OPTZYME: A COMPUTATIONAL TOOL FOR ALTERING ENZYMATIC SPECIFICITY

A Thesis in
Chemical Engineering
by
Matthew Jeffery Grisewood

© 2013 Matthew Jeffery Grisewood

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

May 2013
The thesis of Matthew Jeffery Grisewood was reviewed and approved* by the following:

Costas D. Maranas
Donald B. Broughton Professor of Chemical Engineering
Thesis Co-Advisor

Michael Janik
Associate Professor of Chemical Engineering
Thesis Co-Advisor

Howard M. Salis
Assistant Professor of Chemical Engineering

Thomas K. Wood
Endowed Biotechnology Chair Professor of Chemical Engineering

Andrew Zydney
Professor of Chemical Engineering
Head of the Department of Chemical Engineering

* Signatures are on file in the Graduate School.
ABSTRACT

OptZyme is a new computational procedure that is designed to improve enzymatic activity (i.e., \( k_{\text{cat}} \) or \( k_{\text{cat}}/K_M \)) with a novel substrate. The key concept here is to use transition state analogue (TSA) compounds, which are known for many enzymatic reactions, as substitutes for the typically unknown transition state (TS) structures. Mutations that minimize the interaction energy of the enzyme with its TSA, rather than with its substrate, are identified that lower the transition state energy barrier. Using results of *Escherichia coli* β-glucuronidase as a benchmark, we confirm that \( K_M \) correlates (\( R^2 = 0.960 \)) with the interaction energy between *Escherichia coli* β-glucuronidase and para-nitrophenyl- β D-glucuronide, \( k_{\text{cat}}/K_M \) correlates (\( R^2 = 0.864 \)) with the interaction energy of the transition state analogue molecule, D-glucaro-1, 5-lactone, and \( k_{\text{cat}} \) correlates (\( R^2 = 0.854 \)) with a weighted combination of interaction energies with para-nitrophenyl- β, D-glucuronide and D-glucaro-1, 5-lactone. OptZyme is subsequently used to identify mutants with improved \( K_M \), \( k_{\text{cat}} \) and \( k_{\text{cat}}/K_M \), respectively for the new substrate, para-nitrophenyl- β, D-galactoside. Differences between the three libraries shed light onto structural differences that underpin improved \( K_M \), \( k_{\text{cat}} \) or \( k_{\text{cat}}/K_M \). Mutants predicted to enhance activity of pNP-GAL were mostly those that indirectly or directly helped create hydrogen bonds with the altered sugar ring conformation or its substituents, namely H162S, L361G, W549R, and N550S. Therefore, OptZyme is an inexpensive tool that can predict mutations for promoting catalysis of a reaction of a related but new substrate, where a natural enzyme may be unknown. **Note: This manuscript is being prepared for publication.**
# TABLE OF CONTENTS

LIST OF ABBREVIATIONS ........................................................................................................ v

LIST OF FIGURES .................................................................................................................. vii

LIST OF TABLES .................................................................................................................... ix

ACKNOWLEDGEMENTS .......................................................................................................... x

Chapter 1. INTRODUCTION .................................................................................................... 1
  
  Background on Computational Enzyme Redesign .......................................................... 1
  QM/MM-Based Enzyme Design ......................................................................................... 3
  OptZyme-Based Enzyme Redesign ..................................................................................... 4

Chapter 2. METHODS .......................................................................................................... 7
  
  Redesign of *Escherichia coli* β-glucuronidase (GUS) .................................................. 7

Chapter 3. RESULTS ........................................................................................................... 12
  
  Testing of TSA-based redesign paradigm using $k_{\text{cat}}$ and $K_M$ literature data ....... 12
  Redesign of GUS for improving activity with pNP-GLU .................................................. 18
  Redesign of GUS for introducing catalytic activity with the new substrate
    pNP-GAL ......................................................................................................................... 21

Chapter 4: DISCUSSION ..................................................................................................... 39

Appendix: Derivation of Equations 4 & 5 ........................................................................ 41

References .............................................................................................................................. 43
LIST OF ABBREVIATIONS

· : Bound To
ΔG‡ : Gibb’s Free Energy Barrier
κ : Transmission Coefficient
BE : Binding Energy
C: Arbitrary Constant #1
C1 : Arbitrary Constant #2
C2 : Arbitrary Constant #3
CHARMM : Chemistry At Harvard Molecular Mechanics
E : Enzyme
G : Gibb’s Free Energy
GUS : Escherichia coli β-glucuronidase
h : Planck’s Constant
I1 : Intermediate 1
I2 : Intermediate 2
IE : Interaction Energy
IPRO : Iterative Protein Redesign And Optimization Procedure
k : Reaction Rate Constant
k_B : Boltzmann Constant
MD : Molecular Dynamics
min : Energy-Minimized Structure
MM : Molecular Mechanics
NOE : Nuclear Overhauser Effect
P : Product

PDB : Protein Data Bank

pNP-GAL : para-nitrophenyl- β, D-galactoside

pNP-GLU : para-nitrophenyl- β, D-glucuronide

QM : Quantum Mechanics

R : Ideal Gas Constant

RMSD : Root Mean Square Deviation

S : Substrate

T : Temperature

TS : Transition State

TSA : Transition State Analogue

WT : Wild Type
LIST OF FIGURES

Figure 1: Native Reaction for GUS.................................................................9
Figure 2: Proposed Catalytic Reaction Mechanism of GUS Determined from in 
vacuo QM Calculations..............................................................10
Figure 3: Structural Comparison Between Ground State, Hypothetical Transition State, 
and Transition State Analogue for pNP-GLU.............................11
Figure 4: Active Site Geometry and Restrained Catalytic Contacts..................24
Figure 5: Qualitative GUS Free Energy Diagram Based Upon in vacuo QM Calculations.....26
Figure 6: Ground State Computational Interaction Energy (IE_s) versus the Natural 
Logarithm of Experimental $K_M$.........................................................27
Figure 7: Computationally-determined Transition State Analogue Interaction Energy (IE_TSA) 
versus Experimental $\ln(k_{cat}/K_M)$.......................................................28
Figure 8: Scaled Difference Between Interaction Energy with the Transition State Analogue 
(IE_TSA) and Interaction Energy with pNP-GLU (IE_s) versus the Natural 
Logarithm of $k_{cat}$..............................................................................29
Figure 9: Ground State Interaction Energy (IE_s) Correlation with Catalytic Efficiency.......30
Figure 10: Distribution of Amino Acids in a Sequence Alignment for all 
$\beta$-glucuronidases..........................................................................31
Figure 11: Distribution of Amino Acids at Each Design Position for Top 50 GUS Mutants to 
Enhance $K_M$, $k_{cat}/K_M$, and $k_{cat}$ for pNP-GLU..............................34
Figure 12: Distribution of Amino Acids at Each Design Position for Top 50 GUS Mutants to 
Enhance $K_M$, $k_{cat}/K_M$, and $k_{cat}$ for pNP-GAL..............................36
Figure 13: Structural Comparison Between pNP-GLU, 1,5-glucarolactone, pNP-GAL, 1,5-galactonolactone…………………………………………………………………………………………………………..37

Figure 14: Ramachandran Plot of Top pNP-GAL Mutants. 50 of the top mutants from each of the pNP-GAL libraries were examined……………………………………………………………………..38
LIST OF TABLES

Table 1: Successes Within the Field of Computational Enzyme Design........................................6
Table 2: NOE Restraints Applied During CHARMM Energy Minimization............................. 25
Table 3: Permitted Amino Acids at each Design Position.......................................................... 32
Table 4: Top 10 Mutants Identified Using OptZyme for Optimizing $K_M$, $k_{cat}/K_M$, and $K_M$ for
pNP-GLU........................................................................................................................................ 33
Table 5: Top 10 Mutants Identified Using OptZyme for Optimizing $K_M$, $k_{cat}/K_M$, and $K_M$ for
pNP-GAL........................................................................................................................................ 35
ACKNOWLEDGEMENTS

I would like to acknowledge the National Science Foundation [CBET-0967062] for the funding of this work. I would also like to thank Nate Gifford for his work on the quantum mechanics calculations and Dr. Patrick Cirino (University of Houston) for his experimental testing of the designed mutants.
Chapter 1. INTRODUCTION

Background on Computational Enzyme Redesign

Enzymes are highly-specific, biomolecular catalysts with extraordinary reaction rate enhancements under mild conditions [1]. Enzyme activity is of paramount importance in the economics of cellulosic ethanol (and other biofuel) production [2, 3]. Improving enzymatic activity is currently carried out using primarily experimental techniques (e.g. random mutagenesis and DNA recombination) relying on screening large combinatorial libraries [4]. The vast sequence space (a 100-residue peptide encompasses $20^{100}$ or $1.27 \times 10^{130}$ possible distinct sequences, permitting residues that are included in the original sequence) cannot be exhaustively searched using even the most advanced experimental screening techniques (~$10^{15}$ sequences) [5]. Experiments can be synergistically coupled with efficient computational screening (i.e., fine-tuning of in silico mutants with random mutagenesis) protocols that can identify mutants within promising regions of the sequence space to construct enriched libraries. Reliable computational techniques for identifying mutations that lead to enzymatic activity improvements would have a cross-cutting impact on many fronts ranging from biofuel production and biomass pretreatment to pro-drug activation and the design of new therapeutics [6-9].

Various computational tools utilizing primary, secondary, and/or tertiary protein structural information have been tried to discover promising enzyme redesigns. Comparative modeling focuses on employing primary sequence information to resolve the three-dimensional structure of a query sequence by using
structures of homologous proteins [10, 11]. Comparative models, such as MODELLER [12] and SWISS-MODEL [13], build a model for the unknown structure based on the structures of the set of aligned templates [13]. *Kluveromyces lactis* UDP-galactose-4-epimerase [14] and *Triticum aestivum* oxalate oxidase [15] are two examples of enzymes whose three-dimensional structure has been approximated using homology modeling. The determination of a tertiary structure establishes a basis for the rational redesign of proteins [16] as recently demonstrated for a *Theobroma cacao* cysteine protease [17]. However, despite its simplicity and relative ease of calculation, homology modeling cannot always be relied upon for the reliable determination of a modeled structure. This can be due to a variety of factors ranging from misalignment of sequences, errors in regions without a template, and distortions in correctly aligned regions [10].

The use of structural information in addition to sequence alignment can lead to improvements in prediction. Scoring-based models aim to identify and eliminate unfavorable contacts between residues (e.g. SCHEMA, SIRCH, residue clash maps, etc.) by implementing scoring functions that disfavor steric, charge and hydrophobicity clashes encountered during protein library formation through recombination [18]. These algorithms have been used to distinguish experimentally active from inactive enzymes, including β-lactamase [19], cytochromes P450 [20], and dihydrofolate reductase [21]. The use of ad-hoc scoring functions offers several advantages in terms of low computational cost and reasonable accuracy, but limitations include neglecting steric relief by rotamer movements, restraining
diversity to within protein parental sequence space, and overlooking of hybrid
directionality in recombination protocols.

The use of rigorous potential models [22-25] allow for the exploration of
novel enzyme designs that are not simple concatenations through recombination of
a set of parental sequences. Enzymes, such as a proteinase, aldolase [26], esterase
[27], and kemp eliminase [28], have been designed using such force fields by finding
a beneficial set of residues for binding at the transition state. Methods deployed for
enzyme design making use of detailed potential models, such as CHARMM [23],
AMBER [22], and GROMOS [24], include molecular statics [26, 29] and molecular
dynamics (MD) [27]. The use of scoring functions that combine rigorous and
statistics-based potential models provide a compromise between biophysical rigor
and utilization of known structural information (e.g., Rosetta uses both potential
functions and conformational sampling [30]).

**QM/MM-Based Enzyme Design**

Ultimately, quantum mechanics (QM) promise to improve accuracy by
capturing interactions at the reacting center at the expense of significant
computational cost. QM and molecular mechanics (MM) are typically used together
to computationally model enzymes by using QM for the active site of the enzyme
and MM for the remainder of the enzyme [31]. QM/MM methods have been
previously used to design active enzymes *de novo*, as in the cases of kemp eliminase
and retro-aldolase [1, 32, 33]. Additionally, QM/MM can aid in the redesign of
naturally evolved enzymes to yield active mutants with transformed specificity (e.g.,
ammelide deamination by guanine deaminase [34] and organophosphate hydrolysis by adenosine deaminase [35]). QM/MM allows for direct tracking of reaction intermediates (a molecule that is produced and subsequently consumed) and transition states (an unstable reaction intermediate situated at the saddle point of the potential energy surface) during the chemical reaction. These models have the potential to provide an atomic-level understanding of the system, including interactions responsible for transition state stabilization [36]. Unfortunately, the QM description of the enzyme and transition state structure remains a computationally expensive task.

**OptZyme-Based Enzyme Redesign**

Despite significant effort so far, the computational design of enzymes remains a formidable task with only isolated successes [1, 27-29, 32-35, 37, 38] verified by experiment. These successes are summarized in Table 1. In general, the enzymes identified that had the largest improvements were mutants that were fine-tuned using directed evolution approaches. Here, we introduce a new enzyme design method, termed OptZyme, which uses binding calculations in combination with information about the limiting transition state using as proxy TSA analogues. OptZyme implements as surrogates for the typically unknown transition state (TS) structure, transition state analogue (TSA) compounds which are known for many enzymatic reactions. By identifying mutations that minimize the interaction energy of the enzyme with its TSA rather than with its substrate, transition state energy barrier lowering is presumably achieved. The concept of using TSAs for enzyme redesign has been previously explored [29, 39]. However, OptZyme is unique as it
provides a theoretical framework for making use of TSA binding calculations to inform enzyme design while also integrating preliminary QM information such as rate-limiting step identification and ligand partial charge information. Theoretical analysis shows that we can relate $K_M$ as a function of interaction energy with the substrate and $k_{cat}/K_M$ as a function of interaction energy with the transition state analogue. Using these computationally accessible interaction energies as proxies for $K_M$, $k_{cat}$, and $k_{cat}/K_M$, *Escherichia coli* β-glucuronidase mutants of H162S, D163K, L361R, L361E, W549R, and N550S are identified that optimized the above enzyme parameters for para-nitrophenyl- β, D-galactoside, relative to para-nitrophenyl- β, D-glucuronide. Different library designs are identified for optimizing $K_M$, $k_{cat}$, and $k_{cat}/K_M$, respectively. OptZyme is capable of discriminating between para-nitrophenyl- β, D-galactoside, para-nitrophenyl- β, D-glucuronide, and their corresponding TSAs largely through the customization of hydrogen bonding networks. Mutations that either indirectly or directly construct hydrogen bonds with the altered geometry of the TSA of the new substrate are those that are predicted to improve enzymatic activity towards pNP-GAL.
Table 1: Successes Within the Field of Computational Enzyme Design. The enzyme studied, its improvement in activity over the reference catalyst (WT enzyme for all redesigned enzymes, see source for de novo designed enzymes), and enzyme design approach are reported.

<table>
<thead>
<tr>
<th>Enzyme</th>
<th>Activity Improvement</th>
<th>de novo</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>kemp eliminase</td>
<td>200</td>
<td>Y</td>
<td>1</td>
</tr>
<tr>
<td>butyrylcholinesterase</td>
<td>1800</td>
<td>N</td>
<td>27</td>
</tr>
<tr>
<td>thioredoxin</td>
<td>180</td>
<td>N</td>
<td>29</td>
</tr>
<tr>
<td>retro-aldolase</td>
<td>1000</td>
<td>Y</td>
<td>32</td>
</tr>
<tr>
<td>kemp eliminase</td>
<td>5.9 x 10^5</td>
<td>Y</td>
<td>33</td>
</tr>
<tr>
<td>guanine deaminase</td>
<td>100</td>
<td>N</td>
<td>34</td>
</tr>
<tr>
<td>adenosine deaminase</td>
<td>1 x 10^7</td>
<td>N</td>
<td>35</td>
</tr>
<tr>
<td>phenol oxidase</td>
<td>6300</td>
<td>Y</td>
<td>37</td>
</tr>
<tr>
<td>diels-alderase</td>
<td>&lt; 1</td>
<td>Y</td>
<td>38</td>
</tr>
</tbody>
</table>
Chapter 2. METHODS

Redesign of *Escherichia coli* β-glucuronidase (GUS)

This paper addresses the redesign of *Escherichia coli* β-glucuronidase (GUS) to accept para-nitrophenyl-β, D-galactoside (pNP-GAL), instead of the para-nitrophenyl-β, D-glucuronide (pNP-GLU) substrate. The native reaction is shown in Figure 1. The native substrate for GUS, a glucuronic acid-containing glycosaminoglycan [40, 41], more closely resembles pNP-GLU than pNP-GAL because of the nature of the sugar moiety within the substrate. pNP-GLU was used as the original substrate instead of a glycosaminoglycan because its leaving group (pNP) is easy to monitor spectrophotometrically. Altering the GUS substrate specificity for pNP-GAL requires maximizing catalytic efficiency (k\textsubscript{cat}/K\textsubscript{M}) [42, 43] at moderate substrate concentrations (<<K\textsubscript{M}) or maximizing the turnover number (k\textsubscript{cat}) at high substrate concentrations (>>K\textsubscript{M}).

The transition state structure for the glycosidic hydrolysis of pNP-GLU is, to the best of our knowledge, unknown. However, there is information available on transition state analogues for GUS (i.e., D-glucaro-1, 5-lactone) [44, 45]. TSAs are potent inhibitors with a stable enzyme-bound complex that closely resemble the TS of the enzymatic reaction [46, 47]. TSAs manage to interfere with the enzyme catalytic activity by mimicking the geometry of the TS of the substrate and preferentially binding with the enzyme over the substrate, thus preventing the reaction from proceeding. The TS used for TSA selection was based on the proposed reaction mechanism (see Figure 2) for GUS, supported by QM calculations. The
design approach explored here is to attempt to lower the transition state barrier by optimally redesigning the enzyme so as to improve the binding affinity with the TSA.

1,5-glucarolactone is a known TSA for GUS based on reported strong inhibition and close structural similarity of the carbohydrate moiety with the postulated TS structure \([44, 45]\). The TSA differs from the true TS by the replacement of the glycosidic bond with an ester functional group, resulting in an altered ring conformation due to the sp\(^2\)-hybridized carbonyl (see Figure 3). The TSA resembles the TS by having similar partial charges and stereochemistry within the carbohydrate moiety.
**Figure 1:** Native Reaction for GUS. GUS catalyzes the hydrolysis of a glucuronic acid-containing glycosaminoglycan to form two products, glucuronic acid and an amino sugar (acetylglucosamine in this reaction). pNP-GLU was used as the substrate instead of a glycosaminoglycan because of its fluorescence.
Figure 2: Proposed Catalytic Reaction Mechanism of GUS Determined from *in vacuo* QM Calculations (unpublished). (a) Protonated glutamic acid at position 413 (E413) hydrogen bonds to the glycosidic oxygen. (b) The hydrogen bond with the glycosidic oxygen causes partial positive charge character to appear on the anomeric carbon due to induction. (c) Eventually, the electrons shared between the anomeric carbon and glycosidic oxygen shift entirely to the oxygen atom, resulting in a carbocation intermediate. (d) The carbocation intermediate is then stabilized by the negative charge of the unprotonated glutamic acid at position 504 (E504). (e) This stabilization results in the formation of a covalent intermediate involving E504. Hydrolysis then results in the formation of the reaction products and regeneration of the two catalytic residues.
Figure 3: Structural Comparison Between Ground State (i), Hypothetical Transition State (ii), and Transition State Analogue for pNP-GLU (iii). The (iii) transition state analogue, 1,5-glucarolactone, for pNP-GLU resembles the (ii) proposed transition state in terms of charge distribution and stereospecificity of the carbohydrate. Differences include $sp^2$ hybridization of the analogue compared to $sp^3$ for the transition state and lack of the nitrophenyl (pNP) moiety from the TSA (red triangles). In addition, the TSA differs from (i) pNP-GLU by assuming a more flattened sugar ring geometry (indicated by the dihedral angles) and partial positive charge at the anomeric carbon (green circles/ green arrows).
Chapter 3. RESULTS

Testing of TSA-based redesign paradigm using $k_{\text{cat}}$ and $K_M$ literature data

Before proceeding with the redesign of GUS to accept the new substrate, we used existing $k_{\text{cat}}$ and $K_M$ data from literature to assess the validity of the proposed computationally accessible metrics [48-50]. We first assessed whether the substrate binding calculations (using Chemistry at Harvard Molecular Mechanics; CHARMM [23]) at the ground state for the wild-type enzyme and a handful of mutants were consistent with the calculated $K_M$ values. We subsequently tested whether the reported $k_{\text{cat}}$/$K_M$ values were consistent with the binding calculations of the TSA with the GUS enzyme.

The structure for GUS was computationally assembled from its unbound crystal structure (PDB: 3K46) and an inhibitor-bound structure (PDB: 3LPF) that helped to pinpoint the binding site for pNP-GLU and included a bacterial loop that was not elucidated in the unbound structure [51]. The CHARMM force field was used to minimize the energy of the system while Nuclear Overhauser Effect (NOE) restraints were imposed between important catalytic residues (Table 2, Figure 4) to ensure the conservation of the optimal catalytic geometry [52]. Bond, angle, dihedral, improper dihedral, van der Waal, Urey-Bradley, electrostatic, NOE, Wesson/ Eisenberg solvation, and Generalized Born using Molecular Volume solvation energy terms were used during the CHARMM minimization step. Instead of directly calculating the binding energy ($BE$, Equation 1, where $G$ is the Gibb’s free energy, $E$-$S$ is the Michaelis complex, $E$ is the unbound enzyme, $S$ is the substrate, and $min$ indicates that the molecular structure is at the energy minimum) of the enzyme-substrate complex, we calculated its interaction energy ($IE$, Equation 2).

$$BE = G_{ES}^{\text{min}} - G_E^{\text{min}} - G_S^{\text{min}}$$

(1)
IE is a good surrogate for BE in cases where binding is not conditional on significant conformation rearrangements (no induced fit [53]). In addition, IE is less dependent on the force field as the energetics of any conformational rearrangements do not need to be tracked. Interaction energies were approximated using the iterative protein redesign and optimization procedure (IPRO) [54]. IPRO iteratively randomly perturbs the protein backbone and subsequently assigns optimal rotamers for all the residues followed by an energy relaxation step. Different IPRO trajectories may converge in alternate low energy conformations. To remedy the run-dependent nature of the results, the average of the best IE for each one of 25 IPRO runs was recorded. The calculated IE values were then related to the literature obtained K_M values as follows.

Michaelis-Menten kinetics for GUS enzymatic catalysis (based on mechanism shown in Figure 2) is depicted through Equation 3, where E is GUS, S is pNP-GLU, E·S is GUS bound to pNP-GLU, E·I_1 is the bound carbocation intermediate, E·I_2 is the E504- covalent adduct, E·P is bound D-glucuronic acid, P is the product of the reaction, D-glucuronic acid, and k represents reaction rate constants.

\[ IE_S = G_{E:S}^{min} - G_E - G_S \]  

\[ E + S \rightleftharpoons E \cdot S \rightleftharpoons E \cdot I_1 \rightleftharpoons E \cdot I_2 \rightleftharpoons E \cdot P \rightleftharpoons E + P \]  

QM calculations in vacuo show that the reaction of the Michaelis complex to form the carbocation intermediate is rate limiting in the absence of water (see Figure 5). In addition, assuming a fast rate of hydrolysis of the covalent adduct (i.e., E·I_2) and that the equilibrium constant of product release (i.e., E+P) after the rate limiting step lies far to the right
\[
\left(\frac{[E][P]}{[E \cdot I_1]} \gg 1\right),
\]
Equations 4 and 5 describe the kinetic parameters of the overall reaction (See Appendix for proof).

\[
K_M = \frac{k_{-1}}{k_1} \quad (4)
\]

\[
k_{cat} = k_2 \quad (5)
\]

Using the relationship between Gibb’s free energy and equilibrium concentrations (see Appendix, Equation A11), Equation 6 links the Michaelis constant, \(K_M\), to the binding energy between the enzymatic substrate complex (\(E\cdotI\)) and the unbound reactants, \(B_{E_S}\) (see Equation 1).

\[
B_{E_S} = RT \ln(K_M) \quad (6)
\]

The entropic component of the free energy is accounted for using an accessible area solvent model [55], where the change in solute entropy upon binding is assumed to be negligible relative to the other terms [56]. We find that the root mean square deviation (RMSD) between unbound (E) and bound (E\cdotS) GUS is only 0.22 Å, implying that there is minimal conformational rearrangement in GUS upon binding with pNP-GLU. Therefore, in the remainder, we approximate \(B_{E_S}\) with \(I_{E_S}\) (interaction energy with the substrate, pNP-GLU), where \(B_{E_S}\) is equal to the \(I_{E_S}\) plus deformation energy due to the geometric distortion of the ligand and enzyme upon binding (see Equations 1 and 2) [57]. The reduction in computational time allows for more redesign cycles within IPRO. Using Equation 6 and the assumption that \(B_{E_S}=I_{E_S}\), we deduce that \(K_M\) and \(I_{E_S}\) for the mutant/wild-type enzymes are related through Equation 7.

\[
I_{E_S} = RT \ln(K_M) \quad (7)
\]
Equation 5 implies a linear correlation between \( \ln(K_M) \) and \( IE_S \). Figure 6 depicts the measured \( K_M \) values [48-50] and corresponding calculated \( IE_S \)'s for the wild-type GUS and five mutants. The correlation coefficient of 0.960 implies that the calculated interaction energy is a good descriptor of the GUS/pNP-GLU binding energy. While the actual magnitude of the energy terms in the y-axis is not quantitatively accurate, the relative ordering of the mutants in terms of their \( K_M \) values is consistent with the data.

Unlike \( K_M \), which depends on binding at the ground state, \( k_{cat} \) is directly related to the reaction rate. The rate constant of a reaction is related to the change in the Gibb's free energy between the ground and transition states, based on the Eyring-Polanyi equation derived from transition state theory (Equation 8) (see also Figure 5).

\[
k = \frac{\kappa k_B T}{h} e^{\left(\frac{-\Delta G^\ddagger}{RT}\right)}
\]  

In Equation 8, \( k \) is the rate constant, \( h \) is Planck's constant, \( \kappa \) is the transmission coefficient (assumed invariant among all mutants), \( k_B \) is the Boltzmann constant, and \( \Delta G^\ddagger \) is the change in Gibb's free energy between the ground and transition states (Equation 9).

\[
\Delta G^\ddagger = G_{E-TS}^{\min} - G_{E-S}^{\min}
\]

We cannot directly computationally assess \( \Delta G^\ddagger \) because the transition state structure is unknown. Since the structure of the transition state is unavailable, we postulate that mutations that lead to beneficial interactions of the enzyme with its TSA would produce similar contacts with the unresolved transition state. Equation 10 expresses this postulate mathematically by implying that the difference between the minimized free energy of the transition state and the TSA is invariant with respect to mutations introduced on the enzyme.
\[ G_{E-TS}^{\text{min}} - G_{E-TSA}^{\text{min}} = C \neq f(\text{mutations}) \]  \hspace{1cm} (10)

The unknown (for IPRO trajectories where the TSA is the ligand) free energy of the Michaelis complex can be eliminated by combining Equations 1 and 9, yielding Equation 11.

\[ BE_S + \Delta G^\ddagger = G_{E-TS}^{\text{min}} - G_S^{\text{min}} - G_E^{\text{min}} \]  \hspace{1cm} (11)

Equation 10 can be used to eliminate the unknown free energy of the bound transition state \(G_{E-TS}^{\text{min}}\), giving Equation 12.

\[ BE_S + \Delta G^\ddagger = G_{E-TSA}^{\text{min}} + C - G_S^{\text{min}} - G_E^{\text{min}} \]  \hspace{1cm} (12)

\(G_{TSA}^{\text{min}}\) and \(G_S^{\text{min}}\) are both invariant with respect to mutations to the enzyme.

\[ G_{TSA}^{\text{min}} - G_S^{\text{min}} = C1 \neq f(\text{mutations}) \]  \hspace{1cm} (13)

\(G_S^{\text{min}}\) in Equation 12 can thus be rewritten in terms of \(G_{TSA}^{\text{min}}\) using Equation 13 and combining the two constants \((C\) and \(C1\)) into a single constant \((C2)\), resulting in Equation 14.

\[ BE_S + \Delta G^\ddagger = G_{E-TSA}^{\text{min}} - G_{TSA}^{\text{min}} - G_E^{\text{min}} + C2 \]  \hspace{1cm} (14)

We have already shown that computationally approximated \(\text{IE}_S\) provides a good approximation for \(\text{BE}_S\). We assume that similarly the interaction energy with the TSA (\(\text{IE}_{TSA}\)) is a good approximation for \(\text{BE}_{TSA}\).

\[ G_{TSA}^{\text{min}} = G_{TSA} \]  \hspace{1cm} (15)

\[ G_E^{\text{min}} = G_E \]  \hspace{1cm} (16)

Substitution of \(G_{TSA}^{\text{min}}\) from Equation 15 and \(G_E^{\text{min}}\) from Equation 16 into Equation 14 results in Equation 17.

\[ BE_S + \Delta G^\ddagger = G_{E-TSA}^{\text{min}} - G_{TSA}^{\text{min}} - G_E + C2 \]  \hspace{1cm} (17)

Equation 17 is further simplified by substituting the definition of \(\text{IE}_{TSA}\) (see Equation 2, where the substrate in this case is the TSA).
\[ BE_S + \Delta G^\ddagger = IE_{TSA} + C2 \]  

(18)

Constant \( C2 \) can be eliminated from Equation 18 by expressing it as the difference in the interaction energies between mutant and wild-type (\( WT \)) enzymes,

\[ \Delta IE_{TSA} = \Delta BE_S + \Delta \Delta G^\ddagger \]  

(19)

where \( \Delta IE_{TSA} = IE_{TSA} - IE_{TSA,WT} \), \( \Delta BE_S = BE_S - BE_{S,WT} \), and \( \Delta \Delta G^\ddagger = \Delta G^\ddagger - \Delta G^\ddagger_{WT} \). \( \Delta BE_S \) and \( \Delta \Delta G^\ddagger \) can be recast using Equation 6 (at constant temperature).

\[ \Delta BE_S = RT[\ln(K_M) - \ln(K_M^{WT})] \]  

(20)

\[ \Delta \Delta G^\ddagger = -RT[\ln(k_{cat}) - \ln(k_{cat}^{WT})] \]  

(21)

Substituting \( \Delta BE_S \) from Equation 20 and \( \Delta \Delta G^\ddagger \) from Equation 21 into Equation 19 yields

\[ \Delta IE_{TSA} = -RT \left[ \ln \left( \frac{k_{cat}}{K_M} \right) - \ln \left( \frac{k_{cat}^{WT}}{K_M^{WT}} \right) \right] \]  

(22)

Equation 22 can be used to relate computationally accessible metrics to \( k_{cat}/K_M \), which dictates the catalytic efficiency of the enzyme under substrate limiting conditions ([\( S \]) \( << \) \( K_M \)). Equations 7 and 22 can be combined to directly link \( k_{cat} \) to computationally accessible metrics (Equation 23) providing a design criterion under conditions where substrate concentration is in excess ([\( S \]) \( >> \) \( K_M \)).

\[ \frac{\Delta IE_{TSA}}{(RT)_{TSA}} - \frac{\Delta IE_S}{(RT)_S} = \ln(k_{cat}^{WT}) - \ln(k_{cat}) \]  

(23)

In Equation 23, \( \Delta IE_S = IE_S - IE_{S,WT} \), \( (RT)_{TSA} \) is the \( RT \) term in Equation 22, and \( (RT)_S \) is the \( RT \) term in Equation 7. In general, the temperature in the correlating expressions does not match the experimental ones, therefore \( (RT)_{TSA} \) and \( (RT)_S \) are allowed to differ. For example, for GUS, \( (RT)_{TSA} = 3.66 \text{ kcal/mol} \) (\( T = 4.65 \times 10^4 \text{ K} \)) while \( (RT)_S = 92.36 \text{ kcal/mol} \) (\( T = 1840 \text{ K} \)). Similarly high temperatures are observed in the quantification of RNA-ribosome binding calculations in RBSCalculator [58].
A strong correlation ($R^2=0.864$) is observed between $IE_{TSA}$ and the natural logarithm of $k_{\text{cat}}/K_M$ (see Figure 7), suggesting that $IE_{TSA}$ is a good descriptor of $k_{\text{cat}}/K_M$. The same strong correlation (i.e., $R^2=0.854$) is observed between $IE_{TSA}/(RT)_{TSA}-IE_S/(RT)_{S}$ and the natural logarithm of $k_{\text{cat}}$ (see Figure 8). As a control, we verified that the energy difference between the ground state and unbound reactants does not correlate with the catalytic efficiency (see Figure 9).

**Redesign of GUS for improving activity with pNP-GLU**

OptZyme was first used to identify beneficial mutations that improve $K_M$ ($IE_S$), $k_{\text{cat}}/K_M$ ($IE_{TSA}$), and $k_{\text{cat}}$ ($0.27 \times IE_{TSA} - 0.01 \times IE_S$) with pNP-GLU. The energy terms within parenthesis provide the objective functions that were minimized by OptZyme in each case by identifying appropriate residue mutations. In addition, constraints that ensure that both the substrate and TSA favorably bind GUS (i.e., $IE_S<0$, $IE_{TSA}<0$) were included in OptZyme runs. Design positions (mutable amino acid positions) were selected in locations that are likely to impact active site geometry and directly mediate interactions with the substrate. The same set of design positions was chosen for all sets of calculations (H162, D163, F164, V355, G356, L361, G362, W549, N550).

A high frequency of mutations to glycine by OptZyme was initially observed presumably to avoid steric clashes within the highly-packed active site of GUS. To remedy this bias, we first performed multiple sequence alignments to extract natural amino acid usage patterns for the enzyme under study. The first family alignment was performed using PFAM [59] between GUS and the family 2 glycosyl hydrolases, and the second alignment was performed between GUS and all other β-glucuronidases (as identified in BRENDA [60]).
using Clustal-Omega [61]. If an amino acid was observed at least once in the alignment of all β-glucuronidases (181 sequences, see Figure 10 for distribution of residues within alignment) or if it was observed at least 5% of the time for the family 2 glycosyl hydrolases (excluding gaps, 3975 sequences), then this residue was permitted at the design position (see Table 3 for permissible mutations). Additionally, the total number of glycine residues throughout all design positions was restricted to be at most 2 (matching the glycine utilization frequency in wild-type).

50 independent trajectories of OptZyme were run to optimize $K_M$, $k_{cat}/K_M$, and $k_{cat}$ for GUS. NOE restraints were again used to maintain the optimal catalytic geometry of GUS (Table 2, Figure 4). Each trajectory of OptZyme consisted of 5000 iterations, and simulated annealing was used after 100 cycles (using $T = 7268K$, which corresponds to an acceptance rate of about 50% of redesigns within 10 kcal/mol of the best mutant) to avoid premature convergence to local minima of the GUS free energy landscape. The CHARMM energy terms used were identical to those used in the testing of the TSA-based redesign paradigm, and the backbone-dependent Dunbrack rotamer library was used for side chain optimization [62].

OptZyme was used to identify three libraries of mutants that were computationally predicted to exhibit enhanced kinetic parameters relative to WT (see Table 4, Figure 11). The observed mutants seemed to lower the relevant interaction energy predominantly through improving flexibility in the active site, providing increased solvation stabilization, or improving the electrostatics interaction energy. In all three libraries, mutations to small amino acids, such as glycine or alanine, were frequently observed. The increased flexibility of the backbone enabled by the small residue allowed for nearby backbone amides to come
within hydrogen bonding distance of a hydrogen bond acceptor, as in the case of W549A (K_m), V355A (K_M), and W549A (k_cat). Mutations to polar side chains were also quite prevalent for solvent accessible side chains (e.g., G362R (k_cat/K_M), G362S (K_M), or L361R (k_cat)), providing an energetic benefit because of the hydrophobic effect [63]. Electrostatic interactions, mostly through hydrogen bonding with the hydroxyl groups of the sugar moiety were ubiquitous in each library. For example, W549R in the k_cat/K_M optimization is involved in two hydrogen bonds with D-glucaro-1, 5-lactone. A systematic cutoff is defined for identifying mutations that are representative of the K_M- or k_cat/K_M- optimized libraries. A mutation is considered as representative of a library if it occurs at least 15% of the time for a given design position and at the same time 10% more frequently than in the other libraries. H162A and H162G (extra flexibility of the protein backbone), D163S (enhanced solvation), and G362R (hydrogen bonding/ solvation effects) were mutations representative of the pNP-GLU k_cat/K_M-optimized library. While OptZyme cannot replicate the WT aromatic side chains (because pi-stacking is excluded in the energy calculations), aromatic residues would be unlikely to enhance specificity between the largely unchanged ring conformations of pNP-GLU and pNP-GAL. OptZyme stabilizes the mutant’s tertiary structure through other beneficial energetic contributions (i.e., solvation, hydrogen bonding networks).

Experimental validation of the mutants can be carried out using a high-throughput assay, where the absorbance of the leaving group, para-nitrophenolate, is measured based on its high absorbance at 405 nm [64]. The design of mutants for GUS is handicapped as WT GUS is already very active and the scope for identifying significantly improved mutants is limited. However, GUS activity with pNP-GAL is ~10^7 lower than for pNP-GLU [64].
Therefore, the entire gamut of beneficial interactions leading to switch of specificity from pNP-GLU to pNP-GAL would be detectable using a high-throughput assay.

**Redesign of GUS for introducing catalytic activity with the new substrate pNP-GAL**

The same methodology that was used for improving the kinetic parameters of pNP-GLU was also used to introduce catalytic activity with a new substrate, pNP-GAL. As before, three libraries were constructed that were predicted to enhance $K_M$, $k_{cat}/K_M$, or $k_{cat}$ of GUS towards pNP-GAL (see Table 5, Figure 12). The design of the $k_{cat}/K_M$- and $k_{cat}$-optimized libraries required the use of the TSA for pNP-GAL, D-galactono-1, 5-lactone (see Figure 13). The constructed mutants were stabilized in a similar manner as those described for pNP-GLU. The only representative mutant in the pNP-GAL $K_M$-optimized library was L361N (electrostatic interactions with pNP-GAL C5 substituent/solvation enhancement). L361G (extra flexibility of GUS backbone), W549R (hydrogen bonding with pNP-GAL C2 hydroxyl group), and N550S (solvation enhancement) were enriched mutants for the pNP-GAL $k_{cat}/K_M$-optimized library.

Mutations enriched in the pNP-GAL libraries were also identified. The analysis revealed only one such additional mutation, H162N (electrostatic interaction with the C4 substituent). Structural analysis also revealed that the backbone carbonyl of F161 formed a hydrogen bond in 97.5% of the examined structures (each mutant in Table 5; 2 examined structures for each $k_{cat}$-optimized mutant) with the C5 substituent of pNP-GAL, and this interaction was not observed for pNP-GLU. Thus, the composition of the residue at design position 162 may directly promote the backbone interaction with pNP-GAL. Although the
frequency of the H162S mutation is also prevalent in the pNP-GLU $K_M$-optimized library (see Figure 11), this mutation is observed in 13.3% of the top 10 mutants for each library (30 total mutants; Table 4) for pNP-GLU and 36.7% (Table 5) of the pNP-GAL libraries. Therefore, H162S may be important for the interaction of F161 with the ligand. Among the top 10 mutants for each library (see Table 5), D163K, L361R, and L361E were the mutations at the design positions that were observed to hydrogen bond with either the C4 or C5 substituent of the ligand.

Several mutations were found that make direct contact with the novel ligand. Since the differences between pNP-GLU and pNP-GAL lie at the C4 and C5 substituents of the carbohydrate moiety, mutations that create contacts with these substituents are expected. Indeed, this is the case for the D163K, L361R, and L361E mutations, as well as the contact by F161. However, at least one of the mutants, W549R, representative of the pNP-GAL-optimized libraries forms a contact with the ligand but at the unchanged portion of the carbohydrate. The reason that this mutation is much more common in the pNP-GAL libraries than in the pNP-GLU libraries is because of a slightly deformed sugar ring of pNP-GAL relative to pNP-GLU. This small difference was able to create a noticeable divergence between the libraries. The results show that despite the similar structures used in the library design, OptZyme is sensitive enough to detect even these minor structural variances between substrates.

Figures 11 and 12 indicate that there is a strong tendency for OptZyme to mutate to glycine across each of the six libraries. This tendency can be partially explained by the backbone flexibility of smaller amino acids. Figure 14 shows the Ramachandran plot for the top 50 mutants in each pNP-GAL library (200 total mutants; pNP-GLU results were similar,
data not shown), with the statistically preferred regions for residues that include a $C_\beta$, indicated [65]. Figure 14 shows that position 356 lies within the disallowed region of the map so glycine must be selected to avoid a steric clash. Similarly, mutating to glycine, alanine, or serine can alleviate the steric clash caused by the backbone conformation at position 550. Due to the highly-packed active site of GUS, mutation of one residue to a larger side chain should correspond with another residue’s side chain becoming more compact to avoid a steric clash. These two effects can explain the majority of the glycine and alanine (and perhaps to a lesser extent, serine) mutants in the data.
**Figure 4:** Active Site Geometry and Restrained Catalytic Contacts. The active site is depicted in a ball-and-stick representation (C = black, O = red, N = blue, H = white). The nonbonded interactions seen reflect the bonds restrained (as listed in Table 2). Key catalytic residues are labeled by their one-letter amino acid abbreviation followed by their position number (PNP = pNP-GLU). Atoms involved in restraints are labeled, along with interatomic distances.
Table 2: NOE Restraints Applied During CHARMM Energy Minimization. Restraints were placed on key catalytic contacts, determined from previous experimental and preliminary quantum mechanical information [52]. Distances between atoms were selected based on typical nonbonded interaction lengths, and spring constants were determined iteratively so that the distances were properly restrained while not over-constraining the system. \( k_{\text{min}} \) was the harmonic constant implemented if the interatomic distance was too small, and \( k_{\text{max}} \) was the harmonic constant used if the interatomic distance was too large. \( k_{\text{min}} < k_{\text{max}} \) because catalytic contacts would remain intact at smaller distances. Entries are shown in Figure 4.

<table>
<thead>
<tr>
<th>Atom 1</th>
<th>Atom 2</th>
<th>Minimum (Å)</th>
<th>Maximum (Å)</th>
<th>( k_{\text{min}} )</th>
<th>( k_{\text{max}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>GLU 413: HE2</td>
<td>PNP: O6</td>
<td>1.7</td>
<td>1.8</td>
<td>75.0</td>
<td>100.0</td>
</tr>
<tr>
<td>GLU 504: OE2</td>
<td>PNP: C1</td>
<td>2.5</td>
<td>2.6</td>
<td>75.0</td>
<td>100.0</td>
</tr>
<tr>
<td>ARG 467: HH12</td>
<td>PNP: O7</td>
<td>1.7</td>
<td>1.8</td>
<td>75.0</td>
<td>100.0</td>
</tr>
<tr>
<td>ARG 467: HH22</td>
<td>PNP:O8</td>
<td>1.7</td>
<td>1.8</td>
<td>75.0</td>
<td>100.0</td>
</tr>
</tbody>
</table>
Figure 5: Qualitative GUS Free Energy Diagram Based Upon *in vacuo* QM Calculations. The free energy of each intermediate within the dashed box is grounded on its potential energy, as calculated using QM. Transition states and intermediates found using QM are also labeled according to Figure 2 (*italicized, above curve*). The energy barrier between states $c$ and $d$ is nearly barrier-less. The free energy values along the remainder of the curve are purely hypothetical. Each intermediate is labeled according to the convention used in Equation 3. Based on the known and hypothesized free energies, the reaction of the Michaelis complex to form the first intermediate ($k_2$, as written in Equation 3) is rate-limiting. Thus, the transition state for the entire reaction is identified ($E$·$T$S), as well as its corresponding energy barrier ($\Delta G^\ddagger$).
Figure 6: Ground State Computational Interaction Energy ($\text{IE}_S$) versus the Natural Logarithm of Experimental $K_M$. Interaction energies were calculated using CHARMM, and experimental data was obtained from literature [48-50]. Data points are labeled by the mutations (1: I12V, F365S, W529L, S557P, I560V; 2: S557P, N566A, K568F; 3: S557P, N566S, K568Q; 4: S557P, N566S, K568Q, V473A, G601S; 5: N66D, D151N, A219V, I396T, T480A, Q498R, D508E, K567R) that deviate from wild-type (WT). CHARMM calculated interaction energies at the ground state explain the observed changes in $K_M$ for GUS mutants ($R^2 = 0.960$).
Figure 7: Computationally-determined Transition State Analogue Interaction Energy ($E_{TSA}$) versus Experimental $\ln(k_{cat}/K_M)$. Data was collected as described in Figure 6 (also see for mutant sequence). Enzyme mutants with higher catalytic efficiency ($k_{cat}/K_M$) have a stronger affinity for 1,5-glucarolactone ($R^2 = 0.864$).
Figure 8: Scaled Difference Between Interaction Energy with the Transition State Analogue ($E_{TSA}$) and Interaction Energy with pNP-GLU ($E_S$) versus the Natural Logarithm of $k_{cat}$. Data was obtained as detailed in the caption of Figure 6 (also see for mutant sequence). Scaling is required because the non-quantitative nature of the energy calculations ($\beta = (RT)^{-1}$). With scaling, it is apparent that the turnover number increases as the difference becomes more negative. These results suggest that as the enzyme binds the transition state more tightly, the turnover number increases ($R^2 = 0.854$).
Figure 9: Ground State Interaction Energy (IE$_S$) Correlation with Catalytic Efficiency. Data was obtained as described in Figure 6 (also see for mutant sequence). No discernible correlation between the interaction energy with pNP-GLU and the turnover number is observed ($R^2 = 0.545$).
**Figure 10:** Distribution of Amino Acids in a Sequence Alignment for all β-glucuronidases. The sequence alignment was performed over all β-glucuronidases (as identified using BRENDA) using the Clustal-Omega algorithm. 181 unique sequences were used during the alignment. Design position numbers indicate the position within *E. coli* β-glucuronidase, and the one-letter abbreviation for WT *E. coli* β-glucuronidase is provided at each position. Only amino acids observed >1% of the time at a given position are shown since smaller bars were difficult to decipher the color. With the exception of H162, the *E. coli* WT residue is the amino acid most frequently observed in the alignment.
Table 3: Permitted Amino Acids at each Design Position. Using multiple sequence alignments at both the family and enzyme level, a certain subset of amino acids were allowable at each position. If the amino acid appeared at least once in the β-glucuronidase alignment or was observed at least 5% of the time in the family 2 glycosyl hydrolases, then it was permitted. This table contains the list of permitted amino acids (using one-letter abbreviations) at each design position.

<table>
<thead>
<tr>
<th>Design Position</th>
<th>Permitted Amino Acids</th>
</tr>
</thead>
<tbody>
<tr>
<td>162</td>
<td>A, C, F, G, H, I, L, M, N, Q, S, T, V, Y</td>
</tr>
<tr>
<td>163</td>
<td>C, D, F, G, K, M, Q, R, S, T, W</td>
</tr>
<tr>
<td>164</td>
<td>F, M, Q, W, Y</td>
</tr>
<tr>
<td>550</td>
<td>A, E, F, G, I, L, N, Q, S, T, V, Y</td>
</tr>
</tbody>
</table>
**Table 4:** Top 10 Mutants Identified Using OptZyme for Optimizing $K_M$, $k_{cat}/K_M$, and $K_M$ for pNP-GLU. The one-letter amino abbreviation is provided for each design position, as well as the WT residue. Energy values ($K_M$: IE, $k_{cat}/K_M$: IE$_{TSA}$, $k_{cat}$: 0.27 IE$_{TSA}$ – 0.01 IE) are calculated using CHARMM.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Energy</th>
<th>Design Positions</th>
<th>162</th>
<th>163</th>
<th>164</th>
<th>355</th>
<th>356</th>
<th>361</th>
<th>362</th>
<th>549</th>
<th>550</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K_M$</td>
<td>WT</td>
<td>-116.96</td>
<td>H</td>
<td>D</td>
<td>F</td>
<td>V</td>
<td>G</td>
<td>L</td>
<td>G</td>
<td>W</td>
<td>N</td>
</tr>
<tr>
<td>1</td>
<td>-370.16</td>
<td>Q     D</td>
<td>Q</td>
<td>A</td>
<td>A</td>
<td>G</td>
<td>D</td>
<td>R</td>
<td>G</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>-361.76</td>
<td>A     D</td>
<td>Q</td>
<td>A</td>
<td>G</td>
<td>G</td>
<td>R</td>
<td>R</td>
<td>A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>-360.86</td>
<td>G     S</td>
<td>Q</td>
<td>A</td>
<td>G</td>
<td>K</td>
<td>D</td>
<td>A</td>
<td>S</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>-354.36</td>
<td>Q     D</td>
<td>Q</td>
<td>A</td>
<td>A</td>
<td>G</td>
<td>D</td>
<td>A</td>
<td>G</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>-353.16</td>
<td>A     S</td>
<td>Q</td>
<td>A</td>
<td>G</td>
<td>G</td>
<td>D</td>
<td>A</td>
<td>A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>-347.96</td>
<td>S     S</td>
<td>Q</td>
<td>A</td>
<td>A</td>
<td>G</td>
<td>D</td>
<td>A</td>
<td>G</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>-347.66</td>
<td>A     S</td>
<td>F</td>
<td>A</td>
<td>A</td>
<td>G</td>
<td>D</td>
<td>A</td>
<td>G</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>-344.16</td>
<td>A     S</td>
<td>Q</td>
<td>A</td>
<td>G</td>
<td>G</td>
<td>R</td>
<td>A</td>
<td>S</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>-342.76</td>
<td>S     D</td>
<td>Q</td>
<td>A</td>
<td>G</td>
<td>G</td>
<td>R</td>
<td>R</td>
<td>A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>-341.76</td>
<td>A     S</td>
<td>Q</td>
<td>C</td>
<td>G</td>
<td>K</td>
<td>D</td>
<td>A</td>
<td>G</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$k_{cat}/K_M$</td>
<td>WT</td>
<td>-90.29</td>
<td>H</td>
<td>D</td>
<td>F</td>
<td>V</td>
<td>G</td>
<td>L</td>
<td>G</td>
<td>W</td>
<td>N</td>
</tr>
<tr>
<td>1</td>
<td>-375.39</td>
<td>N     D</td>
<td>Q</td>
<td>A</td>
<td>G</td>
<td>G</td>
<td>D</td>
<td>A</td>
<td>A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>-373.19</td>
<td>N     D</td>
<td>Q</td>
<td>A</td>
<td>G</td>
<td>G</td>
<td>K</td>
<td>A</td>
<td>A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>-372.89</td>
<td>A     D</td>
<td>Q</td>
<td>A</td>
<td>G</td>
<td>G</td>
<td>K</td>
<td>A</td>
<td>A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>-370.79</td>
<td>G     K</td>
<td>Q</td>
<td>A</td>
<td>G</td>
<td>S</td>
<td>K</td>
<td>A</td>
<td>A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>-365.99</td>
<td>A     R</td>
<td>Q</td>
<td>A</td>
<td>G</td>
<td>S</td>
<td>G</td>
<td>A</td>
<td>A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>-363.49</td>
<td>Q     D</td>
<td>Q</td>
<td>A</td>
<td>G</td>
<td>G</td>
<td>D</td>
<td>A</td>
<td>A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>-362.89</td>
<td>A     D</td>
<td>Q</td>
<td>A</td>
<td>G</td>
<td>K</td>
<td>D</td>
<td>A</td>
<td>G</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>-361.89</td>
<td>A     D</td>
<td>Q</td>
<td>A</td>
<td>G</td>
<td>G</td>
<td>D</td>
<td>A</td>
<td>A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>-359.19</td>
<td>S     D</td>
<td>Q</td>
<td>A</td>
<td>G</td>
<td>G</td>
<td>S</td>
<td>A</td>
<td>A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>-357.39</td>
<td>S     S</td>
<td>Q</td>
<td>E</td>
<td>G</td>
<td>K</td>
<td>G</td>
<td>A</td>
<td>A</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| $k_{cat}$ | WT | -31.01 | H | D | F | V | G | L | G | W | N |
| 1    | -90.97  | G     D           | Q   | A   | G   | R   | D   | A   | A   |     |     |
| 2    | -89.65  | G     D           | Q   | A   | G   | K   | D   | A   | A   |     |     |
| 3    | -89.50  | Q     D           | Q   | A   | G   | G   | R   | A   | A   |     |     |
| 4    | -86.28  | G     D           | Q   | A   | G   | E   | K   | A   | A   |     |     |
| 5    | -84.97  | N     D           | Q   | V   | G   | S   | G   | A   | A   |     |     |
| 6    | -83.50  | C     D           | Q   | V   | G   | S   | G   | A   | A   |     |     |
| 7    | -82.87  | G     K           | Q   | E   | G   | S   | R   | A   | A   |     |     |
| 8    | -81.58  | G     K           | Q   | A   | G   | A   | K   | A   | A   |     |     |
| 9    | -81.44  | Q     D           | Q   | A   | G   | G   | K   | A   | A   |     |     |
| 10   | -80.33  | L     D           | Q   | A   | G   | G   | K   | A   | A   |     |     |
Figure 11: Distribution of Amino Acids at Each Design Position for Top 50 GUS Mutants to Enhance $K_M$, $k_{cat}/K_M$, and $k_{cat}$ for pNP-GLU. Design position numbers indicate the position within GUS, and the one-letter abbreviation for WT GUS is provided.
Table 5: Top 10 Mutants Identified Using OptZyme for Optimizing $K_M$, $k_{cat}/K_M$, and $K_M$ for pNP-GAL. The one-letter amino abbreviation is provided for each design position, as well as the WT residue. Energy values ($K_M$: IE$S$, $k_{cat}/K_M$: IE$TSA$, $k_{cat}$: 0.27 IE$TSA$ – 0.01 IE$S$) are calculated using CHARMM.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Energy</th>
<th>Design Positions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>162</td>
</tr>
<tr>
<td>$K_M$</td>
<td>WT</td>
<td>-16.89</td>
</tr>
<tr>
<td>1</td>
<td>-365.39</td>
<td>S</td>
</tr>
<tr>
<td>2</td>
<td>-361.89</td>
<td>S</td>
</tr>
<tr>
<td>3</td>
<td>-359.69</td>
<td>S</td>
</tr>
<tr>
<td>4</td>
<td>-351.89</td>
<td>A</td>
</tr>
<tr>
<td>5</td>
<td>-347.29</td>
<td>N</td>
</tr>
<tr>
<td>6</td>
<td>-344.79</td>
<td>S</td>
</tr>
<tr>
<td>7</td>
<td>-343.09</td>
<td>S</td>
</tr>
<tr>
<td>8</td>
<td>-340.29</td>
<td>S</td>
</tr>
<tr>
<td>9</td>
<td>-337.89</td>
<td>S</td>
</tr>
<tr>
<td>10</td>
<td>-336.69</td>
<td>G</td>
</tr>
</tbody>
</table>

| $k_{cat}/K_M$ | WT | 21.55 | H    | D    | F    | V    | G    | L    | G    | W    | N    |
|              | 1  | -248.85 | A    | D    | Q    | A    | G    | R    | G    | R    | S    |
|              | 2  | -247.35 | Q    | D    | Q    | A    | G    | K    | D    | R    | G    |
|              | 3  | -240.05 | S    | D    | Q    | A    | G    | G    | S    | R    | S    |
|              | 4  | -233.05 | A    | D    | Q    | A    | G    | G    | S    | R    | S    |
|              | 5  | -231.75 | N    | D    | Q    | A    | G    | G    | D    | K    | S    |
|              | 6  | -228.65 | N    | D    | Q    | A    | G    | G    | D    | R    | S    |
|              | 7  | -224.85 | S    | D    | Q    | A    | G    | G    | K    | R    | S    |
|              | 8  | -222.55 | Q    | D    | Q    | A    | G    | D    | K    | A    | G    |
|              | 9  | -222.35 | A    | D    | Q    | A    | G    | G    | D    | K    | S    |
|              | 10 | -221.05 | A    | D    | Q    | A    | G    | G    | K    | K    | S    |

| $k_{cat}$ | WT | 6.07  | H    | D    | F    | V    | G    | L    | G    | W    | N    |
|           | 1  | -60.21 | G    | D    | Q    | A    | G    | E    | K    | A    | A    |
|           | 2  | -58.63 | A    | D    | Q    | A    | G    | G    | R    | T    | A    |
|           | 3  | -58.37 | A    | D    | Q    | A    | G    | G    | A    | T    | A    |
|           | 4  | -57.35 | A    | D    | Q    | A    | G    | G    | K    | T    | A    |
|           | 5  | -55.97 | A    | D    | Q    | A    | G    | E    | K    | A    | A    |
|           | 6  | -55.85 | S    | D    | Q    | A    | G    | D    | K    | A    | G    |
|           | 7  | -55.61 | A    | D    | Q    | A    | G    | G    | K    | K    | S    |
|           | 8  | -55.56 | A    | D    | M    | A    | G    | G    | S    | R    | S    |
|           | 9  | -54.46 | G    | D    | Q    | A    | G    | E    | K    | R    | S    |
|           | 10 | -52.65 | S    | D    | Q    | A    | G    | S    | G    | K    | S    |
Figure 12: Distribution of Amino Acids at Each Design Position for Top 50 GUS Mutants to Enhance $K_M$, $k_{cat}/K_M$, and $k_{cat}$ for pNP-GAL. Design position numbers indicate the position within GUS, and the one-letter abbreviation for WT GUS is provided.
Figure 13: Structural Comparison Between pNP-GLU (i), 1,5-glucarolactone (ii), pNP-GAL (iii), 1,5-galactonolactone (iv). 1,5-glucarolactone is identical to 1,5-galactonolactone with the exception of the same differences that exist between pNP-GAL and pNP-GLU (green circles). In going from pNP-GLU to pNP-GAL, the chirality of the C4 carbon is reversed, and the C5 substituent is changed from a carboxylic acid to an alcohol.
Figure 14: Ramachandran Plot of Top pNP-GAL Mutants. 50 of the top mutants from each of the pNP-GAL libraries were examined. "Core" (white), "allowed" (off white), "generous" (gray), and "outside" (dark gray) regions of the Ramachandran plot were determined by Morris et al. [65]. Results show that glycine residues (crosses) are frequently observed in the "generous" or "outside" regions of the map. Alternatively, the other 19 standard amino acids (squares) are much less frequently observed in the "generous" or "outside" regions. Glycine residues can avoid some of the steric repulsion that is more difficult to avoid for residues with a Cβ. While other amino acid can undergo contortions in the side chain to avoid a strong steric clash, mutation to a glycine residue is more favorable.
Chapter 4: DISCUSSION

Amongst the pNP-GAL libraries, the KM-optimized library is enriched with smaller amino acids. Although this observation could be an artifact due to the larger size of pNP-GAL relative to its TSA, the design positions were chosen at the edge of the active site further away from the pNP substituent. Thus, the smaller side chains in the KM-optimized library are more likely a reflection of the chair conformation of the sugar ring, which has a larger excluded volume than the planar geometry of the TSA. By mutation of the WT side chains to the large, polarizable side chains that are representative of the kcat/KM-optimized library (H162Q, L361K, G362D, W549R), the planar form of the molecule is stabilized through efficient packing of the enzyme and beneficial electrostatic interactions. Therefore, OptZyme suggests that mutation to these large, polar side chains forces pNP-GAL into its more planar form, which stabilizes its transition state and thus promotes catalytic activity.

We have theoretically derived a set of equations that allow us to measure enzymatic properties using computationally-accessible binding calculations. Using a QM-derived hypothetical reaction mechanism for GUS, we were able to show that interaction energy with the substrate (pNP-GLU) correlates with KM, IE_S = RT ln(KM), and interaction energy with the TSA correlates with kcat/KM, IE_TSA ∝ -RT ln(kcat/KM). kcat can be measured through a weighted combination of these two interaction energies, [IE_TSA/(RT)_TSA] - [IE_S/(RT)_S] ∝ ln(kcat). By employing OptZyme, a computational tool used to discover mutations that improve KM, kcat, or kcat/KM, we were able to identify mutations that were predicted to exhibit enhanced enzymatic activity for pNP-GLU. The identified mutants stabilized the ligand mainly through hydrogen bonding networks, improved solvation, and efficient packing within the active site. Additionally, we utilized OptZyme to generate a library of
mutants in favor of kinetic parameters for a similar substrate, pNP-GAL, which was shown to exhibit poor activity in GUS [64]. Though these substrates are quite similar, OptZyme was able to identify novel contacts with the ligand in the pNP-GAL libraries that were not observed in the pNP-GLU libraries. Several mutations were enriched in the pNP-GAL libraries, namely those that interact with the distorted sugar ring conformation or its altered substituents. In comparison of the $K_M$- and $k_{cat}/K_M$-optimized libraries for pNP-GAL, we found that large, polar side chains were more common in the $k_{cat}/K_M$-optimized library. This was a reflection of the more planar geometry of the TSA molecule, suggesting that mutations to large, polar side chains can stabilize the transition state through interactions with the hydroxyl substituents and efficient packing, thereby improving enzymatic activity.

In conclusion, OptZyme is an inexpensive computational tool that can be used to predict activity for a substrate with an unknown natural counterpart, as long as an enzyme with the proper functionality on a somewhat similar substrate exists. OptZyme is available for download at maranas.che.psu.edu. This proof of concept system for the testing of OptZyme can allow us to predict mutations that enhance enzymatic activity for other systems.
Appendix: Derivation of Equations 4 & 5

Equation 1 can be re-written by treating the fast (not rate-limiting) reaction rates as equilibrated species (Equation A1), and the reaction rate can then be written (Equation A2).

\[
\begin{align*}
K_A & \quad k_2 \quad K_P \\
E + S & \rightleftharpoons E \cdot S \rightleftharpoons E \cdot I_1 \rightleftharpoons E + P \\
k_{-2}
\end{align*}
\]  
(A1)

\[
v = k_2 [E \cdot S] - k_{-2} [E \cdot I_1]
\]

(A2)

Equilibrium constants, \(K_A\) and \(K_P\), are defined using Equations A3 and A4, respectively.

Equation A2 can also be re-written in terms of \(K_A\) and \(K_P\) (Equation A5).

\[
K_A = \frac{[E \cdot S]}{[E][S]}
\]

(A3)

\[
K_P = \frac{[E][P]}{[E \cdot I_1]}
\]

(A4)

\[
v = k_2 K_A [E][S] - \frac{k_{-2} [E][P]}{K_P}
\]

(A5)

Using a site balance on the enzyme, Equation A6, the concentration of unbound enzyme, \([E]\), can be expressed (Equation A7) through the substitution of the expressions for \(K_A\) and \(K_P\) from Equations A3 and A4, respectively.

\[
[E_0] = [E] + [E \cdot S] + [E \cdot I_1]
\]

(A6)

\[
[E_0] = [E] + K_A [E][S] + \frac{[E][P]}{K_P}
\]

\[
[E_0] = [E] \left( \frac{[E_0]}{1 + K_A [S] + \frac{[P]}{K_P}} \right)
\]

\[
[E] = \frac{[E_0]}{1 + K_A [S] + \frac{[P]}{K_P}}
\]

(A7)
Substituting the expression for $[E]$ from Equation A7 into Equation A5 allows for the expression of the rate in terms of measurable quantities (Equation A8).

\[

v = \left( \frac{[E_0]}{1 + K_A[S] + \frac{[P]}{K_P}} \right) \left( k_2 K_A[S] - \frac{k_{-2}[P]}{K_P} \right)

\]

\[

v = \frac{k_2 K_A[E_0][S]}{1 + K_A[S]}

\]

\[

v = \frac{k_A(K_A)[E_0][S]}{K_A[1 + K_A[S]]}

\]

\[

v = \frac{k_2[E_0][S]}{K_A + [S]}

\]

(A8)

The definition of equilibrium dictates that the forward reaction rate must equal the reverse reaction rate (Equation A9).

\[

k_1 [E][S] = k_{-1} [E \cdot S]

\]

\[

\frac{[E][S]}{[E \cdot S]} = \frac{k_{-1}}{k_1}

\]

(A9)

Equations A11 (Equation 4) and A12 (Equation 5) can be expressed by comparison of Equation A8 with the standard Michaelis-Menten equation (Equation A10). Equation A11 is simplified to show how $K_M$ can be expressed in terms of individual rate constants by re-writing $K_A$ in terms of equilibrium concentrations (Equation A3) and replacing the ratio of concentrations with a ratio of rate constants (Equation A9).

\[

v = \frac{k_{cat}[E_0][S]}{K_M + [S]}

\]

(A10)

\[

K_M = \frac{1}{K_A} = \frac{[E][S]}{[E \cdot S]} = \frac{k_{-1}}{k_1}

\]

(A11)

\[

k_{cat} = k_2

\]

(A12)
References


15. Khobragade CN, Beedkar SD, Bodade RG, Vinchurkar AS: Comparative structural modeling and docking studies of oxalate oxidase: Possible


Catalytic efficiency determines the in vivo efficacy of PON1 for...
46.

detoxifying organophosphorus compounds. Pharmacogenetics 2000, 10(9):767-779.


57. Kitaura K, Fedorov DG: The FRAGMENT MOLECULAR ORBITAL METHOD PRACTICAL APPLICATIONS TO LARGE MOLECULAR SYSTEMS


