positioning the substrate and the protein sidechains to enable catalysis ("orbital steering").[15] 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.

While many of the selected features are consistent with well-established mechanisms of enzyme catalysis, there were several unexpected observations. One unexpected trend is that several features are selected as predictive of k_{cat}/K_M but not either k_{cat} or K_M. Further analysis of k_{cat} and K_M revealed that there is no significant correlation between two parameters in this dataset (Supplemental Figure 3). This suggests that k_{cat} and K_M are independent parameters for BglB, and it

is therefore not unexpected that features found to be predictive of k_{cat}/K_M are not predictive of either k_{cat} or K_M independently.

A second unexpected observation is that the most common metric used for evaluating designs, interface energy,[2,3,4,5,6] is not selected by the algorithm to be predictive of any kinetic constant. Ideally this would be the single metric optimally correlated with either k_{cat} or k_{cat}/K_M. This likely stems from training the enzyme design algorithm on indirect measures of function, further supporting the need to train force-field based algorithms on direct experimental measurements.

DISCUSSION

The Rosetta Molecular Modeling Suite has been successfully used to guide the engineering of a wide range of enzyme functions. However, there has been a limited ability to benchmark its predictive power 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 data generated here uncovered new structure-function relationships in BglB, and provides a 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 almost as important to catalysis as the two residues which are directly involved in the chemistry. This highlights the underlying interdependence of the entire active site

to catalyze the reaction. This is consistent with a recent report exploring the interconnectedness of a network of five residues in alkaline phosphatase.[16]

The large dataset of kinetic constants generated enabled the use of machine learning techniques to select structural features that are predictive of function. It was unexpected to observe that the calculated interface energy is not found to be predictive of any kinetic parameter, and was not a feature selected by machine learning as predictive of function. This has significant implications for future design strategies since interface energy is one of the most common metrics currently used to evaluate enzyme designs. It may be pertinent to develop additional training datasets, such as we have done for BglB, in order to further quantify 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 hydrolases. 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. Using readily calculated structural features machine learning protocols were employed to select a subset of features that are highly predictive of each measured kinetic parameter. 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.[17] 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 materials.

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: DAC and RWC contributed equally.

ACKNOWLEDGMENTS

This work was supported by ARO #201121557 and NSF #1254205 (IT) and Sloan #BR2014-012 and UC Davis Startup Funds (JBS). We are grateful to the Siegel Lab, Dr. David Wilson, and Jeremy H. Mills for insightful comments and discussions that helped shape this manuscript.

ABBREVIATIONS

pNPG, *p*-nitrophenyl-ß-D-glucoside, RMSD root-mean-square deviation

FIGURES

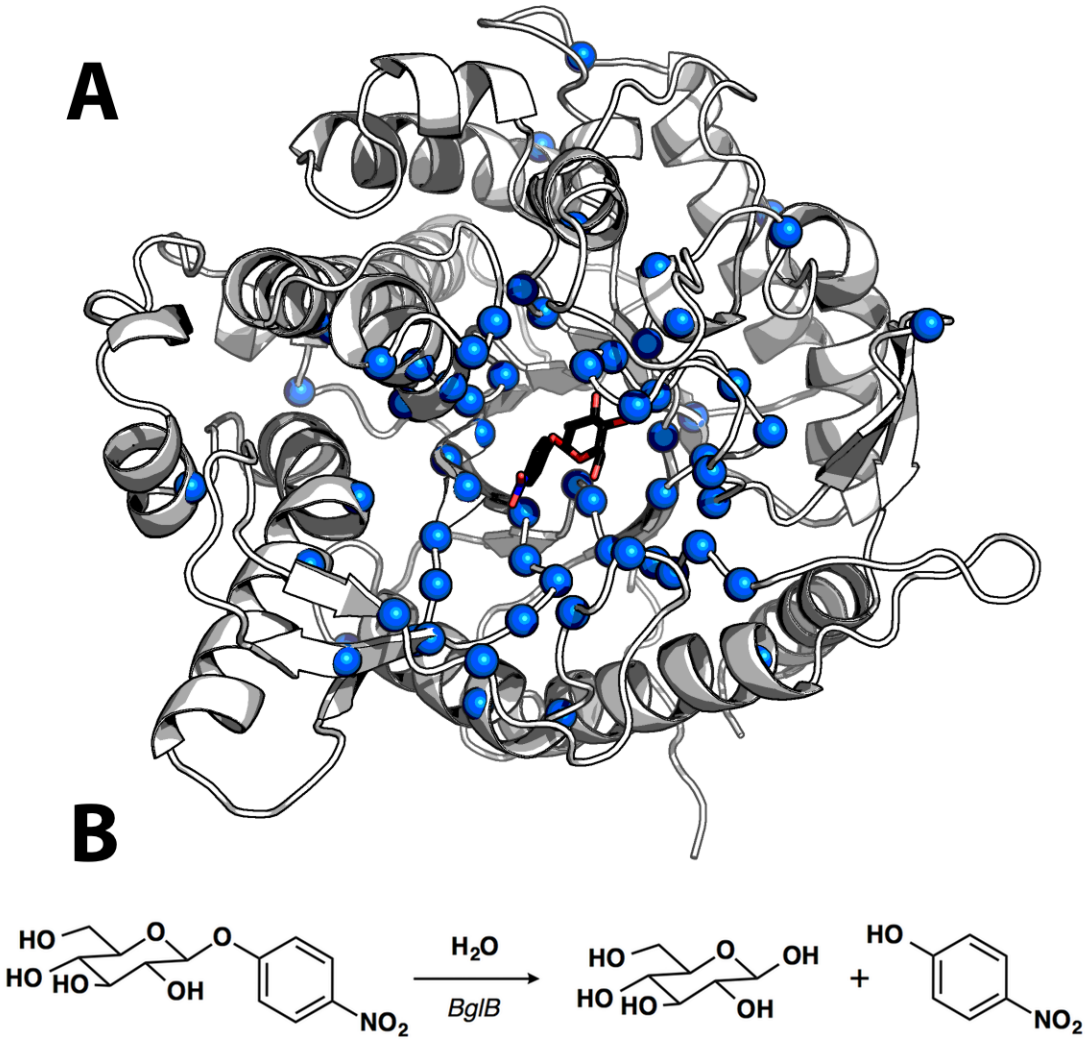

Figure 1. Structure and catalyzed reaction of BglB

(A) Structure of BglB in complex with the modeled *p*-nitrophenyl-ß-D-glucoside used for design drawn with PyMOL.[18] 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

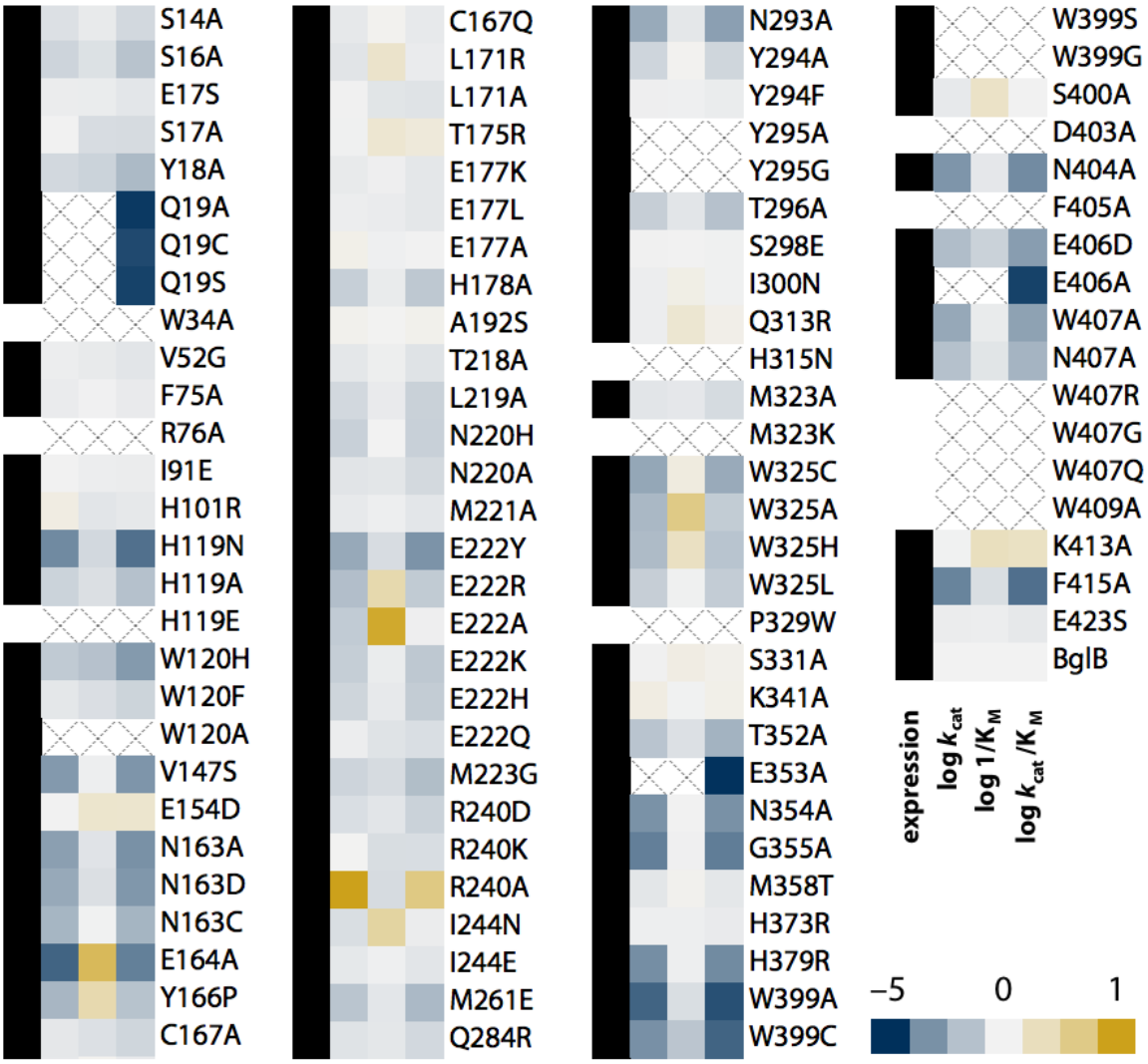

Figure 2. Log scale relative kinetic constants of 104 BglB mutants

The heatmap depicts the effect of each mutation on each kinetic constant relative to native BglB, normalized at 0. 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 $M^{-1}min^{-1}$ (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.

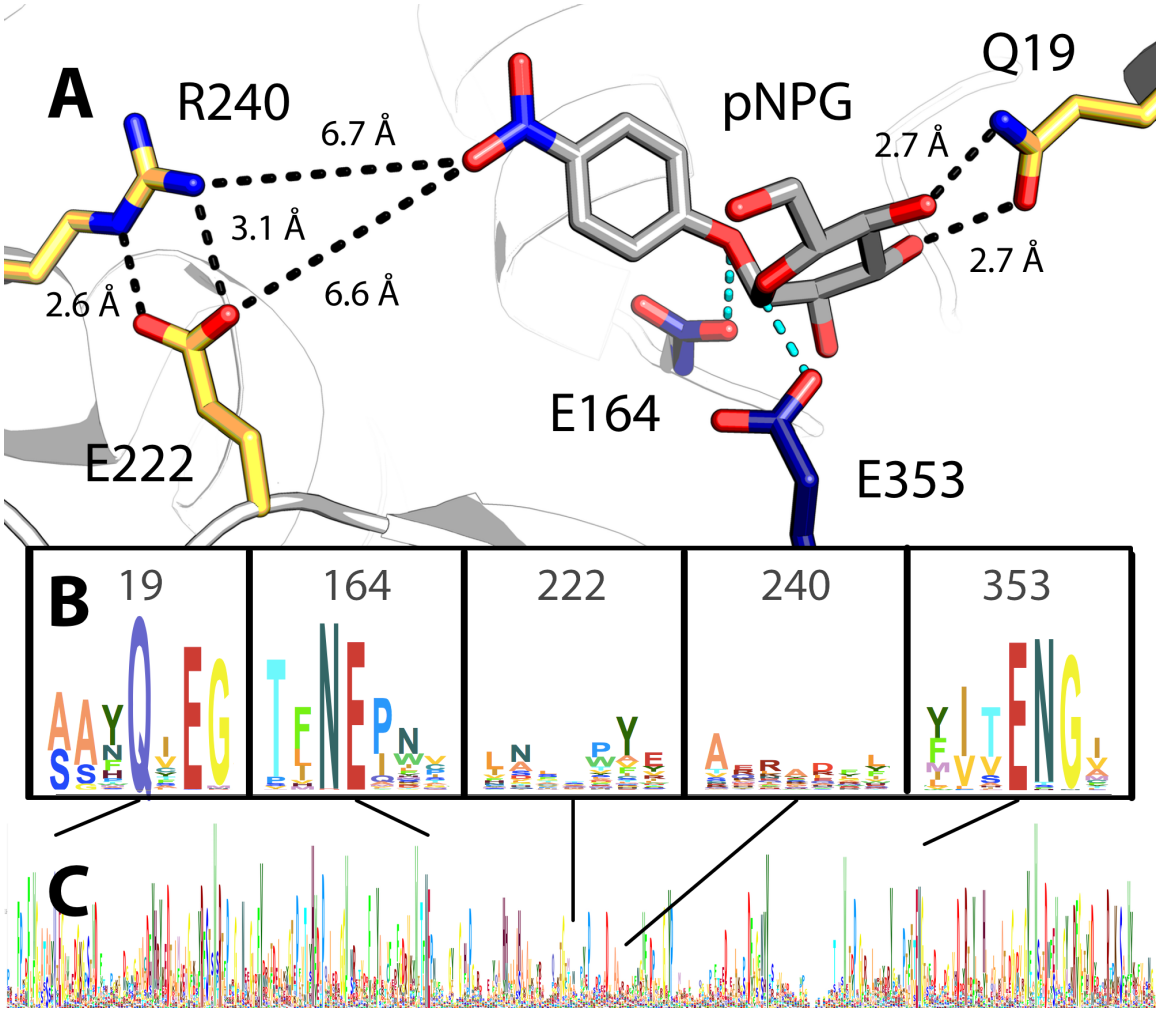

Figure 3. Active site model and conservation analysis of BglB

(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.

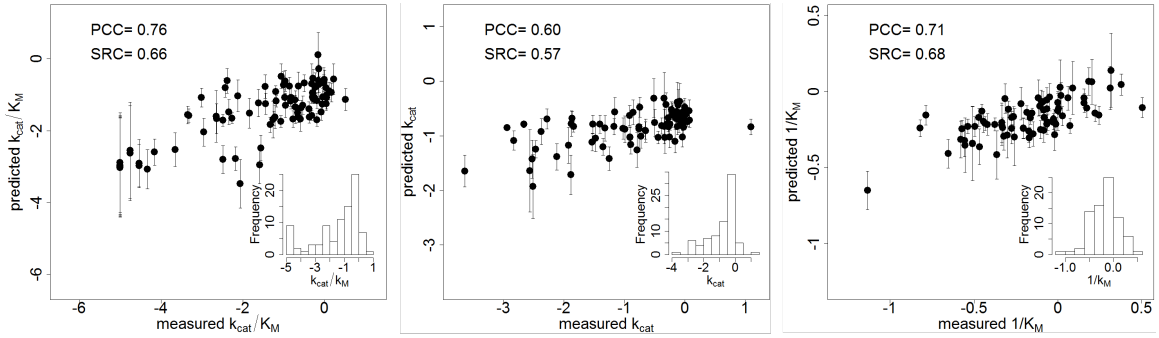

Figure 4. Correlation between machine learning predictions and experimentally-determined kinetic constants

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).

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

Table 1. Most informative structural features predicting each kinetic constant

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 2. *ns=feature not selected by the algorithm*

REFERENCES

1. Mak, W. S.; Siegel, J. B., Computational enzyme design: Transitioning from catalytic proteins to enzymes. *Current opinion in structural biology* **2014,** *27*, 87-94.

2. Siegel, J. B.; Smith, A. L.; Poust, S.; Wargacki, A. J.; Bar-Even, A.; Louw, C.; Shen, B. W.; Eiben, C. B.; Tran, H. M.; Noor, E.; Gallaher, J. L.; Bale, J.; Yoshikuni, Y.; Gelb, M. H.; Keasling, J. D.; Stoddard, B. L.; Lidstrom, M. E.; Baker, D., Computational protein design enables a novel one-carbon assimilation pathway. In *PNAS*, National Acad Sciences: 2015; Vol. 112, pp 3704-3709.

3. Damborsky, J.; Brezovsky, J., Computational tools for designing and engineering enzymes. In *Current Opinion in Chemical Biology*, 2014; Vol. 19, pp 8-16.

4. Gordon, S. R.; Stanley, E. J.; Wolf, S.; Toland, A.; Wu, S. J.; Hadidi, D.; Mills, J. H.; Baker, D.; Pultz, I. S.; Siegel, J. B., Computational Design of an α-Gliadin Peptidase. In *J. Am. Chem. Soc.*, American Chemical Society: 2012; Vol. 134, pp 20513-20520.

5. Marcheschi, R. J.; Li, H.; Zhang, K.; Noey, E. L.; Kim, S.; Chaubey, A.; Houk, K. N.; Liao, J. C., A Synthetic Recursive "+1" Pathway for Carbon Chain Elongation. In *ACS Chem. Biol.*, American Chemical Society: 2012; Vol. 7, pp 689-697.

6. Khare, S. D.; Kipnis, Y.; Greisen, P. J.; Takeuchi, R.; Ashani, Y.; Goldsmith, M.; Song, Y.; Gallaher, J. L.; Silman, I.; Leader, H.; Sussman, J. L.; Stoddard, B. L.; Tawfik, D. S.; Baker, D., Computational redesign of a mononuclear zinc metalloenzyme for

organophosphate hydrolysis. In *Nature Chemical Biology*, Nature Publishing Group: 2012; Vol. 8, pp 294-300.

7. Kumar, M. S.; Bava, K. A.; Gromiha, M. M.; Prabakaran, P.; Kitajima, K.; Uedaira, H.; Sarai, A., ProTherm and ProNIT: thermodynamic databases for proteins and protein–nucleic acid interactions. *Nucleic Acids Research* **2006,** *34* (suppl 1), D204-D206.

8. Kellogg, E. H.; Leaver‐Fay, A.; Baker, D., Role of conformational sampling in computing mutation‐induced changes in protein structure and stability. *Proteins: Structure, Function, and Bioinformatics* **2011,** *79* (3), 830-838.

9. Guerois, R.; Nielsen, J. E.; Serrano, L., Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *Journal of molecular biology* **2002,** *320* (2), 369-387.

10. Isorna, P.; Polaina, J.; Latorre-García, L.; Cañada, F. J.; González, B.; Sanz-Aparicio, J., Crystal Structures of Paenibacillus polymyxa β-Glucosidase B Complexes Reveal the Molecular Basis of Substrate Specificity and Give New Insights into the Catalytic Machinery of Family I Glycosidases. In *Journal of Molecular Biology*, 2007; Vol. 371, pp 1204-1218.

11. Wu, S. J.; Eiben, C. B.; Carra, J. H.; Huang, I.; Zong, D.; Liu, P.; Wu, C. T.; Nivala, J.; Dunbar, J.; Huber, T., Improvement of a potential anthrax therapeutic by computational protein design. *Journal of Biological Chemistry* **2011,** *286* (37), 32586-32592.

12. Kunkel, T. A., Rapid and efficient site-specific mutagenesis without phenotypic selection. *Proceedings of the National Academy of Sciences* **1985,** *82* (2), 488-492.

13. Warshel, A.; Sharma, P. K.; Kato, M.; Xiang, Y.; Liu, H., Electrostatic basis for enzyme catalysis. In *Chemical …*, 2006.

14. McCarter, J. D.; Withers, S. G., Mechanisms of enzymatic glycoside hydrolysis. *Curr Opin Struct Biol* **1994,** *4* (6), 885-92.

15. Mesecar, A. D.; Stoddard, B. L.; Koshland, D. E., Jr., Orbital steering in the catalytic power of enzymes: small structural changes with large catalytic consequences. *Science (New York, N.Y.)* **1997,** *277* (5323), 202-6.

16. Sunden, F.; Peck, A.; Salzman, J.; Ressl, S.; Herschlag, D.; Kuriyan, J., Extensive site-directed mutagenesis reveals interconnected functional units in the Alkaline Phosphatase active site. In *eLife Sciences*, eLife Sciences Publications Limited: 2015; Vol. 4, p e06181.

17. Gibson, D. G.; Young, L.; Chuang, R.-Y.; Venter, J. C.; Hutchison, C. A.; Smith, H. O., Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nature methods* **2009,** *6* (5), 343-345.

18. DeLano, W. L., The PyMOL molecular graphics system. **2002**.