The Pennsylvania State University

The Graduate School

Department of Chemical Engineering

COMPUTATIONAL DESIGN TO SWITCH PROTEIN COFACTOR SPECIFICITY AND CREATE ENZYMATIC ACTIVITY

A Thesis in

Chemical Engineering

by

George A. Khoury

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

May 2010

The thesis of George A. Khoury was reviewed and approved* by the following:

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

Patrick C. Cirino
Assistant Professor of Chemical Engineering

Michael J. Janik
Assistant Professor of Chemical Engineering

Andrew L. Zydney
Head of the Department of Chemical Engineering
Walter L. Robb Chair and Professor of Chemical Engineering

*Signatures are on file in the Graduate School

ABSTRACT

In the first part of this thesis we introduce a computationally-driven enzyme redesign workflow for altering cofactor specificity from NADPH to NADH. By compiling and comparing data from previous studies involving cofactor switching mutations, we show that their effect cannot be explained as straightforward changes in volume, hydrophobicity, charge, or BLOSUM62 scores of the residues populating the cofactor binding site. Instead, we find that the use of a detailed cofactor binding energy approximation is needed to adequately capture the relative affinity towards different cofactors. The implicit solvation models Generalized Born with molecular volume integration and Generalized Born with simple switching were integrated in the Iterative Protein Redesign and Optimization (IPRO) framework to drive the redesign of *Candida boidinii* xylose reductase (CbXR) to function using the non-native cofactor NADH. We identified ten variants, out of the 8,000 possible combinations of mutations, that improve the computationally assessed binding affinity for NADH by introducing mutations in the CbXR binding pocket. Experimental testing revealed that seven out of ten possessed significant xylose reductase activity utilizing NADH, with the best experimental design (CbXR-GGD) being 27-fold more active on NADH. The NADPH-dependent activity for eight out of ten predicted designs was either completely abolished or significantly diminished by at least 90%, yielding a greater than 10^4-fold change in specificity to NADH (CbXR-REG). The remaining two variants (CbXR-RTT and CBXR-EQR) had dual cofactor specificity for both nicotinamide cofactors. The modified IPRO software is available at http://maranas.che.psu.edu.

In the second part of this thesis we present the first steps to developing an enzyme design workflow. We first explored whether changes in interaction energy at the ground state or transition state or both as a result of mutations can explain experimental activity. We chose as our test system the cytochrome $P450_{BM-3}$ monooxygenase to catalyze the hydroxylation of ethane. For the design of this system, first, we had to identify the rate limiting step, and calculate the ground and transition states using quantum mechanical methods. Next, we parameterized these calculated states into the CHARMM forcefield. Thirdly, we explored whether mutations identified by directed evolution always maximize the interaction energy or not. Finally, we systematically selected design positions and explored redesigns via the IPRO algorithm.

Table of Contents

List of Figures vi

List of Tables ix

Acknowledgements x

Chapter 1: Literature Review and Introduction 1

- Section 1.1: Motivation and Objectives 1
- Section 1.2: Background 2
  - Section 1.2.1: Directed Evolution Approaches 2
  - Section 1.2.2: Current Methods and Challenges in Computational Protein Design 5
  - Section 1.2.3: Iterative Protein Redesign and Optimization 7
- Section 1.3: Thesis Overview 11

Chapter 2: Computational Design of *Candida boidinii* Xylose Reductase for Altered Cofactor Specificity 12

- Section 2.1: Introduction and Background 12
- Section 2.2: Materials and Methods 18
  - Section 2.2.1: Modified IPRO Computational Procedure 18
  - Section 2.2.2: Experimental Procedure 20
- Section 2.3: Results 21
  - Section 2.3.1: Analysis of Results from Previous Cofactor Engineering Studies 21
  - Section 2.3.2: Comparison of Calculated Interaction Energies of Enzyme-NAD(P)(H) Complexes vs. Affinities 25
  - Section 2.3.3: Computational Predictions using IPRO 28
  - Section 2.3.4: Experimental Results 31
- Section 2.4: Discussion and Summary 40

Chapter 3: Ground and Transition State Design of Cytochrome $P450_{BM3}$ for Altered Substrate Specificity 42

- Section 3.1: Introduction 42
- Section 3.2: Background on $P450_{\mathbf{BM\text{-}3}}$ 43
- Section 3.3: QM Calculations of Ground and Transition States 48
- Section 3.4: Computational Saturation Mutagenesis Procedure and Results 50
- Section 3.5: Systematic Selection of Design Positions and Iterative Protein Redesign at the Ground and Transition States 54

Chapter 4: Future Work 61

Section 4.1: Future Perspectives on Cofactor Engineering....61
Section 4.2: Future Perspectives/Work on Enzyme Design....61
4.2.1: Determination of Optimal Designs Improving Reactant Binding and Product Off-Rate....61
4.2.2: Use of QM/MM methods to explore transition state stabilization....63
4.2.3: Experimental Assessment of Predicted Designs....65

Bibliography....68

Appendix A: CHARMM Saturation Mutagenesis Script with Solvation in Parallel....83

Appendix B: CHARMM Alanine Scanning Mutagenesis Script with Solvation in Parallel....90

List of Figures

Figure 1 - 1: The primary objective of directed evolution is to create a protein with a specific function through iterative rounds of mutation and screening starting from a parent sequence with associated function. First, a parent sequence is chosen from its closeness to a desired function and its evolvability. A library of new sequences is constructed via site-directed, saturation, or random mutagenesis, or with recombination to introduce mutations from other functional sequences. The mutants are next screened for their ability to carry out their desired function. The fittest mutants are selected and are subsequently used as the parent for the next iteration of directed evolution. This scheme is repeated until the design objective is met (typically after 5-10 generations).[2,3] 3

Figure 1 - 2: Pictorial illustration of how family shuffling searches sequence space vs. single sequence shuffling to create sequence diversity. Family shuffling yields chimeras that have a much larger sequence divergence since it causes sequence block exchange. Single shuffling yields clones with only a few point mutations. When the library sizes are similar, the family shuffling causes increased sequence diversity of the resulting chimeric libraries allowing increased sample diversity. [19] 4

Figure 1 - 3: A schematic representation of IPRO algorithm implemented in this work utilizing the generalized born implicit solvent models. First, a local region of the protein (1-5 consecutive residues around the targeted ligand) is randomly chosen for perturbation (A). The φ and ψ angles of each targeted position are perturbed by up to 5˚ (B). All amino acid rotamers consistent with these torsion angles are selected at each position from the Dunbrack and Cohen rotamer library.[28,29,120] Rotamer-backbone and rotamer-rotamer energies are calculated for all the selected rotamers. The binding energy is minimized using a MILP formulation to select the optimal rotamer at each of the positions (C). The backbone of the protein is relaxed through energy minimization with GBSW to allow it to adjust to the new side-chains. (D). The ligand position is readjusted with respect to the modified backbone and side chains (E). The binding score/interaction energy of the protein-ligand complex is evaluated with the GBMV implicit solvation model (F), and the move is accepted or rejected using the Metropolis criterion (G) until the iteration limit is met. 8

Figure 2 - 1: (A) The structure of the homology modeled CbXR with NADPH bound and D-xylose situated in a deep cavity inside the $(\alpha/\beta)_8$ barrel. (B) The cofactor binding pocket of wild-type CbXR containing NADH with no hydrogen bonding interactions near the 2'-hydroxyl group. (C) The cofactor binding pocket of CbXR containing hydrogen bonding interactions within 2.5 Å of the 2'-phosphate of NADPH. These hydrogen bonding interactions are important for the specificity of CbXR for NADPH over NADH. This figure was made using PyMOL (Delano Scientific). 16

Figure 2 - 2: Comparison of average hydrophobicity, volume, charge, and BLOSUM62 score for all design positions. Error bars are shown for a 95% confidence interval. No statistically significant signal was found except for charge in position 2, where NADH-preferring residues were found to be more negative than NADPH-preferring residues, which is consistent with previous reports in the literature. 23

Figure 2 - 3: Changes in experimental ground state binding energies from Petschacher et al.[141] vs. our calculated changes in interaction energies. Shown are the changes in interaction energy with solvation showing reasonable correlation with the experimental data (R^2=67%), whereas changes in interaction energy without solvation correlated significantly less with the experimental data (R^2=24%) (data not shown). 27

Figure 2 - 4: CbXR-EDS binding pocket containing NADH. The mutated residues Glu-272, Asp-273, and Ser-274 are labeled. Hydrogen bonding interactions are observed within 2.5 Å between the negative Glu-272 and the 3'-OH from NADH. This figure was made using PyMOL (Delano Scientific). 30

Figure 2 - 5: Michaelis-Menten plot for (A) wild type CbXR with NADPH and (B) three tested variants of engineered CbXR with NADH. 35

Figure 2 - 6: Structures of redesigned NAD(P)H binding pockets. (A) CbXR-GGD and (B) CbXR-MGD establish new hydrogen bond interactions between the mutated residues in CbXR and the bridging phosphates in NADH. The net charge change of these mutations is negative which may serve to compensate for the lack of negative 2'-phosphate in NADH. The mutations to glycine may serve to add conformational flexibility in the backbone to allow proper positioning of the NADH. CbXR-RTT, the mutation predicted by IPRO that was experimentally found to have dual cofactor specificity, bound to NADH (C) and NADPH (D). New hydrogen bond interactions are shown stabilizing the 3'-phosphate in NADH and NADPH from Arg-272, which may be the cause of the dual cofactor specificity. In NADPH, new hydrogen bonds are found to stabilize the 2'-phosphate group from Arg-272 and Thr-274. A neutral net change in charge is thought to contribute to dual cofactor specificity as well. All hydrogen bonds shown are within 2.5 Å. This figure was made using PyMOL (Delano Scientific). 37

Figure 2 - 7: Plots of the natural log of specific activity toward NADPH (A) or NADH (B) versus interaction energy for CbXR mutants described in this study. The correlation coefficient for mutants yielding activity for NADPH is 79%, whereas the correlation is only 30% for NADH. 39

Figure 3 - 1: P450BM-3 catalyzed hydroxylation of a substrate. 44

Figure 3 - 2: Consensus catalytic cycle for oxygen activation and transfer by Cytochrome P450.[208] 45

Figure 3 - 3: Calculated key transition state equilibrium bond lengths and angles used in reparameterization in CHARMM in conjunction with the charges calculated. All distances shown are in Angstroms. 50

Figure 3 - 4: Formulation for computational saturation mutagenesis procedure. 51

Figure 3 - 5: Interaction energy improvement (-ΔΔG) compared to the wild-type P450BM-3 upon single amino acid mutations at the 14 positions changed in mutant 535-h for the binding of the ground state (ethane) structure. The x-axis value represents the mutated position in the enzyme. The blue (top) amino-acid abbreviations represent the computationally determined optimal mutation at that position, whereas in cases the experimental and computationally optimal mutants differ, red values (bottom) indicate the experimental mutation. 52

Figure 3 - 6: Improvement in interaction energy (-ΔΔG), compared to the wild-type P450BM-3, upon single amino acid mutations at the 14 positions changed in mutant 535-h for the transition state structure. Mutations were found that significantly improve the interaction energy between the protein and the transition state structure that were not found to improve the binding of the reactant state (ethane). 54

Figure 3 - 7: Schematic of Design Position Selection Protocol. Design positions were selected based on sequence, structure, and energetic factors. 57

Figure 3 - 8: Visual depiction of best ground and transition state binding pockets relative to the wild-type binding pocket. The best designs improved the number of contacts while still allowing the substrate to bind/unbind. 60

List of Tables

Table 2 - 1: Summary[1] of NAD(P)(H) cofactor engineering studies extending from Marohnic et al.[136] 14

Table 2 - 2: NAD(P)(H) binding pockets structurally aligned with combinatorial extension. 22

Table 2 - 3: Residues used in calculating average properties of NAD(P)H-binding residues. 24

Table 2 - 4: Computational and Experimental Results 29

Table 2 - 5: Michaelis-Menten constants for wild-type and mutant CbXR. 36

Table 3 - 1: Design positions selected from sequence, structure, and energetic factors. 56

Table 3 - 2: IPRO generated designs optimizing the interaction energy between the ground and transition states. 58

Table 3 - 3: Comparison of IPRO designs using systematically and experimentally selected design positions. 59

Acknowledgements

I would like to foremost thank my thesis advisor, Dr. Costas Maranas, for his excellent ideas, scientific guidance, and unwavering support and faith. Through his guidance, I have a new appreciation for applied engineering, and a sincere love for optimization. While working with him, I have really grown and appreciate being pushed to new limits. Through my experiences researching, writing, and presenting under his guidance, I have learned far more than I could have imagined when I began as an undergraduate. I was also very fortunate to be working on the exciting projects within this thesis. I also want to thank him for passing on his logical and farsighted approach to scientific research, his ability to present complicated ideas with clarity, his attention to detail, and his patience.

I also want to thank Dr. Michael Janik for his constant and patient help with the P450/DHFR project. We consistently had many long and thoughtful conversations where he made a sincere effort to make me understand all of the details being asked of. Thank you for all of the one-on-one tutorials in quantum mechanical methods.

Thank you to Dr. Patrick Cirino who I constantly asked for guidance with biochemistry/mechanistic questions throughout these projects and his help with the construction of the mutants designed. Thank you also for involving me in the iGEM team several years ago.

Thank you to Dr. Ping Lin, who spent numerous hours helping me with my QM codes and always pointing me in the right direction.

I would like to thank Hossein Fazelinia, Robert Pantazes, and Dr. Patrick Suthers for their guidance throughout this project formally and informally. I have benefitted greatly through many discussions and tutorials on both scientific expertise and career advice.

Thanks to everyone in the Maranas lab with whom I've had the chance to interact with. The combination of researchers with many diverse backgrounds has made it an enjoyable place to learn. In particular I would like to thank Hossein Fazelinia for being my research mentor as an undergraduate, teaching me so much about protein engineering in so little time. The idea of teaching one to fish is applicable in my learning with Hossein as he taught me "how to fish" in protein engineering from day 1, and was never too busy to help even in the toughest and busiest of times. I would like to thank Bob Pantazes for being an excellent programmer, roommate, and an absolutely brilliant colleague. He always patiently guided me in learning to script to simplify repetitive tasks, and always would teach me as he was doing so. He also spent many hours helping me debug code, and always stressed the importance of intelligent programming.

Thank you to my dad for always pushing me to learn math and science. Being a teacher by profession, he took time to make sure I spent a significant amount of time focusing on studying. Thank you to my mom for bringing me into this world and pushing me to always get highest honors in everything I did. Thank you to both my mom and dad for being encouraging and providing me with everything I have ever needed since I was a baby. To them I dedicate this thesis.

Chapter 1: Literature Review and Introduction

Section 1.1: Motivation and Objectives

Advances in technology in the past several years have greatly increased mankind's ability to control life at its most basic level – the genetic. Over millions of years, the struggle for continued existence has resulted in proteins supplying creative, diverse, and efficient solutions to a plethora of problems.[1] These solutions include being able to harvest energy from the environment, to replicating and repairing their own genetic code. Since nature only provides a limited array of proteins, protein engineers seek to broaden known protein function to new environments and tasks (i.e. requiring enhanced thermostability, operating in non-aqueous environments, or binding non-native substrates, and combinations of these),[2,3] or creating new functions.[4] This is done by a series of targeting mutations.

Recently, the literature has reported a diverse set of biotechnological applications of protein engineering including enzymes with improved protein thermostability,[5-7] genetic circuits,[8] biosensors,[9,10] chiral separations,[11] the creation of gene switches[12] and signal transduction pathways,[13,14] and enzymes with improved catalytic activity.[15-17] An extensive literature review on engineering for alternate cofactor switching is provided in Chapter 2 of this thesis. Even with these successes, an open challenge is understanding at a molecular-level why one protein carries out a task better than another.[1] This challenge is compounded by the large combinatorial design space in the active site of many proteins ($20^{\#\text{of Design Positions}}$) as well as the fact that mutations far away from an active site can influence protein function.[18] Engineering for enzymatic activity is especially challenging since minute changes in chemical properties or structure can have large effects on catalytic activity.[1] Therefore, correctly predicting the changes in amino acid sequence

that can produce a specific behavior is difficult. To this end, the underlying objective of this research is to develop new computational tools and to improve existing ones aimed at modeling and subsequently optimizing the enzyme/substrate or enzyme/cofactor interactions in systems of industrial interest. This research will have an impact on protein engineering by using energy functions that have improved correlation with experimentally derived metrics (i.e. k_{cat}, K_m, and binding energy) by including highly accurate implicit solvation models as part of the energetic objective function, as well as introducing the use of quantum mechanical energy functions to compliment molecular mechanics which is currently used in many approaches. The hypothesis is that by using higher accuracy representations of enzymatic systems (at the expense of computational time) the probability of success in design will improve. Further, introducing the ability for time-intensive steps (i.e. structural refinements, backbone perturbations, backbone relaxations) in the computational design to run in parallel will significantly reduce the time necessary to produce a design of interest. In this chapter, I provide a general overview on progress in computational protein design to date and the necessary background needed on the algorithms used and modified to understand the subsequent chapters.

Section 1.2: Background

Section 1.2.1: Directed Evolution Approaches

Directed evolution is the use of subsequent rounds of mutation and artificial selection/screening to alter the properties of biological molecules and systems.[1] Figure 1-1 provides a general overview of directed evolution.

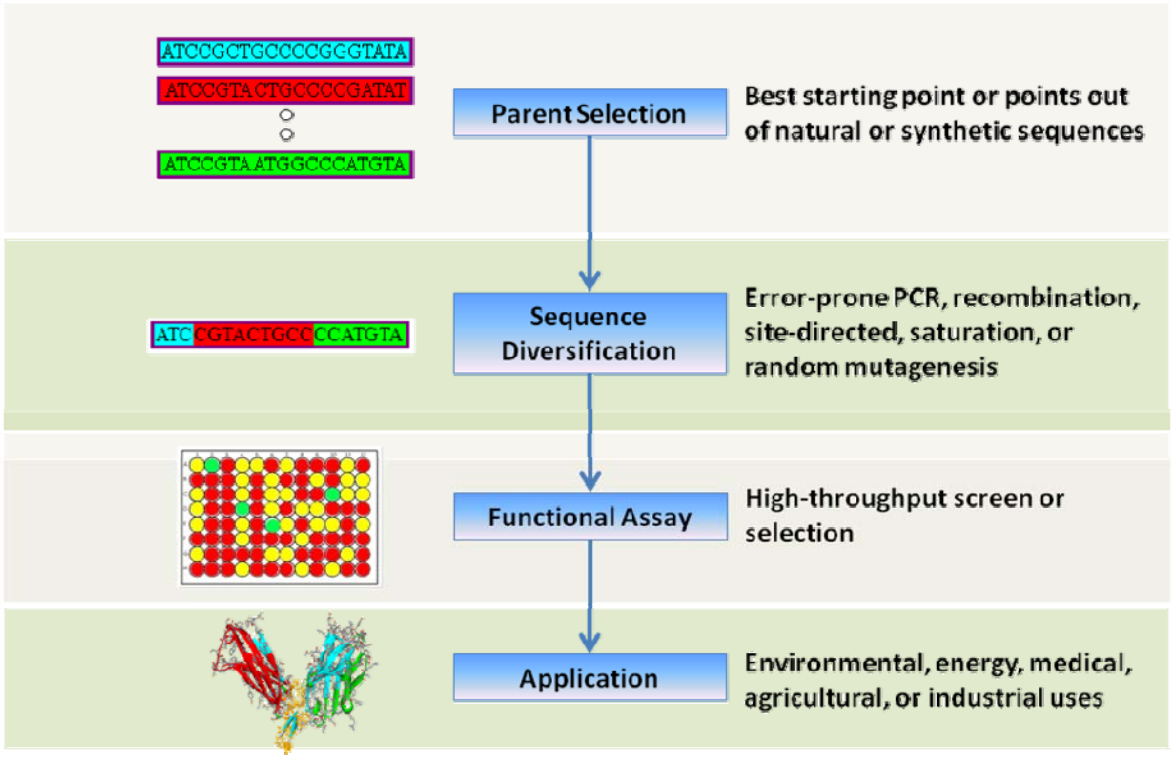

Figure 1 - 1: The primary objective of directed evolution is to create a protein with a specific function through iterative rounds of mutation and screening starting from a parent sequence with associated function. First, a parent sequence is chosen from its closeness to a desired function and its evolvability. A library of new sequences is constructed via site-directed, saturation, or random mutagenesis, or with recombination to introduce mutations from other functional sequences. The mutants are next screened for their ability to carry out their desired function. The fittest mutants are selected and are subsequently used as the parent for the next iteration of directed evolution. This scheme is repeated until the design objective is met (typically after 5-10 generations).[2,3]

Notable works in directed evolution include the DNA shuffling of the genes encoding class C cephalosporinases. In this work, Crameri et al. compared the use of single shuffling and family shuffling to recombine *Citrobacter freundii, Enterobacter cloacae, Klebsiella pneumonia,* and *Yersinia enterocolitica* on the catalytic efficiency of obtaining moxalactamase activity. They found that single shuffling yielded only eightfold improvements from the four genes evolved separately, compared to the 270- to 540-fold improvement in activity from the four genes shuffled together. [19] DNA shuffling is a powerful is tool used in directed evolution. This method generates diversity by combining useful mutations from individual genes (also termed genetic

recombination). Figure 1–2 pictorially illustrates the sequence diversity obtained via single and family gene shuffling.

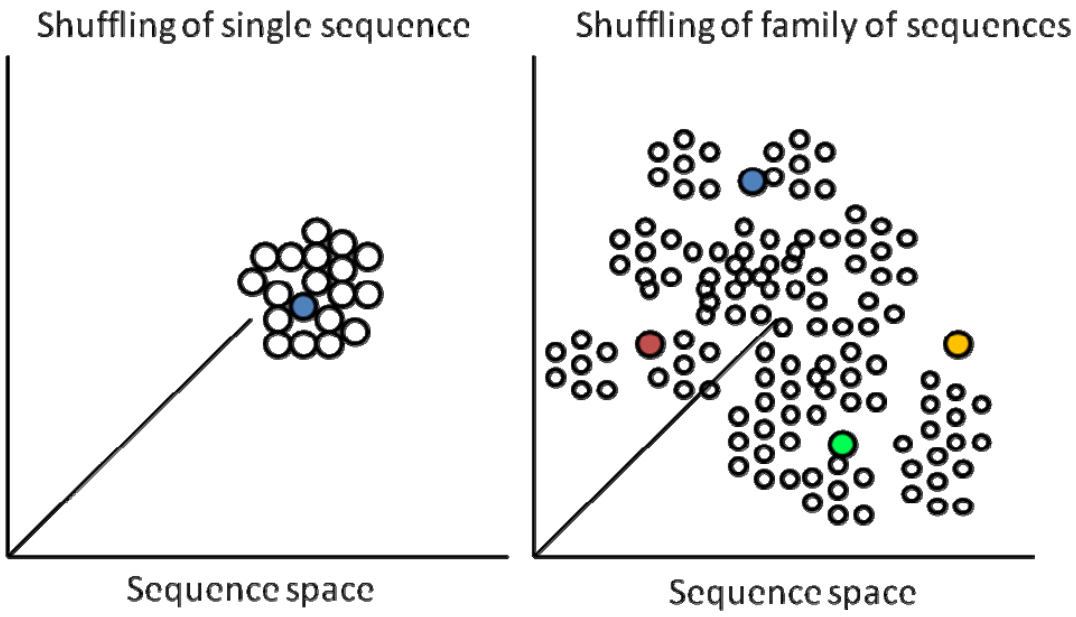

Figure 1 - 2: Pictorial illustration of how family shuffling searches sequence space vs. single sequence shuffling to create sequence diversity. Family shuffling yields chimeras that have a much larger sequence divergence since it causes sequence block exchange. Single shuffling yields clones with only a few point mutations. When the library sizes are similar, the family shuffling causes increased sequence diversity of the resulting chimeric libraries allowing increased sample diversity. [19]

Beaudry et al.[20] created a library of 10^{13} variants of the *Tetrahymena* ribozyme, which naturally catalyzes the sequence-specific cleavage of RNA via a phosphoester transfer mechanism. This enzyme cannot naturally cleave DNA, therefore in their paper they described how they introduced a selection constraint to the population of ribozyme variants created, resulting in variants with a 100x improvement in DNA cleavage activity. Despite numerous successes via directed evolution, genomic diversity is not simple to generate in the lab, and especially not within a practical period of time. The Church Lab recently addressed this challenge by creating a multiplex automated genome engineering (MAGE) approach for the large-scale programming and evolution of cells.[21] This approach concurrently targets multiple locations on a chromosome to modify a cell or

population of cells, therefore producing combinatorial genomic variety. MAGE was applied to optimize the 1-deoxy-D-xylulose-5-phosphate (DXP) biosynthesis pathway to in *Escherichia coli* to overproduce lycopene. By modifying 24 genetic components in the DXP pathway, 4.3 billion combinatorial variants per day were created, and those variants with more than 5-fold increase in lycopene production were isolated. This approach was a significant approach over existing metabolic and protein engineering approaches, and can be applied to expediting the design and evolution of organisms with novel and enhanced performance characteristics.[21] These experimental methods are the cutting edge in directed evolution, but are still relatively slow and costly. Therefore I will now introduce the recent advances in computational methods available.

Section 1.2.2: Current Methods and Challenges in Computational Protein Design

Computational protein design provides a method for the efficient generation of protein catalysts for any chemical reaction. The use of computational methods has recently led to many protein redesign successes [22-26] by optimizing protein-ligand and/or enzyme-cofactor interactions using static and molecular dynamics calculations. These methods have focused primarily on rotamer, geometry, and energetic optimization. Major developments have been made towards the *de novo* design problems which include the creation of novel protein folds, enzymatic activity, and binding interfaces. [27] *Ab initio* design of proteins entails identifying the amino acid choices that best fit into a protein fold. The Cartesian coordinates of a protein's backbone atoms define the a protein's structure. At atomistic detail, candidate protein designs are produced by selecting amino acid side chains, or statistically preferred rotamers [28,29] to fit in the backbone design. Therefore protein design problem formulations involve both residue and a rotamer assignment. Rotamer/rotamer and rotamer/backbone energies are calculated for all rotamers in a chosen

library to determine how well possible designs fit into a given fold. Methods utilizing molecular mechanics-based potential energy functions are estimated with a number of different force-fields (i.e. CHARMM [30], DREIDING [31], AMBER [32], GROMOS [33]). Custom scoring/energy function as have also been created specifically for protein design.[34-37]

A limitation in design is the choice of the objective function, as few problems can be sufficiently addressed by the straightforward energy minimization of a single protein state. Activity level is very difficult to assess computationally. Instead, multi-objective optimization searches are necessary for designing for improved binding affinity (increase interactions while maintaining fold stability), specificity (stabilize a state relative to another), and designing *de novo* proteins (avoid aggregation and alternative structures). [27] Stability is often used as a proxy for a design's fitness for alternative functions/substrates since it is a prerequisite, although not necessarily a monotonic descriptor of function. Since this surrogate is an indirect descriptor of activity, it is necessary to design not just one design, but an entire combinatorial library.[37]

The enormity of the design space is vast, and therefore deterministic and stochastic methods have been employed. Stochastic methods explore the feasible space by making a series of random and/or directed moves. [37] Mayo and colleagues recently released two enhancements to the stochastic optimizer FASTER that resulted in 100x improvement in convergence speed.[38] Monte Carlo methods,[35,39,40] genetic algorithms, [41-43] and simulated annealing [44,45] methods also exist and have been used in computational protein design with different success rates. Stochastic methods are typically used for problems with an enormous design space with relatively small computing resources, but are not guaranteed to converge to the best variant.[37]

Deterministic methods on the other hand are advantageous since they are guaranteed to converge to a global minimum in a given objective function. The key disadvantage of using these methods is their intractability for large systems. The most commonly used method is dead-end elimination,[46] where rotamers and residues determined to not be optimal are eliminated over a

series of computational cycles. Xie and Sahinidis reformulated the hierarchy used in discrete searches by explicitly considering residue elimination on top of the traditional rotamer elimination methods, which led to a 100x increase in convergence speed.[47]

For an excellent review of the advances in methodology (energy functions, search and optimization procedures, solvation effects) and applications (specificity, affinity, catalytic efficiency) see Lippow et al.[27] Other reviews on current protein design methods and viewpoints include the references within.[27,48-55] The leading well-developed methods available include ORBIT[56-66] by the Mayo lab, which given a backbone structure designs a sequence such that it folds to that of the backbone. The Rosetta suite of software developed by the David Baker Lab is one program to treat diverse structure prediction and design problems.[67-116] This thesis research utilized and advanced the Iterative Protein Redesign and Optimization (IPRO) algorithm[117-119] developed by the Maranas Lab, which uses a mixed-integer linear programming formulation to design combinatorial libraries, and will be covered in the next section.

Section 1.2.3: Iterative Protein Redesign and Optimization

The Iterative Protein Redesign and Optimization (IPRO) algorithm developed by Saraf et al. is an iterative framework used to design combinatorial libraries with a mixed-integer linear programming formulation. IPRO is capable of determining the optimal combination of residue/rotamer combinations that minimize the interaction energy between a protein and desired substrate for a given set of design positions. It is pictorially illustrated in Figure 1–3.

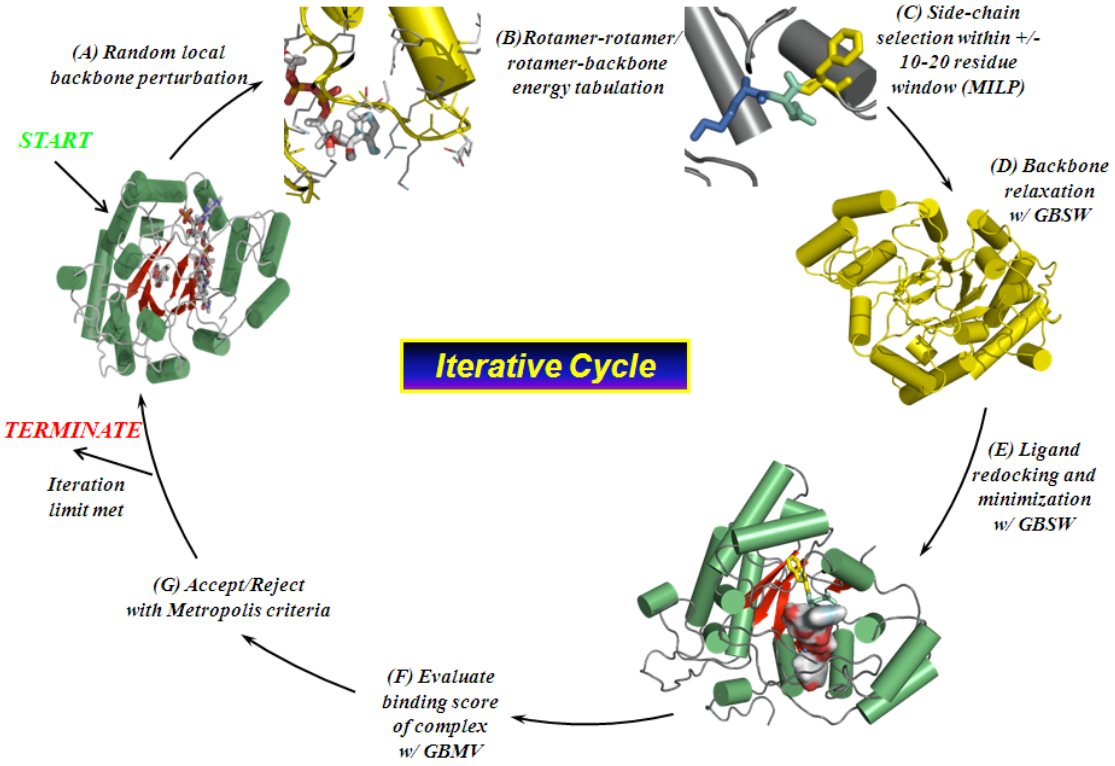

Figure 1 - 3: A schematic representation of IPRO algorithm implemented in this work utilizing the generalized born implicit solvent models. First, a local region of the protein (1-5 consecutive residues around the targeted ligand) is randomly chosen for perturbation (A). The φ and ψ angles of each targeted position are perturbed by up to 5˚ (B). All amino acid rotamers consistent with these torsion angles are selected at each position from the Dunbrack and Cohen rotamer library.[28,29,120] Rotamer-backbone and rotamer-rotamer energies are calculated for all the selected rotamers. The binding energy is minimized using a MILP formulation to select the optimal rotamer at each of the positions (C). The backbone of the protein is relaxed through energy minimization with GBSW to allow it to adjust to the new side-chains. (D). The ligand position is readjusted with respect to the modified backbone and side chains (E). The binding score/interaction energy of the protein-ligand complex is evaluated with the GBMV implicit solvation model (F), and the move is accepted or rejected using the Metropolis criterion (G) until the iteration limit is met.

IPRO was utilized and extended to generating novel libraries of DHFR and β-lactamase enzymes [119], changing substrate [14,117] and cofactor [118] specificities, and adding a new Ca^{+2} binding site to a domain [26]. In the author of this thesis' opinion, it has not received the attention it deserves by the scientific community, namely because of its original complexity in use. Recent efforts by the Maranas lab (via graduate student Robert Pantazes) to make it modular and reusable have helped

considerably, and it is now available for download and use through the group webpage. The original algorithm utilized the Baker energy function[121] for the rotamer selection step and the CHARMM[30,122,123] energy function decomposed below for all energy minimization steps.

$$Epot = Ebond + Eangle + Edihedrals + Eimpropers + Eelec + EvdW + Erestraint$$

$$EvdW = \sum_{ij}\left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}}\right)$$

$$Eelec = \sum_{ij}\frac{q_i q_j}{\varepsilon 4\pi\varepsilon_0 r_{ij}}$$

$$Ebonds = \sum k_b (r - r_0)^2$$

$$Eangles = \sum k_\theta (\theta - \theta_0)^2$$

$$Edihedrals = \sum \left|k_\phi\right| - k_\phi \cos(n\phi)$$

$$Eimpropers = \sum k_\chi (\chi - \chi_0)^2$$

$$Erestraint = \sum k_i m_i (r_i - r_{0i})^2$$

In Chapter 2,[118] I show that the standard CHARMM energy function was insufficient to be able to correlate calculated interaction energies with experimental binding energies. Instead, we found that the introduction of solvation effects was necessary to achieve a qualitative correlation between computational and experimental results. Therefore, we introduced solvation effects into IPRO via the Generalized Born with a simple switching implicit solvent model (GBSW)[124,125] for all minimization steps in IPRO, and the Generalized Born with molecular volume integration (GBMV) implicit solvent model[126,127] for all interaction energy calculations. Explicit solvent calculations are by far the most accurate means of incorporating solvation, but greatly reduce the viable system size due to high computational cost. Implicit solvation models utilize estimations to

the solvation effects and are far more efficient than explicit solvation methods.[128] GBSW was chosen for all energy minimizations to ensure proper packing of hydrophobic cores, and GBMV was used to approximate the solvation component for the binding energy calculations. In the GBMV method, the effective Born radius is computed by numerical integration of the molecular volume. The Coulomb field approximation includes a higher order correction term to improve agreement with the radii calculated from solving the Poisson-Boltzmann equation. With respect to absolute solvation energies, the model achieved an overall 1.3% error compared to the highly accurate converged Poisson solutions for a sample of 3000+ proteins previously tested from the PDB .[124,129] This method was used for the binding energies as it was highly accurate, yet more tractable in an iterative form than solving the Poisson-Boltzmann equation. The minimizations utilized the GBSW model, as the GBMV model may utilize a sharp molecular surface representation for some systems, which would lead to large fluctuations in energy and cause stability problems in the simulations. GBSW is very similar to GBMV, but is 2-3 times faster than the GBMV method since it replaces the computationally expensive molecular surface calculation with a simple smoothing function at the dielectric boundary. GBSW's use of a smoothed dielectric boundary allows the change in polarization forces to vary more smoothly compared to GBMV.

The introduction of these implicit solvent models increased the computational cost for designing a protein. Therefore, I introduced the ability of running IPRO in parallel through CHARMM by using a different build and adjusting the code accordingly to utilize parallel CHARMM for all energy minimizations and interaction energy calculations. This significantly increased the speed of being able to redesign a protein – previously taking almost two weeks of computation whereas now only taking two days assuming everything is properly parameterized for the design of *Candida boidinii* xylose reductase.

Section 1.3: Thesis Overview

The remainder of this thesis is organized as follows. Chapter 2 describes the computational design of the enzyme *Candida boidinii* xylose reductase and its corresponding experimental verification. Chapter 3 describes the integration of quantum mechanical calculations as part of the palette of tools available to redesign enzymes, and its application on the redesign of Cytochrome $P450_{BM-3}$ monooxygenase. Finally, Chapter 4 concludes by offering some perspectives on future work in cofactor engineering and the development of an enzyme design workflow.

Chapter 2: Computational Design of *Candida boidinii* Xylose Reductase for Altered Cofactor Specificity

Section 2.1: Introduction and Background

The ability of enzymes to catalyze chemical reactions with great specificity, efficiency, and selectivity provides the basis of metabolism in all living organisms. By carefully redesigning metabolism through enzyme modification, many desired biocatalytic transformations can be efficiently carried out in a variety of microbial production hosts. Proteins have been previously computationally designed to bind new ligands,[24] proteins,[130] and nucleic acids,[131] to improve protein stability,[121,132] as well as to introduce novel enzymatic activity,[133,134] demonstrating that the fundamental rudiments of molecular recognition can adequately be captured via computational design. The systematic fine-tuning of molecular recognition between proteins and ligands finds many biotechnological applications ranging from improved catalytic activity,[16] improved protein thermostability,[5-7] genetic circuits,[8] biosensors,[9,10] chiral separations,[11] the construction of novel enzymes with alternative functionality,[56,135] the creation of gene switches[12] and signal transduction pathways.[13,14] Many of the aforementioned applications require the enzymes to operate under unnatural conditions (e.g., at elevated temperatures or in nonaqueous environments), and/or possess altered cofactor or substrate specificity.[3] Even with these successes, predictably changing a protein's cofactor specificity has not been reported via a systematic computational workflow.

In the past few years, there have been many reported successes of enzyme redesign for altered cofactor specificity utilizing structural analysis with site-directed mutagenesis as their method for redesign. Table 2-1 summarizes the best identified mutations involved in changing cofactor

specificity (extending an earlier compilation).[136] Key successful redesigns include the work of Woodyer et al.[137] that succeeded in relaxing the cofactor specificity of *Pseudomonas stutzeri* phosphite dehydrogenase from 100-fold in favor of nicotinamide adenine dinucleotide (NAD^+) to 3-fold in favor of nicotinamide adenine dinucleotide phosphate ($NADP^+$) using homology modeling and site-directed mutagenesis to identify and construct a double mutant. This double mutant showed potential as an efficient *in vitro* NAD(P)(H) regeneration system for reductive biocatalysis.[137] Watanabe et al.[138] used site-directed mutagenesis to change cofactor specificity of a *Pichia stipitis* NAD^+-dependent xylitol dehydrogenase (PsXDH) from NAD^+ to $NADP^+$ as part of an efficient biomass-ethanol conversion system. Their designs yielded greater activity for $NADP^+$ than NAD^+ after redesign. Kostrzynska et al.[139] found that in the aldo-keto reductase family of enzymes, the IPKS (Ile-Pro-Lys-Ser) motif is strictly conserved. They utilized site-directed mutagenesis at a conserved Lys-270 in *P. stipitis* xylose reductase (PsXR) to conclude that it binds to the 2'-phosphate of the NADPH (reduced form of $NADP^+$). Site-directed mutagenesis-based studies also successfully pinpointed sets of mutations leading to complete reversal of *Candida tenuis* xylose reductase (CtXR) cofactor specificity from NADPH to NADH (reduced form of NAD^+).[140,141] Similarly, Liang et al.[142] used a semi-rational approach called combinatorial active site saturation (CASTing) to switch cofactor preference from NADPH to NADH in PsXR.

Table 2 - 1: Summary[1] of NAD(P)(H) cofactor engineering studies extending from Marohnic et al.[136]

Source	Enzyme	Specificity Change	Mutation(s)[2]	Reference(s)
Candida tenuis	xylose reductase	NADPH → NADH	K274R, K274G, K274M, S275A, N276D, R280H, K274R/N276D	140,141
Corynebacterium	2,5-diketo-D-gluconic acid	NADPH → NADH	K232G, R235G, R238H & F22Y/RS233T/R235E/A272G	143,144
Escherichia coli	glutathione reductase	NADPH → NADH	A179G/A183G/V197E/ R198M/K199F/H200D/R204P	145
Escherichia coli	ketol acid reductoisomerase	NADPH → NADH	R68D, K69L, K75V, R76D	146
Neurospora crassa	nitrate reductase	NADPH → NADH	S920D/R932S	147
Pichia stipitis	xylose reductase	NADPH→ NADH	K270M, K270S/ S271G/N272P/R276F	139,142
Pseudomonas fluorescens	*p*-hydroxybenzoate hydroxylase	NADPH → NADH	R33S/Q34R/P35R/D36A/Y37E	148
Rattus norvegicus	cytochrome p450 reductase	NADPH → NADH	S596D	149
Saccharomyces. cerevisiae	17ß-hydroxysteroid dehydrogenase	NADPH → NADH	Y49D	150
Sinorhizobium morelense	1,5-anhydro-D-fructose reductase	NADPH → NADH	A13G/S33D	151
Anabaena. sp. (strain PCC 7119)	ferredoxin:NADP+ reductase	NADP+ → NAD+	S223D	152
Escherichia coli	isocitrate dehydrogenase	NADP+ → NAD+	C201I/C332Y/K344D/Y345I/V351A/ Y391K/R395S	153
Thermus thermophilus	isocitrate dehydrogenase	NADP+ → NAD+	K283D/Y284I/N287G/V288I/I290A	154
Vibrio harveyi	aldehyde dehydrogenase	NADP+ → NAD+	T175D, T175E, T175S, T175N, T175Q	155
Bacillus stearothermophilus	L-lactate dehydrogenase	NADH → NADPH	I51K/D52S	156
Rattus norvegicus	cytochrome b5 reductase	NADH → NADPH	D239T	136
Spinacia oleracea	nitrate reductase	NADH → NADPH	E864S/F876R	157
Thermus thermophilus	ß-isopropylmalate dehydrogenase	NADH → NADPH	D236R/D289K/I290A/A296V/G337Y	158
Bacillus stearothermophilus	D-lactate dehydrogenase	NAD+ → NADP+	D175A	159
Bacillus stearothermophilus	glyceraldehyde-3-phosphate dehydrogenase	NAD+ → NADP+	D32A/L187A/P188S	160
Gluconobacter oxydans	xylitol dehydrogenase	NAD+ → NADP+	D38S/M39R	161
Homo sapien	human mitochondrial NAD(P)-dependent malic enzyme	NAD+ → NADP+	Q362K	162
Pichia stipitis	xylitol dehydrogenase	NAD+ → NADP+	D207A/I208R/F209S/N211R	138
Pseudomonas stutzeri	phosphite dehydrogenase	NAD+ → NADP+	E175A/A176R	137
Saccharomyces cerevisiae	formate dehydrogenase	NAD+ → NADP+	D196A/Y197R	163
Thermus thermophilus	isopropylmalate dehydrogenase	NAD+ → NADP+	S226R/D278K/I279Y/A285V/P324T/ P325Y/G328E/G329R/S330L	164
Tramitichromis intermedius	leucine dehydrogenase	NAD+ → NADP+	D203A/I204R/D210R	165

[1]In all studies, structural analysis was used to determine residues to be mutated. Mutations were introduced by site-directed mutagenesis except for Liang et al.,[142] who used a combinatorial saturation mutagenesis approach.

[2]The best mutants reported in each study are summarized in this table. Multiple mutations occurring in a single mutant are separated by "/". Commas are used to separate individual mutants.

Purely experimental design efforts relying on combinatorial library constructxon and screening have been successful for a number of cofactor alteration studies (see Table 2-1), however, the lessons learned do not easily generalize to other systems. To address the lack of a systematic procedure, we introduce a generally applicable computational workflow based on the Iterative Protein Redesign and Optimization algorithm (IPRO).[119] The approach is tested for the xylose reductase enzyme from the yeast *Candida boidinii* (CbXR). Xylose reductase catalyzes the reduction of the open chain form of D-xylose to xylitol.

Xylose reductase belongs to the aldo-keto reductase (AKR) superfamily.[166,167] The AKR superfamily shares a common $(\alpha/ß)_8$-barrel fold without a Rossmann-fold motif and their members show varied preferences for NADPH over NADH.[168] The active site, conserved in both structure and sequence in nearly all AKRs, is situated in a deep cavity inside the $(\alpha/ß)_8$ barrel, and is defined by a tetrad of catalytic residues. In CtXR, these residues are Asp-46, Tyr-51, Lys-80, and His-113,[169] and are homologous to Asp-45, Tyr-50, Lys-79, and His-112 in CbXR. Previous studies[169-172] of AKRs have identified the functional role these residues have on the catalytic mechanism, but because they are further than 12 Å from the residues involved in determining cofactor specificity, they have minimal effect on cofactor binding. Structures of the apo- and holo- forms of CtXR have been determined to 2.2 Å resolution.[168] This enzyme selectively binds NADPH over NADH by roughly 20-fold.[141] In contrast, CbXR (62% homologous to CtXR) is strictly an NADPH-dependent enzyme. The structure of the homology modeled CbXR is shown in Figure 2-1A, with NADPH bound and D-xylose situated inside the $(\alpha/ß)_8$ barrel. In Figure 2-1B, the cofactor binding pocket is shown with no hydrogen bonding interactions observed between wild-type CbXR and NADH. In Figure 2-1C, hydrogen bond interactions are shown between the 2'-phosphate in NADPH and the surrounding residues Lys-272, Ser-273, and Asn-274. Alignment of AKRs reveals a conserved Lys residue near position 274 (amino acid position 274 in CtXR; position 272 in CbXR), which plays a critical role in cofactor binding.[140] One

notable exception is the presence of an Arg residue rather than Lys at position 276 of the XR from *C. parapsilosis*, which prefers NADH as its cofactor.[173] Leitgeb et al. showed that replacement of Lys-274 with Arg in CtXR results in reversal of cofactor specificity for NADH over NADPH [140].

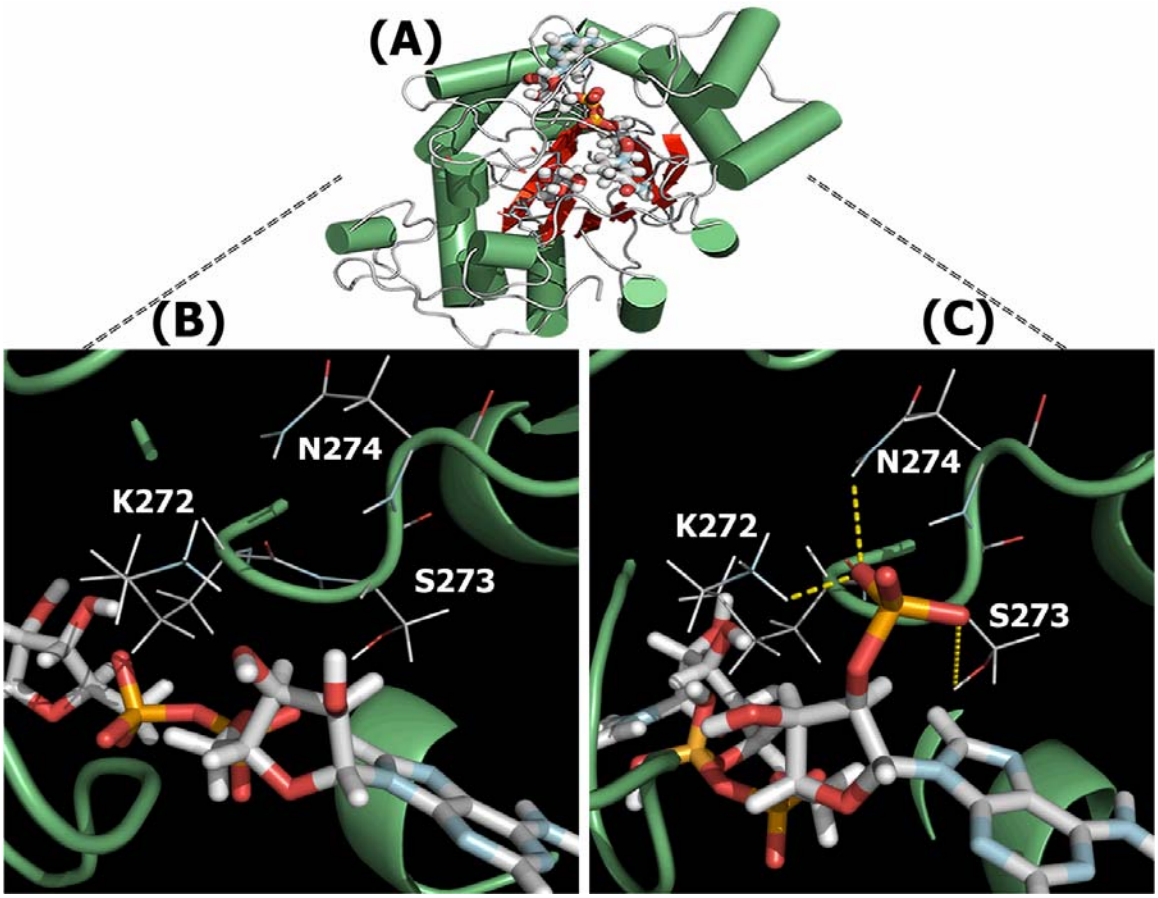

Figure 2 - 1: (A) The structure of the homology modeled CbXR with NADPH bound and D-xylose situated in a deep cavity inside the $(\alpha/ß)_8$ barrel. (B) The cofactor binding pocket of wild-type CbXR containing NADH with no hydrogen bonding interactions near the 2'-hydroxyl group. (C) The cofactor binding pocket of CbXR containing hydrogen bonding interactions within 2.5 Å of the 2'-phosphate of NADPH. These hydrogen bonding interactions are important for the specificity of CbXR for NADPH over NADH. This figure was made using PyMOL (Delano Scientific).

Xylitol has been listed among the top value-added platform chemical products of biomass refining.[174] The production of xylitol from xylose by engineered *Escherichia coli* growing on glucose and expressing a xylose reductase from either *C. boidinii*, *C. tenuis*, *P. stipitis*, or

Saccharomyces cerevisiae was recently studied.[175] Of the enzymes tested, functional expression of CbXR in *E. coli* resulted in the highest titers of xylitol production. It is unclear whether this is related to its strict requirement for NADPH or whether it is more a function of its expression characteristics. We explored whether xylitol production could be improved by enabling the use of NADH for xylose reduction. In addition to a number of other strategies recently explored,[176] we sought a variant of CbXR with either dual cofactor specificity or specificity toward NADH (which could then be co-expressed with wild-type CbXR). In addition, due to the higher stability of NADH relative to NADPH,[143] and the higher cost of NADPH regeneration compared to NADH generation,[177] a NADH-utilizing CbXR variant may prove industrially useful. We initially constructed the K272R mutation in CbXR and found this mutant to be active on NADH, while NADPH activity was weakened by five-fold. However, NADH-utilizing activity was less than 5% of the wild-type enzyme's activity with NADPH. We therefore sought to use computational design to more effectively engineer mutants with activity toward NADH.

The goal of this work was to explore the computational design of CbXR to bind (and subsequently oxidize) NADH as its cofactor. We first extracted and analyzed data from various cofactor usage alteration studies to pinpoint key interactions, factors, and trends that are discernable when performing cofactor switches between these particular substrates. We next validated the use of a computationally-derived interaction energy as a reasonable objective function and binding free energy surrogate by correlating it to published experimental binding results. This surrogate of cofactor affinity was found to correlate (R^2=72%) with experimental activities for a system previously designed using IPRO.[119] Our working hypothesis was that computationally generated sets of mutations that improve binding of NADH to CbXR will lead to mutants that exhibit enzymatic activity on NADH. Next, we modified the Iterative Protein Redesign and Optimization (IPRO) framework as presented by Saraf et al.[119] to improve modeling accuracy by adding implicit solvation models to drive the identification of sets of

mutations that have increased affinity for NADH as evidenced by improved interaction energies, as well as increased stability for the CbXR mutants relative to the wild-type. Lastly, we constructed and experimentally tested the best variants predicted by IPRO to assess the value of computations to drive redesign.

Section 2.2: Materials and Methods

Section 2.2.1: Modified IPRO Computational Procedure

The Iterative Protein Redesign and Optimization (IPRO) framework, which was previously developed by our group,[117,119] performs enzyme redesign by optimally identifying mutations in the protein sequence using energy-based scoring functions. The modified IPRO algorithm is available for download at http://maranas.che.psu.edu. In this effort, we added the implicit solvation models GBSW and GBMV to the minimization and interaction energy calculation steps, respectively. The steps of the algorithm are as follows. First, design positions are selected, and the torsion angles in a small region around a design position of the backbone are perturbed by up to ±5 degrees. The vast majority of evolutionary engineering studies over the past ten years involve simple uphill walks on the plot of fitness versus sequence.[178] As a result, the positions chosen for redesign of CbXR were Lys-272, Ser-273, and Asn-274 after structurally aligning CbXR with CtXR using Combinatorial Extension[179] between residues 200-290 and based on previous cofactor engineering studies on CtXR.[140,141] Next, all amino acid rotamers consistent with these torsion angles are selected at each position from the Dunbrack rotamer library.[29,180] For the design positions, the rotamers considered include all amino acids, whereas for non-design positions, the possible rotamers are only those from the native amino acid. Next, rotamer-rotamer and rotamer-backbone energies are calculated for all of the selected rotamers in the previous step

using the energy function presented in Kuhlman et al.[121] A mixed-integer linear programming formulation is then used to select the optimal combination of rotamers in the design window such that the energy is minimized for the torsion angles considered. The backbone of the protein is next relaxed through energy minimization with the GBSW implicit solvation model to allow the backbone to adjust to the new side chains. The ligand position is then readjusted in the next step with respect to the modified backbone and side chains using the Fast-Fourier Transform ZDOCK docking software[181] (version 2.3) with constraints added to block residues 8Å from the binding pocket from being considered in the docking step. The interaction energy of the protein-ligand complex is next evaluated with the GBMV implicit solvation model and the move is accepted or rejected based on whether the interaction energy has been improved relative to the best design thus far with the Metropolis criteria[182] to escape local minima. Please refer to Saraf et al.[119] for further details of the algorithm. Here IPRO was used to identify the optimal set of rotamers or residues on CbXR in the NADPH binding pocket necessary to increase the affinity for NADH over NADPH.

Although a high-resolution crystal structure of CbXR has not been determined, the amino acid sequence of CtXR[183] is sufficiently similar to that of CbXR[184] to act as a plausible model for CbXR (62% sequence similarity). The model structure of CbXR was constructed by homology modeling through Modeller using defined geometrical restraints between the conserved atoms of binding pocket residues and the cofactor obtained from the homologous CtXR crystal structure with NADPH bound (PDB: 1K8C).[168]

IPRO was performed with the modifications for solvation on a Linux PC cluster using eight 3.06GHz Xeon CPUs with 4GB RAM for 2 CPU days to improve the interaction energies of CbXR for NADH. In each iteration, interaction energy calculations took approximately 6 seconds of CPU time per evaluated mutant.

Section 2.2.2: Experimental Procedure

The redesigned proteins were constructed using standard site-directed mutagenesis techniques[185] and all sequences were verified by DNA sequencing. Proteins were then expressed in *E. coli* BL21 as follows: Seed cultures (10 ml in LB medium containing 50μg/ml kanamycin) were grown at 37°C to an OD_{600} of ~ 2.0 and were used to inoculate cultures by dilution to a final OD_{600} of 0.1 in 125 ml of LB (50μg/ml kan). When the cultures (at 37ºC) reached an OD_{600} of 0.6-0.7, protein expression was initiated by adding 1.0 mM IPTG and transferring the cell cultures to 30ºC. After 9 hours of induction, cells were pelleted by centrifugation at 3200 g for 20 min, washed twice with 25 ml of 50mM potassium phosphate buffer (pH 7.5). Cell pellets were stored at -20ºC until use. The cell pellets were resuspended to a final OD_{600} of 100 in ice-cold lysis buffer (50mM potassium phosphate buffer (pH 7.5), 4mM $MgCl_2$, 3.3μg/ml DNase I). Cells were lysed by three passes through a French Pressure cell press, and centrifuged at 4ºC, 3750 g for 25 min to separate cellular debris. The resulting supernatant contained the soluble xylose reductase.

Xylose reductase activity was measured in 96-well microtiter plates using a Spectra Max Plus384 plate reader. A typical enzymatic reaction contained 300mM xylose, 300μM β-NADPH or 300μM β-NADH, 50mM potassium phosphate buffer (pH 7.5), 40μl cell lysate supernatant and 5mM KCN (to reduce background dehydrogenase activity) in 200 μl total final volume. Reduction in the β-NADH or β-NADPH concentration was monitored by the decrease in absorbance at 340 nm (extinction coefficient ~ 6.2 $(mM\text{-}cm)^{-1}$). Reactions were initiated by adding reduced cofactor and measurements were taken every 3 seconds for 90 seconds. One unit is defined as the enzyme activity that consumes one μmol of NADH or NADPH in one minute (background activity in the absence of xylose is subtracted). Total protein concentration was

measured using the Quick Start™ Bradford protein assay protocol (Bio-Rad laboratories) based on binding of Coomassie Blue dye to proteins. Bovine serum albumin was used as a standard.

Section 2.3: Results

Section 2.3.1: Analysis of Results from Previous Cofactor Engineering Studies

We first explored whether the experimentally observed binding affinities for NAD(P)(H) and/or enzymatic activities requiring these cofactors can be explained by using simple metrics such as residue volume, charge and hydrophobicity. Net charge,[186] hydrophobicity,[187] and side-chain volume[188] data for all amino acids were collected. A structural alignment was performed for the nicotinamide binding pockets targeted by mutational studies of the following proteins: glutathione reductase (GR),[145] ketol acid reductoisomerase (KARI),[146] *p*-hydroxybenzoate hydroxylase (PHBH),[148] 2,5-diketo-D-gluconic acid (2,5-DKG),[143,144] 1,5-anhydro-D-fructose reductase (1,5-AFR),[151] isocitrate dehydrogenase (IDH),[153] glyceraldehyde-3-phosphate dehydrogenase (GAPDH),[160] *P. stipitis* xylitol dehydrogenase (PsXDH),[138] ferrodoxin:NADP+ reductase,[152] and L-lactate dehydrogenase (L-LDH).[156] The nicotinamide cofactor binding pockets of these proteins were aligned to the NADPH binding pocket of CtXR using Combinatorial Extension.[179] These proteins were chosen as they are well characterized and most had high resolution crystal structures available. The structural alignments used RCSB Protein Data Bank (PDB) [129] crystal structures to provide the atomic coordinates for all structures except for PsXDH, which was constructed via the SWISS-MODEL first approach method.[189,190] The results of the different nicotinamide binding pockets structurally aligned with Combinatorial Extension are shown in Table 2-2. Significant structural similarity in the nicotinamide cofactor binding pockets was found across the enzymes used based on their root of mean square deviation (RMSD) values.

In contrast, no significant sequence alignment occurred in the binding pockets of the sampled proteins and that of CtXR, except in 2,5-DKG, where there is 50% sequence similarity.

Table 2 - 2: NAD(P)(H) binding pockets structurally aligned with combinatorial extension.

Protein	PDB Code	Sequence Positions Aligned	RMSD Binding Pocket (Å)	Sequence Identity of Binding Pocket (%) (20-50 residues)	Overall RMSD (Å)
CtXR	1MI3	270-290	0.0	100.0	0.0
GR	1GER	170-210	1.9	6.2	4.9
KARI	1YRL	60-80	2.9	12.5	4.4
PHBH	1PBB	20-50	2.5	12.5	4.5
2,5-DKG	1A80	220-250	0.8	50	1.3
1,5-AFR	2GLX	2-49	1.7	18.8	3.7
IDH	2D1C	280-300	3.25	12.5	4.7
GAPDH	1GD1	180-200	3.26	6.2	5.28
PsXDH	[1]	200-220	1.42	15.8	4.9
ferrodoxin:NADP+ reductase	1QUF	220-240	3.31	6.2	5.15
L-LDH	1LDB	46-70	1.88	0	5.31

[1]This structure was generated utilizing the SWISS-MODEL first approach homology modeling method as there was no initial PDB crystal structure available.[189,190]

Based on previous mutational studies performed on CtXR, positions Lys-274, Ser-275, and Asn-276 emerged as key locations to mutate to increase cofactor specificity for NADH over NADPH.[141,166,191] We defined Positions 1-3 as the residues that are aligned to K274, S275, and N276, respectively. Positions 1-3 are nearby the phosphate group in NADPH, but are over 12Å from the hydride transfer site in the catalytic mechanism, highlighting that these positions affect cofactor specificity and affinity but are not directly involved in the reaction. Next, a statistical analysis on charge, hydrophobicity, and volume was performed for each design position structurally aligned to CtXR in both their NADP(H) and NAD(H)-preferring forms for the residues listed in Table 2-3. This allowed us to discern whether any of those metrics played an identifiable role in cofactor specificity. For each position, we calculated the average value of each parameter, as depicted in Figure 2-2 with error bars representing 95% confidence intervals.

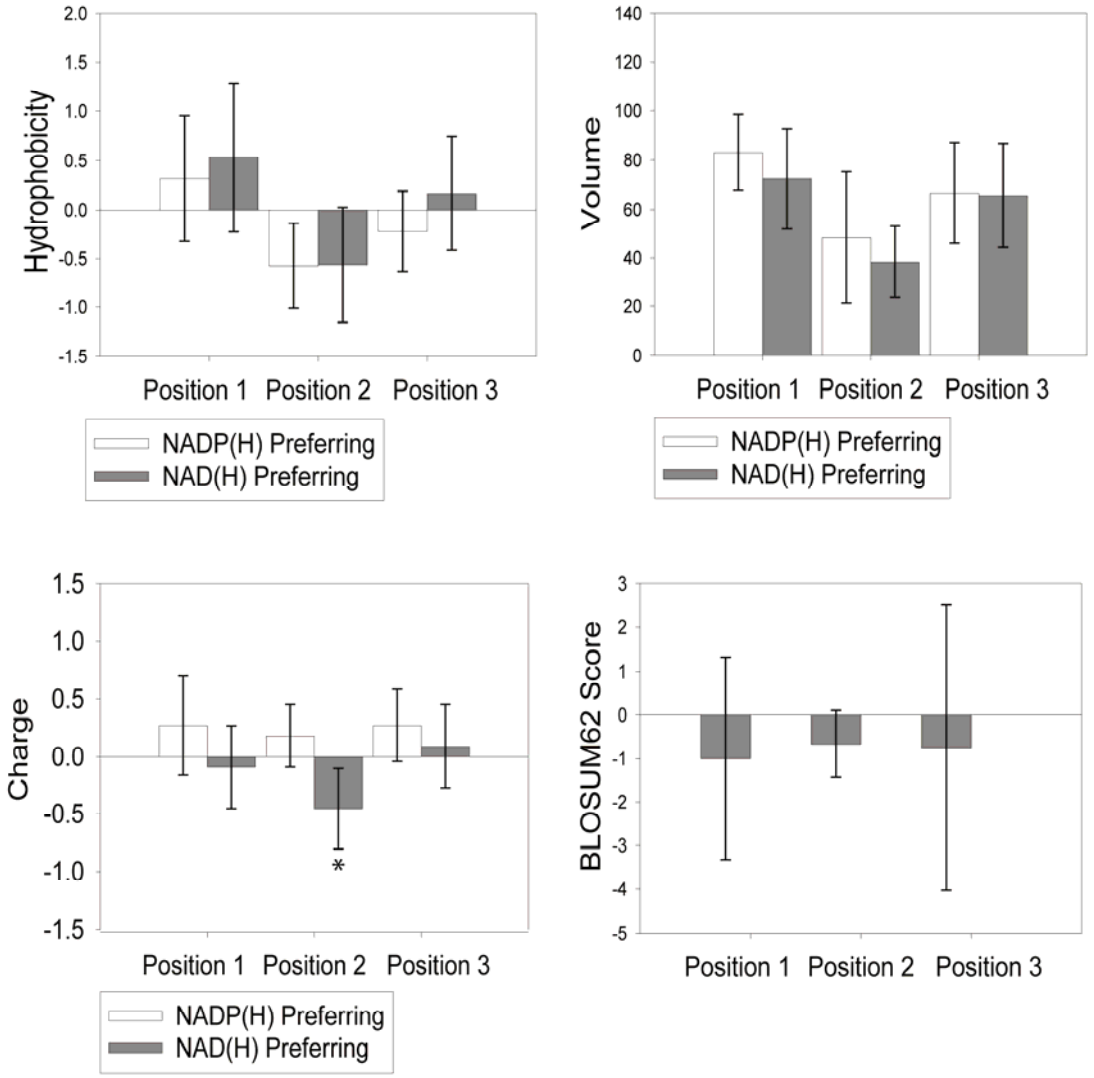

Figure 2 - 2: Comparison of average hydrophobicity, volume, charge, and BLOSUM62 score for all design positions. Error bars are shown for a 95% confidence interval. No statistically significant signal was found except for charge in position 2, where NADH-preferring residues were found to be more negative than NADPH-preferring residues, which is consistent with previous reports in the literature.

Table 2 - 3: Residues used in calculating average properties of NAD(P)H-binding residues.

	NADP(H) → NAD(H) Preferring		
Protein	***Position 1***	***Position 2***	***Position 3***
CtXR	Lys→Arg	Ser→Ala	Asn→Asp
GR	Val	Gly	Ala
KARI	Leu	Arg→Asp	Lys→Leu
PHBH	Glu	Arg→Ser	Gln→Arg
2,5-DKG	Lys→Gly	Ser	Val
1,5-AFR	Met	Ser→Asp	Thr
IDH	Lys→Asp	Tyr→Iso	Ala
GAPDH	Ala→Leu	Ser→Pro	His
PsXDH	Val	Ala→Asp	Arg→Iso
ferrodoxin:NADP+ reductase	Iso	Ser→Asp	Arg
L-LDH	Lys→Iso	Ser→Asp	Ala

While differences exist between the average values for charge, hydrophobicity, and volume, the average values are well within their confidence intervals for the mean, indicating no statistically significant signal. In position 2, enzymes preferring NAD(H) over NADP(H) were on average more negative compared with NADP(H)-preferring enzymes, which is consistent with what would be expected. The more positively charged residues electrostatically interact with the negatively charged phosphate of the adenosine ribose in NADP(H). The residues that are more negative in the NAD(H)-preferring enzymes may be compensating for the lack of the negative 2'-phosphate present in NADP(H) and are stabilizing the 2'-OH in the enzymes' NAD(H)-bound form.[146,148] In addition, we performed a similar analysis using the BLOSUM62[192] scores of the mutations in each position leading to altered cofactor specificity. The BLOSUM62 scores reported are based on the change in amino acid when going from NADPH-preferring to NADH-preferring residues. There was no statistically significant difference in the average scores per position. Notably, the mutations considered resulted, on average, in negative BLOSUM62 scores, indicating generally non-conservative mutations.[193] These results do not mean that charge, hydrophobicity, volume, and BLOSUM62 scores do not have an effect on affinity for different

cofactors. Instead, they imply that the effect of each factor separately is not monotonic or even discernible in isolation of all other metrics. Therefore, straightforward approaches using size, charge or hydrophobicity observations to suggest successful enzymatic redesigns cannot be successfully applied. Given the insufficiency of simple metrics to drive redesign, we next explored whether calculated binding affinities could be used to support enzyme redesign.

Section 2.3.2: Comparison of Calculated Interaction Energies of Enzyme-NAD(P)(H) Complexes vs. Affinities

Here we explore whether cofactor interaction energy is an adequate surrogate of cofactor specificity to drive computational cofactor alteration. To test this, we contrasted calculated interaction energy values (through CHARMM[30,194]) with published kinetic parameter data from a study aimed at changing specificity from NADPH to NADH in CtXR.[141] We compare the results of interaction energy changes calculated with and without solvation effects to determine whether the substantially increased computational cost needed for solvation is necessary.

The crystal structure of CtXR with NADH bound (PDB:1MI3) provided the starting coordinates for this analysis.[129] For this complex, we imposed a harmonic restraint to all non-hydrogen atoms with a force constant of 0.1 and mass weighting enabled. The CHARMM force field was applied and the complexes were energy minimized using the Adopted Basis-set Newton-Raphson (ABNR)[194] method with a Generalized Born with a simple switching implicit solvent model (GBSW).[124,125] The energy function in CHARMM accounts for forces from van der Waals interactions, bond stretching, bond angles, dihedral (torsion) angles, improper dihedral angles, electrostatics, and solvation. All minimizations converged successfully within the iteration limit. The interaction energy for the minimized wild-type complex was calculated using the Generalized Born with molecular volume integration (GBMV)[126,127] implicit solvent model as:

Interaction Energy = Energy of Complex – Energy of Apo Enzyme – Energy of Cofactor (1)

The interaction energy represents the intermolecular component of the total energy. The minimized wild-type structure was then mutated in positions relevant to altering cofactor specificity for NAD(P)H as reported in the literature.[141] Each mutated structure was minimized and had its interaction energy calculated using the same methods applied to the wild-type structure. ΔInteraction Energy was then calculated as:

$$\Delta Interaction\ Energy = Interaction\ Energy_{(Mutant)} - Interaction\ Energy_{(Wild\text{-}type)} \quad (2)$$

Here we used GBSW in all energy minimizations to ensure proper packing of hydrophobic cores while GBMV was used to approximate the solvation component for the interaction energy calculations. In the GBMV method,[126,127] the effective Born radius is computed by numerical integration of the molecular volume. The Coulomb field approximation includes a higher order correction term to improve agreement with the radii calculated from solving the Poisson-Boltzmann equation. The GBMV method was used for the interaction energy calculations because it is highly accurate but still more tractable in an iterative form than other options, such as solving the Poisson-Boltzmann equation. The minimizations utilized the GBSW model, as the GBMV model may utilize a sharp molecular surface representation for some systems, which would lead to large fluctuations in energy and cause stability problems in the simulations. GBSW is very similar to GBMV, but it is 2-3 times faster since it replaces the computationally expensive molecular surface calculation with a simple smoothing function at the dielectric boundary. GBSW's use of a smoothed dielectric boundary allows the change in polarization forces to vary more smoothly compared to GBMV.

For mutations changing cofactor specificity from NADPH to NADH, Figure 2-3 shows the calculated changes in interaction energy from (wild-type to mutant) including solvation against experimental ground state binding energy data outlined in Petschacher et al.[141] The correlation coefficient value is equal to 67%. This implies the calculated interaction energy explains 67% of

the variance in the experimental binding data. Although not a fully quantitative description, this is generally sufficient for rank ordering of different enzyme redesigns. When eliminating the implicit solvation model GBMV from the energy calculations, the correlation was reduced to 24% (data not shown), implying the need to include solvation effects in enzyme redesign. With these observations, we next modified and used the Iterative Protein Redesign and Optimization (IPRO)[119] framework to account for solvation based on the GBSW and GBMV models to explore redesigns for CbXR.

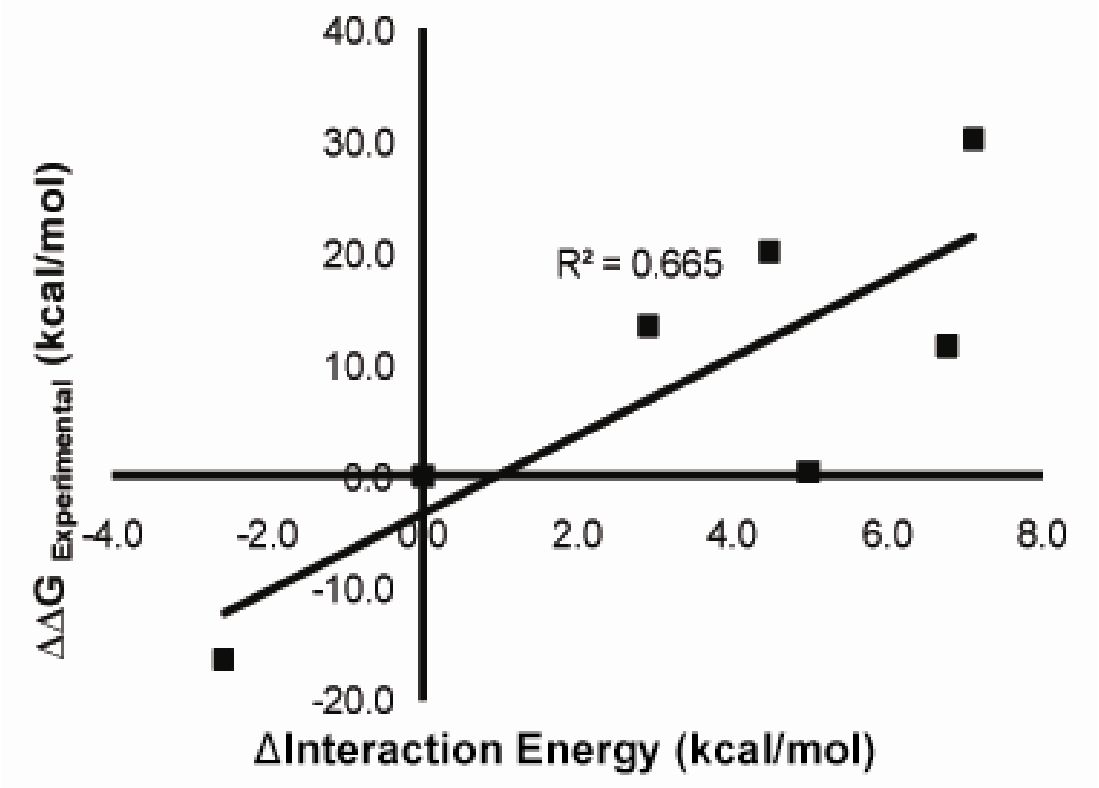

Figure 2 - 3: Changes in experimental ground state binding energies from Petschacher et al.[141] vs. our calculated changes in interaction energies. Shown are the changes in interaction energy with solvation showing reasonable correlation with the experimental data (R^2=67%), whereas changes in interaction energy without solvation correlated significantly less with the experimental data (R^2=24%) (data not shown).

Section 2.3.3: Computational Predictions using IPRO

Using the modified IPRO, we were able to generate variants of CbXR with improved interaction energies for NADH by targeting the design positions Lys-272, Ser-273, and Asn-274 in the NADPH binding pocket.

The wild-type interaction energies of CbXR-NADH and CbXR-NADPH were calculated to be -232 kcal/mol and -339 kcal/mol, respectively, and the interaction energy improvements towards NADH as a result of mutations predicted by IPRO for the top 10 designs are reported in Table IV. The mutants generated improvements in interaction energies for NADH by up to 78% relative to the wild-type and were selected among the 20^3 (=8,000) possible combinations of mutations. The interaction energies of the redesigned variants with the native cofactor NADPH were also calculated to assess the effect of the NADH binding improving mutations on the retention or abolishment of affinity for NADPH. Notably, we found that mutations in position 272 to methionine to be most effective at suppressing binding affinity based on an increase in interaction energy for the native cofactor. This is in agreement with the experimental results derived by Petschacher et al.[141] who found that the K274M mutation in the homologous CtXR increases NADPH dissociation and reduces the catalytic efficiency of CtXR utilizing NADPH. The increased hydrophobicity of the methionine side chain relative to lysine may imply that the orientation of the methionine side chain with respect to bulk water is not favored.[141,166]

Table 2 - 4: Computational and Experimental Results

Mutations	*ΔInteraction Energy$_{NADH}$ (kcal/mol)*	*ΔInteraction Energy$_{NADPH}$ (kcal/mol)*	*Activity with NADH (mU/mg protein)*	*Activity with NADPH (mU/mg protein)*	*Specificity (NADH/NADPH)*	*ΔCharge*
Wild-Type	0	0	0.7 ± 2.2	78.6 ± 4.7	0.01	0
EDS	-181	42	12.8 ± 2.1	5.5 ± 2.7	2.4	-3
EDR	-133	29	1.9 ± 0.8	< 0.1	> 19	-2
MGD	-129	181	17.1 ± 1.1	1.9 ± 0.6	8.9	-2
GGD	-126	195	19.0 ± 0.8	4.3 ± 0.8	4.4	-2
EQR	-103	-10	4.1 ± 0.6	72.3 ± 5.2	0.06	-1
RTT	-102	-30	14.4 ± 2.3	109.3 ± 16.5	0.13	0
MES	-99	180	1.81 ± 0.4	< 0.1	> 18	-2
MAE	-92	266	5.4 ± 1.3	< 0.1	> 54	-2
REG	-79	135	11.2 ± 3.1	< 0.1	> 112	-1
RSE	-70	15	10.8 ± 1.2	29.5 ± 0.4	0.37	-1
R	-60	-9	7.4 ± 2.5	14.7 ± 5.4	0.5	0
Negative Controls						
RNI	-45	74	< 0.1	ND	ND	0
KKG	101	232	< 0.1	ND	ND	1
RHC	-73	127	< 0.1	ND	ND	0

[1]The top designs predicted by IPRO with their changes in interaction energy are reported. The mutation labels (e.g. EDS) correspond to positions 272, 273, and 274 respectively in CbXR. The wild-type interaction energy with NADH was calculated as -232 kcal/mol and with NADPH as -339 kcal/mol.
[2]The NADH and NADPH-linked activities of the CbXR variants are reported in this table for comparison with the computational predictions. Values of < 0.1 indicate activity could not be detected above the background value in the absence of xylose. ND indicates the activity was not determined.
[3]The net local change in charge was calculated as a result of mutation and provided since charge was found to be important in determining cofactor specificity.

In Figure 2-1B, no hydrogen bonding interactions were present between NADH and the design positions chosen in CbXR. In contrast, as shown in Figure 2-4, the best computationally-derived design, CbXR-EDS (involving three point mutations K272E, S273D, and N274S) improved the interaction energy by -181 kcal/mol while forming a number of new hydrogen bonds between CbXR and the NADH. These newly formed hydrogen bonds likely explain the acquired affinity for NADH. Note that a hydrogen bond with the glutamic acid at position 272 stabilizes the 3'-OH of NADH (near the 2'-phosphate position of NADPH).

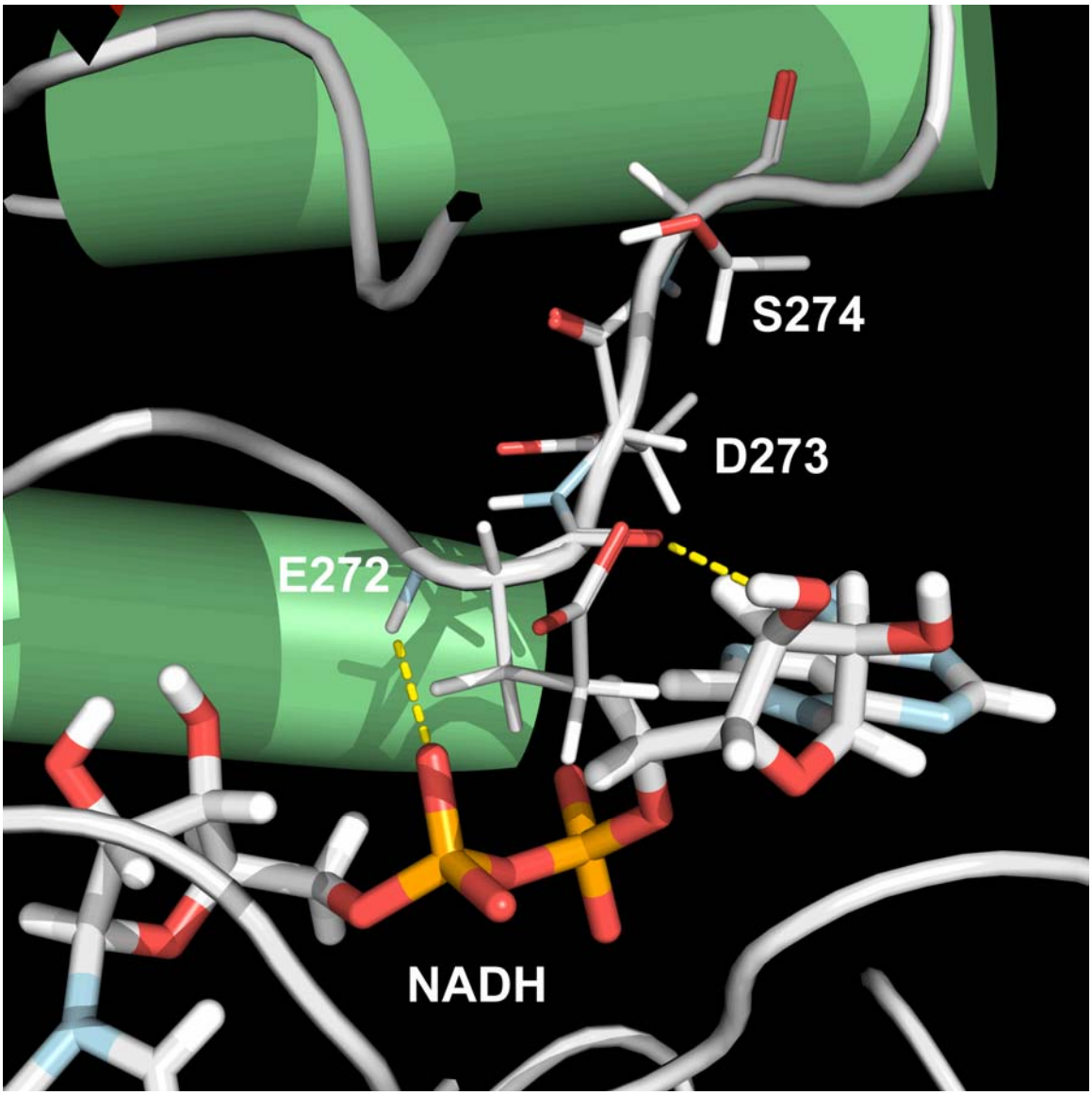

Figure 2 - 4: CbXR-EDS binding pocket containing NADH. The mutated residues Glu-272, Asp-273, and Ser-274 are labeled. Hydrogen bonding interactions are observed within 2.5 Å between the negative Glu-272 and the 3'-OH from NADH. This figure was made using PyMOL (Delano Scientific).

In nine out of the ten variants generated as shown in Table 2-4, including CbXR-EDS, the net charge change of the residues in the three positions considered is negative relative to the wild-type, with the change in CbXR-EDS being greatest (-3). This is in agreement with the results of the statistical analysis presented above. Presumably, this is because the more positively charged

residues in the NADP(H)-bound enzymes electrostatically interact with the negatively charged phosphate of the adenosine ribose. The residues with a higher net negative charge change in the NAD(H)-preferring enzymes, specifically the Asp and Glu residues, are thought to provide a significant portion of substrate specificity for NAD(H) by hydrogen bonding to one or both of the 2'- and 3'-OH and to compensate for the lack of a partially negative 2'-phosphate present in NADP(H).[137,146,148,195] Also, in three of the top five designs, position 272 was mutated to glutamic acid, indicating that this may be a critical mutation in changing the cofactor specificity of this enzyme.

Interestingly, CbXR-EQR and CbXR-RTT increased binding affinity for NADH, as required by IPRO, but also increased binding affinity for the original cofactor NADPH. Of the mutants generated, CbXR-RTT was the only design in which the net charge change as a result of mutation in the three design positions did not change. Comparing CbXR-RTT to the wild-type, there is no significant change in hydrophobicity or side-chain volume in any of the residues compared to the wild-type. Conservative increases in side-chain volume as a result of the mutations may slightly increase van der Waals and hydrogen bonding interactions to fine-tune the enzyme to bind NADH as well, without disrupting the original hydrogen bonding network and positive charge preference of the 2'-phosphate of NADPH

With these computationally-predicted designs, we next experimentally assessed the effect of the predicted mutations on cofactor preference to assess the efficacy of our computational predictions.

Section 2.3.4: Experimental Results

We experimentally constructed the top ten predicted designs to test the computational procedure and also shed light onto the functional significance of mutations in the binding pocket of CbXR.

One additional mutant (CbXR-R) was also constructed by mutating Lys-272 to Arg. The wild-type lysine in this position provides a positive charge for NADPH binding and the mutation of this residue to Arg was previously shown to change the cofactor specificity of CtXR from NADPH to NADH.[140,141] As negative controls, we also constructed three mutants not predicted by IPRO (CbXR-RNI, -KKG, -RHC).

Specific activities (μmol/min/mg) of clarified cell lysates containing the engineered CbXR mutants in the presence of 300 μM NADH and 300 mM D-xylose were measured and are presented in Table 2-4. Wild-type CbXR, as expected, clearly showed activity for NADPH (78.6 mU/mg protein) and less than 1 mU/mg protein of activity for NADH. Interestingly, while all top ten predicted designs clearly displayed some levels of NADH-linked enzymatic activity, all three negative controls exhibited a complete loss of reductase activity. Interaction energy calculations were performed on the negative controls for completeness and are reported in Table 2-4. Notably, CbXR-KKG was calculated to have worse affinity for both cofactors, which is consistent with the observation from results presented above for mutants having a net positive charge change. CbXR-RNI and CbXR-RHC were calculated to have increased affinity for NADH and decreased affinity for NADPH, with zero net charge change. The lack of NADPH activity for these mutants with no local charge change bolsters the importance of charge in determining specificity and affinity for cofactor.

Experimental results for redesigning the cofactor binding pocket of CbXR for NADH specificity confirmed a number of important computationally predicted redesign trends. Cofactor specificity of this enzyme is markedly influenced by different amino acid substitutions in three design positions. Replacement of Lys-272 by Arg, which was previously shown to completely reverse nicotinamide cofactor specificity in CtXR,[140] also yielded NADH activity in CbXR while weakening the NADPH-linked catalytic activity (by ~5 fold; see Table 2-4). While this mutant did not make the top ten predicted designs, off-line interaction energy calculations showed a

significant -60 kcal/mol (26%) improvement in interaction energy toward NADH relative to the wild-type CbXR. The effect of this mutation on NADH binding is clearly dependent on amino acids in positions 273 and 274. In the presence of Arg at position 272, mutation of Ser273 and Asn274 to larger and more hydrophobic amino acids in CbXR-RNI and CbXR-RHC resulted in a complete loss of reductase activity, whereas smaller and more hydrophilic amino acids at these positions in CbXR-REG, CbXR-RTT, and CbXR-RSE exhibited improved enzymatic activity.

In agreement with the computational results, methionine in position 272 is found to improve binding and activity for NADH while abolishing activity for NADPH. It appears that more negatively charged residues in the design positions help to explain the observed cofactor affinity alterations, as the net charge change for the three residues in CbXR-MGD, CbXR-MAE, and CbXR-MES is negative relative to wild-type. This may serve to compensate for the lack of a partially negative 2'-phosphate in NADP(H).[137,146,148,195]

Of the mutants experimentally tested, only CbXR-RTT showed activity toward NADH and also increased activity for NADPH. This is consistent with the computational results in that the binding affinity for both cofactors was increased for this mutant (Table 2-4). CbXR-EQR was predicted computationally to have a small increase in affinity for NADPH while also binding NADH. Experimental results revealed a slight decrease (~8%) in activity for NADPH while introducing novel activity for NADH. Cofactor specificity of the designed mutants was measured as the ratio of activity on NADH vs. NADPH. Seven of the ten predicted mutations exhibited specificity values greater than one, indicating greater specificity for NADH. Four mutants (EDR, MES, MAE, REG) exhibited completely diminished (< 0.1mU/mg) activity on NADPH, most likely as a result of local charge repulsion between the 2'-phosphate and the more negative residues in the design region. Mutant CbXR-REG exhibited a greater than 10^4-fold change in substrate specificity from NADPH to NADH.

The variants that showed the highest activity toward NADH, (i.e., CbXR-GGD, CbXR-MGD, and CbXR-RTT) were further analyzed by determining their Michaelis kinetic parameters for NADH and NADPH in the presence of saturating concentrations of D-xylose (300mM). Data were fitted to the Michaelis-Menten equation for a single substrate using non-linear least squares regression as shown in Figure 2-5. K_m and V_{max} values are listed in Table 2-5. The K_m values for the mutant enzymes and wild-type CbXR are comparable, however, the V_{max} values for these mutants are approximately one order of magnitude lower than the one for the wild-type enzyme. This suggests that NADH binding strength for these mutants is comparable to that of NADPH to the wild-type, and that IPRO successfully improved substrate binding. Figure 2-6 highlights the computationally-predicted enzyme-cofactor interactions for the best three mutant enzymes. NADH binding is suggested to be stabilized by a network of hydrogen bonds, absent in the wild-type enzyme, as well as van der Waals interactions between the side chains of residues in the design positions and the 2'- and 3'-OH groups in NADH. In CbXR-GGD (Figure 2-6A) and CbXR-MGD (Figure 2-6B), new hydrogen bonding interactions were established between the new residues and bridging phosphate groups in NADH. It is interesting that mutations to glycine were selected, perhaps to introduce conformational flexibility that allows better placement of the new cofactor in the binding pocket. In CbXR-RTT, new hydrogen bonding interactions appear to stabilize the 3'-hydroxyl group both for NADH (Figure 2-6C) and NADPH (Figure 2-6D). New hydrogen bonds from Arg-272 and Thr-274 are found to stabilize the 2'-phosphate group in NADPH. These mutations yield a net neutral charge change, which may be why both cofactors can be bound without substantial electrostatic resistance.

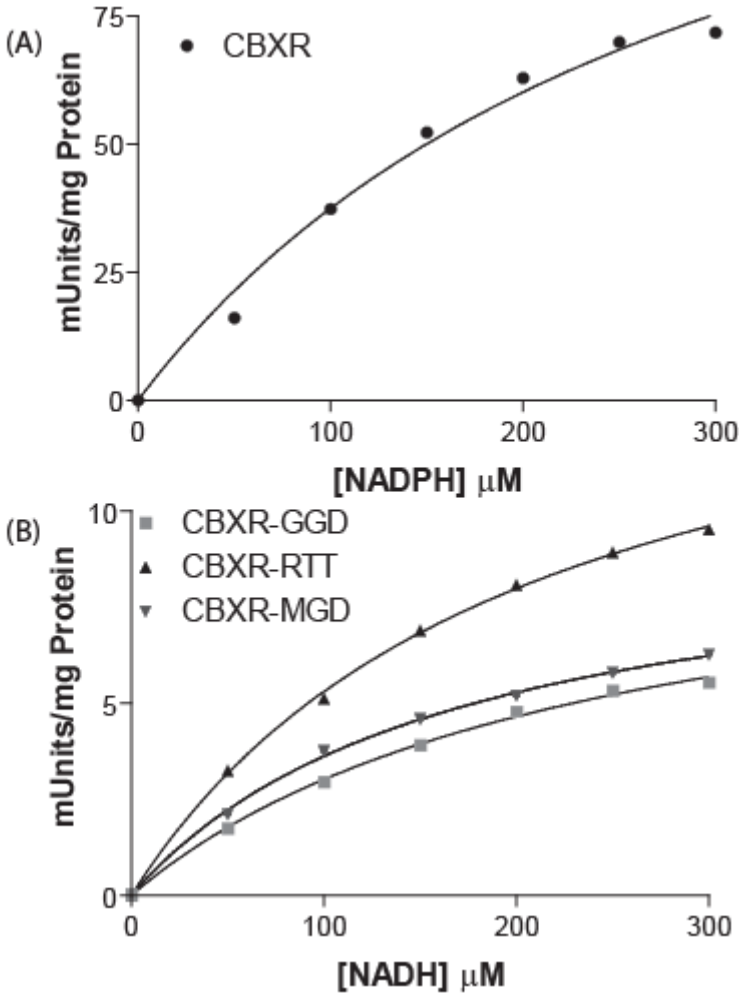

Figure 2 - 5: Michaelis-Menten plot for (A) wild type CbXR with NADPH and (B) three tested variants of engineered CbXR with NADH.

Table 2 - 5: Michaelis-Menten constants for wild-type and mutant CbXR.

Engineered CbXR with NADH				**Wild-type CbXR with NADPH**
	CbXR-GGD	***CbXR-RTT***	***CbXR-MGD***	***CbXR***
K_m (μM)	238 ± 24	205 ± 12	169 ± 12	307± 87
V_{max} (mUnits/mg)	10 ± 0.6	16 ± 0.5	9 ± 0.3	152 ± 25

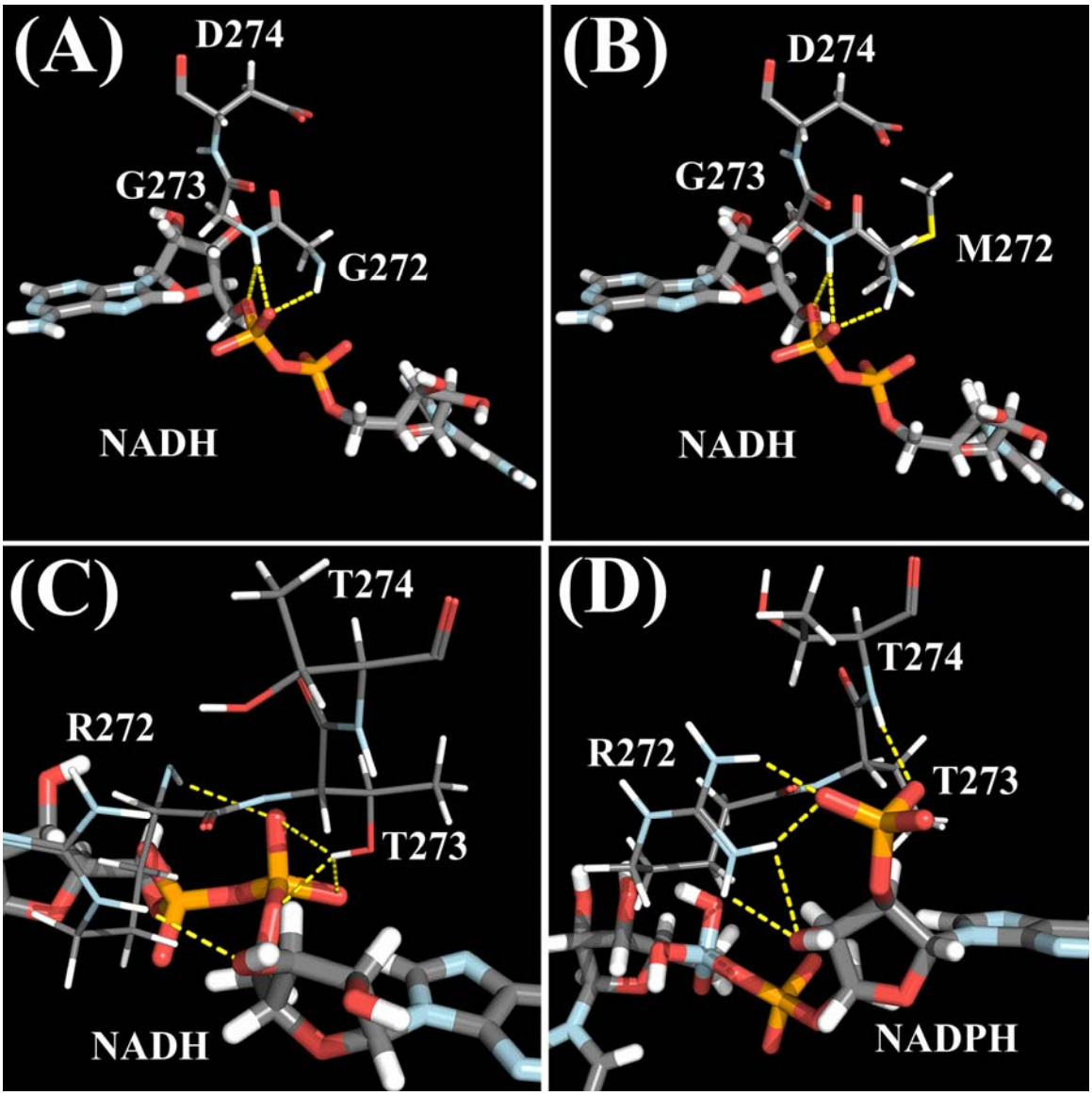

Figure 2 - 6: Structures of redesigned NAD(P)H binding pockets. (A) CbXR-GGD and (B) CbXR-MGD establish new hydrogen bond interactions between the mutated residues in CbXR and the bridging phosphates in NADH. The net charge change of these mutations is negative which may serve to compensate for the lack of negative 2'-phosphate in NADH. The mutations to glycine may serve to add conformational flexibility in the backbone to allow proper positioning of the NADH. CbXR-RTT, the mutation predicted by IPRO that was experimentally found to have dual cofactor specificity, bound to NADH (C) and NADPH (D). New hydrogen bond interactions are shown stabilizing the 3'-phosphate in NADH and NADPH from Arg-272, which may be the cause of the dual cofactor specificity. In NADPH, new hydrogen bonds are found to stabilize the 2'-phosphate group from Arg-272 and Thr-274. A neutral net change in charge is thought to contribute to dual cofactor specificity as well. All hydrogen bonds shown are within 2.5 Å. This figure was made using PyMOL (Delano Scientific).

Figure 2-7 plots the natural log of specific activity against interaction energy for all mutants. For NADPH (Figure 2-7A) there was a 79% correlation, and only 30% for NADH (Figure 2-7B). The difference in the ability of the interaction energy to predict differences in activity toward the two cofactors may be related to the fact that the position of NADPH is based on crystallographic data,[168] while NADH was computationally docked using ZDOCK (version 2.3),[74] causing some of the catalytic atoms to be positioned sub-optimally for the reaction to occur. An alternate explanation for this difference in correlations is that mutations that improve NADH binding may also disrupt xylose reduction to some extent, in which case activity will not necessarily correlate with interaction energy.

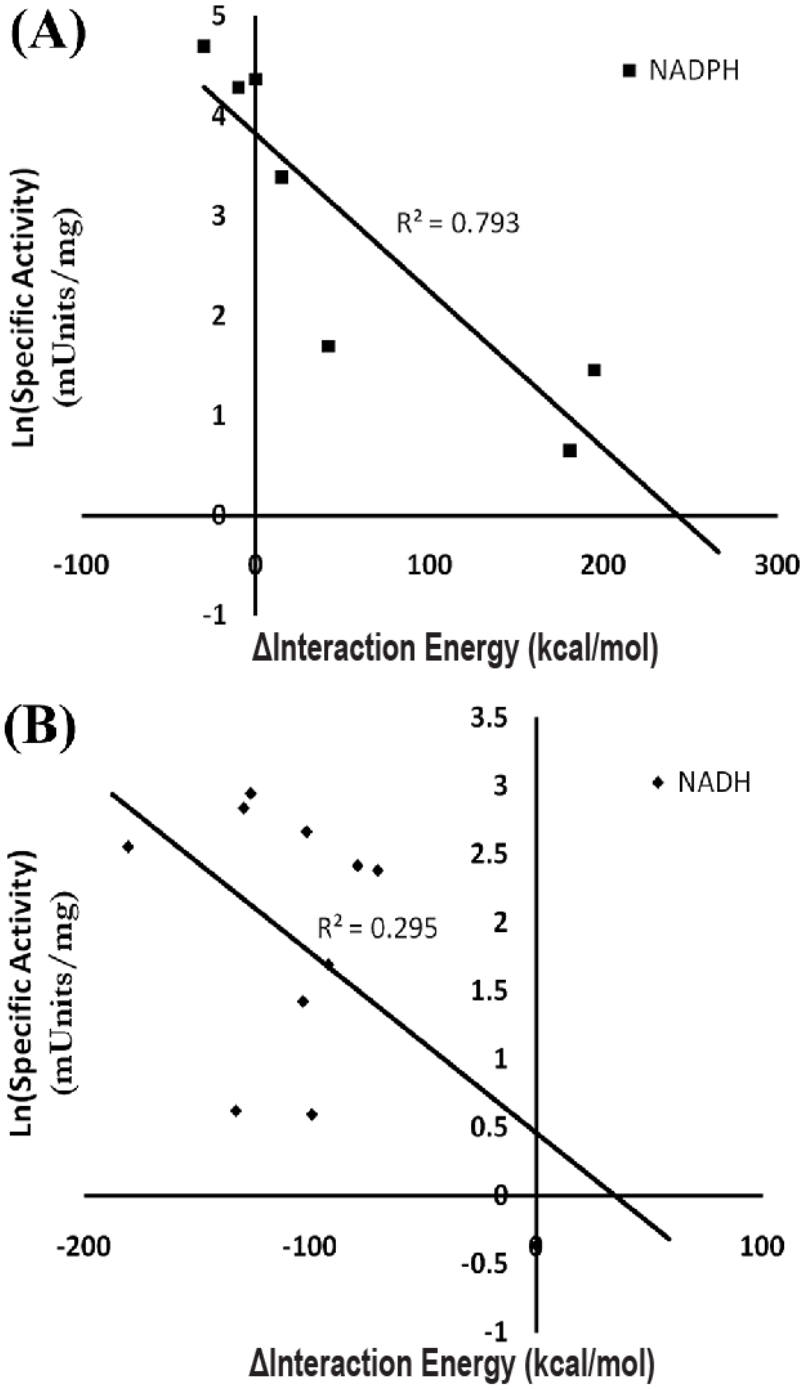

Figure 2 - 7: Plots of the natural log of specific activity toward NADPH (A) or NADH (B) versus interaction energy for CbXR mutants described in this study. The correlation coefficient for mutants yielding activity for NADPH is 79%, whereas the correlation is only 30% for NADH.

Section 2.4: Discussion and Summary

Redox enzyme variants with dual or switched cofactor preference are useful choices in metabolic engineering studies to better understand the role of cofactor utilization in strain performance. To date, strategies to engineer nicotinamide cofactor specificity have mainly relied on structural analysis and site-directed mutagenesis (see Table I). Despite a number of successes, a systematic computational workflow to drive design of cofactor specificity has been absent.

In this paper, using a modified IPRO workflow we identified sets of mutations that changed the nicotinamide cofactor specificity of CbXR from its physiological preference for NADPH, to the alternate cofactor NADH. We used calculated interaction energies to determine the increased or decreased affinities of CbXR variants for both nicotinamide cofactors, which were verified by our experimental results. Modifying the computational framework to account for implicit solvation[124,125,196] effects, we conclude that the increased computational expense needed to account in detail for solvation using GBSW and GBMV was warranted, as manifested by the successful experimental redesigns. Seven out of ten mutants proposed computationally to have increased affinity toward NADH were verified experimentally to bind and show significant activity toward NADH. Two variants identified by IPRO (i.e., CbXR-EQR and CbXR-RTT) led to dual cofactor specificity with preference for NADPH. Our results suggest interaction energies can successfully serve to introduce activity towards a new cofactor. Nevertheless, reaching the activity levels of the wild-type enzyme using the native cofactor for the redesigned enzymes using the new cofactor remains a challenge.[137] For example, for CtXR, Petschacher et al.[141] through site-directed mutagenesis was able to achieve increased catalytic efficiency for the alternate cofactor, but only at 27% of the native cofactor's efficiency (mutant K274R). Additional engineering efforts are therefore necessary to further increase activity toward NADH by

expanding the list of positions for mutation. Specifically, it may be necessary to proactively design the catalytic atoms in the binding pocket.

Given there were only three design positions, we believe the reported top ten designs are a good representation of the top performing ones. The rotamer/residue selection step in IPRO converges to the globally optimal solution for the randomly perturbed φ and ψ angles, however, a rigorous mathematical proof is not possible given the reliance on a simulated annealing step after every backbone relaxation/redocking step. Our computational results showed that the CbXR variants binding NADH are characterized by a net negative charge change in the binding pocket. We suggest that this net negative charge change coupled with the predicted new hydrogen bonding interactions between the mutants and NADH are important factors in ushering the change in CbXR's cofactor specificity. This is consistent with what has been observed in the literature: more negative residues in the binding pocket of NAD(H)-preferring enzymes compensate for the lack of partially negative 2'-phosphate of the NADP(H).[146,148,195] In summary, the computational procedure presented here can serve as a powerful tool for introducing enzyme activity toward a non-native cofactor. It can be applied to other enzyme-cofactor systems, and the methodology can be extended to engineer specificity toward oxidized or reduced nicotinamide cofactors, as well as to non-nicotinamide cofactors of interest such as AMP and GMP.

Chapter 3: Ground and Transition State Design of Cytochrome $P450_{BM3}$ for Altered Substrate Specificity

Section 3.1: Introduction

Enzymes are remarkably versatile structures exquisitely tuned by nature to selectively carry an incredible array of catalytic functions. Their immense potential to provide solutions to challenges in biomass treatment, biofuels production, biosensing, wastewater and environmental pollutants treatment[24,26,56,117,135,197-202] has long been recognized. Regrettably, enzymes with potential commercial utility often suffer from poor stability under the desired reaction conditions, inadequate catalytic activity or a lack of specificity for non-native substrates. Protein engineering efforts are increasingly effective at identifying mutations and/or recombinations that create enzymes with improved performance characteristics.[22,121,130,134,135,200,203] Purely experimental library screening approaches cannot predictably lead to optimized designs due to the difficulty/cost of screening and the enormity of the combinatorial design space. The use of computational methods has recently led to many protein redesign successes[22-26] by optimizing protein-ligand or and/or enzyme-cofactor interactions using static and molecular dynamics calculations. We have previously successfully computationally redesign proteins that were experimentally verified to use a different cofactor[118] and constructed novel calcium binding sites.[26] However, the rational design of enzymes for improved or novel catalytic activity remains an open challenge, in part because catalytic efficiency depends on not only ground state (GS) but also transition state (TS) interactions and energy barriers. A challenge in enzyme design is to improve substrate specificity, active site access, and binding while maintaining or even improving transition state stabilization. In this work we take the first steps to address this challenge by

performing calculations at multiple energy scales (QM and MM) and offering designs for further experimental validation. We choose as our system the cytochrome $P450_{BM-3}$ monooxygenase, which is functionally expressed at high levels in *E. coli* and has become a prime target for hydroxylase engineering of small alkanes towards alcohols. It is an ideal system on which to test our enzyme design work-flow because the reaction mechanism is well established, experimental design attempts exist for comparison, and the system is computationally tractable. From a practical viewpoint, the selective oxidation of light alkanes can produce liquid fuels or value-added chemicals from remote natural gas sources or less valuable refinery by-products.

In this work, we take the first steps to create a general computational workflow that can create enzymatic activity for a non-natural substrate. First, density functional theory (DFT) quantum mechanical (QM) calculations were employed to converge on the ground and transition states of the rate-limiting step. Next we parameterized the ground and transition states with QM into a molecular mechanics (MM) parameterization in CHARMM,[30,123] and explored its application against experimental mutant data prior to moving forward with computational design with a new computational saturation mutagenesis protocol. Arnold and coworkers[204] used directed evolution to identify a mutant of $P450_{BM-3}$, 535-h, which was capable of hydroxylating ethane to ethanol with 14 amino acid substitutions and 3 mutations occurring in the active site region (Positions 78, 82, 328). Finally IPRO was used to find the optimal rotamer/residue combination in predefined design positions that are systematically selected to offer designs for further experimental study.

Section 3.2: Background on $P450_{BM-3}$

We choose as our system for design the cytochrome $P450_{BM-3}$ monooxygenase from the bacterium *Bacillus megaterium.*[205,206] This soluble enzyme utilizes oxygen and the reduced cofactor NADPH to hydroxylate fatty acids, as depicted in Figure 3-1.

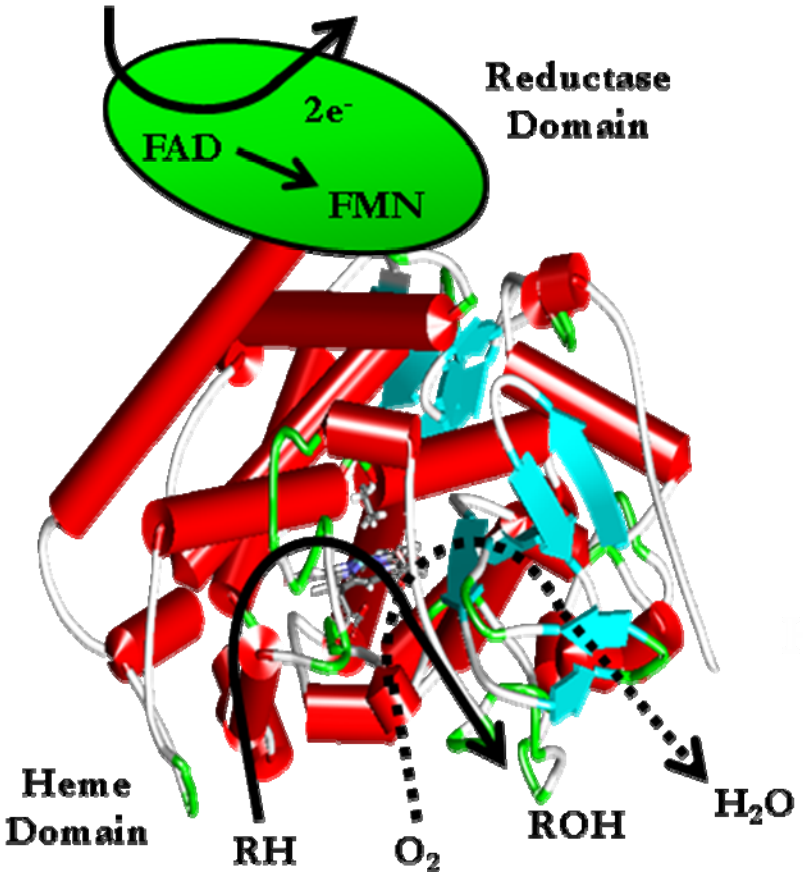

Figure 3 - 1: P450BM-3 catalyzed hydroxylation of a substrate.

$P450_{BM-3}$ is a single peptide chain composed of a heme-containing domain fused to a reductase domain (Figure 3-1) responsible for supplying electrons from NADPH to the heme via FAD and FMN moieties. The catalytic center in P450 involves iron(Fe(III)) equatorially coordinated by four nitrogens of prottoporphyrin IX and axially coordinated below the heme (proximal) by a cysteine thiolate.

The general reaction mechanism (Figure 3-2) of substrate oxidation begins with the substrate accessing the binding pocket and excluding a distal coordinating water molecule. This triggers transfer of one electron to the iron center (forming Fe(II)), which then allows O_2 to coordinate and form superoxo-Fe(III). Delivery of a second electron to the heme enables heterolytic O-O cleavage resulting in the formation of H_2O and the strongly oxidizing iron-oxo intermediate,

which then abstracts hydrogen from the substrate. The likely mechanism creates an organic radical in this hydrogen abstraction step that then ”rebounds” to the iron-hydroxyl site and hydroxylates the substrate.[207] Based on earlier studies on the cytochrome P450 catalytic mechanism, ripping the hydrogen from the substrate is energetically rate limiting since the bond strength of a terminal C-H bond is strong at 101 kcal/mol.

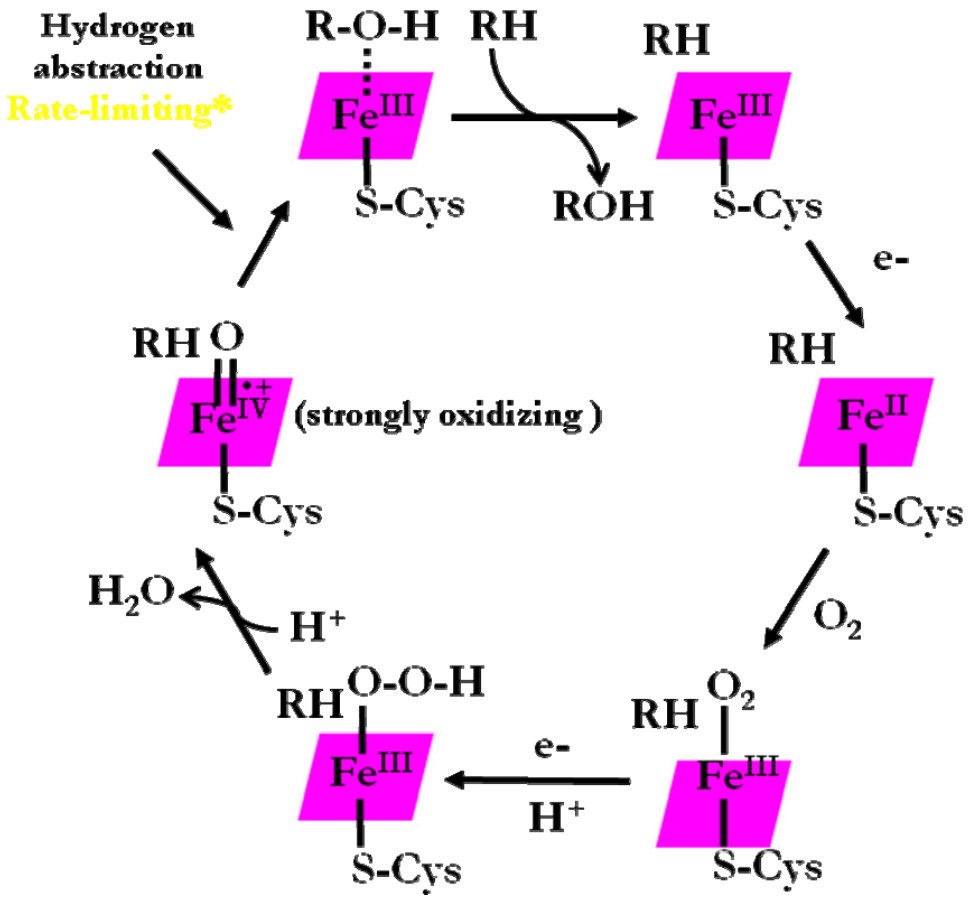

Figure 3 - 2: Consensus catalytic cycle for oxygen activation and transfer by Cytochrome P450.[208]

Wild-type $P450_{BM-3}$ naturally hydroxylates subterminal, saturated C-H bonds (bond strength ~99 kcal/mol). The low hydroxylation activity toward terminal C-H bonds (~101 kcal/mol) suggests that the BM-3 heme iron-oxo reactive intermediate does not have sufficient reactivity to efficiently perform terminal oxidation, a property generally associated with more complex, di-

iron core enzymes such as methane monooxygenase (MMO). The relatively complicated molecular organization of MMO and accompanying difficulties associated with obtaining large quantities of this enzyme complex make it a poor candidate in applied biocatalysis.[209] P450's on the other hand have been the subject of extensive characterization, engineering, and biotechnological implementation for more than 30 years. The ability of $P450_{BM-3}$ to be functionally expressed at high levels in the bacterium *E. coli* has made it a prime target for hydroxylase applications and engineering.[205] The heme domain of this P450 can be functionally expressed separately from the reductase portion. Conveniently, the heme domain alone is sufficient for catalyzing the "peroxygenase" or "peroxide-shunt" reaction, in which hydrogen peroxide (H_2O_2) reacts with the ferrous heme species to generate a hydroperoxo intermediate leading to iron-oxo.[210]

This work addresses the timely issue of controlled hydroxylation of small gaseous hydrocarbons (methane, ethane and propane) and terminal C-H bonds, and of elucidating determinants of C-H bond activation by mononuclear iron centers. In general, it is not clear whether the protein environment influencing the chemistry of P450 heme-catalyzed oxidations could support an iron-oxo species having sufficient strength to oxidize terminal C-H bonds or even methane (~104 kcal/mol) at biologically or biotechnologically useful turnover rates. Recent engineering efforts by Arnold and coworkers have provided encouraging results by showing that the mononuclear heme moiety is capable of oxidizing propane and ethane and even iodomethane, which has a bond strength of 103 kcal/mol.[211] Using a variety of mutagenesis and directed evolution techniques, they followed an evolutionary progression of BM-3 mutants that first hydroxylated long-chain alkanes (octane), followed by shorter chain alkanes (propane and ethane) but with poor coupling efficiency, and finally a propane monooxygenase variant having high coupling efficiency on propane and the ability to oxidize iodomethane.

These previous research efforts collectively demonstrate how important a role the protein environment plays in dictating not only substrate access but also heme chemistry and oxidative strength. These results are encouraging, particularly in light of the potential applications relating to conversion of small gaseous alkanes (methane, ethane and propane) into liquid alcohols via monohydroxylation. However, the activities or coupling efficiencies of the reported mutants are low and far from finding practical applications. Engineering activity towards progressively smaller substrates was met with increasing difficulty, and further attempts at using evolutionary techniques to isolate more active and/or efficient variants acting on propane and ethane have been difficult.[211,212] Reasons for these catalytic difficulties relating to substrate size are likely due to a combination of factors that include poorer substrate affinity, reduced substrate exposure to the heme face (resulting in increased uncoupling), reduced interaction- or free energy-dependent conformational changes that mediate electron transfer, and elevated terminal C-H bond energies. We postulate that this is not due to inherent limitations in the mononuclear heme moiety, but rather it is due to an increase in the complexity of the binding pocket chemistry required to achieve the desired functionality. This translates to progressively fewer protein engineering solutions that can support the desired reactivity. Hence evolutionary or combinatorial methods become less likely to identify the rare solutions.

We propose that by combining strengthening interaction energy at the reactant and transition states coupled with computational protein optimization and rational protein design, the $P450_{BM-3}$ protein environment can be fine-tuned to improve upon the results achieved by Arnold and coworkers using evolutionary design methods.

While the accumulation of additively acting point mutations throughout the protein structure is an effective approach to gradually evolving enzyme activity, a major limitation to this technique is its ineffectiveness in identifying cooperative mutations. Active-site mutations can improve alkane binding and proximity to the reactive iron-oxo species as well as stabilize the reaction transition

state, leading to both improved cofactor coupling efficiency and increased rates of C-H bond activation. This applies to increasing existing activities on small substrates (propane, ethane) as well as identifying solutions enabling "new" (or detectable) activity on methane.

The availability of crystal structures for both the wild-type P450 BM-3 heme domain[213] and one of its evolved propane-oxidizing mutants ("139-3")[211] enables us to use reliable active site structural information for the computational methods outlined in this work. Importantly, the tertiary structure of mutant 139-3 is changed very little relative to the wild-type enzyme structure (average Cα RMSD = 0.5 Å), indicating that accumulation of mutations conserving function during in vitro evolution (*albeit modified function*) correspondingly maintain the important structural features required for catalysis. In accordance, we expect successful mutagenesis strategies revealed through our studies to cause little perturbation to the backbone structure, further justifying our use of the available structure as a platform for engineering.

In summary, this enzyme system was selected because of its potential for computational redesign and its technologically important oxidation chemistry. The P450 catalytic mechanism is complex, requiring a mononuclear reaction center. Modeling and engineering this system therefore addresses a relevant and realistic enzyme engineering challenge. $P450_{BM-3}$ is a well-studied model system and engineering target, whose catalysis can be driven directly by NADPH or by H_2O_2 using the heme domain alone. Now that we have discussed the necessary background information on our system of interest and have identified the rate-limiting step (RLS), we proceed to calculate the ground and transition state structures for the RLS.

Section 3.3: QM Calculations of Ground and Transition States

We initiated preliminary computational efforts toward method validation and design of $P450_{BM-3}$ mutants for enhanced alkane oxidation activity. We calculated the GS and TS structures using

Gaussian03[214] using the density functional theory (DFT) method with the hybrid UB3LYP exchange-correlation functional.[215,216] The Lanl2DZ basis set was also invoked. The QM calculations included the Fe(IV) oxo, porphyrin, truncated cysteinate anion ($CysS^-$) and alkane substrate. Our transition state search procedure considered the "rebound" mechanism proposed earlier. The computational model system consisted of a cluster model including the ethane molecule and iron-oxo porphyrin complex and neglected the extended enzyme structure. The transition state was calculated with a linear scan method, and optimized and confirmed by vibrational frequency analysis. The QM calculated activation barrier for the hydrogen abstraction step was 33.3 kcal/mol calculated as:

$$Activation\ Barrier = Energy_{Transition\ State} - Energy_{Ground\ State}$$

The DFT UB3LYP/Lanl2DZ method is known to have a ~7.73kcal/mol mean deviation from experimental values of the activation barrier, which should be sufficient for this study. The calculated TS equilibrium bond lengths, angles, dihedrals, and Mullikan charges were extracted and used to parameterize a new molecule type in CHARMM representing the GS/TS as shown in Figure 3-3. Strict constraints were implemented into CHARMM for the bond lengths to preserve the integrity of the calculated transition state. A strict constraint on the iron, oxygen, hydrogen, carbon dihedral angle was implemented as well, which allowed for the ethane intermediate portion of the transition state to rotate. With a calculated GS and TS structure properly parameterized in CHARMM, we next performed a computational saturation mutagenesis on the positions previously determined experimentally by Arnold and coworkers to be able to catalyze the hydroxylation of ethane.

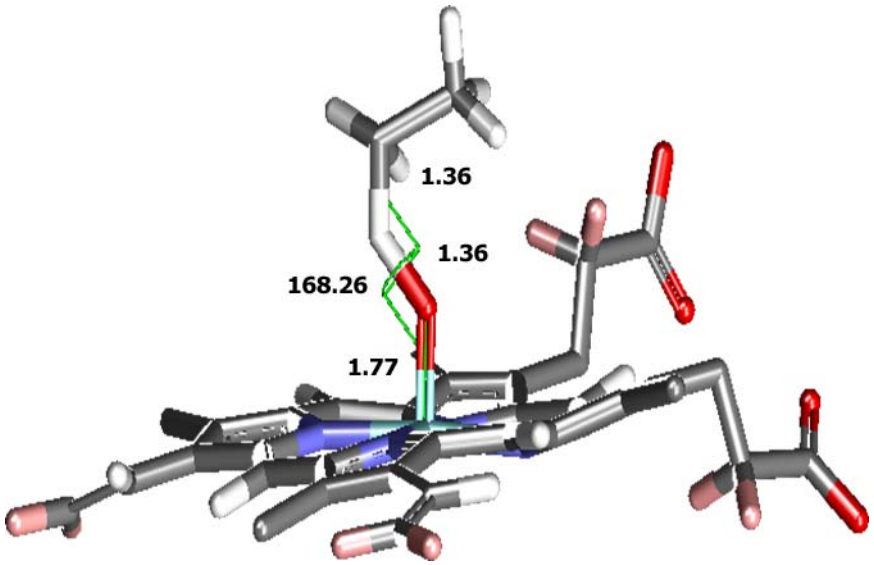

Figure 3 - 3: Calculated key transition state equilibrium bond lengths and angles used in reparameterization in CHARMM in conjunction with the charges calculated. All distances shown are in Angstroms.

Section 3.4: Computational Saturation Mutagenesis Procedure and Results

We developed a computational saturation mutagenesis procedure to computationally evaluate the effects of mutations in the 14 positions indentified in mutant 535-h.[204] The formulation presented in Figure 3-4 includes the CHARMM[30,123] energy function (van der Waals, bonds, angles, dihedrals, impropers, and electrostatic energy terms as well as solvation). We utilized the Generalized Born with simple switching (GBSW)[124,125] implicit solvent model for all energy minimization steps and the Generalized Born with molecular volume integration (GBMV)[126,127] model for all interaction energy calculations. The nested for loop calls CHARMM, selects predetermined design positions by the user, mutates each position to every amino acid one at a time, minimizes the energy of the system, and then calculates the interaction energy. The procedure was written in Python and currently is capable of running on one node or in parallel. The full code is available as is in Appendix A.

For {Pos1, Pos2, Pos3,…PosN}

For Amino Acid = {Ala, Arg, Asn,…M}

Mutate

Calculate Interacation Energy

$$Minimize \sum_i \sum_j Epot_{ij}$$

where:

$Epot_{ij} =$

$$\sum_{ij}\left(\frac{A_{ij}}{r_{ij}^{12}}-\frac{B_{ij}}{r_{ij}^{6}}\right)+\sum_{ij}\frac{q_i q_j}{\varepsilon 4\pi\varepsilon_0 r_{ij}}+\sum k_b(r-r_0)^2+\sum k_\theta(\theta-\theta_0)^2+\sum\left|k_\phi\right|-k_\phi\cos(n\phi)+\sum k_\chi(\chi-\chi_0)^2$$

Evdw + Eelec + Ebond + Eangles + Edihedrals + Eimpropers

Figure 3 - 4: Formulation for computational saturation mutagenesis procedure.

Arnold and coworkers [204] used directed evolution to identify a mutant of $P450_{BM-3}$, 535-h, which was capable of hydroxylating ethane to ethanol. This mutant involved 14 amino acid substitutions relative to the wild-type, with 3 mutations occurring in the active site region (Positions 78, 82, 328). The method outlined above was run to explore whether the 535-h mutant performance, identified by Meinhold et al., [204] can be explained by improvements in enzyme-ethane binding and enhanced transition state stabilization. Each run took 7 real hours of compute time distributed across 4 3.0 Ghz Intel Xeon Processors. The BM-3 crystal structure of P450 bound to the heme domain and the palmitoleic acid substrate was downloaded from the Protein Data Bank (PDB).[129] The QM calculated and MM parameterized ground and transition state structures were inserted to the enzyme by tethering the atomic coordinates to the coordinates of the heme in the crystal structure.

In Figure 3-5, we plot the interaction energy improvement ($-\Delta\Delta G_{calculated}$) compared to the wild-type enzyme for every position and single mutation choice. A positive value in Figure 5 indicates stronger binding of ethane to the mutant as compared to the wild-type enzyme. Looking at the 53-5h mutations (one at a time) arrived at through the directed evolution procedure we find that they are sometimes but not always the most energetically beneficial. In particular, for residues 78 and 328 that are in contact with the substrate (but not for position 82) the identified mutations are near at the interaction energy optimum. Mutations found by directed evolution in some cases appear to be near optimal in interaction energy, but in other cases they are not. Clearly, something else is needed to explain the identified mutations. Therefore we carry out exactly the same study at the transition state.

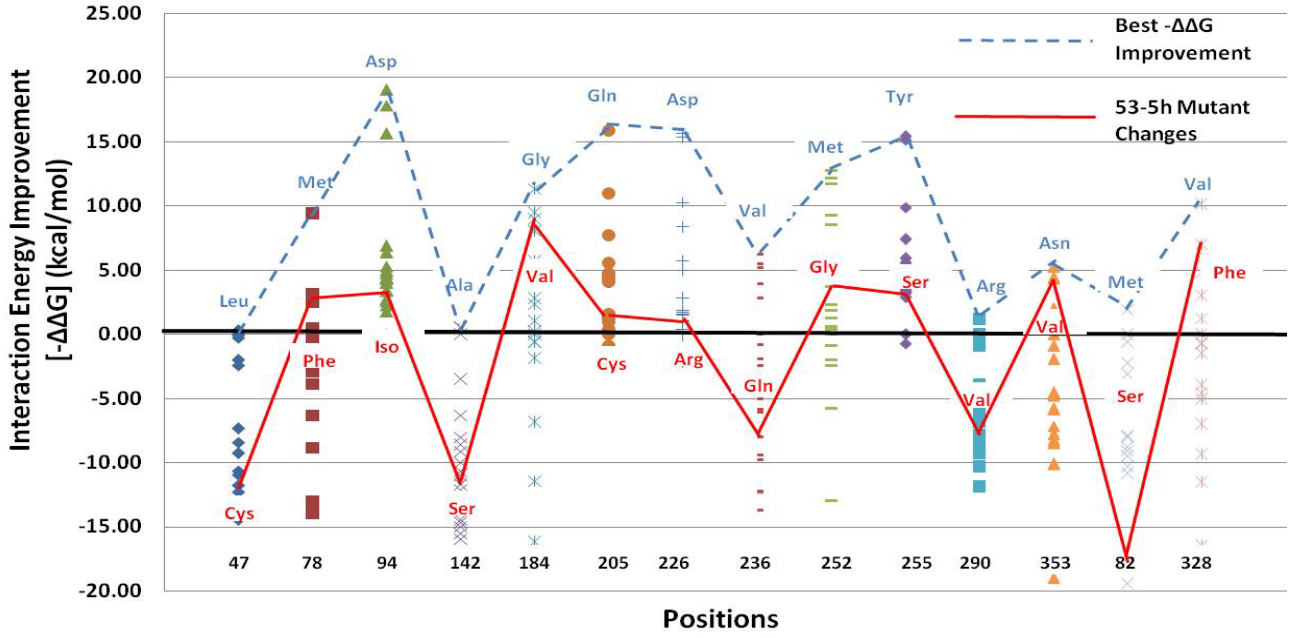

Figure 3 - 5: Interaction energy improvement (-ΔΔG) compared to the wild-type P450BM-3 upon single amino acid mutations at the 14 positions changed in mutant 535-h for the binding of the ground state (ethane) structure. The x-axis value represents the mutated position in the enzyme. The blue (top) amino-acid abbreviations represent the computationally determined optimal mutation at that position, whereas in cases the experimental and computationally optimal mutants differ, red values (bottom) indicate the experimental mutation.

Figure 3-6 illustrates the results of the computational saturation mutagenesis procedure applied to the transition state, where the interaction energy was calculated exactly the same way as in the ground state calculations. Interaction energy improvements at the TS are significantly higher on average than the corresponding ones at the ground state. This is largely due to the difference in charge distribution between the ground and transition states. The experimentally found mutations were able to track the energetically optimal residues at the transition state much better than the ground state, with the exception of those in the active site. The trends seen between the ground and transition states are that some mutations are improving substrate binding, while mutations in other positions are important for improving transition state stabilization, and presumably lowering the activation barrier. We don't know if directed evolution is optimal in terms of activity, or if other mutations exist that have not been identified. These results demonstrate the complementary nature of ground state and transition state calculations for explaining substrate binding and improving enzymatic activity levels and tell us when designing for activity, mutations must be selected that improve interactions at both the ground and transition states. Based on these findings, next we switch gears from the analysis of previous results to design at both the GS/TS with IPRO.

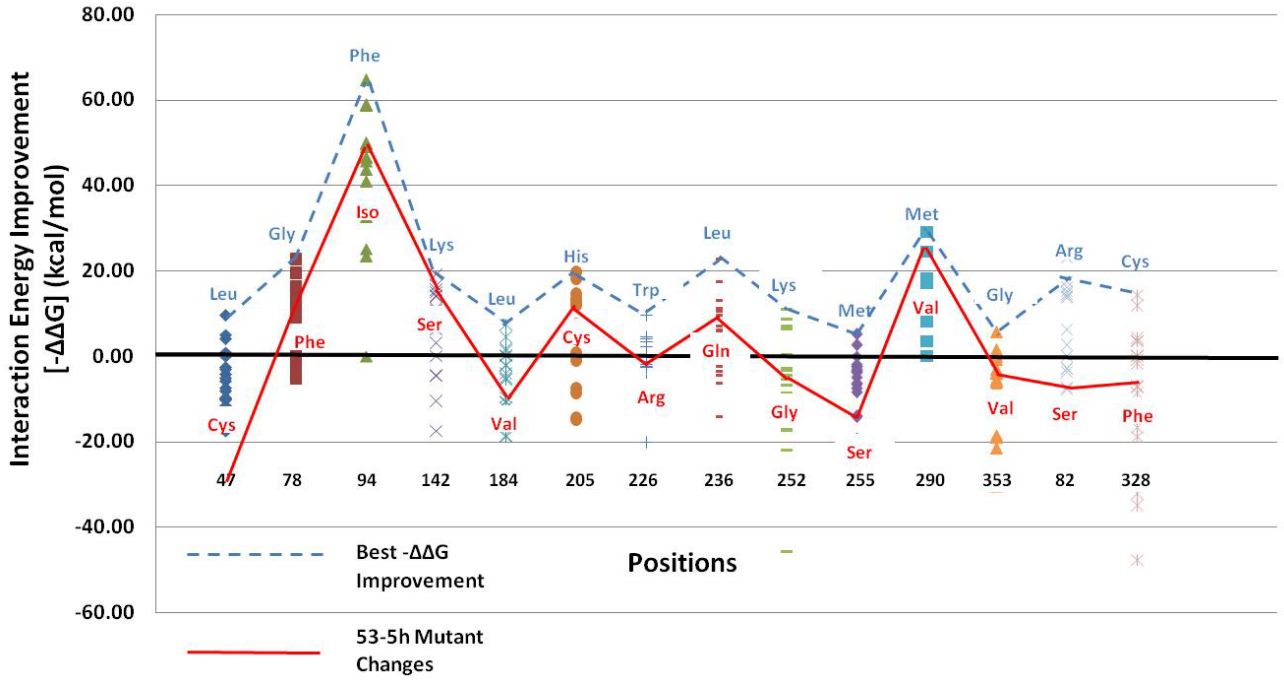

Figure 3 - 6: Improvement in interaction energy (-ΔΔG), compared to the wild-type P450BM-3, upon single amino acid mutations at the 14 positions changed in mutant 535-h for the transition state structure. Mutations were found that significantly improve the interaction energy between the protein and the transition state structure that were not found to improve the binding of the reactant state (ethane).

Section 3.5: Systematic Selection of Design Positions and Iterative Protein Redesign at the Ground and Transition States

In this next section, we describe the methods used to redesign P450 computationally to hydroxylate ethane using calculations at the ground and transition states. We first had to systematically identify design positions for design. Experimental methods usually involve performing random or saturation mutagenesis on positions known to influence the active site, but computational methods on the other hand usually rely upon rational selection of design positions identified by visualizing the active site residues and their proximity to the target substrate.

We therefore propose a new approach to systematically select design positions. In our approach we employ sequence, structural, and energetic factors. Shannon entropy analysis is possibly the

most sensitive tool to estimate the diversity of a system.[217] Therefore, the sequence of $P450_{BM-3}$ was aligned with the P450 superfamily, and 169 positions with intermediate Shannon entropies were identified. Intermediate Shannon entropies are defined as those with values between 2 and 3. Positions with entropies lower than 2 were not selected since this indicates that sequence conservation is probable and mutation of such a position may cause the enzyme to lose function. Values greater than 3 indicate that the sequence positions are approaching the limit where every amino acid choice is equally probably and therefore random.[218] Therefore, we selected positions with entropic values between 2 and 3.

Next, the distances of the previously identified positions to the ethane were calculated, and only those within 8A were selected as part of the final group of design positions. Lastly, we developed and performed a computational alanine scanning mutagenesis mutating every sequence position of P450BM3 to alanine, and identifying which of the positions affected the interaction energy with the ethane most drastically. The average change in interaction energy was 1.64 kcal/mol with a standard deviation of 1.02kcal/mol as a result of mutation to alanine. Therefore design positions that changed the interaction energy by more than 2.5 kcal/mol, or 1 standard deviation, were considered in the final pool of design positions. Based on the sequence, structure, and energetic factors, as well as knowledge of the active site residues, we refined our final # of design positions to 16 positions in Table 3-1. A pictorial illustration of how design positions were selected is shown in Figure 3-7. The code for the alanine scanning procedure can be found in Appendix B.

Table 3 - 1: Design positions selected from sequence, structure, and energetic factors.

Design Position	***WT Amino Acid***
47	Arg
69	Lys
75	Leu
78	Val
82	Ala
85	Gly
88	Thr
94	Lys
142	Pro
177	Met
182	Asp
200	Glu
260	Thr
327	Thr
328	Ala
329	Pro

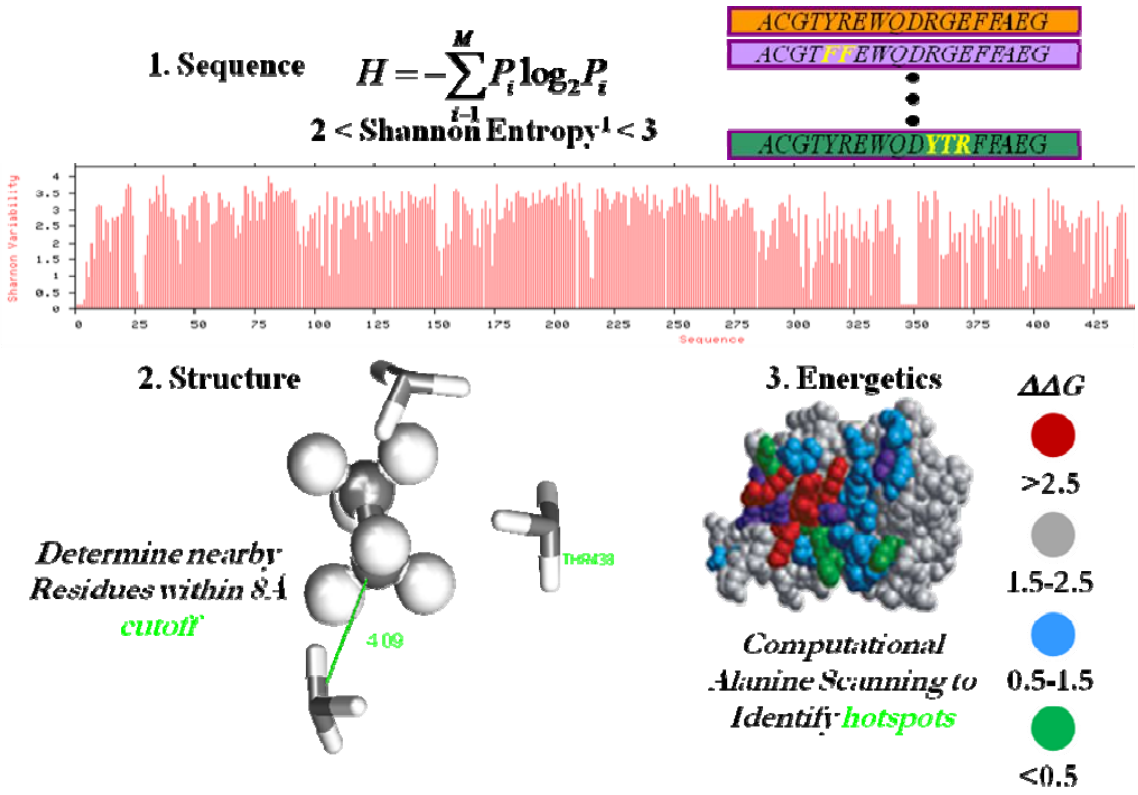

Figure 3 - 7: Schematic of Design Position Selection Protocol. Design positions were selected based on sequence, structure, and energetic factors.

We next used the IPRO[117-119] framework running in parallel and with solvation and optimized the interaction energy between the P450 and the ground and the transition states calculated previously. IPRO was performed on a Linux PC cluster using 4 3.06GHz Xeon CPUs with 4GB RAM for 8.3 CPU days.

IPRO generated 8 ground state and 6 transition state solutions that optimized the interaction energy between the P450 and the substrates presented in Table 3-2.

Table 3 - 2: IPRO generated designs optimizing the interaction energy between the ground and transition states.

Ground State Designs	*Transition State Designs*
260G	75D, 78K, 82G
88G, 260G	75D, 78K, 82G, 260G
88G, 260G, 327G, 328G	75D, 78K, 82G, 260G, 327G, 328G
88G, 200K, 260G, 327G, 328G	75D, 78K, 82G, 177G, 182K, 260G, 327G, 328G
88G, 177K, 182G, 200K, 260G, 327G, 328G	75D, 78K, 82G, 177G, 182K, 200K, 260G, 327G, 328G
47K, 88G, 177K, 182G, 200K, 260G, 327G, 328G	47H, 75D, 78K, 82G, 177G, 182K, 200K, 260G, 327G, 328G
47K, 88G, 177K, 182G, 200E, 260G, 327G, 328G	
47K, 94R, 88G, 177K, 182G, 200E, 260G, 327G, 328G	

At this stage in the design process, we cannot describe any specific designs in detail without experimental results. Instead, we will highlight some of the general trends found. What we are seeing is that IPRO predicted more positive and more hydrophobic residues at the ground state. The change in charge can be explained by the partial negative charge on the oxygen portion of the iron-oxo species. The increase in hydrophobicity can be explained by the reaction mechanism, whereas the mechanism can exclude the water molecules more easily after the substrate has accessed the pocket. For the transition state, the residues predicted were net smaller than the wild-type. Mutations to glycine can be rationalized by the backbone needing more flexibility to conform around the smaller ethane substrate.

We next employed IPRO using the design positions found by Meinhold et al. while employing directed evolution approaches.[204] The goal of this was to compare whether the experimentally-

found positions would improve interaction energy and the number of stabilizing residue contacts within 3 angstroms to the ethane relative to the design position selection procedure outlined above. The results are presented in Table 3-3.

Table 3 - 3: Comparison of IPRO designs using systematically and experimentally selected design positions.

Design	***Meinhold et al. Positions***	***Our Positions***
State	Ground	Ground
# of Contacts within 3 Angstroms	18	17
Design	***Meinhold et al. Positions***	***Our Positions***
State	Transition	Transition
# of Contacts within 3 Angstroms	16	22

The wild-type P450 has 11 contacts within 3 angstroms of the ethane. Our best ground state design improved the number of contacts to 17, whereas the experimentally determined positions improved the number of contacts to 18. IPRO using the Meinhold et al. positions improved the interaction energy by 25.6% relative to the best designs predicted by our systematically determined design positions. At the transition state, we observed just the opposite. The IPRO designs using our design positions improved the number of contacts to 22 from 11, whereas the designs predicted with Meinhold et al. design positions improved the number of contacts to 16. Our best design improved the interaction energy by 58.1% relative to the best design predicted with the Meinhold et al. design positions at the transition state. The design positions experimentally found computationally improved the interaction energy the best at the ground state, whereas the systematically selected design positions improved the interaction energy the best at the transition state. This conflicts with the results found by the saturation mutagenesis procedure, where we found that the transition state mutations better tracked directed evolution. Finally, with several designs found to improve the ground and transition state interactions, let us make sure the substrate is still capable of entering the binding pocket. Figure 3-8 shows the binding pockets of the best ground and transition state designs using our systematically selected

design positions, relative to the wild-type binding pocket. Clearly the substrate can still access the binding pocket to bind/unbind.

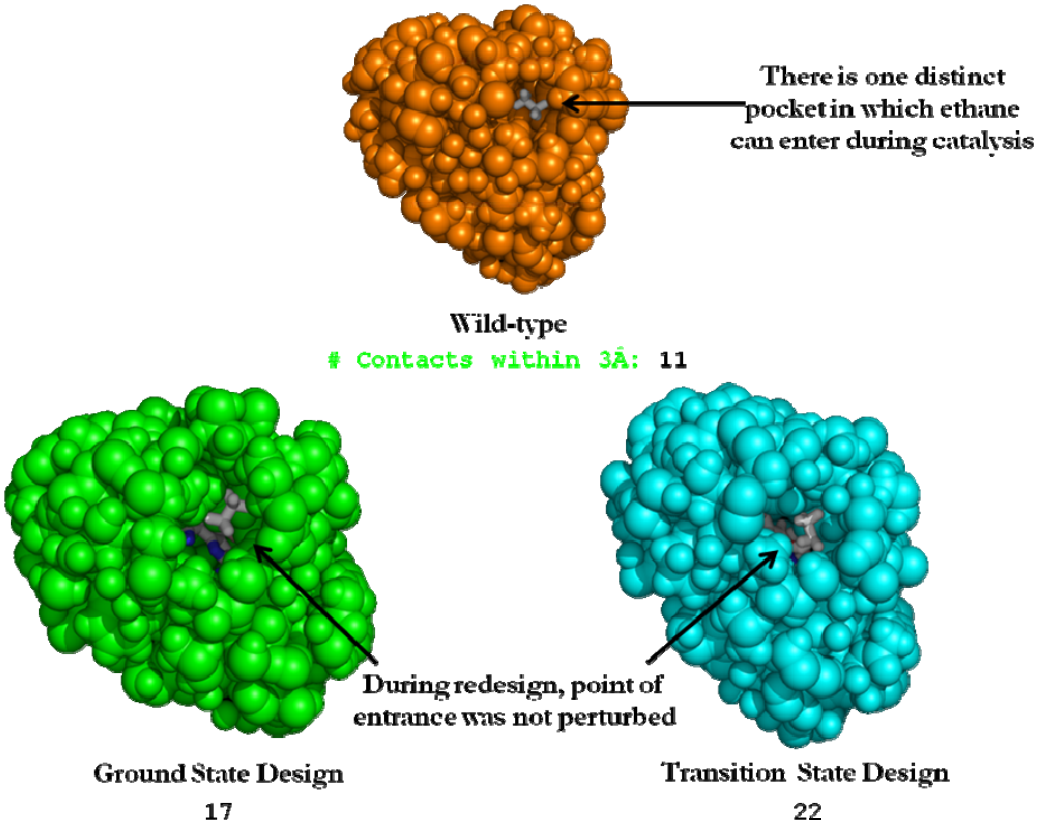

Figure 3 - 8: Visual depiction of best ground and transition state binding pockets relative to the wild-type binding pocket. The best designs improved the number of contacts while still allowing the substrate to bind/unbind.

With the shortage of experimental data for this system, the next steps would be to construct the designs predicted both by Meinhold et al.'s design positions as well as our systematically selected design positions for experimental quantitative comparison. These next steps will be taken when developing a comprehensive enzyme design workflow followed by experimental verification. These limited number of sequence designs are offered for further experimental study

Chapter 4: Future Work

Section 4.1: Future Perspectives on Cofactor Engineering

Cofactor engineering as a subset of protein engineering has a plethora of applications in metabolic engineering, since many metabolic pathways are limited by cofactor preference. Being able to systematically change an enzyme's cofactor preference will allow pathways that previously could not be explored to be entered. Future development of computational protein design will more likely progress as a result of understanding failures rather than the rare successes, the development of more accurate energy functions, the determination and understanding of a computational surrogate for enzymatic activity, and the adoption of fully-automated protein design.

Section 4.2: Future Perspectives/Work on Enzyme Design

We recently have been approved for funding to develop a computational enzyme design workflow followed by experimental verifications by the NSF. The preliminary results presented in Chapter 3 were utilized in the proposal, and our future work entails producing the following deliverables.

4.2.1: Determination of Optimal Designs Improving Reactant Binding and Product Off-Rate

At the ground states level we plan to deploy and customize our previously developed computational protein procedure Iterative Protein Redesign and Optimization (IPRO)[117-119] to

explore what mutations in the wild-type enzymes are likely to lead to enzyme variants which improve the interaction energy between the reactant/enzyme as well as disrupt the interaction energy between the product/enzyme. We will target candidate designs that bind the reactant more strongly (improve substrate binding) while also allowing for product molecule release (improved product off-rate). We will develop a systematic approach for design position selection based on sequence variability, structure factors, as well as energetic. To this end, we plan to make use of the efficient amino acid selection scheme embedded within IPRO to drive amino design that relies on the sequential solution of mixed integer linear programming (MILP) optimization problems to global optimality. This framework has been shown to be effective at identifying mutations that change enzymatic cofactor specificity [219] as well as graft new binding sites onto existing protein scaffolds.[26] Currently IPRO performs binding calculations using CHARMM at only the ground state. We plan to extend IPRO by including not just the binding energy optimization at the ground states but also optimizing the interaction energy at the QM calculated transition state.

We intend to explore different ways of melding together all these objectives within IPRO. This includes minimizing interaction energy at the ground state while imposing a lower limit on the minimum allowable improvement in the TS stabilization energy and an upper bound on the interaction energy at the product ground state. Conversely, we will explore the direct minimization of TS stabilization energy while imposing bounds on how high/low the interaction energy is allowed to be for the reactant/product ground states, respectively. We will use known good/poor binders[220-225] for the DHFR system to arrive at good threshold values for successful binding. Similarly to the original IPRO procedure, the outer optimization problem will be solved using a Metropolis criterion to update amino acid choices after each iteration. We plan to start with these two choices as performance targets and evolve it in response to the experimental results. For example, we will explore whether backbone relaxation and substrate redocking steps are needed after each time the inner rotamer optimization problems are solved. This work will seamlessly

integrate with QM/MM calculations by making use of QM results to re-parameterize the employed force-field. The improved transition state stabilization indicated by IPRO will also be validated by directly computing activation barriers using QM/MM methods with a subset of encouraging mutants.

We intend to focus our design efforts on the well-known rate-limiting steps for both DHFR reduction (benchmark/test case) and P450 hydroxylation (design case). The novelty here will be to identify a manageable number of enzyme redesigns with full backbone flexibility that allow for proper reactant/product placement and energetic interactions at the ground states.

In addition, we plan to use molecular dynamics MD simulations using CHARMM on the set of identified promising mutants to assess the structural changes upon mutation and their dynamic features and confirm the ability of the substrate to access the active site. Accurate equilibrated structures are critical to obtain the correct reaction path and accurate reaction barrier using the subsequent QM/MM method. MD simulations will aid in validating the assumption of a "static" backbone structure and limited tertiary structure rearrangement, or identify structural changes that must be considered and added within the IPRO framework. Simulations will be carried out for the designed protein-substrate complexes to obtain more accurate complex structure and accurate binding free energy using techniques such as MM-PBSA.[226]

4.2.2: Use of QM/MM methods to explore transition state stabilization

In recent years, multi-scale modeling techniques using combined quantum-mechanics/molecular-mechanics (QM/MM) methods are making rapid progress both methodologically and with respect to their range of application, especially in mechanistic studies of enzymes.[227-231] In the QM/MM approach, quantum mechanics (QM) is applied to the reactive center to properly describe interactions, and classical molecular mechanics (MM) is used to treat the extended enzyme

environment. The development of QM/MM methods has enabled the modeling of complex chemical and biological processes for a reasonable computational effort at the necessary accuracy. In the area of enzyme design, the QM/MM approach has been use to evaluate the potential performance of design variations, and obtain detailed information about the impact of composition alterations on the catalytic mechanisms and activation barriers.[232] We proposed to apply the QM/MM method in three different stages of the computational enzyme design process: (1) to obtain more accurate ground state and transition state structures and force field parameters taking into account the polarization effect due to the enzyme environment for input into IPRO, force-field based enzyme optimization protocols; (2) to validate the IPRO/MM approach to determining transition-state stabilization by assuring that trends in enthalpic activation barriers with mutation are preserved between the IPRO/MM and QM/MM methods, and (3) to accurately evaluate the change of the transition state barrier for selected mutants with advanced free-energy techniques. We will apply the QM/MM implementation in CHARMM[30,122] and Q-Chem[233] software packages to obtain the reaction profile. This choice provides the best integration with the IPRO/MM part of our design scheme, and utilizes a flexible software package with multiple transition state search methods, facile construction of QM/MM input, and ease of parallelization across available Linux computational clusters.

We will use the DFT method with hybrid B3LYP exchange-correlation functional[215,216] and Lanl2DZ, 6-31G* basis set for the QM calculations and CHARMM or AMBER force field for MM calculations. The B3LYP functional performs well for accurately reproducing a number of properties, including enthalpies of formation. The transition states will be obtained by direct optimization to the transition state, or by constraining the reaction coordinate in terms of the distances between key atoms, or finding the minimum-energy path (MEP) using the nudged-elastic-band (NEB) approach as implemented by Yang and coworkers.[234] We will also explore the calculation of free energies of activation using QM/MM methods, to determine whether

qualitative enthalpic trends among mutants are preserved on a free-energy landscape. As long-time QM/MM/MD simulation is needed for the free-energy calculation to allow for proper molecular-dynamics sampling of the enzyme, semi-empirical QM methods, such as PDDG/PM3, or SCC-DFTB (self-consistent charge density-functional tight-binding) methods, will be adopted to save computational costs.[235]

4.2.3: Experimental Assessment of Predicted Designs

Throughout our studies we will successively construct, express, and characterize select DHFR (benchmark/test case) and $P450_{BM-3}$ (design case) mutants suggested by the design predictions. DHFR was chosen as a testing ground for our computational methods due to the wealth of previous kinetic studies, including the influence of various mutations on kinetic parameters in the reduction of DHF.[220-224,236] However it will still be necessary to perform similar studies in-house under a uniform set of reaction conditions. As described below, a variety of kinetic parameters will be determined for activities of DHFR on DHF and $P450_{BM-3}$ on ethane. We will then compare experimental results from the computational predictions, and assess whether individual mutants (or which mutants) qualitatively reflect the computationally determined binding affinity and activity. Data accumulated from these mutant studies will also help to identify relationships between the kinetic parameters, e.g. Michaelis constant (K_m), rate constant (k_{cat}), and coupling efficiency. This in turn will help to identify which computational design determinants (binding energy, transition state energy, active site hydrophobicity, etc.) are more or less important for improving coupling efficiency or K_m values. Results from these experimental studies will provide insights onto which predictions (residue positions or combinations thereof) better reflect the expected computational/kinetic results. Discrepancies between experiments and computational predictions will pinpoint interventions within the computational workflow. For example, a lack of

qualitative agreement between k_{cat} and QM/MM computed activation barrier trends will necessitate an increase in the QM-described part of the enzyme, extension to explicit solvation models, and/or advances in the level of theory applied. Discrepancies between trends in K_m and IPRO-evaluated binding energy trends will target force-field evaluation or a relaxing of structural constraints.

We can readily construct and assay select DHFR variants using standard procedures well-documented for this enzyme.[237] For the case of P450 activity assays, approximately 10 – 20 individual mutants (each containing a variety of point mutations) can be characterized from clarified cell lysates in a matter of 4-5 months. More promising or interesting mutants will be fully purified for more detailed characterization. Cirino has previously worked extensively on the P450 systems with Arnold and co-workers, and protocols similar to those already published will be used here.[204,210,238]Briefly, we use plasmid pCwori to express in *E. coli* strain BL-21 the full-length $P450_{BM-3}$, or only the heme domain carrying a C-terminal 6-Histidine tag fusion to facilitate rapid nickel affinity purification. Standard protein engineering protocols will be used to add mutations to the $P450_{BM-3}$ heme domain [185]. Soluble P450 concentrations in cell lysates or purified samples can be quantified via the characteristic reduced heme CO-binding spectrum (450 nm). For reactions involving the reductase domain, NADPH oxidation rates will be measured spectrophotometrically (6.22 mM^{-1} cm^{-1} at 340 nm). Alkane hydroxylation reactions will be performed in buffer under one atmosphere of dioxygen, nitrogen, and different mole fractions of ethane or propane (provided by a balloon filled with the gases). An NADPH regeneration system (using sodium isocitrate and isocitrate dehydrogenase) will be used for the reactions. Ethane and propane hydroxylation reactions will be performed in 25-mL Schlenk flasks topped with a balloon and containing enzyme (either purified or in cell lysate) and alkane-containing buffer. The hydroxylation products will be derivatized to alkyl nitrites, samples will be worked-up as described [238] and analyzed by GC.

From the initial rates of NADPH oxidation and alcohol formation data, coupling efficiencies can readily be determined. Reaction rate constants (k_{cat}) and hence activation energies can be estimated based on the quantified P450 expression level (by CO-binding). By varying the partial pressure of ethane or propane, the Michaelis constant (Km) for each mutant and substrate will be estimated. Similarly, the binding constants (K_d) for these substrates will be estimated spectrophotometrically by monitoring the heme iron spin shift (from 418 nm to 390 nm) accompanying substrate binding. Stabilities of select mutants compared to wild-type P450 BM-3 will also be established as described [239]. Finally, some mutants may suffer perturbed electron transfer to the heme, but may still bind substrate and stabilize the hydroxylation reaction transition state upon formation of the iron-oxo species. Thus, it will be important to test whether mutants with poor activity on NADPH+O_2 show improved kinetics using the H_2O_2 shunt pathway. Therefore we will also screen all mutants for peroxygenase activity.[210] Heme domain mutants with notable kinetic properties will be purified and their peroxygenase kinetics characterized more thoroughly.

Steps 1-3 put forth a roadmap that we plan to follow to identify improved DHFR/folate reductases and $P450_{BM-3}$ ethane hydroxylases to establish a generalized computational workflow for enzyme engineering. We anticipate that as this future undertaking unfolds we may modify this roadmap and add additional computational and/or experimental steps.

Bibliography

1. Shuler ML, Kargi F. Bioprocess engineering. Upper Saddle River, NJ: Prentice Hall; 2002. xx, 553 p. p.
2. Arnold FH. Combinatorial and computational challenges for biocatalyst design. Nature 2001;409(6817):253-257.
3. Moore GL, Maranas CD. Computational Challenges in Combinatorial Library Design for Protein Engineering. Aiche Journal 2004;50(2):262-272.
4. Romero PA, Arnold FH. Exploring protein fitness landscapes by directed evolution. Nat Rev Mol Cell Biol 2009;10(12):866-876.
5. Korkegian A, Black ME, Baker D, Stoddard BL. Computational thermostabilization of an enzyme. Science 2005;308(5723):857-860.
6. Johannes TW, Woodyer RD, Zhao H. Directed evolution of a thermostable phosphite dehydrogenase for NAD(P)H regeneration. Appl Environ Microbiol 2005;71(10):5728-5734.
7. McLachlan MJ, Johannes TW, Zhao H. Further improvement of phosphite dehydrogenase thermostability by saturation mutagenesis. Biotechnol Bioeng 2008;99(2):268-274.
8. Hasty J, McMillen D, Collins JJ. Engineered gene circuits. Nature 2002;420(6912):224-230.
9. Vercillo NC, Herald KJ, Fox JM, Der BS, Dattelbaum JD. Analysis of ligand binding to a ribose biosensor using site-directed mutagenesis and fluorescence spectroscopy. Protein Sci 2007;16(3):362-368.
10. Hellinga HW, Marvin JS. Protein engineering and the development of generic biosensors. Trends Biotechnol 1998;16(4):183-189.
11. Maier NM, Franco P, Lindner W. Separation of enantiomers: needs, challenges, perspectives. J Chromatogr A 2001;906(1-2):3-33.
12. Chockalingam K, Chen ZL, Katzenellenbogen JA, Zhao HM. Directed evolution of specific receptor-ligand pairs for use in the creation of gene switches. Proc Natl Acad Sci U S A 2005;102(16):5691-5696.
13. Koh JT. Engineering selectivity and discrimination into ligand-receptor interfaces. Chem Biol 2002;9(1):17-23.
14. Tang SY, Fazelinia H, Cirino PC. AraC regulatory protein mutants with altered effector specificity. J Am Chem Soc 2008;130(15):5267-5271.
15. Griswold KE, Kawarasaki Y, Ghoneim N, Benkovic SJ, Iverson BL, Georgiou G. Evolution of highly active enzymes by homology-independent recombination. Proc Natl Acad Sci U S A 2005;102(29):10082-10087.
16. Varadarajan N, Gam J, Olsen MJ, Georgiou G, Iverson BL. Engineering of protease variants exhibiting high catalytic activity and exquisite substrate selectivity. Proc Natl Acad Sci U S A 2005;102(19):6855-6860.
17. Rui L, Cao L, Chen W, Reardon KF, Wood TK. Protein engineering of epoxide hydrolase from Agrobacterium radiobacter AD1 for enhanced activity and

enantioselective production of (R)-1-phenylethane-1,2-diol. Appl Environ Microbiol 2005;71(7):3995-4003.
18. Spiller B, Gershenson A, Arnold FH, Stevens RC. A structural view of evolutionary divergence. Proc Natl Acad Sci U S A 1999;96(22):12305-12310.
19. Crameri A, Raillard SA, Bermudez E, Stemmer WPC. DNA shuffling of a family of genes from diverse species accelerates directed evolution. Nature 1998;391(6664):288-291.
20. Beaudry AA, Joyce GF. Directed Evolution of an Rna Enzyme. Science 1992;257(5070):635-641.
21. Wang HH, Isaacs FJ, Carr PA, Sun ZZ, Xu G, Forest CR, Church GM. Programming cells by multiplex genome engineering and accelerated evolution. Nature 2009;460(7257):894-898.
22. Jiang L, Althoff EA, Clemente FR, Doyle L, Rothlisberger D, Zanghellini A, Gallaher JL, Betker JL, Tanaka F, Barbas CF, III, Hilvert D, Houk KN, Stoddard BL, Baker D. De Novo Computational Design of Retro-Aldol Enzymes. Science 2008;319(5868):1387-1391.
23. Allert M, Rizk SS, Looger LL, Hellinga HW. Computational design of receptors for an organophosphate surrogate of the nerve agent soman. Proc Natl Acad Sci U S A 2004;101(21):7907-7912.
24. Looger LL, Dwyer MA, Smith JJ, Hellinga HW. Computational design of receptor and sensor proteins with novel functions. Nature 2003;423(6936):185-190.
25. Yin H, Slusky JS, Berger BW, Walters RS, Vilaire G, Litvinov RI, Lear JD, Caputo GA, Bennett JS, DeGrado WF. Computational design of peptides that target transmembrane helices. Science 2007;315(5820):1817-1822.
26. Fazelinia H, Cirino PC, Maranas CD. OptGraft: A computational procedure for transferring a binding site onto an existing protein scaffold. Protein Sci 2009;18(1):180-195.
27. Lippow SM, Tidor B. Progress in computational protein design. Curr Opin Biotechnol 2007;18(4):305-311.
28. Dunbrack RL. Rotamer libraries in the 21(st) century. Current Opinion in Structural Biology 2002;12(4):431-440.
29. Dunbrack RL, Jr., Karplus M. Backbone-dependent rotamer library for proteins. Application to side-chain prediction. J Mol Biol 1993;230(2):543-574.
30. MacKerell J, A. D., Brooks, B., Brooks, III, C.L., Nilsson, L., Roux, B., Won, Y., and Karplus, M. CHARMM: The Energy Function and Its Parameterization with an Overview of the Program. In: al. PvRSe, editor. The Encyclopedia of Computational Chemistry. Volume 1: John Wiley & Sons: Chichester; 1998. p 271-277.
31. Mayo SL, Olafson BD, Goddard WA. DREIDING: a generic force field for molecular simulations. The Journal of Physical Chemistry 1990;94(26):8897-8909.
32. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. Journal of the American Chemical Society 1995;117(19):5179-5197.
33. Scott WRP, Hunenberger PH, Tironi IG, Mark AE, Billeter SR, Fennen J, Torda AE, Huber T, Kruger P, van Gunsteren WF. The GROMOS biomolecular simulation program package. Journal of Physical Chemistry A 1999;103(19):3596-3607.
34. Chiu TL, Goldstein RA. Optimizing potentials for the inverse protein folding problem. Protein Eng 1998;11(9):749-752.

35. Kuhlman B, Baker D. Native protein sequences are close to optimal for their structures. Proc Natl Acad Sci U S A 2000;97(19):10383-10388.
36. Looger LL, Hellinga HW. Generalized dead-end elimination algorithms make large-scale protein side-chain structure prediction tractable: implications for protein design and structural genomics. J Mol Biol 2001;307(1):429-445.
37. Moore GL. Modeling and optimization in directed evolution protocols and protein engineering; 2005.
38. Allen BD, Mayo SL. Dramatic performance enhancements for the FASTER optimization algorithm. J Comput Chem 2006;27(10):1071-1075.
39. Dantas G, Kuhlman B, Callender D, Wong M, Baker D. A large scale test of computational protein design: folding and stability of nine completely redesigned globular proteins. J Mol Biol 2003;332(2):449-460.
40. Kuhlman B, O'Neill JW, Kim DE, Zhang KY, Baker D. Accurate computer-based design of a new backbone conformation in the second turn of protein L. J Mol Biol 2002;315(3):471-477.
41. Johnson EC, Lazar GA, Desjarlais JR, Handel TM. Solution structure and dynamics of a designed hydrophobic core variant of ubiquitin. Structure 1999;7(8):967-976.
42. Desjarlais JR, Handel TM. De novo design of the hydrophobic cores of proteins. Protein Sci 1995;4(10):2006-2018.
43. Raha K, Wollacott AM, Italia MJ, Desjarlais JR. Prediction of amino acid sequence from structure. Protein Sci 2000;9(6):1106-1119.
44. Xu Z, Farid RS. Design, synthesis, and characterization of a novel hemoprotein. Protein Sci 2001;10(2):236-249.
45. Jiang X, Farid H, Pistor E, Farid RS. A new approach to the design of uniquely folded thermally stable proteins. Protein Sci 2000;9(2):403-416.
46. Desmet J, Demaeyer M, Hazes B, Lasters I. The Dead-End Elimination Theorem and Its Use in Protein Side-Chain Positioning. Nature 1992;356(6369):539-542.
47. Xie W, Sahinidis NV. Residue-rotamer-reduction algorithm for the protein side-chain conformation problem. Bioinformatics 2006;22(2):188-194.
48. Rosenberg M, Goldblum A. Computational protein design: a novel path to future protein drugs. Curr Pharm Des 2006;12(31):3973-3997.
49. Poole AM, Ranganathan R. Knowledge-based potentials in protein design. Curr Opin Struct Biol 2006;16(4):508-513.
50. Ambroggio XI, Kuhlman B. Design of protein conformational switches. Curr Opin Struct Biol 2006;16(4):525-530.
51. Koder RL, Dutton PL. Intelligent design: the de novo engineering of proteins with specified functions. Dalton Trans 2006(25):3045-3051.
52. Baker D. Prediction and design of macromolecular structures and interactions. Philos Trans R Soc Lond B Biol Sci 2006;361(1467):459-463.
53. Butterfoss GL, Kuhlman B. Computer-based design of novel protein structures. Annu Rev Biophys Biomol Struct 2006;35:49-65.
54. Vizcarra CL, Mayo SL. Electrostatics in computational protein design. Curr Opin Chem Biol 2005;9(6):622-626.
55. Chica RA, Doucet N, Pelletier JN. Semi-rational approaches to engineering enzyme activity: combining the benefits of directed evolution and rational design. Curr Opin Biotechnol 2005;16(4):378-384.
56. Bolon DN, Mayo SL. Enzyme-like proteins by computational design. Proc Natl Acad Sci U S A 2001;98(25):14274-14279.

57. Charbonneau D, Brown TM, Latham DW, Mayor M. Detection of Planetary Transits Across a Sun-like Star. Astrophys J 2000;529(1):L45-L48.
58. Darancet P, Wipf N, Berger C, de Heer WA, Mayou D. Quenching of the quantum Hall effect in multilayered epitaxial graphene: the role of undoped planes. Phys Rev Lett 2008;101(11):116806.
59. Marshall SA, Morgan CS, Mayo SL. Electrostatics significantly affect the stability of designed homeodomain variants. J Mol Biol 2002;316(1):189-199.
60. Mayo GL, Long LS, Fitzsimmons T, Scribbick FW, Carlquist S. Delayed orbital foreign body reaction to dicotyledon (hardwood) libriform fibers. Arch Ophthalmol 2002;120(12):1770-1771.
61. Mazeh T, Naef D, Torres G, Latham DW, Mayor M, Beuzit JL, Brown TM, Buchhave L, Burnet M, Carney BW, Charbonneau D, Drukier GA, Laird JB, Pepe F, Perrier C, Queloz D, Santos NC, Sivan JP, Udry S, Zucker S. The Spectroscopic Orbit of the Planetary Companion Transiting HD 209458. Astrophys J 2000;532(1):L55-L58.
62. Mooers BH, Datta D, Baase WA, Zollars ES, Mayo SL, Matthews BW. Repacking the Core of T4 lysozyme by automated design. J Mol Biol 2003;332(3):741-756.
63. Ross SA, Sarisky CA, Su A, Mayo SL. Designed protein G core variants fold to native-like structures: sequence selection by ORBIT tolerates variation in backbone specification. Protein Sci 2001;10(2):450-454.
64. Ruben M, Landa A, Lortscher E, Riel H, Mayor M, Gorls H, Weber HB, Arnold A, Evers F. Charge transport through a cardan-joint molecule. Small 2008;4(12):2229-2235.
65. Sarisky CA, Mayo SL. The beta-beta-alpha fold: explorations in sequence space. J Mol Biol 2001;307(5):1411-1418.
66. Shifman JM, Mayo SL. Modulating calmodulin binding specificity through computational protein design. J Mol Biol 2002;323(3):417-423.
67. Raman S, Lange OF, Rossi P, Tyka M, Wang X, Aramini J, Liu G, Ramelot T, Eletsky A, Szyperski T, Kennedy M, Prestegard J, Montelione GT, Baker D. NMR Structure Determination for Larger Proteins Using Backbone-Only Data. Science.
68. Blum B, Jordan MI, Baker D. Feature space resampling for protein conformational search. Proteins 2009.
69. Raman S, Huang YJ, Mao B, Rossi P, Aramini JM, Liu G, Montelione GT, Baker D. Accurate automated protein NMR structure determination using unassigned NOESY data. J Am Chem Soc;132(1):202-207.
70. Shen Y, Bryan PN, He Y, Orban J, Baker D, Bax A. De novo structure generation using chemical shifts for proteins with high-sequence identity but different folds. Protein Sci;19(2):349-356.
71. Das R, Andre I, Shen Y, Wu Y, Lemak A, Bansal S, Arrowsmith CH, Szyperski T, Baker D. Simultaneous prediction of protein folding and docking at high resolution. Proc Natl Acad Sci U S A 2009;106(45):18978-18983.
72. Krieger E, Joo K, Lee J, Raman S, Thompson J, Tyka M, Baker D, Karplus K. Improving physical realism, stereochemistry, and side-chain accuracy in homology modeling: Four approaches that performed well in CASP8. Proteins 2009;77 Suppl 9:114-122.
73. Kidd BA, Baker D, Thomas WE. Computation of conformational coupling in allosteric proteins. PLoS Comput Biol 2009;5(8):e1000484.
74. Raman S, Vernon R, Thompson J, Tyka M, Sadreyev R, Pei J, Kim D, Kellogg E, DiMaio F, Lange O, Kinch L, Sheffler W, Kim BH, Das R, Grishin NV, Baker D. Structure prediction for CASP8 with all-atom refinement using Rosetta. Proteins 2009;77 Suppl 9:89-99.

75. Kim DE, Blum B, Bradley P, Baker D. Sampling bottlenecks in de novo protein structure prediction. J Mol Biol 2009;393(1):249-260.
76. DiMaio F, Tyka MD, Baker ML, Chiu W, Baker D. Refinement of protein structures into low-resolution density maps using rosetta. J Mol Biol 2009;392(1):181-190.
77. Davis IW, Raha K, Head MS, Baker D. Blind docking of pharmaceutically relevant compounds using RosettaLigand. Protein Sci 2009;18(9):1998-2002.
78. Das R, Baker D. Prospects for de novo phasing with de novo protein models. Acta Crystallogr D Biol Crystallogr 2009;65(Pt 2):169-175.
79. Davis IW, Baker D. RosettaLigand docking with full ligand and receptor flexibility. J Mol Biol 2009;385(2):381-392.
80. Shen Y, Vernon R, Baker D, Bax A. De novo protein structure generation from incomplete chemical shift assignments. J Biomol NMR 2009;43(2):63-78.
81. Ramelot TA, Raman S, Kuzin AP, Xiao R, Ma LC, Acton TB, Hunt JF, Montelione GT, Baker D, Kennedy MA. Improving NMR protein structure quality by Rosetta refinement: a molecular replacement study. Proteins 2009;75(1):147-167.
82. Cho HD, Sood VD, Baker D, Weiner AM. On the role of a conserved, potentially helix-breaking residue in the tRNA-binding alpha-helix of archaeal CCA-adding enzymes. RNA 2008;14(7):1284-1289.
83. Keeble AH, Joachimiak LA, Mate MJ, Meenan N, Kirkpatrick N, Baker D, Kleanthous C. Experimental and computational analyses of the energetic basis for dual recognition of immunity proteins by colicin endonucleases. J Mol Biol 2008;379(4):745-759.
84. Das R, Baker D. Macromolecular modeling with rosetta. Annu Rev Biochem 2008;77:363-382.
85. Shen Y, Lange O, Delaglio F, Rossi P, Aramini JM, Liu G, Eletsky A, Wu Y, Singarapu KK, Lemak A, Ignatchenko A, Arrowsmith CH, Szyperski T, Montelione GT, Baker D, Bax A. Consistent blind protein structure generation from NMR chemical shift data. Proc Natl Acad Sci U S A 2008;105(12):4685-4690.
86. Das R, Qian B, Raman S, Vernon R, Thompson J, Bradley P, Khare S, Tyka MD, Bhat D, Chivian D, Kim DE, Sheffler WH, Malmstrom L, Wollacott AM, Wang C, Andre I, Baker D. Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home. Proteins 2007;69 Suppl 8:118-128.
87. Wang C, Bradley P, Baker D. Protein-protein docking with backbone flexibility. J Mol Biol 2007;373(2):503-519.
88. Das R, Baker D. Automated de novo prediction of native-like RNA tertiary structures. Proc Natl Acad Sci U S A 2007;104(37):14664-14669.
89. Malmstrom L, Riffle M, Strauss CE, Chivian D, Davis TN, Bonneau R, Baker D. Superfamily assignments for the yeast proteome through integration of structure prediction with the gene ontology. PLoS Biol 2007;5(4):e76.
90. Wollacott AM, Zanghellini A, Murphy P, Baker D. Prediction of structures of multidomain proteins from structures of the individual domains. Protein Sci 2007;16(2):165-175.
91. Yarov-Yarovoy V, Baker D, Catterall WA. Voltage sensor conformations in the open and closed states in ROSETTA structural models of K(+) channels. Proc Natl Acad Sci U S A 2006;103(19):7292-7297.
92. Misura KM, Chivian D, Rohl CA, Kim DE, Baker D. Physically realistic homology models built with ROSETTA can be more accurate than their templates. Proc Natl Acad Sci U S A 2006;103(14):5361-5366.
93. Yarov-Yarovoy V, Schonbrun J, Baker D. Multipass membrane protein structure prediction using Rosetta. Proteins 2006;62(4):1010-1025.

94. Lacy DB, Lin HC, Melnyk RA, Schueler-Furman O, Reither L, Cunningham K, Baker D, Collier RJ. A model of anthrax toxin lethal factor bound to protective antigen. Proc Natl Acad Sci U S A 2005;102(45):16409-16414.
95. Kim DE, Chivian D, Malmstrom L, Baker D. Automated prediction of domain boundaries in CASP6 targets using Ginzu and RosettaDOM. Proteins 2005;61 Suppl 7:193-200.
96. Chivian D, Kim DE, Malmstrom L, Schonbrun J, Rohl CA, Baker D. Prediction of CASP6 structures using automated Robetta protocols. Proteins 2005;61 Suppl 7:157-166.
97. Bradley P, Malmstrom L, Qian B, Schonbrun J, Chivian D, Kim DE, Meiler J, Misura KM, Baker D. Free modeling with Rosetta in CASP6. Proteins 2005;61 Suppl 7:128-134.
98. Misura KM, Morozov AV, Baker D. Analysis of anisotropic side-chain packing in proteins and application to high-resolution structure prediction. J Mol Biol 2004;342(2):651-664.
99. Kim DE, Chivian D, Baker D. Protein structure prediction and analysis using the Robetta server. Nucleic Acids Res 2004;32(Web Server issue):W526-531.
100. Rohl CA, Strauss CE, Chivian D, Baker D. Modeling structurally variable regions in homologous proteins with rosetta. Proteins 2004;55(3):656-677.
101. Rohl CA, Strauss CE, Misura KM, Baker D. Protein structure prediction using Rosetta. Methods Enzymol 2004;383:66-93.
102. Kuhn M, Meiler J, Baker D. Strand-loop-strand motifs: prediction of hairpins and diverging turns in proteins. Proteins 2004;54(2):282-288.
103. Meiler J, Baker D. Rapid protein fold determination using unassigned NMR data. Proc Natl Acad Sci U S A 2003;100(26):15404-15409.
104. Chivian D, Kim DE, Malmstrom L, Bradley P, Robertson T, Murphy P, Strauss CE, Bonneau R, Rohl CA, Baker D. Automated prediction of CASP-5 structures using the Robetta server. Proteins 2003;53 Suppl 6:524-533.
105. Bradley P, Chivian D, Meiler J, Misura KM, Rohl CA, Schief WR, Wedemeyer WJ, Schueler-Furman O, Murphy P, Schonbrun J, Strauss CE, Baker D. Rosetta predictions in CASP5: successes, failures, and prospects for complete automation. Proteins 2003;53 Suppl 6:457-468.
106. Meiler J, Baker D. Coupled prediction of protein secondary and tertiary structure. Proc Natl Acad Sci U S A 2003;100(21):12105-12110.
107. Wedemeyer WJ, Baker D. Efficient minimization of angle-dependent potentials for polypeptides in internal coordinates. Proteins 2003;53(2):262-272.
108. Sadreyev RI, Baker D, Grishin NV. Profile-profile comparisons by COMPASS predict intricate homologies between protein families. Protein Sci 2003;12(10):2262-2272.
109. Tsai J, Bonneau R, Morozov AV, Kuhlman B, Rohl CA, Baker D. An improved protein decoy set for testing energy functions for protein structure prediction. Proteins 2003;53(1):76-87.
110. Kinch LN, Baker D, Grishin NV. Deciphering a novel thioredoxin-like fold family. Proteins 2003;52(3):323-331.
111. Schueler-Furman O, Baker D. Conserved residue clustering and protein structure prediction. Proteins 2003;52(2):225-235.
112. Saunders CT, Baker D. Evaluation of structural and evolutionary contributions to deleterious mutation prediction. J Mol Biol 2002;322(4):891-901.
113. Bonneau R, Strauss CE, Rohl CA, Chivian D, Bradley P, Malmstrom L, Robertson T, Baker D. De novo prediction of three-dimensional structures for major protein families. J Mol Biol 2002;322(1):65-78.

114. Bonneau R, Ruczinski I, Tsai J, Baker D. Contact order and ab initio protein structure prediction. Protein Sci 2002;11(8):1937-1944.
115. Ruczinski I, Kooperberg C, Bonneau R, Baker D. Distributions of beta sheets in proteins with application to structure prediction. Proteins 2002;48(1):85-97.
116. Rohl CA, Baker D. De novo determination of protein backbone structure from residual dipolar couplings using Rosetta. J Am Chem Soc 2002;124(11):2723-2729.
117. Fazelinia H, Cirino PC, Maranas CD. Extending Iterative Protein Redesign and Optimization (IPRO) in protein library design for ligand specificity. Biophys J 2007;92(6):2120-2130.
118. Khoury GA, Fazelinia H, Chin JW, Pantazes RJ, Cirino PC, Maranas CD. Computational design of Candida boidinii xylose reductase for altered cofactor specificity. Protein Sci 2009;18(10):2125-2138.
119. Saraf MC, Moore GL, Goodey NM, Cao VY, Benkovic SJ, Maranas CD. IPRO: an iterative computational protein library redesign and optimization procedure. Biophys J 2006;90(11):4167-4180.
120. Dunbrack RL, Jr., Cohen FE. Bayesian statistical analysis of protein side-chain rotamer preferences. Protein Sci 1997;6(8):1661-1681.
121. Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, Baker D. Design of a novel globular protein fold with atomic-level accuracy. Science 2003;302(5649):1364-1368.
122. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. Charmm - a Program for Macromolecular Energy, Minimization, and Dynamics Calculations. J Comput Chem 1983;4(2):187-217.
123. Brooks BR, Bruccoleri RE, Olafson DJ, States DJ, Swaminathan S, Karplus M. CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. J Comput Chem 1983;4:187-217.
124. Im W, Lee MS, Brooks CL, 3rd. Generalized born model with a simple smoothing function. J Comput Chem 2003;24(14):1691-1702.
125. Born M. Volumes and hydration warmth of ions. Zeitschrift Fur Physik 1920;1:45-48.
126. Lee MS, Salsbury FR, Brooks CL. Novel generalized Born methods. Journal of Chemical Physics 2002;116(24):10606-10614.
127. Born M. Der aufbau der materie; dreis aufsätze über moderne atomistik und elektronentheorie. Berlin,: J. Springer; 1920. 3 p. \2113., 2181 p. p.
128. Lazaridis T, Karplus M. Effective energy functions for protein structure prediction. Curr Opin Struct Biol 2000;10(2):139-145.
129. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28(1):235-242.
130. Kortemme T, Baker D. Computational design of protein-protein interactions. Curr Opin Chem Biol 2004;8(1):91-97.
131. Ashworth J, Havranek JJ, Duarte CM, Sussman D, Monnat RJ, Jr., Stoddard BL, Baker D. Computational redesign of endonuclease DNA binding and cleavage specificity. Nature 2006;441(7093):656-659.
132. Dahiyat BI, Mayo SL. De novo protein design: Fully automated sequence selection. Science 1997;278(5335):82-87.
133. Jiang L, Althoff EA, Clemente FR, Doyle L, Rothlisberger D, Zanghellini A, Gallaher JL, Betker JL, Tanaka F, Barbas CF, 3rd, Hilvert D, Houk KN, Stoddard BL, Baker D. De novo computational design of retro-aldol enzymes. Science 2008;319(5868):1387-1391.

134. Rothlisberger D, Khersonsky O, Wollacott AM, Jiang L, DeChancie J, Betker J, Gallaher JL, Althoff EA, Zanghellini A, Dym O, Albeck S, Houk KN, Tawfik DS, Baker D. Kemp elimination catalysts by computational enzyme design. Nature 2008;453(7192):190-195.
135. Zanghellini A, Jiang L, Wollacott AM, Cheng G, Meiler J, Althoff EA, Rothlisberger D, Baker D. New algorithms and an in silico benchmark for computational enzyme design. Protein Sci 2006;15(12):2785-2794.
136. Marohnic CC, Bewley MC, Barber MJ. Engineering and characterization of a NADPH-utilizing cytochrome b5 reductase. Biochemistry 2003;42(38):11170-11182.
137. Woodyer R, van der Donk WA, Zhao H. Relaxing the nicotinamide cofactor specificity of phosphite dehydrogenase by rational design. Biochemistry 2003;42(40):11604-11614.
138. Watanabe S, Kodaki T, Makino K. Complete reversal of coenzyme specificity of xylitol dehydrogenase and increase of thermostability by the introduction of structural zinc. J Biol Chem 2005;280(11):10340-10349.
139. Kostrzynska M, Sopher CR, Lee H. Mutational analysis of the role of the conserved lysine-270 in the Pichia stipitis xylose reductase. FEMS Microbiology Letters 1998;159(1):107-112.
140. Leitgeb S, Petschacher B, Wilson DK, Nidetzky B. Fine tuning of coenzyme specificity in family 2 aldo-keto reductases revealed by crystal structures of the Lys-274-->Arg mutant of Candida tenuis xylose reductase (AKR2B5) bound to NAD+ and NADP+. FEBS Lett 2005;579(3):763-767.
141. Petschacher B, Leitgeb S, Kavanagh KL, Wilson DK, Nidetzky B. The coenzyme specificity of Candida tenuis xylose reductase (AKR2B5) explored by site-directed mutagenesis and X-ray crystallography. Biochem J 2005;385(Pt 1):75-83.
142. Liang L, Zhang J, Lin Z. Altering coenzyme specificity of Pichia stipitis xylose reductase by the semi-rational approach CASTing. Microb Cell Fact 2007;6:36.
143. Banta S, Swanson BA, Wu S, Jarnagin A, Anderson S. Alteration of the specificity of the cofactor-binding pocket of Corynebacterium 2,5-diketo-D-gluconic acid reductase A. Protein Eng 2002;15(2):131-140.
144. Banta S, Swanson BA, Wu S, Jarnagin A, Anderson S. Optimizing an artificial metabolic pathway: engineering the cofactor specificity of Corynebacterium 2,5-diketo-D-gluconic acid reductase for use in vitamin C biosynthesis. Biochemistry 2002;41(20):6226-6236.
145. Scrutton NS, Berry A, Perham RN. Redesign of the coenzyme specificity of a dehydrogenase by protein engineering. Nature 1990;343(6253):38-43.
146. Rane MJ, Calvo KC. Reversal of the nucleotide specificity of ketol acid reductoisomerase by site-directed mutagenesis identifies the NADPH binding site. Arch Biochem Biophys 1997;338(1):83-89.
147. Shiraishi N, Croy C, Kaur J, Campbell WH. Engineering of pyridine nucleotide specificity of nitrate reductase: mutagenesis of recombinant cytochrome b reductase fragment of Neurospora crassa NADPH:Nitrate reductase. Arch Biochem Biophys 1998;358(1):104-115.
148. Eppink MH, Overkamp KM, Schreuder HA, Van Berkel WJ. Switch of coenzyme specificity of p-hydroxybenzoate hydroxylase. J Mol Biol 1999;292(1):87-96.
149. Elmore CL, Porter TD. Modification of the nucleotide cofactor-binding site of cytochrome P-450 reductase to enhance turnover with NADH in Vivo. J Biol Chem 2002;277(50):48960-48964.
150. Kristan K, Stojan J, Adamski J, Rizner TL. Rational design of novel mutants of fungal 17 beta-hydroxy steroid dehydrogenase. J Biotechnol 2007;129(1):123-130.
151. Dambe TR, Kuhn AM, Brossette T, Giffhorn F, Scheidig AJ. Crystal structure of NADP(H)-dependent 1,5-anhydro-D-fructose reductase from Sinorhizobium morelense

at 2.2 A resolution: construction of a NADH-accepting mutant and its application in rare sugar synthesis. Biochemistry 2006;45(33):10030-10042.
152. Medina M, Luquita A, Tejero J, Hermoso J, Mayoral T, Sanz-Aparicio J, Grever K, Gomez-Moreno C. Probing the determinants of coenzyme specificity in ferredoxin-NADP+ reductase by site-directed mutagenesis. J Biol Chem 2001;276(15):11902-11912.
153. Chen R, Greer A, Dean AM. A highly active decarboxylating dehydrogenase with rationally inverted coenzyme specificity. Proc Natl Acad Sci U S A 1995;92(25):11666-11670.
154. Yaoi T, Miyazaki K, Oshima T, Komukai Y, Go M. Conversion of the coenzyme specificity of isocitrate dehydrogenase by module replacement. J Biochem (Tokyo) 1996;119(5):1014-1018.
155. Zhang L, Ahvazi B, Szittner R, Vrielink A, Meighen E. Change of nucleotide specificity and enhancement of catalytic efficiency in single point mutants of Vibrio harveyi aldehyde dehydrogenase. Biochemistry 1999;38(35):11440-11447.
156. Holmberg N, Ryde U, Bulow L. Redesign of the coenzyme specificity in L-lactate dehydrogenase from bacillus stearothermophilus using site-directed mutagenesis and media engineering. Protein Eng 1999;12(10):851-856.
157. Barber MJ. Altered pyridine nucleotide specificity of spinach nitrate reductase. Faseb Journal 2000;14(8):A1416-A1416.
158. Miller SP, Lunzer M, Dean AM. Direct demonstration of an adaptive constraint. Science 2006;314(5798):458-461.
159. Bernard N, Johnsen K, Holbrook JJ, Delcour J. D175 discriminates between NADH and NADPH in the coenzyme binding site of Lactobacillus delbrueckii subsp. Bulgaricus D-lactate dehydrogenase. Biochem Biophys Res Commun 1995;208(3):895-900.
160. Clermont S, Corbier C, Mely Y, Gerard D, Wonacott A, Branlant G. Determinants of coenzyme specificity in glyceraldehyde-3-phosphate dehydrogenase: role of the acidic residue in the fingerprint region of the nucleotide binding fold. Biochemistry 1993;32(38):10178-10184.
161. Ehrensberger AH, Elling RA, Wilson DK. Structure-guided engineering of xylitol dehydrogenase cosubstrate specificity. Structure 2006;14(3):567-575.
162. Hsieh JY, Liu GY, Chang GG, Hung HC. Determinants of the dual cofactor specificity and substrate cooperativity of the human mitochondrial NAD(P)(+)-dependent malic enzyme - Functional roles of glutamine 362. J Biol Chem 2006;281(32):23237-23245.
163. Serov AE, Popova AS, Fedorchuk VV, Tishkov VI. Engineering of coenzyme specificity of formate dehydrogenase from Saccharomyces cerevisiae. Biochem J 2002;367(Pt 3):841-847.
164. Chen R, Greer A, Dean AM. Redesigning secondary structure to invert coenzyme specificity in isopropylmalate dehydrogenase. Proc Natl Acad Sci U S A 1996;93(22):12171-12176.
165. Galkin A, Kulakova L, Ohshima T, Esaki N, Soda K. Construction of a new leucine dehydrogenase with preferred specificity for NADP+ by site-directed mutagenesis of the strictly NAD+-specific enzyme. Protein Eng 1997;10(6):687-690.
166. Kavanagh KL, Klimacek M, Nidetzky B, Wilson DK. Structure of xylose reductase bound to NAD+ and the basis for single and dual co-substrate specificity in family 2 aldo-keto reductases. Biochem J 2003;373(Pt 2):319-326.
167. Jez JM, Penning TM. The aldo-keto reductase (AKR) superfamily: an update. Chem Biol Interact 2001;130-132(1-3):499-525.

168. Kavanagh KL, Klimacek M, Nidetzky B, Wilson DK. The structure of apo and holo forms of xylose reductase, a dimeric aldo-keto reductase from Candida tenuis. Biochemistry 2002;41(28):8785-8795.
169. Regina Kratzer DKWBN. Catalytic mechanism and substrate selectivity of aldo-keto reductases: Insights from structure-function studies of Candida tenuis xylose reductase. IUBMB Life 2006;58(9):499-507.
170. Bohren KM, Grimshaw CE, Lai CJ, Harrison DH, Ringe D, Petsko GA, Gabbay KH. Tyrosine-48 is the proton donor and histidine-110 directs substrate stereochemical selectivity in the reduction reaction of human aldose reductase: enzyme kinetics and crystal structure of the Y48H mutant enzyme. Biochemistry 1994;33(8):2021-2032.
171. Barski OA, Gabbay KH, Grimshaw CE, Bohren KM. Mechanism of human aldehyde reductase: characterization of the active site pocket. Biochemistry 1995;34(35):11264-11275.
172. Schlegel BP, Jez JM, Penning TM. Mutagenesis of 3 alpha-hydroxysteroid dehydrogenase reveals a "push-pull" mechanism for proton transfer in aldo-keto reductases. Biochemistry 1998;37(10):3538-3548.
173. Lee JK, Koo BS, Kim SY. Cloning and characterization of the xyl1 gene, encoding an NADH-preferring xylose reductase from Candida parapsilosis, and its functional expression in Candida tropicalis. Appl Environ Microbiol 2003;69(10):6179-6188.
174. Werpy T. PG, Aden A., Bozell J., Holladay J., White J., Manheim A. Top value added chemicals from biomass, Volume I: Results of
screening for potential candidates from sugars and synthesis gas. U S Department of Energy 2004.
175. Cirino PC, Chin JW, Ingram LO. Engineering Escherichia coli for xylitol production from glucose-xylose mixtures. Biotechnol Bioeng 2006;95(6):1167-1176.
176. Chin JW, Khankal R, Monroe CA, Maranas CD, Cirino PC. Analysis of NADPH supply during xylitol production by engineered Escherichia coli. Biotechnol Bioeng 2009;102(1):209-220.
177. Chenault HK, Whitesides GM. Regeneration of nicotinamide cofactors for use in organic synthesis. Appl Biochem Biotechnol 1987;14(2):147-197.
178. Tracewell CA, Arnold FH. Directed enzyme evolution: climbing fitness peaks one amino acid at a time. Curr Opin Chem Biol 2009;13(1):3-9.
179. Shindyalov IN, Bourne PE. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng 1998;11(9):739-747.
180. Dunbrack RL, Jr., Karplus M. Conformational analysis of the backbone-dependent rotamer preferences of protein sidechains. Nat Struct Biol 1994;1(5):334-340.
181. Chen R, Li L, Weng Z. ZDOCK: an initial-stage protein-docking algorithm. Proteins 2003;52(1):80-87.
182. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equation of State Calculations by Fast Computing Machines. J Chem Phys 1953;21(6):1087-1092.
183. Hacker B, Habenicht A, Kiess M, Mattes R. Xylose utilisation: cloning and characterisation of the Xylose reductase from Candida tenuis. Biol Chem 1999;380(12):1395-1403.
184. Kang MH, Ni H, Jeffries TW. Molecular characterization of a gene for aldose reductase (CbXYL1) from Candida boidinii and its expression in Saccharomyces cerevisiae. Appl Biochem Biotechnol 2003;105 -108:265-276.
185. An YF, Ji JF, Wu WF, Lv A, Huang RB, Wei YT. A rapid and efficient method for multiple-site mutagenesis with a modified overlap extension PCR. Appl Microbiol Biotechnol 2005;68(6):774-778.

186. Klein P, Kanehisa M, DeLisi C. Prediction of protein function from sequence properties. Discriminant analysis of a data base. Biochim Biophys Acta 1984;787(3):221-226.
187. Cid H, Bunster M, Canales M, Gazitua F. Hydrophobicity and structural classes in proteins. Protein Eng 1992;5(5):373-375.
188. Krigbaum WR, Komoriya A. Local interactions as a structure determinant for protein molecules: II. Biochim Biophys Acta 1979;576(1):204-248.
189. Arnold K, Bordoli L, Kopp J, Schwede T. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. Bioinformatics 2006;22(2):195-201.
190. Schwede T, Kopp J, Guex N, Peitsch MC. SWISS-MODEL: An automated protein homology-modeling server. Nucleic Acids Res 2003;31(13):3381-3385.
191. Petschacher B, Nidetzky B. Engineering Candida tenuis Xylose reductase for improved utilization of NADH: antagonistic effects of multiple side chain replacements and performance of site-directed mutants under simulated in vivo conditions. Appl Environ Microbiol 2005;71(10):6390-6393.
192. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A 1992;89(22):10915-10919.
193. Jonson PH, Petersen SB. A critical view on conservative mutations. Protein Eng 2001;14(6):397-402.
194. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. Charmm - a Program for Macromolecular Energy, Minimization, and Dynamics Calculations. Journal of Computational Chemistry 1983;4(2):187-217.
195. Carugo O, Argos P. NADP-dependent enzymes. I: Conserved stereochemistry of cofactor binding. Proteins 1997;28(1):10-28.
196. Lee MS, Salsbury FR, Brooks CL. Novel generalized Born methods. J Chem Phys 2002;116(24):10606-10614.
197. Alvizo O, Allen BD, Mayo SL. Computational protein design promises to revolutionize protein engineering. Biotechniques 2007;42(1):31-+.
198. Dwyer MA, Looger LL, Hellinga HW. Computational design of a Zn2+ receptor that controls bacterial gene expression. Proc Natl Acad Sci U S A 2003;100(20):11255-11260.
199. Kaplan J, DeGrado WF. De novo design of catalytic proteins. Proceedings of the National Academy of Sciences of the United States of America 2004;101(32):11566-11570.
200. Korkegian A, Black ME, Baker D, Stoddard BL. Computational thermostabilization of an enzyme. Science 2005;308(5723):857-860.
201. Choi EJ, Mao J, Mayo SL. Computational design and biochemical characterization of maize nonspecific lipid transfer protein variants for biosensor applications. Protein Science 2007;16(4):582-588.
202. Shah PS, Hom GK, Ross SA, Lassila JK, Crowhurst KA, Mayo SL. Full-sequence computational design and solution structure of a thermostable protein variant. Journal of Molecular Biology 2007;372(1):1-6.
203. Gribenko AV, Patel MM, Liu J, McCallum SA, Wang C, Makhatadze GI. Rational stabilization of enzymes by computational redesign of surface charge-charge interactions. Proceedings of the National Academy of Sciences 2009;106(8):2601-2606.
204. Meinhold P, Peters, M. W., Chen, M. M. Y., Takahashi, K., Arnold, F. H. Direct Conversion of Ethane to Ethanol by Engineered Cytochrome P450 BM3. ChemBioChem 2005;6(10):1765-1768.

205. Urlacher VB, Schmid RD. Protein engineering of the cytochrome P450 monooxygenase from Bacillus megaterium. Protein Engineering 2004;388:208-224.
206. Yun CH, Kim KH, Kim DH, Jung HC, Pan JG. The bacterial P450BM3: a prototype for a biocatalyst with human P450 activities. Trends in Biotechnology 2007;25(7):289-298.
207. Shaik S, Kumar D, de Visser SP, Altun A, Thiel W. Theoretical perspective on the structure and mechanism of cytochrome P450 enzymes. Chemical Reviews 2005;105(6):2279-2328.
208. Groves JT. The bioinorganic chemistry of iron in oxygenases and supramolecular assemblies. Proceedings of the National Academy of Sciences of the United States of America 2003;100(7):3569-3574.
209. Leak DJ, Sheldon RA, Woodley JM, Adlercreutz P. Biocatalysts for selective introduction of oxygen. Biocatalysis and Biotransformation 2009;27(1):1-26.
210. Cirino PC, Arnold FH. A self-sufficient peroxide-driven hydroxylation biocatalyst. Angew Chem Int Ed Engl 2003;42(28):3299-3301.
211. Fasan R, Meharenna YT, Snow CD, Poulos TL, Arnold FH. Evolutionary history of a specialized p450 propane monooxygenase. J Mol Biol 2008;383(5):1069-1080.
212. Fasan R, Chen MM, Crook NC, Arnold FH. Engineered alkane-hydroxylating cytochrome P450(BM3) exhibiting nativelike catalytic properties. Angewandte Chemie-International Edition 2007;46(44):8414-8418.
213. Li H, Poulos TL. The structure of the cytochrome p450BM-3 haem domain complexed with the fatty acid substrate, palmitoleic acid. Nat Struct Biol 1997;4(2):140-146.
214. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Montgomery J, J. A., Vreven T, Kudin KN, Burant JC, Millam JM, Iyengar SS, Tomasi J, Barone V, Mennucci B, Cossi M, Scalmani G, Rega N, Petersson GA, Nakatsuji H, Hada M, Ehara M, Toyota K, Fukuda R, Hasegawa J, Ishida M, Nakajima T, Honda Y, Kitao O, Nakai H, Klene M, Li X, Knox JE, Hratchian HP, Cross JB, Bakken V, Adamo C, Jaramillo J, Gomperts R, Stratmann RE, Yazyev O, Austin AJ, Cammi R, Pomelli C, Ochterski JW, Ayala PY, Morokuma K, Voth GA, Salvador P, Dannenberg JJ, Zakrzewski VG, Dapprich S, Daniels AD, Strain MC, Farkas O, Malick DK, Rabuck AD, Raghavachari K, Foresman JB, Ortiz JV, Cui Q, Baboul AG, Clifford S, Cioslowski J, Stefanov BB, Liu G, Liashenko A, Piskorz P, Komaromi I, Martin RL, Fox DJ, Keith T, Al-Laham MA, Peng CY, Nanayakkara A, Challacombe M, Gill PMW, Johnson B, Chen W, Wong MW, Gonzalez C, Pople JA. Gaussian 03, Revision C.02,. Revision C.02. Wallingford CT: Gaussian, Inc.; 2004.
215. Becke AD. Density-functional thermochemistry. III. The role of exact exchange. J Chem Phys 1993;98(7):5648-5652.
216. Stephens PJ, Devlin FJ, Chabalowski CF, Frisch MJ. Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields. The Journal of Physical Chemistry 1994;98(45):11623-11627.
217. Shannon CE. Prediction and entropy of printed English. The Bell System Technical Journal 1951;30:50-64.
218. Paramesvaran J, Hibbert EG, Russell AJ, Dalby PA. Distributions of enzyme residues yielding mutants with improved substrate specificities from two different directed evolution strategies. Protein Eng Des Sel 2009;22(7):401-411.
219. Khoury GA, Fazelinia H, Chin JW, Pantazes RJ, Cirino PC, Maranas CD. Computational design of candida boidinii xylose reductase for altered cofactor specificity. Protein Science 2009;9999(999A):NA.

220. Gekko K, Yamagami K, Kunori Y, Ichihara S, Kodama M, Iwakura M. Effects of point mutation in a flexible loop on the stability and enzymatic function of Escherichia coli dihydrofolate reductase. J Biochem 1993;113(1):74-80.
221. Ohmae E, Iriyama K, Ichihara S, Gekko K. Nonadditive effects of double mutations at the flexible loops, glycine-67 and glycine-121, of Escherichia coli dihydrofolate reductase on its stability and function. J Biochem 1998;123(1):33-41.
222. Ohmae E, Ishimura K, Iwakura M, Gekko K. Effects of point mutations at the flexible loop alanine-145 of Escherichia coli dihydrofolate reductase on its stability and function. J Biochem 1998;123(5):839-846.
223. Ohmae E, Sasaki Y, Gekko K. Effects of five-tryptophan mutations on structure, stability and function of Escherichia coli dihydrofolate reductase. J Biochem 2001;130(3):439-447.
224. Posner BA, Li L, Bethell R, Tsuji T, Benkovic SJ. Engineering specificity for folate into dihydrofolate reductase from Escherichia coli. Biochemistry 1996;35(5):1653-1663.
225. Rajagopalan PT, Zhang Z, McCourt L, Dwyer M, Benkovic SJ, Hammes GG. Interaction of dihydrofolate reductase with methotrexate: ensemble and single-molecule kinetics. Proc Natl Acad Sci U S A 2002;99(21):13481-13486.
226. Kollman PA, Massova I, Reyes C, Kuhn B, Huo SH, Chong L, Lee M, Lee T, Duan Y, Wang W, Donini O, Cieplak P, Srinivasan J, Case DA, Cheatham TE. Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models. Accounts of Chemical Research 2000;33(12):889-897.
227. Friesner RA, Guallar V. Ab initio quantum chemical and mixed quantum mechanics/molecular mechanics (QM/MM) methods for studying enzymatic catalysis. Annual Review of Physical Chemistry 2005;56:389-427.
228. Hu H, Yang WT. Free energies of chemical reactions in solution and in enzymes with ab initio quantum mechanics/molecular mechanics methods. Annual Review of Physical Chemistry 2008;59:573-601.
229. Senn HM, Thiel W. QM/MM studies of enzymes. Current Opinion in Chemical Biology 2007;11(2):182-187.
230. Warshel A. Computer simulations of enzyme catalysis: Methods, progress, and insights. Annual Review of Biophysics and Biomolecular Structure 2003;32:425-443.
231. Hans Martin Senn WT. QM/MM Methods for Biomolecular Systems. Angewandte Chemie International Edition 2009;48(7):1198-1229.
232. Alexandrova AN, RoÌˆthlisberger D, Baker D, Jorgensen WL. Catalytic Mechanism and Performance of Computationally Designed Enzymes for Kemp Elimination. Journal of the American Chemical Society 2008;130(47):15907-15915.
233. Kong J, White CA, Krylov AI, Sherrill D, Adamson RD, Furlani TR, Lee MS, Lee AM, Gwaltney SR, Adams TR, Ochsenfeld C, Gilbert ATB, Kedziora GS, Rassolov VA, Maurice DR, Nair N, Shao YH, Besley NA, Maslen PE, Dombroski JP, Daschel H, Zhang WM, Korambath PP, Baker J, Byrd EFC, Van Voorhis T, Oumi M, Hirata S, Hsu CP, Ishikawa N, Florian J, Warshel A, Johnson BG, Gill PMW, Head-Gordon M, Pople JA. Q-chem 2.0: A high-performance ab initio electronic structure program package. Journal of Computational Chemistry 2000;21(16):1532-1548.
234. Xie L, Liu HY, Yang WT. Adapting the nudged elastic band method for determining minimum-energy paths of chemical reactions in enzymes. Journal of Chemical Physics 2004;120(17):8039-8052.
235. Riccardi D, Schaefer P, Yang Y, Yu HB, Ghosh N, Prat-Resina X, Konig P, Li GH, Xu DG, Guo H, Elstner M, Cui Q. Development of effective quantum mechanical/molecular

mechanical (QM/MM) methods for complex biological processes. Journal of Physical Chemistry B 2006;110(13):6458-6469.
236. Gekko K, Kamiyama T, Ohmae E, Katayanagi K. Single amino acid substitutions in flexible loops can induce large compressibility changes in dihydrofolate reductase. J Biochem 2000;128(1):21-27.
237. Zolli-Juran M, Cechetto JD, Hartlen R, Daigle DM, Brown ED. High throughput screening identifies novel inhibitors of Escherichia coli dihydrofolate reductase that are competitive with dihydrofolate. Bioorg Med Chem Lett 2003;13(15):2493-2496.
238. Peters MW, Meinhold P, Glieder A, Arnold FH. Regio- and enantioselective alkane hydroxylation with engineered cytochromes P450 BM-3. Journal of the American Chemical Society 2003;125(44):13442-13450.
239. Salazar O, Cirino PC, Arnold FH. Thermostabilization of a cytochrome p450 peroxygenase. Chembiochem 2003;4(9):891-893.
240. Rajagopalan PT, Lutz S, Benkovic SJ. Coupling interactions of distal residues enhance dihydrofolate reductase catalysis: mutational effects on hydride transfer rates. Biochemistry 2002;41(42):12618-12628.
241. Ohmae E, Fukumizu Y, Iwakura M, Gekko K. Effects of mutation at methionine-42 of Escherichia coli dihydrofolate reductase on stability and function: implication of hydrophobic interactions. J Biochem 2005;137(5):643-652.
242. Kamiyama T, Gekko K. Effect of ligand binding on the flexibility of dihydrofolate reductase as revealed by compressibility. Biochim Biophys Acta 2000;1478(2):257-266.
243. Ohmae E, Iriyama K, Ichihara S, Gekko K. Effects of point mutations at the flexible loop glycine-67 of Escherichia coli dihydrofolate reductase on its stability and function. J Biochem 1996;119(4):703-710.
244. Gekko K, Kunori Y, Takeuchi H, Ichihara S, Kodama M. Point mutations at glycine-121 of Escherichia coli dihydrofolate reductase: important roles of a flexible loop in the stability and function. J Biochem 1994;116(1):34-41.
245. Wong KF, Selzer T, Benkovic SJ, Hammes-Schiffer S. Impact of distal mutations on the network of coupled motions correlated to hydride transfer in dihydrofolate reductase. Proc Natl Acad Sci U S A 2005;102(19):6807-6812.
246. Wang L, Goodey NM, Benkovic SJ, Kohen A. Coordinated effects of distal mutations on environmentally coupled tunneling in dihydrofolate reductase. Proc Natl Acad Sci U S A 2006;103(43):15753-15758.
247. Sikorski RS, Wang L, Markham KA, Rajagopalan PT, Benkovic SJ, Kohen A. Tunneling and coupled motion in the Escherichia coli dihydrofolate reductase catalysis. J Am Chem Soc 2004;126(15):4778-4779.
248. Pineda P, Kanter A, McIvor RS, Benkovic SJ, Rosowsky A, Wagner CR. Dihydrofolate reductase mutant with exceptional resistance to methotrexate but not to trimetrexate. J Med Chem 2003;46(14):2816-2818.
249. Miller GP, Benkovic SJ. Deletion of a highly motional residue affects formation of the Michaelis complex for Escherichia coli dihydrofolate reductase. Biochemistry 1998;37(18):6327-6335.
250. Cameron CE, Benkovic SJ. Evidence for a functional role of the dynamics of glycine-121 of Escherichia coli dihydrofolate reductase obtained from kinetic analysis of a site-directed mutant. Biochemistry 1997;36(50):15792-15800.
251. Wagner CR, Huang Z, Singleton SF, Benkovic SJ. Molecular basis for nonadditive mutational effects in Escherichia coli dihydrofolate reductase. Biochemistry 1995;34(48):15671-15680.

252. Huang Z, Wagner CR, Benkovic SJ. Nonadditivity of mutational effects at the folate binding site of Escherichia coli dihydrofolate reductase. Biochemistry 1994;33(38):11576-11585.
253. Fierke CA, Benkovic SJ. Probing the functional role of threonine-113 of Escherichia coli dihydrofolate reductase for its effect on turnover efficiency, catalysis, and binding. Biochemistry 1989;28(2):478-486.
254. Adams J, Johnson K, Matthews R, Benkovic SJ. Effects of distal point-site mutations on the binding and catalysis of dihydrofolate reductase from Escherichia coli. Biochemistry 1989;28(16):6611-6618.
255. Chen JT, Mayer RJ, Fierke CA, Benkovic SJ. Site-specific mutagenesis of dihydrofolate reductase from Escherichia coli. J Cell Biochem 1985;29(2):73-82.
256. Andres J, Moliner V, Safont VS, Domingo LR, Picher MT. On Transition Structures for Hydride Transfer Step in Enzyme Catalysis. A Comparative Study on Models of Glutathione Reductase Derived from Semiempirical, HF, and DFT Methods. The Journal of Organic Chemistry 1996;61(22):7777-7783.
257. Pu J, Ma S, Garcia-Viloca M, Gao J, Truhlar DG, Kohen A. Nonperfect synchronization of reaction center rehybridization in the transition state of the hydride transfer catalyzed by dihydrofolate reductase. J Am Chem Soc 2005;127(42):14879-14886.
258. Li L, Falzone CJ, Wright PE, Benkovic SJ. Functional role of a mobile loop of Escherichia coli dihydrofolate reductase in transition-state stabilization. Biochemistry 1992;31(34):7826-7833.
259. Bystroff C, Oatley SJ, Kraut J. Crystal structures of Escherichia coli dihydrofolate reductase: the NADP+ holoenzyme and the folate.NADP+ ternary complex. Substrate binding and a model for the transition state. Biochemistry 1990;29(13):3263-3277.
260. Hammes-Schiffer S, Watney JB. Hydride transfer catalysed by Escherichia coli and Bacillus subtilis dihydrofolate reductase: coupled motions and distal mutations. Philos Trans R Soc Lond B Biol Sci 2006;361(1472):1365-1373.
261. Sergi A, Watney JB, Wong KF, Hammes-Schiffer S. Freezing a single distal motion in dihydrofolate reductase. J Phys Chem B 2006;110(5):2435-2441.
262. Hammes-Schiffer S. Quantum-classical simulation methods for hydrogen transfer in enzymes: a case study of dihydrofolate reductase. Curr Opin Struct Biol 2004;14(2):192-201.

Appendix A: CHARMM Saturation Mutagenesis Script with Solvation in Parallel

```
#!/usr/bin/python
import os
import sys
import random

# define amino acid types
aminos = ["GLY", "GLU", "GLN", "LYS", "ARG", "HSD", "ASP", "ASN", "PHE", "TRP",
"TYR", "ALA", "VAL", "LEU", "ILE", "CYS", "SER", "MET", "THR"]

# define positions to mutate

positions = ["47", "78", "94", "142", "184", "205", "226", "236", "252", "255", "290", "353",
"82", "328"]

#mutation preparation

a = "rename resname "
b = " select segid pep .and. resid "
c = " end\n"
d = "delete atom sele segi pep .and. resid "
e = " .and. .not. (type n .or. type ca .or. type c .or. type ha .or. type hn) end\n\n"

def mutation(position, aa):
	line = a + aa + b + position + c + d + position + e
	return line

#define charmm text used for minimization
def minimization(position, aa):
	minimizationt = """* CHARMm Minimization
*

wrnl -2
!prnl -2
bomb -5

open read unit 1 name /usr/global/cdm/c34b1/toppar/top_all22_prot.inp card
read rtf card unit 1
close unit 1
open read unit 1 name /usr/global/cdm/c34b1/toppar/par_all22_prot.inp card
read para card unit 1
```

```
close unit 1

stream /usr/global/cdm/c34b1/toppar/stream/toppar_all22_prot_heme.str

stream /usr/global/cdm/c34b1/toppar/stream/toppar_all22_prot_fluoro_alkanes.str

! read in primary sequence
open unit 1 read form name p450.pdb
read sequ pdb unit 1
close unit 1

gene pep setup

! read in coordinates
open unit 1 read form name p450.pdb
read coor pdb unit 1
close unit 1

"""

        minimizationt += mutation(position, aa)

        minimizationt += """

open write card name temp.pdb unit 14
write coordinates pdb select all end unit 14
*Temporary
*
close unit 14

delete atom sele segi pep end

open read card name temp.pdb unit 14
read sequence pdb unit 14
generate pep setup
rewind unit 14
read coordinate pdb unit 14
close unit 14

system "rm temp.pdb"

! read in heme
open unit 1 read form name heme.pdb
read sequ pdb unit 1
close unit 1

gene hem setup

! read in coordinates
```

```
open unit 1 read form name heme.pdb
read coor pdb unit 1
close unit 1

open unit 2 read form name ethane.pdb
read sequ pdb unit 2
close unit 2

gene eth setup

open unit 2 read form name ethane.pdb
read coor pdb unit 2
close unit 2

ic fill preserve
ic param
ic build
hbuild

set usepme false
NBONDS CUTNB 14.0 CTOFNB 12.0 CTONNB 10.
SCALER WMAIN = RADIUS
stream "radius.str"
SET EPSW = 80 !
SET EPSP = 1

NBONDS CDIE EPS @EPSP

GBSW EPSP @EPSP EPSW @EPSW sw 0.3 sgamma 0.00542 dgp 1.5 GBenergy

nbon nbxm 5
skip all excl vdw bond angl urey dihe impr elec gbener
mini abnr nstep 2000 tolgrd 0.001

"""

#output file

      minimizationt += """
! write new coordinates to a file \n"""
      minimizationt += "open write unit 1 name p450" + position +  "_" + aa + "refined.pdb card
\n"
      minimizationt += """write coor sele segi pep end pdb unit 1 card
* After backbone relaxation
*

"""
```

```
	minimizationt += "open write unit 2 name ethane" + position + "_" + aa +
"refined.pdbcard\n"
		minimizationt += """write coor sele segi eth end pdb unit 2 card

"""

	minimizationt += "open write unit 3 name heme" + position + "_" + aa + "refined.pdb
card\n"
	minimizationt +="""write coor sele segi hem end pdb unit 3 card
stop

"""

	return minimizationt

def energycalc(position, aa):
	energycalct = """* CHARMm Accurate Energy Calculation
*

wrnl -2
!prnl -2
bomb -5

open read unit 1 name /usr/global/cdm/c34b1/toppar/top_all22_prot.inp card
read rtf card unit 1
close unit 1

open read unit 1 name /usr/global/cdm/c34b1/toppar/par_all22_prot.inp card
read para card unit 1
close unit 1

stream /usr/global/cdm/c34b1/toppar/stream/toppar_all22_prot_heme.str

stream /usr/global/cdm/c34b1/toppar/stream/toppar_all22_prot_fluoro_alkanes.str

"""
	energycalct += "open unit 1 read form name p450" + position + "_" + aa + "refined.pdb\n"
	energycalct += """read sequ pdb unit 1
close unit 1
gene pep setup

"""

	energycalct += "open unit 1 read form name p450" + position + "_" + aa + "refined.pdb\n"
	energycalct += """read coor pdb unit 1
close unit 1

"""
```

```
        energycalct += "open unit 1 read form name heme" + position + "_" + aa + "refined.pdb\n"
        energycalct +="""read sequ pdb unit 1
close unit 1
gene hem setup
"""
        energycalct += "open unit 1 read form name heme" + position + "_" + aa + "refined.pdb\n"
        energycalct += """read coor pdb unit 1
close unit 1

"""
        energycalct +="""
update atom CDIE eps 1 cutnb 14 ctofnb 12 ctonnb 10 switch vswitch
GBMV GRID EPSILON 80 DN 0.2 watr 1.4 GEOM P6 8.0 -
WTYP 0 NPHI 10 SHIFT -0.007998 SLOPE 0.9026 CORR 1 CONV

skip all excl vdw bond angl urey dihe impr harm elec gbener
ener
set totp ?ener

"""

        energycalct +="open unit 1 read form name ethane" + position + "_" + aa + "refined.pdb\n"
        energycalct +="""read sequ pdb unit 1
close unit 1

gene eth setup

"""
        energycalct +="open unit 1 read form name ethane" + position + "_" + aa + "refined.pdb\n"
        energycalct +="""read coor pdb unit 1
close unit 1

GBMV clear
GBMV GRID EPSILON 80 DN 0.2 watr 1.4 GEOM P6 8.0 -
WTYP 0 NPHI 10 SHIFT -0.007998 SLOPE 0.9026 CORR 1 CONV

skip all excl vdw bond angl urey dihe impr harm elec gbener
ener
set tot ?ener

dele atom sele segid pep end
dele atom sele segid hem end

GBMV clear
GBMV GRID EPSILON 80 DN 0.2 watr 1.4 GEOM P6 8.0 -
WTYP 0 NPHI 10 SHIFT -0.007998 SLOPE 0.9026 CORR 1 CONV

skip all excl vdw bond angl urey dihe impr harm elec gbener
ener
```

```
set tots ?ener

set ben 0.0
calc ben @tot - @totp - @tots
open write name energy.txt unit 6 card
write title unit 9
* @ben
* @tot
*


stop

"""

        return energycalct

#open the file that the energies are going to be appended to and create it

def energyoutputs():
        file = open("energies.txt", "w")
        file.close()

energyoutputs()

def energyfile(position, aa):
        file = open("energy_file", "w")
        file.write(energycalc(position, aa))
        file.close()

def minimizationfile(position, aa):
        file = open("minimization_file", "w")
        file.write(minimization(position, aa))
        file.close()

def minimizationrun():
        command =
os.system("""/usr/global/bin/mpirun/usr/global/cdm/c34b1.xj/exec/gnu/charmm.mpi.large
<minimization_file > minimizationout""")
                #       command = os.syste("""charmm <minimization_file> minimization_out""")
                        return comman

def energyrun():
        command =
os.system("""/usr/global/bin/mpirun/usr/global/cdm/c34b1.xj/exec/gnu/charmm.mpi.large
<energy_file > energy_out"""
                #       command = os.ystem("""charmm <energy_file > energy_out""")
                        return comman
```

```
#for looping
for position in positions:
	for aa in aminos:
		minimizationfile(position, aa)
		energyfile(position, aa)
		print position + "_" + aa + "\n"
		minimizationrun()
		energyrun()
		energy = []
		energy_output_file = open("energy.txt","r")
		for line in energy_output_file:
			energy.append(line)
		energy_output_file.close()
		file = open("energies.txt", "a")
		information = "\n" + position + "_" + aa + "\n"
		file.write(information)
		for line in energy:
			if line.startswith(" RDTITL>"):
				file.write(line)
		file.close()
\		os.system("""rm fort.9""")
```

Appendix B: CHARMM Alanine Scanning Mutagenesis Script with Solvation in Parallel

```
#!/usr/bin/python

import os
import sys
import random

# define amino acid types


aminos = ["ALA"]

# define positions to mutate

pos = range(1,456)
positions = []
for elem in pos:
	positions.append(str(elem))

#mutation prepp

a = "rename resname "
b = " select segid pep .and. resid "
c = " end\n"
d = "delete atom sele segi pep .and. resid "
e = " .and. .not. (type n .or. type ca .or. type c .or. type ha .or. type hn) end\n\n"

def mutation(position, aa):
		line = a + aa + b + position + c + d + position + e
		return line

#define charmm text used for minimization
def minimization(position, aa):
		minimizationt = """* CHARMm Minimization
*

wrnl -2
!prnl -2
bomb -5

open read unit 1 name /usr/global/cdm/c34b1/toppar/top_all22_prot.inp card
read rtf card unit 1
close unit 1
```

```
open read unit 1 name /usr/global/cdm/c34b1/toppar/par_all22_prot.inp card
read para card unit 1
close unit 1

stream /usr/global/cdm/c34b1/toppar/stream/toppar_all22_prot_heme.str

stream /usr/global/cdm/c34b1/toppar/stream/toppar_all22_prot_fluoro_alkanes.str

! read in primary sequence
open unit 1 read form name p450.pdb
read sequ pdb unit 1
close unit 1

gene pep setup

! read in coordinates
open unit 1 read form name p450.pdb
read coor pdb unit 1
close unit 1

"""

        minimizationt += mutation(position, aa)

        minimizationt += """

open write card name temp.pdb unit 14
write coordinates pdb select all end unit 14
*Temporary
*
close unit 14

delete atom sele segi pep end

open read card name temp.pdb unit 14
read sequence pdb unit 14
generate pep setup
rewind unit 14

read coordinate pdb unit 14
close unit 14

system "rm temp.pdb"

! read in heme
open unit 1 read form name heme.pdb
read sequ pdb unit 1
close unit 1
```

```
gene hem setup

! read in coordinates
open unit 1 read form name heme.pdb
read coor pdb unit 1
close unit 1

open unit 2 read form name ethane.pdb
read sequ pdb unit 2
close unit 2

gene eth setup

open unit 2 read form name ethane.pdb
read coor pdb unit 2
close unit 2

ic fill preserve
ic param
ic build
hbuild

set usepme false
NBONDS CUTNB 14.0 CTOFNB 12.0 CTONNB 10.
SCALER WMAIN = RADIUS
stream "radius.str"
SET EPSW = 80 !
SET EPSP = 1

NBONDS CDIE EPS @EPSP

GBSW EPSP @EPSP EPSW @EPSW sw 0.3 sgamma 0.00542 dgp 1.5 GBenergy

nbon nbxm 5
skip all excl vdw bond angl urey dihe impr harm elec gbener
mini abnr nstep 2000 tolgrd 0.001

"""

#output file

        minimizationt += """
! write new coordinates to a file \n"""
        minimizationt += "open write unit 1 name p450" + position +  "_" + aa + "refined.pdb
card \n"
        minimizationt += """write coor sele segi pep end pdb unit 1 card
* After backbone relaxation
*
```

```
"""
        minimizationt += "open write unit 2 name ethane" + position + "_" + aa + "refined.pdb
card\n"
        minimizationt += """write coor sele segi eth end pdb unit 2 card

"""

        minimizationt += "open write unit 3 name heme" + position + "_" + aa + "refined.pdb
card\n"
        minimizationt +="""write coor sele segi hem end pdb unit 3 card
stop

"""

        return minimizationt

def energycalc(position, aa):
        energycalct = """* CHARMm Accurate Energy Calculation
*

wrnl -2
!prnl -2
bomb -5

open read unit 1 name /usr/global/cdm/c34b1/toppar/top_all22_prot.inp card
read rtf card unit 1
close unit 1

open read unit 1 name /usr/global/cdm/c34b1/toppar/par_all22_prot.inp card
read para card unit 1
close unit 1

stream /usr/global/cdm/c34b1/toppar/stream/toppar_all22_prot_heme.str

stream /usr/global/cdm/c34b1/toppar/stream/toppar_all22_prot_fluoro_alkanes.str

"""
        energycalct += "open unit 1 read form name p450" + position + "_" + aa +
"refined.pdb\n"
        energycalct += """read sequ pdb unit 1
close unit 1
gene pep setup

"""

        energycalct += "open unit 1 read form name p450" + position + "_" + aa +
"refined.pdb\n"
```

```
		energycalct += """read coor pdb unit 1
close unit 1

"""

		energycalct += "open unit 1 read form name heme" + position + "_" + aa +
"refined.pdb\n"
		energycalct +="""read sequ pdb unit 1
close unit 1
gene hem setup
"""
		energycalct += "open unit 1 read form name heme" + position + "_" + aa +
"refined.pdb\n"
		energycalct += """read coor pdb unit 1
close unit 1

"""

		energycalct +="""
update atom CDIE eps 1 cutnb 14 ctofnb 12 ctonnb 10 switch vswitch
GBMV GRID EPSILON 80 DN 0.2 watr 1.4 GEOM P6 8.0 -
WTYP 0 NPHI 10 SHIFT -0.007998 SLOPE 0.9026 CORR 1 CONV

skip all excl vdw bond angl urey dihe impr harm elec gbener
ener
set totp ?ener

"""

		energycalct +="open unit 1 read form name ethane" + position + "_" + aa +
"refined.pdb\n"
		energycalct +="""read sequ pdb unit 1
close unit 1

gene eth setup

"""
		energycalct +="open unit 1 read form name ethane" + position + "_" + aa +
"refined.pdb\n"
		energycalct +="""read coor pdb unit 1
close unit 1

GBMV clear
GBMV GRID EPSILON 80 DN 0.2 watr 1.4 GEOM P6 8.0 -
WTYP 0 NPHI 10 SHIFT -0.007998 SLOPE 0.9026 CORR 1 CONV

skip all excl vdw bond angl urey dihe impr harm elec gbener
ener
set tot ?ener
```

```
dele atom sele segid pep end
dele atom sele segid hem end

GBMV clear
GBMV GRID EPSILON 80 DN 0.2 watr 1.4 GEOM P6 8.0 -
WTYP 0 NPHI 10 SHIFT -0.007998 SLOPE 0.9026 CORR 1 CONV

skip all excl vdw bond angl urey dihe impr harm elec gbener
ener
set tots ?ener

set ben 0.0
calc ben @tot - @totp - @tots
open write name energy.txt unit 6 card
write title unit 9
* @ben
* @tot
*


stop

"""

        return energycalct

#open the file that the energies are going to be appended to and create it

def energyoutputs():
        file = open("energies.txt", "w")
        file.close()

energyoutputs()

def energyfile(position, aa):
        file = open("energy_file", "w")
        file.write(energycalc(position, aa))
        file.close()

def minimizationfile(position, aa):
        file = open("minimization_file", "w")
        file.write(minimization(position, aa))
        file.close()

def minimizationrun():
        command = os.system("""/usr/global/bin/mpirun
/usr/global/cdm/c34b1.xj/exec/gnu/charmm.mpi.large <minimization_file > minimization_out""")
#       command = os.system("""charmm <minimization_file> minimization_out""")
```

```
        return command

def energyrun():
        command = os.system("""/usr/global/bin/mpirun
/usr/global/cdm/c34b1.xj/exec/gnu/charmm.mpi.large <energy_file > energy_out""")
#       command = os.system("""charmm <energy_file > energy_out""")
        return command

#for looping
for position in positions:
    for aa in aminos:
                minimizationfile(position, aa)
                energyfile(position, aa)
                print position + "_" + aa + "\n"
                minimizationrun()
                energyrun()
                energy = []
                energy_output_file = open("energy.txt","r")
                for line in energy_output_file:
                        energy.append(line)
                energy_output_file.close()
                file = open("energies.txt", "a")
                information = "\n" + position + "_" + aa + "\n"
                file.write(information)
                for line in energy:
                        if line.startswith(" RDTITL>"):
                                file.write(line)
                file.close()
                os.system("""rm fort.9""")
```

VITA

George A. Khoury

M.S. in Chemical Engineering, Pennsylvania State University, expected May 2010

Thesis Title: Computational Design to Switch Protein Cofactor Specificity and Create Enzymatic Activity

Advisor: Professor Costas D. Maranas (Co-Advised by Professors Mike Janik and Patrick Cirino)

(35 Credits)

B.S. in Chemical Engineering with High Distinction and Honors, Schreyer Honors College

Pennsylvania State University, May 2009

GPA: 3.93/4.00 (Class Marshal, 158 Credits)

This thesis research resulted in the following publications and research presentations:

Publications:

Khoury, G.A., H. Fazelinia, J.W. Chin, R.J. Pantazes, P.C. Cirino and C.D. Maranas, "Computational Design of *Candida boidinii* Xylose Reductase for Altered Cofactor Specificity," *Protein Science*, 2009, 18(10): p.2125-38, **featured on cover**

Oral and Poster* Presentations:**

"Ground and Transition State Binding Calculations to Improve Cytochrome $P450_{BM3}$ Reactivity and Specificity."**George A. Khoury,** Ping Lin, Michael J. Janik, Patrick C. Cirino, and Costas D. Maranas, Genomic Science Contractor-Grantee and Knowledgebase Workshop. Feb 7-10, 2010, Arlington, VA.***

"Engineering the Cofactor Specificity of Candida boidinii Xylose Reductase."**George A. Khoury**, Hossein Fazelinia, Jonathan W. Chin, Patrick C. Cirino, and Costas D. Maranas., AICHE National Meeting. November 8-13, 2009, Nashville, TN.

"Ground and Transition State Computations on Cytochrome P450 BM-3 For Understanding Mutant Reactivity and Selectivity." **George A. Khoury**, Ping Lin, Michael J. Janik, Patrick C. Cirino, and Costas D. Maranas., AICHE National Meeting. November 8-13, 2009, Nashville, TN.

"Computational Enzyme Redesign for Cofactor/Substrate Specificity." **George A. Khoury,** Hossein Fazelinia, Patrick C. Cirino, and Costas D. Maranas. ACS National Meeting. August 16-19, 2009. Washington, D.C.

"Computational Analysis and Design for Altered Cofactor Specificity." **George A. Khoury**, Hossein Fazelinia, Patrick Cirino, and Costas Maranas. AICHE National Meeting. November 2008, Philadelphia, PA