A kinetically characterized library of over 100 mutant enzymes to benchmark *in silico* enzyme redesign

AUTHOR NAMES (Word Style "BB_Author_Name"). Include in the byline all those who have made substantial contributions to the work, even if the paper was actually written by only one person. Use first names, initials, and surnames (e.g., John R. Smith) or first initials, second names, and surnames (e.g., J. Robert Smith). Do not use only initials with surnames (e.g., J. R. Smith) because this causes indexing and retrieval difficulties and interferes with unique identification of an author. Do not include professional or official titles or academic degrees. At least one author must be designated with an asterisk as the author to whom correspondence should be addressed. †, ⊥, ∥, ∇, *

AUTHOR ADDRESS (Word Style "BC_Author_Address"). The affiliation should be the institution where the work was conducted. If the present address of an author differs from that at which the work was done, indicate with a symbol and give the Present Address under Author Information. If more than one address, use symbols to match author names to address(es).

KEYWORDS (Word Style "BG_Keywords"). If you are submitting your paper to a journal that requires keywords, provide significant keywords to aid the reader in literature retrieval.

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 [Cites]. 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

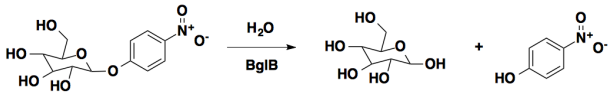
The ability to rationally reengineer enzyme functions has the potential to allow the development of highly efficient and specific catalysts tailored for needs beyond what was selected for during natural evolution. A rapidly growing route for engineering enzyme catalysts is the use of computational tools to evaluate mutations in silico before experimentally characterizing the mutants in the lab. Using the Rosetta Molecular Modeling Suite novel functions encompassing reengineering both specificity and chemistry have been accomplished. However, the design efforts often require screening hundreds of designs many of which introduce the intended functional effect. This has led to significant efforts to improve the design and modeling protocols to improve predictive capabilities.

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 protein thermostability, and have played a critical role in evaluating and improving computational algorithms to accurately model and design these 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 this level. 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. 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 from [Isorna] 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]

Scheme 1. BglB on 4-nitrophenyl-ß-D-glucoside



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. Greater than 95% 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 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.

▩ 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 4-nitro-ß-D-glucoside, 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 among 4-nitrophenyl-ß-D-glucoside, the acid-base E164, the nucleophile E353, and Y295, which stabilizes the attacking species. The angle between the attacking oxygen from Glu 353, the anomeric carbon, and the phenolic oxygen was constrained to 180˚, in accordance with an SN2-like mechanism [Cite].

Two methods were used for selecting mutants to generate and kinetically characterize. 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. A total of 69 positions were covered over the 102 mutants made.

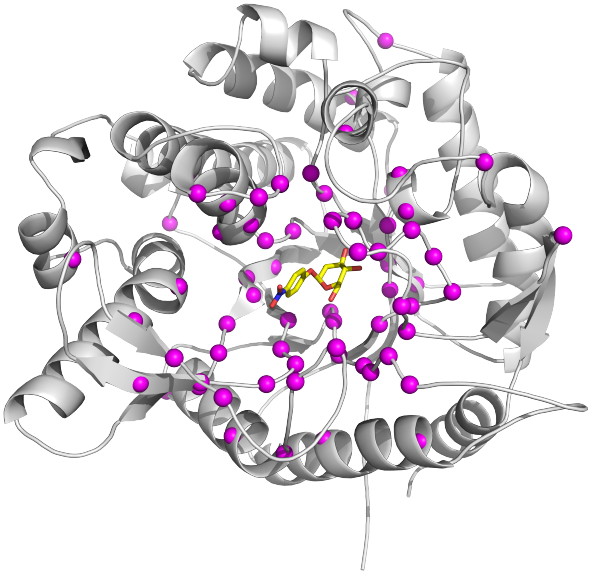


Figure 1. Every site mutated in BglB in our study (alpha carbons shown as pink spheres)

Protein production and purification

Each of the 102 site-specific mutants was generated via Kunkel mutagenesis and sequence verified using the Transcriptic cloud laboratory platform. All mutants 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, XXX 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 figure XXX the distribution of yields for all 102 mutants are illustrated. Greater than XXX% maintained the yields obtained for native BglB, and XXX% were not expressed and purified as a soluble protein above our limit of detection (0.2 mg/mL).

Kinetic characterization of mutants

The Michaelis-Menten kinetic constants for each of the 104 mutants was measured using the colorimetric assay of 4-nitrophenyl-ß-D-glucoside hydrolysis. Ten biological replicates of

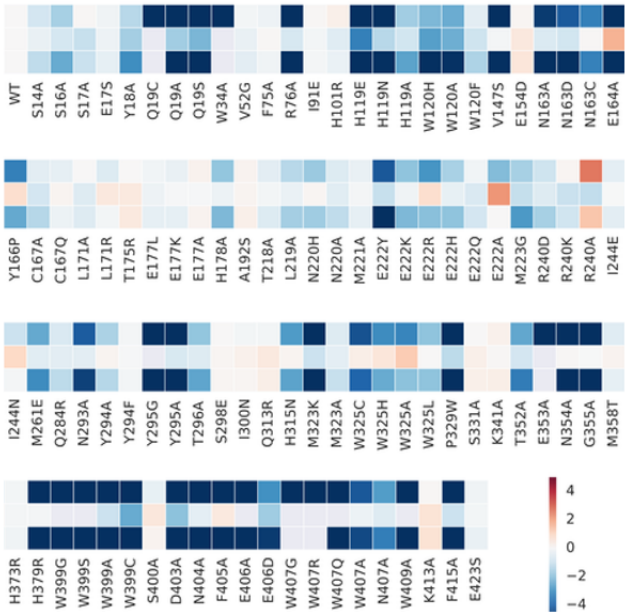


Figure 2. Heatmap of kinetic constants. Data are relative to the wild type enzyme and log-transformed.

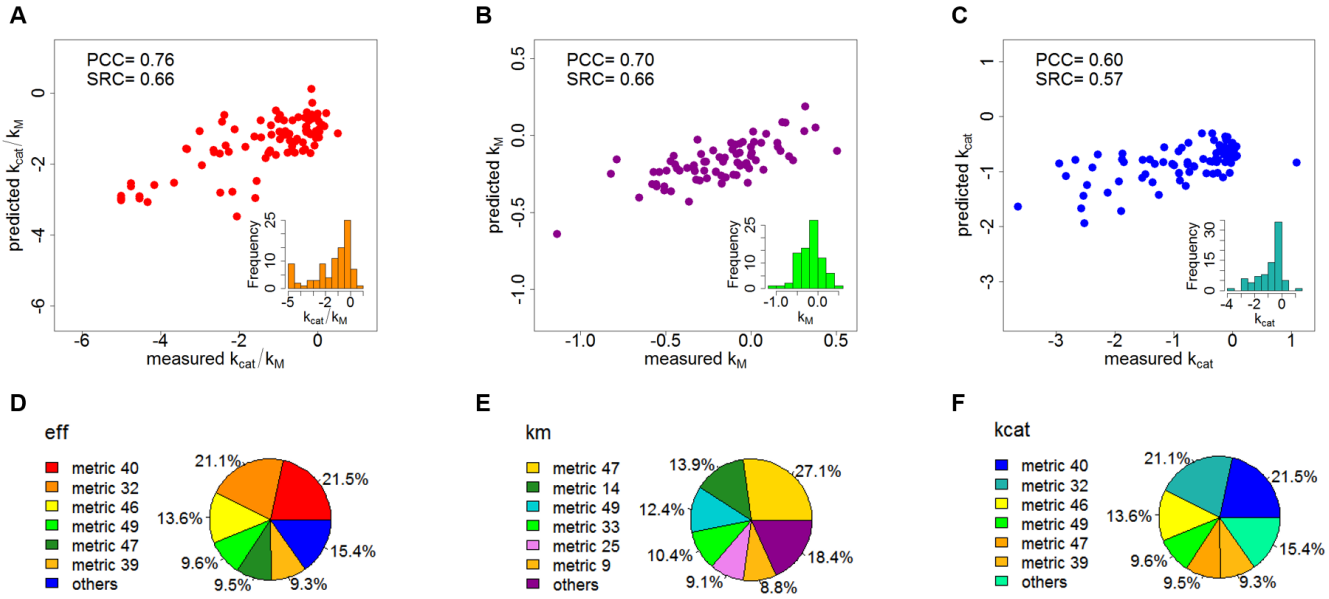
the wild type enzyme had an average kcat of 880 ± 10 M/min, an average KM of 5 ± 0.2 mM, and kcat/KM of 171,000 ± 0.05 M/min. Kinetic constants for each mutant with the substrate 4-nitrophenyl-ß-D-glucoside are represented as a heatmap in Figure 2. Observed rates at 8 substrate concentrations were fit to the Michaelis-Menten equation using SciPy. The values for each metric are in log scale and normalized relative to wild type BglB. 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 XXX solubly purified mutants YYY 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 XXX.

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 4-nitrophenyl-ß-D-glucoside 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, side chain repacking. The command line and an example set of input files for wild type BglB are provided in Supplemental XXX.

The lowest 10% in total energy models were selected to represent each mutant, and 59 structural metrics for each mutants were calculated. 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 using an elastic net algorithm. The correlations between structural features and kcat, KM, and kcat/KM are given in Figure 4.

Figure 4. Plot of predicted value and measured value by mining the new dataset. The most informative features were selected by considering all the models constructed in the 1,000 times of randomization. (Below)



▩ 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.

Through the use of the large dataset we were able to utilize machine learning techniques that identified structural features correlated with function. We found that kcat correlated with weight 0.918 to the change in solvent-accessible area of the ligand upon binding, KM correlated with weight 1 to packing of active site residues in the absence of the ligand, and overall efficiency correlated with weight 1 to the number of hydrogen bonds between the ligand and protein side chains. Full results are given in table XXX.

The metric with the best correlation to overall efficiency, , ... While the metrics identified by machine learning are consistent with chemical principles, more data sets of standardized kinetic constants on enzyme classes beyond the glycosyl hydrolase family will be needed to determine if these features are system-specific or generally correlated with function among many enzyme classes.

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 the reaction chemistry are as important to catalysis as the three residues which are directly involved in the chemistry. Additionally, we found a mutation not previously observed in any natural variant in the glycosyl hydrolase 1 family which resulted in an 10-fold increase in kcat. This emphasizes the importance of not limiting design efforts to changes previously observed in nature when engineering function towards a non-natural substrate. However, that observation that XXX% of all the highly conserved residues (i.e. >90%) within the active site contributed >1000-fold to the naturally occurring catalytic efficiency supports the widely held assumption that conservation is indicative of functional importance.

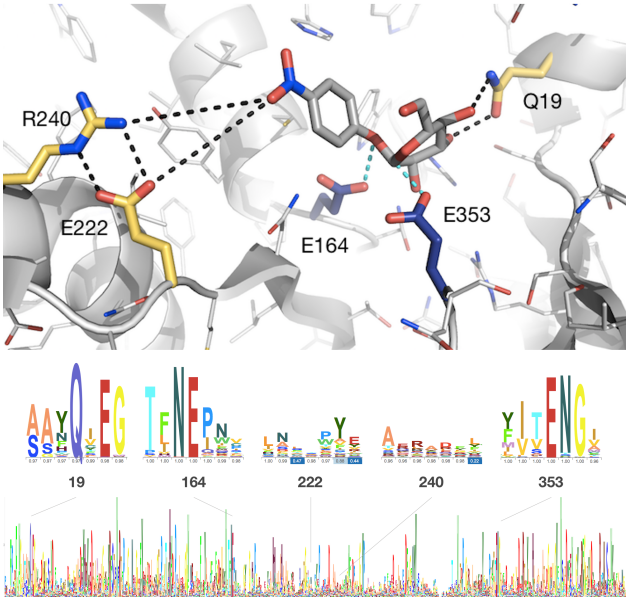


Figure 3. Drawing of the active site of BglB

▩ 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 kinetic properties. Finally, by implement 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 reaction scheme, figure X). 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 106 individual proteins were fit to the Michaelis-Menten equation using the Python module SciPy.

TABLES. Each table must have a brief (one phrase or sentence) title that describes its contents. The title should follow the format "Table 1. Table Title" (Word Style "VD_Table_Title"). The title should be understandable without reference to the text. Put details in footnotes, not in the title (use Word Style "FE_Table_Footnote"). Do NOT modify the amount of space before and after the title as this allows for the space above and below the table to be inserted upon editing.

Use tables (Word Style "TC_Table_Body") when the data cannot be presented clearly as narrative, when many precise numbers must be presented, or when more meaningful interrelationships can be conveyed by the tabular format. **Do not use Word Style "TC_Table_Body" for tables containing artwork.** Tables should supplement, not duplicate, text and figures. Tables should be simple and concise. It is preferable to use the Table Tool in your word-processing package, placing one entry per cell, to generate tables.

ASSOCIATED CONTENT

(Word Style "TE_Supporting_Information"). **Supporting Information**. A brief statement in nonsentence format listing the contents of material supplied as Supporting Information should be included, ending with "This material is available free of charge via the Internet at http://pubs.acs.org." For instructions on what should be included in the Supporting Information as well as how to prepare this material for publication, refer to the journal's Instructions for Authors.

AUTHOR INFORMATION

Corresponding Author

* (Word Style "FA_Corresponding_Author_Footnote"). Give contact information for the author(s) to whom correspondence should be addressed.

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

CCR2, CC chemokine receptor 2; CCL2, CC chemokine ligand 2; CCR5, CC chemokine receptor 5; TLC, thin layer chromatography.

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.