An iterative approach to *de novo* computational enzyme design and the successful application to the Kemp elimination

Thesis by

Heidi Kathleen Privett

In Partial Fulfillment of the Requirements for the
Degree of Doctor of Philosophy

California Institute of Technology

Pasadena, California

2009

(Defended May 13, 2009)

ACKNOWLEDGEMENTS

I offer thanks to my advisor, Stephen Mayo, for his support and encouragement, for allowing me the opportunity to work on an exciting project that at times seemed impossible, and for not letting me give up when my research appeared hopeless.

I would also like to thank all of the past and present members of the Mayo Lab who contributed to a fun, congenial, and collaborative work environment and were always there to provide advice and encouragement.

I would like to thank all of my official collaborators, who have all been acknowledged individually in the relevant chapters and appendices.

In addition, I thank all of the Caltech students, postdocs, and staff who generously gave their time to teach me new techniques, trained me on equipment housed in their labs, and helped me troubleshoot problems in my projects even though my research had little bearing on their own. These generous individuals include Jonas Oxgaard (Goddard Lab), Amanda Cashin (Dougherty Lab), Ariele Hanek (Dougherty Lab), Dan Caspi (Stoltz Lab), Doug Behenna (Stoltz Lab), Robert Dirks (Pierce Lab), Jens Kaiser (Rees Lab), Jost Veilmetter (Protein Expression Facility), Rich Olson (Bjorkman Lab), Adrian Rice (Bjorkman/Rees Labs), Hernan Garcia (Phillips Lab), Brendan Mack (Davis Lab), and Jenn Stockdill (Stoltz Lab).

I also thank my many Caltech and non-Caltech friends, without whom I would have never made it, and my family for telling me that I could be anything I wanted to be when I grew up.

ABSTRACT

The development of reliable methods for the "on demand" *de novo* design of an enzymatic catalyst for an arbitrary chemical reaction has been an elusive goal of the computational protein design community. Recent successful results of *de novo* computational enzyme design have been encouraging, but the activity of the enzymes produced so far is still well below that of natural enzymes and the generalizability of these methods has yet to be established.

Presented in this thesis are methods that we have developed for the computational design of enzyme active sites as well as results from the evaluation of these methods through a test case, the Kemp elimination. Initial Kemp elimination designs were shown to be inactive. However, in the course of refining these design procedures, we carried out extensive theoretical and experimental evaluation of several of these inactive designs, which allowed us to identify the causes of the inactivity and led to adjustments of our design procedure. These modified methods were then successfully used to design four distinct enzymes for this reaction in three inert scaffolds including the scaffold that housed the previously inactive designs. In addition, we demonstrate that molecular dynamics simulations can accurately predict the activity of designed Kemp elimination enzymes and can be used as a reliable prescreening step, allowing us to focus our experimental efforts on designs that are most likely to be active.

The work presented here demonstrates that the cyclic evaluation and redesign of both active and inactive enzymes was instrumental in the identification and resolution of deficiencies in our computational methods and directly resulted in *de novo* designed enzymes with novel and increased activity.

TABLE OF CONTENTS

TABLES AND FIGURES

ABBREVIATIONS

AFU	arbitrary fluorescence units
AR	androgen receptor
CD	circular dichroism
C. fimi	*Cellullomonas fimi*
CLEARSS	combinatorial libraries emphasizing and reflecting scored sequences
CV	column volume
Da	Dalton (1 g/mol)
DEE	dead end elimination
DHT	dihydrotestosterone
DMF	dimethylformamide
DNA	deoxyribonucleic acid
E. coli	*Escherichia coli*
FASTER	fast and accurate side-chain topology and energy refinement
FFH	L-2-phenyl-4-benzylphenyloxazolin-5-one hydrolase
FMEC	faster minimum energy conformation
FOX	L-2-phenyl-4-benzylphenyloxazolin-5-one
GdnHCl	guanidine hydrochloride
GMEC	global minimum energy conformation
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfornic acid
HPLC	high-pressure liquid chromatography
IPTG	isopropyl β-D-1-thiogalactopyranoside
IR	infrared
k_{cat}	catalytic constant
KE	Kemp elimination
K_M	Michaelis constant
k_{uncat}	rate constant for an uncatalyzed reaction
LB	Luria-Bertani broth
LK	Lazaridis-Karplus (solvent exclusion model)

MBP	maltose binding protein
MC	Monte Carlo
MD	molecular dynamics
MES	2-(*N*-morphilino)ethanesulfonic acid
MME	monomethyl ether
MR	molecular replacement
MS	mass spectrometry
MUG	4-methylumbelliferyl-β-D-glucopyranoside
MUX	4-methylumbelliferyl-β-D-xylopyranoside
MWCO	molecular weight cut off
NBT	5-nitrobenzotriazole
NBZ	5-nitrobenzoxizole
NBX	transition state of the 5-nitrobenzoxizole Kemp elimination reaction
Ni-NTA	nickle-nitrilotriacetic
NMR	nuclear magnetic resonance
NPT	constant number of particles, pressure, and temperature
OD	optical density at a specific wavelength
ORBIT	optimization of rotamers by iterative techniques
PBS	phosphate buffered saline
PDB	protein data bank
PEG	polyethylene glycol
PNK	polynucleotide kinase
*p*NP	para-nitrophenol
PCR	polymerase chain reaction
P*i*	inorganic phosphate
PPMAL	Protein/Peptide MicroAnalytical Laboratory (Caltech)
RMSD	root mean squared deviation
S. avidinii	*Streptomyces avidinii*
S. cerevisiae	*Saccharomyces cerevisiae*
SDS-PAGE	sodium dodecyl sulfate polyacrylamide gel electrophoresis

S. solfataricus	*Sulfolobus solfataricus*
T. aurantiacus	*Thermoascus* aurantiacus
TAX	*T. aurantiacus* xylanase
TES	testosterone
T_m	midpoint of thermal denaturation curve
T. maritima	*Thermotoga maritima*
Tris	tris(hydroxymethyl)aminomethane)
TS	transition state
UV-vis	ultraviolet-visible
VDW	van der Waals

Chapter I

Introduction

Enzymes are awesome

Enzymes are extremely efficient catalysts, accelerating chemical reaction rates up to 10^{19} times that of the uncatalyzed reaction.[1] In addition to the large rate enhancements that can be observed, enzymes can also carry out reactions with extreme regio- and stereospecificity, eliminating the need for protecting groups and reliably producing a single product.[2,3] In the face of concern over the environmental impact of chemical synthesis, enzymes have emerged as attractive alternatives to chemical catalysts because they work under mild, aqueous conditions, reducing the generation of hazardous wastes that are often associated with organic synthesis.[2] The enzymes themselves are, of course, biodegradable and can usually be produced in large quantities via recombinant expression in bacteria or fungi.

Despite their promise, significant challenges prevent the widespread use of enzymes as industrial catalysts. The applicability of enzymes can be limited by their instability in conditions appropriate for industrial processes, including high temperatures, extreme pHs, and organic solvents.[2] In addition, the scope of reactions that can be catalyzed by enzymes is limited to those found in natural metabolic pathways, although some enzymes, including lipases, have been shown to be somewhat promiscuous in their substrate specificity.[4] Directed evolution has been used to improve stability, optimize efficiency, and modify the substrate specificity of many enzymes.[5-7] However, this

method requires existing activity toward the reaction of interest, so it cannot be used to introduce truly novel chemistries.[5]

Catalytic antibodies have shown promise in the catalysis of a wide variety of chemical transformations including stereoselective and novel reactions.[8] However, the catalytic efficiencies of antibodies have traditionally been modest due in part to selections that are based on binding to a synthetic transition state analog instead of on enzymatic turnover.[8,9] Reactive immunization has been used to overcome these limitations, producing catalytic antibodies with reaction rates approaching those of the wild-type enzyme, but this method can only be applied to reactions whose transition states are known and can be readily mimicked with a reactive transition state analog that is accessible by current synthetic methods.[10,11] In addition, Xu *et al*. has suggested that the immunoglobulin scaffold itself may limit the scope of reactions amenable to catalysis with antibodies.[8]

While directed evolution and catalytic antibodies have both been successfully used for enzyme engineering, both have features that keep them from being applied generally for the engineering of novel enzymatic activities. In contrast, computational protein design does not suffer from these limitations, and can be envisioned as a solution to many complex synthetic organic chemistry problems.

Computational protein design

Computational protein design has shown great promise for developing novel functions in proteins. The general approach to solving a computational design problem is cyclic (Figure 1-1). Beginning with the backbone coordinates of a high-resolution

protein crystal structure, an optimization algorithm is used to search through combinations of side-chain identities and conformations for the sequence and geometries of amino acids that will best stabilize the protein fold.[12,13] The extent to which a particular sequence might stabilize the desired fold is evaluated through a force field scoring function, which calculates an energy for the sequence that should correlate to the protein's free energy of folding. The sequence and conformation of amino acids that will best stabilize the backbone structure (i.e., fold into the desired conformation) is assumed to be the one with the lowest energy. This global minimum energy conformation (GMEC) must then be experimentally validated by structural and/or thermodynamic comparison of the designed sequence to that of the native protein.[12-14] Information gained from evaluating the deviations of theory from experiment can then be used to readjust the force field parameters, thus completing the design cycle.[12]

For a small 59-residue protein, there are about 10^{77} possible sequences (assuming that all 20 amino acids are allowed at all positions). If a single molecule of each of these sequences were to actually be synthesized, their combined mass would be approximately 7.6×10^{56} g, which according to some estimates, approaches the mass of the observable universe. When the conformational flexibility of the side chains is also taken into account, the number of possible solutions explodes even further. To reduce the combinatorial complexity of the problem to a reasonable size, we limit our designs to use a library of discrete sidechain conformations called rotamers, which represent the statistically significant amino acid sidechain conformations found in protein crystal structures.[15] In results described later in this text, we used rotamer libraries based on

those developed by Dunbrack and Karplus,[16] as well as sidechain conformer libraries developed in the Mayo lab.[17]

The ORBIT (Optimization of Rotamers By Iterative Techniques) software suite is a computational protein design package developed in the Mayo lab.[13] Standard implementations of ORBIT use a scoring function based on physical principles and apply the DREIDING force field,[18] which incorporates four empirically based potential functions to calculate the total energy (E_{total}) of a structure:

$$E_{total} = E_{vdw} + E_{h\text{-}bond} + E_{elect} + E_{as} \quad .$$

(1) A van der Waals (VDW) interaction energy (E_{vdw}) is calculated for each pair of rotamers using a Lennard-Jones 12-6 potential.[18]

(2) A hydrogen bond potential ($E_{h\text{-}bond}$) is used that is angle-, distance-, and hybridization-dependent.[19]

(3) Electrostatic interactions (E_{elect}) are calculated based on Coulomb's Law incorporating a distance-dependent dielectric of 40r, where r is the interatomic distance.[19]

(4) A solvation term (E_{as}) is used that employs a solvation potential based either on the protein's surface area or the occlusion of one atom by another. Both of these solvation models give an energy benefit to buried nonpolar regions of the protein and penalize exposed nonpolar and buried polar regions.[20,21]

The optimization algorithms provided by ORBIT apply a variety of methods to establish the optimal sequence or set of sequences to stabilize a given fold. Algorithms based on the Dead-End Elimination theorem (DEE)[22-25] are used to quickly identify and remove amino acid rotamers and pairs of rotamers that cannot exist in the GMEC.

ORBIT also supplies stochastic methods of sequence optimization such as those based on the Fast and Accurate Side-Chain Topology and Energy Refinement (FASTER) algorithm[15] and Monte Carlo[26,27] either alone or in combination with DEE to both decrease calculation time and to sample sequence space around the GMEC.

ORBIT has been used to design proteins in a wide variety of systems. Successful implementations include the full sequence design of a protein that adopts a zinc finger fold independent of zinc binding,[28] the redesign of calmodulin to increase its binding specificity for a single target peptide,[29] and the *de novo* design of a protein-protein interface.[30]

Computational enzyme design

Promisingly, ORBIT has also been used to design an enzyme with modest catalytic activity.[31] This "protozyme" with p-nitrophenol acetate hydrolysis activity with a k_{cat}/k_{uncat} of 10^2 was one of the first examples of a *de novo* computationally designed enzyme. Other labs have also employed computational tools to design enzymes, including transplanting reactive metalloenzyme active sites into inert proteins.[32,33] In addition, computational enzyme design methodologies have been used to switch the specificity of existing enzymes.[34,35] More recent dramatic successes from one of these labs include the *de novo* design of retroaldolases as well as enzymes that catalyze the Kemp elimination, a reaction for which no natural enzyme exists.[36,37]

The promise of computational enzyme design has been clearly established. However, the generalizability of these methods for other chemical transformations has not yet been demonstrated. In addition, while some of the enzymes designed so far have

had impressive catalytic activity, the computational design of enzymes with true native-like efficiency still presents a challenge.[36,37] The goal of engineering "designer enzymes" for any reaction remains extremely attractive as it will allow the scope of enzymatic reactions to extend beyond the limits of natural cellular metabolism, broadening the range of possible substrates and products, especially those with stereogenic centers. Once this challenge has been met and the generality of computational protein design techniques has been established, rapid, on-demand engineering of enzymes will be possible for many important chemical reactions.

Towards this goal, my work has focused on the introduction and evaluation of new enzyme design capabilities both in ORBIT and in a related program, Phoenix. Force field parameterization in protein design has historically been carried out with respect to protein stability and overall fold without regard to specific function. In the case of ORBIT, parameters for the DREIDING force field were optimized through the sequence design of a small protein and the subsequent experimental evaluation of changes in overall thermodynamic stability.[12,38] The resulting parameters were weighted to emphasize VDW contacts and buried hydrophobic surface area. While these parameters were successfully used to design many hyperstable proteins, they are not necessarily well suited for designing enzymes because most natural enzymes are not evolved for optimum stability.[39,40]

According to the transition state theory, enzymes achieve such large rate enhancements through specific tight binding and stabilization of the reaction transition state (TS).[41] Computational simulations of enzyme active sites have suggested that polar and nonpolar residues that contact the TS but are not directly involved in the reaction

chemistry can provide additional stabilization to the TS through electrostatics and VDW interactions, thus promoting catalysis.[42] To design an effective enzyme *de novo*, we must first determine the nature of the interactions between the protein and the rate limiting TS that promote catalysis; information about these interactions (e.g., sidechain functional groups and contact geometries) can be gained from *ab initio* calculations, analogy to existing enzymes, or chemical intuition.[17,43] These protein-TS contacts can then be incorporated into the scoring function.[17,44]

The major changes made to ORBIT to accommodate enzyme design include the introduction of a geometry biasing term that allows an energetic benefit to be added to those sequences that can make specified stabilizing contacts to a TS model present in the active site. In addition, we have implemented various methods for the creation of libraries of transition state poses within the active site that can be sampled during the sequence search.[17]

Chapter II of this thesis is a journal article that I co-authored describing our computational enzyme design methodology in detail. Here, our methods were evaluated through recapitulation of the active site configurations of three natural enzymatic/binding protein systems: *Escherichia coli* chorismate mutase, *Saccharomyces cerevisiae* triosephosphate isomerase, and *Streptomyces avidinii* streptavadin.

As a result of our previous *de novo* design experience, we chose to focus on a well-studied chemical system: the general base-catalyzed Kemp elimination (KE) of 5-nitrobenzisoxazole (Figure 1-2). The KE has been used since the 1970s as a physical organic model for proton transfer from carbon[45,46] and more recently as a model system for enzymatic proton transfer reactions.[47-49] Other attractive features of this reaction are

that it is irreversible with a single transition state and that it has a product that can be observed spectrophotometrically (λ_{max} = 405 nm).[45] In addition, multiple catalytic antibodies have been created that can catalyze this reaction with rate accelerations up to 10^6 times faster than the background reaction.[50,51] The crystal structure of one of these antibodies has given us clues as to how transition state stabilization can be achieved for this reaction (Figure 1-3).[52] In this catalytic antibody, stabilization occurs through a combination of a carboxylate general base, extensive π-stacking above and below the plane of the ring system, and hydrogen bond contacts to the base. In addition, there is precedent for the amenability of this reaction to computational enzyme design methodologies, as Röthlisberger *et al.* were able to computationally introduce catalytic activity for the KE into three separate inert scaffolds, creating multiple active enzymes.[37]

In Chapter III, the computational and experimental details of this system are described and one of the resulting inactive designs is discussed. In the protein design cycle (Figure 1-1), the design procedure cannot be adjusted without some information from the initial inactive design indicating the possible cause of inactivity. To complete the design cycle, we first had to determine why this initial design was inactive. This chapter also includes details of crystallographic analysis carried out in collaboration with the Molecular Observatory at Caltech and molecular dynamic (MD) simulation studies carried out in collaboration with Ken Houk's lab at the University of California, Los Angeles that were used to analyze the inactive design. The X-ray crystal structure of this inactive design confirmed that the actual active site of the design was very similar to the predicted structure. Thus, the inactivity was not due to gross misplacement of the active site residues or disruption of the overall protein fold. MD analysis of the design helped

us to determine possible causes of the inactivity including an active site that was too flexible and solvent exposed.

The lessons learned from our first KE designs pointed us in the direction of more buried, less polar active sites. Using the same scaffold as the initial design, we moved the active site away from the natural, solvent-exposed active site and located it farther into the barrel of the protein. In Chapter IV, I discuss this new design, HG-2, which was predicted to have activity by the blind MD simulations due to the drier, less flexible active site. This activity was confirmed by experimental characterization of this enzyme, also discussed in Chapter IV.

Because of the expense and time associated with experimental evaluation of designed enzymes, strategies for the *a priori* differentiation of active designs from inactive ones are needed to make the process of enzyme design more efficient. In Chapter IV, additional designs are described which were carried out using two scaffolds that have been used to produce successful KE designs in David Baker's lab at the University of Washington.[37] Of the six enzymes synthesized, four showed activity and three of these resulted from the evaluation our enzyme design methods through redesigning the active site of scaffolds used in the active KE designs from the Baker lab. The fourth active design is unique. In most cases, blind MD analysis of these designs was successful in distinguishing active designs from inactive ones. MD analysis could thus serve as an important tool in the computational design procedure, providing an initial screen of sequences predicted by the design procedure to help us determine the designs on which to focus our experimental efforts.

Appendix A describes an early attempt at *de novo* design of enzymatic activity. Here, our goal was to design an enantioselective enzyme for the kinetic resolution of of N-benzoyl-L-phenylalanine through the selective hydrolysis of L-2-phenyl-4-benzylphenyloxazolin-5-one (FOX). Lessons learned from this first unsuccessful attempt at enzyme design led us to focus on less flexible scaffolds and chemical systems that have a smaller background reaction rate. In addition to the *de novo* enzyme design project that has spanned my entire graduate career, I have had the opportunity to work on other computational design projects related to enzymes and binding proteins. In Appendix B, I discuss our ongoing efforts to alter the specificity of an existing thermophilic xylanase. Appendix C presents computational efforts toward changing the specificity of an androgen receptor as part of a collaboration with the Fletterick lab at the University of California, San Francisco. Appendix D is the first description of the recombinant over-expression and purification of a thermophilic xylanase from *Thermoascus aurantiacus* (TAX). A recombinant version of this enzyme was necessary to allow genetic manipulation in the creation of new designs and TAX was used as the scaffold for the initial inactive KE design and one of the subsequent active designs. Because of its ease of expression, thermostability, and ability to tolerate multiple mutations, this enzyme proved to be a useful scaffold for computational design.

In sum, the work presented in this thesis shows that by iterative structural and theoretical evaluation of active and inactive designs, adjustment of our enzyme design procedure, and subsequent redesign, we can identify and address deficiencies in our design methodology, resulting in *de novo* designed enzymes with significant activity for the reaction of interest. The field of *de novo* computational enzyme design is extremely

promising and the work presented here is a significant step towards the goal of a general method for the computational design as enzymes in the Mayo lab. This work will serve as the foundation for future studies, which will be undertaken to obtain reaction rate accelerations and efficiencies comparable to those of natural enzymes and to generalize these methods for a wide variety of chemistries.

References

1. Wolfenden, R.; Snider, M. J., The depth of chemical time and the power of enzymes as catalysts. *Acc. Chem. Res.* **2001**, *34*, 938-945.

2. Wong, C.-H.; Whitesides, G. M., *Enzymes in Synthetic Organic Chemistry*. Elsevier Science & Technology Books: Tarrytown, NY, **1994**; Vol. 12.

3. Schmid, A.; Dordick, J. S.; Hauer, B.; Kiener, A.; Wubbolts, M.; Witholt, B., Industrial biocatalysis today and tomorrow. *Nature* **2001**, *409*, 258-268.

4. Bornscheuer, U.; Kazlauskas, R., Catalytic promiscuity in biocatalysis: using old enzymes to form new bonds and follow new pathways. *Angew. Chem. Int. Ed.* **2004**, *43*, 6032-6040.

5. Shao, Z.; Arnold, F. H., Engineering new functions and altering existing functions. *Curr. Opin. Struct. Biol.* **1996**, *6*, 513-518.

6. Williams, G. J.; Nelson, A. S.; Berry, A., Directed evolution of enzymes for biocatalysis and the life sciences. *Cell. Mol. Life Sci.* **2004**, *61*, 3034-3046.

7. Horsman, G. P.; Liu, A. M. F.; Henke, E.; Bornscheuer, U. T.; Kazlauskas, R. J., Mutations in distant residues moderately increase the enantioselectivity of Pseudomonas fluorescens esterase towards methyl- 3-bromo-2-methylpropanoate and ethyl 3-phenylbutyrate. *Chem. Eur. J.* **2003**, *9*.

8. Xu, Y.; Yamamoto, N.; Janda, K. D., Catalytic antibodies: hapten design strategies and screening methods. *Biorg. Med. Chem.* **2004**, *12*, 5247-5268.

9. Hilvert, D., Critical analysis of antibody catalysis. *Annu. Rev. Biochem.* **2000**, *69*, 751-793.

10. Wirshing, P.; Ashley, J. A.; Lo, C. H. L.; Janda, K. D.; Lerner, R. A., Reactive immunization. *Science* **1995**, *270*, 1775-1782.

11. Barbas III, C. F.; Heine, A.; Zhong, G.; Hoffmann, T.; Gramatikova, S.; Bjornestedt, R.; List, B.; Anderson, J.; Stura, E. A.; Wilson, I. A.; Lerner, R. A., Immune versus natural selection: Antibody aldolases with enzymatic rates but broader scope. *Science* **1997**, *278*, 2085-2092.

12. Dahiyat, B. I.; Mayo, S. L., Protein design automation. *Protein Sci.* **1996**, *5*, 895-903.

13. Street, A. G.; Mayo, S. L., Computational protein design. *Structure* **1999**, *7*, 105-109.

14. Gordon, D. B.; Marshall, S. A.; Mayo, S. L., Energy functions for protein design. *Curr. Opin. Struct. Biol.* **1999**, *9*, 509-513.

15. Dunbrack, R. L. J., Rotamer libraries in the 21st century. *Curr. Opin. Struct. Biol.* **2002**, *12*, 431-440.

16. Dunbrak, R. L.; Karplus, M., Backbone-dependent rotamer library for proteins. *J. Mol. Biol.* **1993**, *230*, 543-574.

17. Lassila, J. K.; Privett, H. K.; Allen, B. D.; Mayo, S. L., Combinatorial methods for small-molecule placement in computational enzyme design. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 16710-16715.

18. Mayo, S. L.; Olafson, B. D.; Goddard III, W. A., DREIDING: A generic force field for molecular simulations. *J. Phys. Chem.* **1990**, *94*, 8897-8909.

19. Dahiyat, B. I.; Gordon, B.; Mayo, S. L., Automated design of the surface positions of protein helices. *Protein Sci.* **1997**, *6*, 1333-1337.

20. Street, A. G.; Mayo, S. L., Pairwise calculation of protein solvent-accessible surface areas. *Fold Des.* **1998**, *3*, 253-258.

21. Lazaridis, T.; Karplus, M., Effective energy functions for protein structure prediction. *Curr. Opin. Struct. Biol.* **2000**, *10*, 139-145.

22. Desmet, J.; De Maeyer, M.; Hazes, B.; Lasters, I., The dead-end elimination theorem and its use in protein side-chain positioning. *Science* **1992**, *356*, 539-542.

23. Gordon, D. B.; Mayo, S. L., Radical performance enhancements for combinatorial optimization algorithms based on the dead-end elimination theorem. *J. Comput. Chem.* **1998**, *19*, 1505-1514.

24. Goldstein, R. F., Efficient rotamer elimination applied to protein side-chains and related spin-glasses. *Biophys. J.* **1994**, *66*, 1335-1340.

25. Pierce, N. A.; Spriet, J. A.; Desmet, J.; Mayo, S. L., Conformational splitting: a more powerful criterion for dead-end elimination. *J. Comput. Chem.* **2000**, *21*, 9999-1009.

26. Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H., Equation of state calculations by fast computing machines. *J. Chem. Phys.* **1953**, *21*, 1087-1092.

27. Kirkpatrick, S.; Gelatt, C. D.; Vecchi, M. P., Optimization by simulated annealing. *Science* **1983**, *220*, 671-675.

28. Dahiyat, B. I.; Mayo, S. L., De novo protein design: fully automated sequence selection. *Science* **1997**, *278*, 82-87.

29. Shifman, J. M., Mayo, S. L., Modulating calmodulin binding specificity through computational protein design. *J. Mol. Biol.* **2002**, *323*, 417-423.

30. Huang, P. S.; Love, J. J.; Mayo, S. L., A de novo designed protein protein interface. *Protein Sci.* **2007**, *16*, 2770-2774.

31. Bolon, D. N.; Mayo, S. L., Enzyme-like proteins by computational protein design. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 14274-14279.

32. Benson, D. E.; Wisz, M. S.; Hellinga, H. W., Rational design of nascent metalloenzymes. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 6292-6297.

33. Benson, D. E.; Haddy, A. E.; Hellinga, H. W., Converting a maltose receptor into a nascent binuclear copper oxygenase by computational design. *Biochemistry* **2002**, *41*, 3262-3269.

34. Ashworth, J.; Havranek, J. J.; Duarte, C. M.; Sussman, D.; Monnat, R. J.; Stoddard, B. L.; Baker, D., Computational redesign of endonuclease DNA binding and cleavage specificity. *Nature* **2006**, *441*, 656-659.

35. Chen, C. Y.; Georgiev, I.; Anderson, A. C.; Donald, B. R., Computational structure-based redesign of enzyme activity. *Proc. Natl. Acad. Sci. USA* **2009**.

36. Jiang, L.; Althoff, E. A.; Clemente, F. R.; Doyle, L.; Röthlisberger, D.; Zanghellini, A.; Gallaher, J. L.; Betker, J. L.; Tanaka, F.; Barbas, C. F.; Hilvert, D.; Houk, K. N.; Stoddard, B. L.; Baker, D., De novo computational design of retro-aldol enzymes. *Science* **2008**, *319*, 1387-1391.

37. Rothlisberger, D.; Khersonsky, O.; Wollacott, A.; Jiang, L.; Dechancie, J.; Betker, J.; Gallaher, J. L.; Althoff, E. A.; Zanghellini, A.; Dym, O.; Albeck, S.; Houk, K. N.; Tawfik, D.; Baker, D., Kemp elimination catalysts by computational enzyme design. *Nature* **2008**, *453*, 190-U194.

38. Dahiyat, B. I.; Mayo, S. L., Probing the role of packing specificity in protein design. *Proc. Natl. Acad. Sci. USA* **1997**, *94*, 10172-10177.

39. Arnold, F. H.; Wintrode, P. L.; Miyazaki, K.; Gershenson, A., How enzymes adapt: lessons from directed evolution. *Trends Biochem. Sci.* **2001**, *26*, 100-106.

40. Giver, L.; Gershenson, A.; Freskgard, P. O.; Arnold, F. H., Directed evolution of a thermostable esterase. *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 12809-12813.

41. Lienhard, G. E., Enzymatic catalysis and transition-state theory. *Science* **1973**, *180*, 149-154.

42. Villa, J.; Warshel, A., Energetics and dynamics of enzymatic reactions. *Journal of Chemical Physics B* **2001**, *105*, 7887-7907.

43. Tantillo, D. J.; Chen, J. G.; Houk, K. N., Theozymes and compuzymes: theoretical models for biological catalysis. *Curr. Opin. Chem. Biol.* **1998**, *2*, 743-750.

44. Zanghellini, A.; Jiang, L.; Wollacott, A.; Cheng, G.; Meiler, J.; Althoff, E.; Rothlisberger, D.; Baker, D., New algorithms and an in silico benchmark for computational enzyme design. *Protein Sci.* **2006**, *15*, 2785-2794.

45. Casey, M. L.; Kemp, D. S.; Paul, K. G.; Cox, D. D., The physical organic chemistry of benzisoxazoles. I. The mechanism of the base-catalyzed decomposition of benzisoxazoles. *J. Org. Chem.* **1973**, *38*, 2294-2301.

46. Kemp, D. S.; Casey, M. L., Physical organic chemistry of benzisoxazoles. II. Linearity of the Bronsted free energy relationship for the base-catalyzed decomposition of benzisoxazoles. *J. Am. Chem. Soc.* **1973**, *95*, 6670-6680.

47. Hollfelder, F.; Kirby, A. J.; Tawfik, D. S.; Kikuchi, K.; Hilvert, D., Characterization of proton-transfer catalysis by serum albumins. *J. Am. Chem. Soc.* **2000**, *122*, 1022-1029.

48. Hu, Y.; Kouk, K. N.; Kikuchi, K.; Hotta, K.; Hilvert, D., Nonspecific medium effects versus specific group positioning in the antibody and albumin catalysis of the base-promoted ring-opening reactions of benzisoxazoles. *J. Am. Chem. Soc.* **2004**, *126*, 8197-8205.

49. Hilvert, D.; Seebeck, F. P., Positional ordering of reacting groups contributes significantly to the efficiency of proton transfer at an antibody active site. *J. Am. Chem. Soc.* **2005**, *127*, 1307-1312.

50. Thorn, S. N.; Daniels, R. G.; Auditor, M. T. M.; Hilvert, D., Large rate accelerations in antibody catalysis by strategic use of haptenic charge. *Nature* **1995**, *373*, 228-230.

51. Mueller, R.; Debler, E. W.; Steinmann, M.; Seebeck, F. P.; Wilson, I. A.; Hilvert, D., Bifunctional catalysis of proton transfer at an antibody active site. *J. Am. Chem. Soc.* **2007**, *129*, 460-461.

52. Debler, E. W.; Ito, S.; Seebeck, F. P.; Heine, A.; Hilvert, D.; Wilson, I. A., Structural origins of efficient proton abstraction from carbon by a catalytic antibody. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 4984-4989.

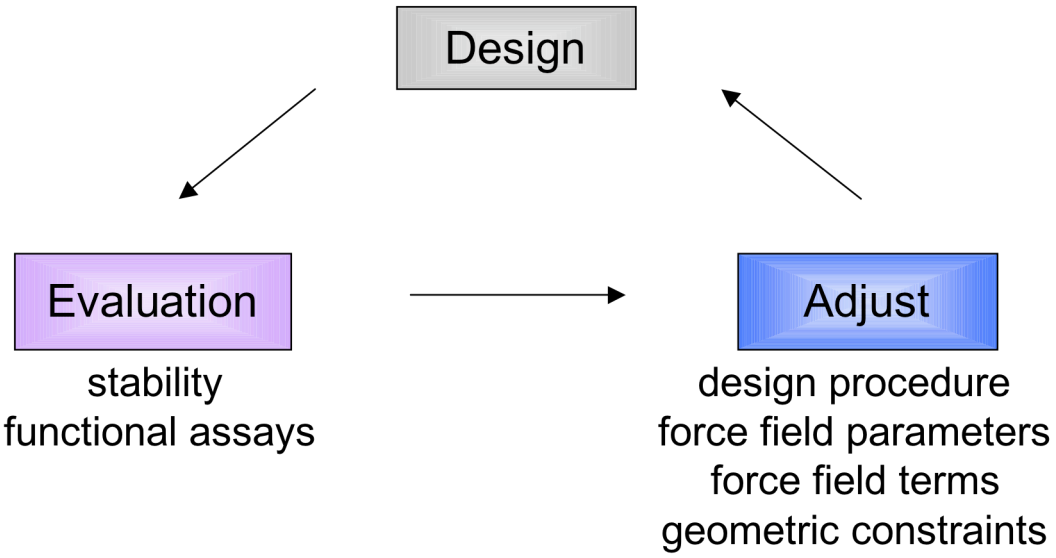


Figure 1-1. The protein design cycle. After design, the sequences predicted by the algorithm are synthesized and experimentally evaluated for their desired characteristics. The correlation of experiment and theory is used to adjust the design procedure for future designs. Adapted from Dahiyat *et al.* 1996.[12]

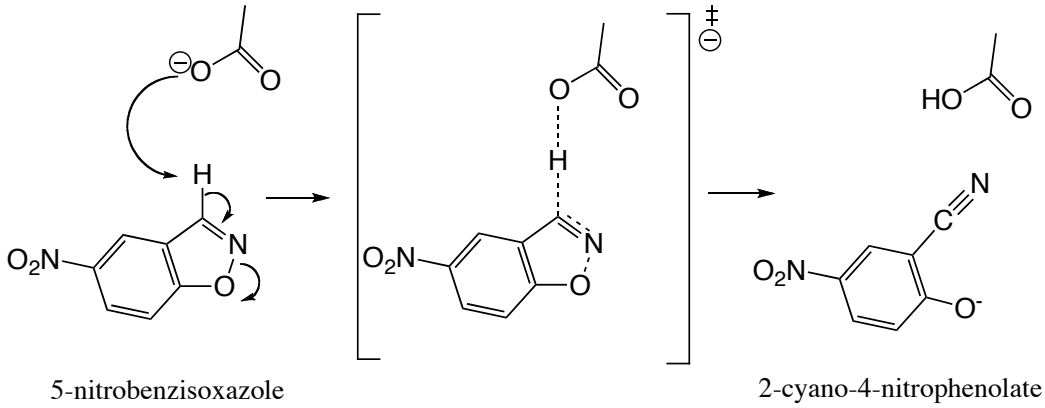


Figure 1-2. The Kemp elimination of 5-nitrobenzisoxazole.

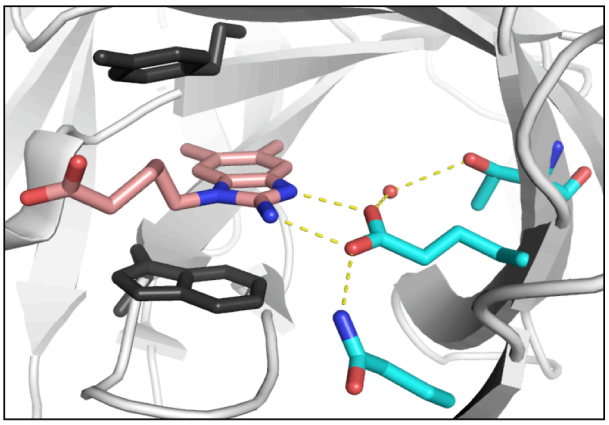


Figure 1-3. Kemp elimination catalytic antibody 34E4.[52] The co-crystallized hapten is shown in pink, the general base is shown in cyan along with two supporting contacts. Hydrogen bonds are indicated with dotted lines.

Chapter II

Combinatorial methods for small molecule placement in computational enzyme design

The text of this chapter was adapted from a manuscript coauthored with Jonathan K. Lassila, Benjamin D. Allen, and Stephen L. Mayo.

Lassila, J.K., Privett, H.K., Allen, B.D., and Mayo, S.L. (2006) Combinatorial methods for small molecule placement in computational enzyme design. *Proceedings of the National Academy of Sciences USA 103,* 16710-16715.

Abstract

The incorporation of small molecule transition state structures into protein design calculations poses special challenges because of the need to represent the added translational, rotational, and conformational freedoms within an already difficult optimization problem. Successful approaches to computational enzyme design have focused on catalytic sidechain contacts to guide placement of small molecules in active sites. We describe a process for modeling small molecules in enzyme design calculations that extends previously described methods, allowing favorable small molecule positions and conformations to be explored simultaneously with sequence optimization. Because all current computational enzyme design methods rely heavily on sampling of possible active site geometries from discrete conformational states, we tested the effects of discretization parameters on calculation results. Rotational and translational step sizes as well as sidechain library types were varied in a series of computational tests designed to

identify native-like binding contacts in three natural systems. We find that conformational parameters, especially the type of rotamer library used, significantly affect the ability of design calculations to recover native binding-site geometries. We describe the construction and use of a crystallographic conformer library, and find that it more reliably captures active-site geometries than traditional rotamer libraries in the systems tested.

Introduction

As catalysts, enzymes offer advantageous properties including dramatic rate enhancements, complete control over absolute stereochemistry, and nontoxic biodegradation. Yet a fundamental limiting factor in the use of enzymes for chemical synthesis, bioremediation, therapeutics, and other applications is the availability of enzymes with the required activities, specificities, and tolerances to reaction conditions. It is therefore a major goal of computational protein design to be able to reliably create completely new protein catalysts with specific properties on demand.

A catalyst by definition must reduce the energy barrier for formation of the transition state. To design transition-state-stabilizing interactions, computational protein design groups have incorporated transition-state or high-energy intermediate state structures into design calculations. These efforts have yielded experimentally verified new catalytic proteins.[1,2] However, substantial challenges still prevent routine or reliable design of enzymes. One major challenge is in finding energy functions that are fast enough for large calculations but that still provide informative approximations of electrostatic and desolvation effects in the protein environment.[3,4] This paper focuses on

another fundamental challenge, the need to represent the large translational, rotational, and conformational freedoms of a small molecule within already astronomically large sequence design calculations.

Here we define protein design as the selection of amino acid sequences such that the resulting protein occupies a given three-dimensional fold and has desired functional properties. Earlier experiments sought to redesign full protein sequences or confer increased thermostability,[5,6] but newer work has successfully introduced other properties, including catalytic activity, conformational specificity, ligand affinity, and even novel protein folds.[1,2,7-9] In these examples, sidechain placement algorithms were used to select from a set of discrete, probable sidechain rotamers using energy functions tuned to produce thermostable proteins. These calculations represent difficult optimization problems[10] and they can also be large—a sample calculation performed on a typical enzyme active site yields more than 10^{65} possible sequence combinations, even when excluding movements of the small molecule.

The computational demands of sequence selection prevent ligand positioning using standard docking procedures, which often approximate or neglect sidechain flexibility.[11] Approaches developed specifically for the purpose of enzyme and binding site design have introduced other schemes to limit the calculation size. Looger *et al.* used stationary, inflexible ligand poses in a large number of individual protein design calculations and demonstrated experimentally that several of the resulting proteins had high ligand affinity.[8] Lilien *et al.* reported and experimentally validated an ensemble-based method that allows ligand translation and rotation simultaneously with sidechain optimization but only permits mutation of two or three amino acid positions at a time.[12]

Chakrabarti *et al*. described a method for sequence design that neglects conformational and positional ligand flexibility and has not been experimentally tested.[13,14]

To design new enzyme active sites, a ligand placement method must be able to select side chains in many positions and must consider rotational, translational, and conformational freedom of the small molecule. The new catalytic proteins of Bolon and Mayo[1] and Dwyer *et al*[2] (this article, by Dwyer *et al*. has been retracted since the publication of this manuscript[40]) were designed by treating high-energy-state structures of the reacting molecules as extensions of contacting amino acid sidechain rotamers. In the latter case, a two-step procedure was utilized, where ligands, anchoring side chains, and other catalytic side chains were placed through a geometric screening procedure and surrounding side chains were designed in a second step.[2,15] We have developed a process for ligand placement in computational protein design calculations that expands upon previous work and that allows ligand rotation, translation, and conformational freedom to be explored combinatorially within the sequence design calculation itself. The implementation of ligand placement procedures within the context of the pairwise-decomposable protein design framework makes it possible to use a single energy function that can be parameterized as needed to reproduce experimental data.

We tested both a simple rotational and translational process for ligand placement as well as the previously used targeted ligand placement approach. A contact-based screening method is described that allows selection of ligand positions and confomations compatible with catalytic contacts. Test calculations in three systems, *E. coli* chorismate mutase, *S. cerevisiae* triosephosphate isomerase, and *S. avidinii* streptavidin, suggest that the success of ligand placement procedures can be quite sensitive to conformational

sampling parameters, including rotational and translational step sizes and the types of rotamer libraries used. We evaluated the efficacy of two standard rotamer libraries and two crystallographic conformer libraries. Traditional rotamers are constructed from canonical χ angles determined by statistical analysis of the Protein Data Bank,[16-18] whereas conformers have Cartesian coordinates taken directly from high-resolution structures. Conformer libraries may allow more accurate modeling because they are not limited to ideal geometries and their sizes can be tuned more easily and naturally.[19,20] In our tests, a backbone-independent conformer library recovered wild-type-like active site geometries more successfully than the other libraries, despite smaller size.

Results and Discussion

We have implemented and tested a process for incorporation of small molecules into computational protein design calculations. The procedure is general and may be used to place ground-state ligands or transition-state structures. It is also amenable to multistate design methods that seek to explicitly reflect the energy difference between reactant and transition states or between alternative ligands.

General Calculation Procedure

Each ligand placement calculation comprised five steps. In the first step, a large number of discrete variations of ligand coordinates was created. Initial sets of orientations were created by one of two methods, either simple rotation and translation or a targeted placement approach, both of which are discussed in more detail in subsequent sections. In the tests described here, each set of ligand variations contained 10^6-10^9

members, reflecting rotational and translational movement as well as internal conformational flexibility.

Next, the large number of substrate orientations was reduced to a manageable number (< ~ 20,000) using both a simple hard-sphere steric potential to check for backbone clashes and a set of user-defined geometric criteria for sidechain/ligand contacts. In this work, geometric criteria were defined to reflect the distances, angles, and torsions characteristic of important catalytic contacts observed in the crystal structures (Figure 2-1). In designing an enzyme with no naturally existing precedent, ideal contact geometries would be based on chemical intuition and/or quantum mechanical calculations. The geometric criteria were applied as follows. For every ligand variation, each of the geometric criteria was tested for satisfaction by contacts from any possible amino acid sidechain conformation in all designed protein positions. If a ligand variation was not able to make at least one of each type of user-specified contact, that ligand variation was discarded from the set. After geometric and steric pruning, the ligand variations remaining were only those theoretically capable of making each of the user-specified contacts.

In the third step, pairwise energies for all sidechain/sidechain, sidechain/backbone, backbone/ligand, and sidechain/ligand interactions were calculated using the full force field. In our work, this normally includes a scaled van der Waals term,[21] hydrogen-bonding and electrostatic terms,[22] and a solvation potential.[23,24]

The fourth step is an optional energy biasing that favors sidechain/ligand contacts deemed necessary for catalysis or binding. This energy biasing step helps to overcome the shortcomings of molecular mechanics energy functions, as well as the inherent

limitation of treating a multi-state design problem—*differential* stabilization of transition state relative to substrate in protein versus solution—using single-state design algorithms. As methods for modeling electrostatics and solvation and for designing over multiple states improve, the need for this biasing step should be reduced. Previous work utilized selective application of solvation energy[1] or an additional search algorithm step[8] for the same purpose. We favor the use of adjustable bias energies that can be tailored for specific purposes and investigated as a design variable.

To implement the bias, user-specified energies were added or subtracted from pairwise sidechain/ligand interaction energies. We use the energy bias under two regimes, one for normal design calculations and another for rapid assessment of catalytic residue arrangements within a protein scaffold. In normal design calculations, a small energy benefit is simply applied to favor specified types of sidechain/ligand contacts. Alternatively, to quickly identify potential catalytic residues, exaggerated energetic benefits and penalties are applied together. A very large energy benefit is given for desired types of pairwise interactions (100 kcal/mol was used in the test cases reported here). An even larger energy penalty (10,000 kcal/mol here) is applied to all other pairwise sidechain/ligand interactions, except when the side chain is alanine or glycine. In other words, the energy penalty forces all designed side chains to alanine or glycine unless they participate in user-specified catalytic contacts with the ligand. Although this process clearly does not yield physically relevant energetics, it offers a useful tool to investigate the catalytic conformational space within a binding pocket. The tests performed here to study the effect of sampling parameters on calculation results took advantage of this second approach. Calculations performed to demonstrate sequence

selection utilized the normal design approach of applying a simple energy benefit to catalytic contacts.

Finally, in the fifth step, optimal sequences were identified using the FASTER[25,26] or HERO[27] search methods. In the test cases described here, the result reported is the lowest-energy sequence with the maximal number of specified contacts.

Rotation-Translation Search

Simple rotation and translation can be used to fill the active site with an initial set of ligand variations in the first step of the process described. Because discrete steps must be used to rotate and translate the ligand, we evaluated the sensitivity of the calculation results to rotational and translational step sizes. A series of calculations was performed using an alanine-containing active-site background, as discussed in step 4 above. We first tested different rotational step sizes using the crystallographic translational starting position with three initial random rotations. Backbone-dependent and backbone-independent rotamer and conformer libraries were tested. Each sidechain library was tested with and without inclusion of the specific crystallographic sidechain rotamers from the structure under examination.

As seen in Table 2-1, the results of these calculations (in terms of both RMSD relative to crystallographic position and number of wild-type contacts) were strongly dependent on the both the rotational step size and the rotamer library used. In the case of chorismate mutase, only the backbone-independent conformer library was able to find nativelike geometry and contacts. Figure 2-2 shows results from this library with the 5° step size. When the crystallographic rotamers were included in the calculation, however,

all four libraries returned native-like results. It should be noted that none of the three test case structures were included in the set of structures used to create the conformer libraries. The backbone-independent conformer library appeared the most consistently successful with the other two test cases as well, although it showed strong dependence on rotational step size in streptavidin.

Next, we tested various combinations of rotational and translational step sizes starting from random initial ligand positions and using only the backbone-independent conformer library (Figure 2-3; Table 2-2). The crystallographic rotamers from the structures under investigation were not included in these calculations. The results show that, subject to the constraints imposed by the geometries defined in the pruning step and the biasing step, more than one combination of rotational and translational step size is viable for each test case and the sensitivity of the result to step size varies among the test cases.

The rotation/translation tests were performed using three initial random starting positions for each system. The starting positions were created by randomly rotating and translating the ligand within a 1 $Å^3$ box around the ligand centroid (or the centroid of the bicyclic ring system in biotin). Using the same atom comparisons as described in the tables, the nine initial positions had RMSDs relative to crystallographic positions of between 2.1 Å and 4.5 Å, with an average of 3.2 Å. These tests do not provide full, unbiased searches of the active sites. Full active site searches could be conducted using this method by performing separate calculations for grid points distributed evenly through the active site. Given the time required to perform these smaller calculations (Table 2-2), searching an entire active site using rotational and translational perturbations

would be computationally expensive. For example, examining a 3.6 x 3.6 x 3.6 Å grid using the 10° and 0.3 Å step sizes would require an estimated 324 hours on a 16-processor cluster for placement of ligands and catalytic side chains in the chorismate mutase active site. Thus, for initial positioning of a ligand within an active site, rotational and translational placement is inefficient. However, the ability to adjust small molecule position and conformation simultaneously with sidechain optimization should be extremely valuable for refining an initial position identified from a coarser search method.

Targeted Ligand Placement

A second approach places the small molecule with reference to a contacting side chain (Figure 2-4). In this approach, one or more small molecule variations are placed for every rotamer of the selected contacting side chain in every putative active-site position. This process has the advantage that ligand poses are targeted more efficiently to orientations that are able to make productive sidechain contacts. Previous computational enzyme design work utilized similar approaches.[1,2] In contrast to previous methods, however, our procedure does not maintain any association between the targeting rotamer and the small molecule—once the set of ligand conformations and orientations is constructed in step 1, the ligand variations are all subjected to pruning, pairwise energy calculations, and optimization as independent entities in the calculation. An implication of this procedure is that a ligand may engage in a catalytic contact with a rotamer, amino acid, or protein position that differs from those of the sidechain rotamer that was originally used to place that ligand.

We tested the effect of four types of sidechain libraries on the ability of a targeted placement process to find wild-type-like ligand positions and contacts. For the three test cases, the following sidechain contacts were used to anchor the ligand: chorismate mutase, C11 carboxylate to arginine; streptavidin, N1 to aspartate; triosephosphate isomerase, O2 and O3 to histidine. For each contact type, variations were allowed in the geometry of the contact, including the contacting atoms (NH1-NH2 versus NE-NH1 for arginine) and variations in defined distances, angles, and dihedrals of the contact.

As with the rotational and translational search, success in achieving native active-site conformations was highly dependent on the sidechain library used (Table 2-3). Only the backbone-independent conformer library yielded results for all three test cases that were comparable to those with crystallographic rotamers included. Using that library, all three systems returned all wild-type contacts with low ligand RMSD relative to the crystallographic position. As with the rotation/translation search, the chorismate mutase case showed the strongest sensitivity to rotamer library. Inspection of the structures revealed that an arginine side chain (Arg 28) occupies a conformation in the inhibitor-bound, active enzyme structure that was not well approximated in the other rotamer libraries.

The targeted placement approach allowed a thorough and directed search of active-site conformational space, including between 10^6 and 10^9 small molecule orientations and conformations spread throughout the active site. In contrast to the rotation/translation method, a full active-site search took between one and eighteen hours to complete using the backbone-independent conformer library and no initial starting position was required. This method offers an efficient first step for defining active-site

geometry in a new protein scaffold. One shortcoming is that it may be difficult to sample the many geometrical variations of a flexible hydrogen-bonding interaction. For example, the 972 variations in guanidino-carboxylate contact geometry sampled in the chorismate mutase case are probably adequate to reflect flexibility in this relatively rigid dual hydrogen-bonding interaction. A less-restrained interaction, however, such as a serine hydrogen bonding with a sterically unrestricted ligand carbonyl oxygen, results in a compromise between maintaining a manageable calculation size and modeling contact flexibility. One solution is to use a targeted method to find an initial ligand position within the binding site and then, in a second calculation, optimize both active-site packing and fine rotational and translational placement of the ligand.

Sequence Design

The computational tests described in the previous sections were designed to evaluate the effects of calculation parameters on recovery of native enzyme geometries, and the design of active-site residues was limited to catalytic side chains. However, the general procedure described here is equally amenable to full active-site design calculations.

In previously published work, 18 active site residues of *E. coli* chorismate mutase were redesigned simultaneously with rotational and translational relaxation of the transition-state structure from the starting crystallographic position.[28] The six predicted mutations were experimentally investigated and some were found to confer increased catalytic efficiency[28] or thermostability. A detrimental mutation predicted in the study underscored the importance of continued work on energy functions. In the calculation that motivated this experimental work, the initial starting position of the small molecule

was taken from the crystal structure and a limited degree of rotational and translational optimization was employed.

We performed a test calculation to demonstrate that small molecules can be placed simultaneously with full active-site sidechain optimization, without reference to any known starting position. In a sample calculation using *E. coli* chorismate mutase, the targeted placement method was used to identify 10^7 small molecule variations. In this example, after the geometric pruning step and elimination of variants with backbone steric clashes, 155 small molecule variations remained. These variants were evaluated combinatorially with ten different side chain identities in twelve active-site positions. Using FASTER for optimization, the calculation took approximately 9 hours to complete on a 16-processor cluster with about 70% of the total calculation time consumed in calculating a surface-area-based solvation term.

Conclusions

The described procedures allow the incorporation of small molecule placement directly into sequence design calculations. The test calculations performed suggest that the results of computational enzyme design processes can be quite sensitive to calculation parameters including the rotamer library used and the coarseness of ligand positioning. These results emphasize that the conformational space of a calculation must be explored before meaningful conclusions can be reached about energy functions.

Given that we still have much to learn about the complex relationship between protein structure and catalytic activity,[29,30] luck and choice of system may continue to play a role in the success of *de novo* computational enzyme design efforts for some time.

However, the power of computational enzyme design to stringently evaluate our understanding of the energetics of catalysis should not be overlooked. Experimental feedback gained from both successful and unsuccessful designs will make it possible to critically examine energy functions for modeling active sites. Employing quality transition-state structures derived from *ab initio* calculations and experimental evidence will help computational design experiments to provide more meaningful information about the effectiveness of energy functions. The use of large sidechain structural libraries and fine movements of transition-state structures will help to reduce errors from conformational sampling. Backbone relaxation and multi-state design will offer other important tools to improve the value of design calculations. Finally, the construction of gene libraries or large numbers of computationally designed variants has great potential for overcoming the shortcomings of enzyme design models,[31] but results from these experiments will be most useful for furthering our understanding of catalysis and design if both active and inactive variants are reported. By critically evaluating current methods for computational enzyme design, we will move closer to a deeper and more practically useful understanding of the sequence determinants of enzyme activity in the future.

Methods

Structures and charges

PDB files were used without minimization (*E. coli* chorismate mutase,[32] 1ecm; *S. avidinii* streptavidin,[33] 1mk5; *S. cerevisiae* triosephosphate isomerase,[34] 1ney). Hydrogens were added with REDUCE.[35]

A library of ligand internal conformations was created for each system as follows. Chorismate mutase: An HF/6-31G* *ab initio* transition-state structure[36] was used with only one variation—the O4 hydroxyl proton was allowed to occupy three positions, 60°, 180°, and -35°, defined by the H-C-O-H dihedral angle. The minima in a torsional profile at the HF/6-31G* level were at approximately 180° and -35°, and 60° was included as an option because hydrogen-bonding patterns in chorismate mutases from other species suggested population of that region of torsional space. Streptavidin: Four rotatable bonds in biotin were allowed to occupy three positions each (60°, -60°, 180° for sp3-sp3 bonds and 30°, 90°, 150° for the symmetric carboxylate group). Thirty-four conformations were excluded because of high internal energy calculated using the van der Waals component of the DREIDING force field.[37] Triosephosphate isomerase: The pdb structure used was the Michaelis complex with the substrate dihydroxyacetone phosphate. In ground-state dihydroxyacetone phosphate, two rotatable bonds (defined by the P-O-C-C and C-C-O-H dihedral angle) were allowed to occupy three positions each (60°, -60°, 180°). Three conformations were excluded because of high internal DREIDING van der Waals energy.

Ligand atomic charges were obtained by fitting charges to electrostatic potential from HF/6-31G* single-point energy calculations using the transition-state structure (chorismate mutase) or crystallographic ground-state structure (biotin, dihydroxyacetone

phosphate). *Ab initio* calculations and charge determinations were performed using Spartan (Wavefunction, Inc.) or Jaguar (Schrödinger, Inc.).

Sidechain rotamer libraries

Standard backbone-dependent and backbone-independent rotamer libraries were used with expansion by one standard deviation about $\chi 1$ and $\chi 2$.[17]

Crystallographic conformer libraries were prepared using coordinates from 149,813 side chains selected from 1,011 unique structures. A clustering algorithm was developed based on ideas described by Shetty *et al*[20] and is described briefly here. Every sidechain conformation from the raw data set is assigned to exactly one cluster. Each cluster is represented by the centroid, which is the member with coordinates closest to the average coordinates of all cluster members. A conformer library consists of a list of all of the cluster representatives and their coordinates. In our clustering algorithm, clusters are assigned through discrete clustering moves: *Switch* allows a single raw conformer to leave one cluster and join another; *Merge* combines two clusters into one; *Split* allows a raw conformer to start a new cluster on its own. These moves are depicted in Figure 2-5.

RMSDs between pairs of conformers are compared to determine whether or not to apply a particular move. *Switch* is applied so that each raw conformer is a member of the cluster whose centroid is closest to it. *Merge* and *Split* are applied based on the value of the clustering parameter *p*: two clusters are merged if their centroids are within *p* of each other, whereas a conformer splits off and starts a new cluster if the closest centroid of any existing cluster is farther than *p* from it. The clustering moves are applied as follows until the number of clusters converges:

1. Start with a small number of clusters (1 was used in this work), and randomly assign a single raw conformer to each as the sole member and cluster representative.
2. Assign each raw conformer in the data set to the cluster whose centroid is closest.
3. While the number of clusters is not converged:
    a. Iteratively attempt to *Merge* pairs of clusters until no cluster can be further merged.
    b. For each conformer C:
        i. Measure the distance *d* between C and the centroid of every existing cluster.
        ii. If the distance *d* to the closest cluster centroid is greater than *p*, *Split* C off as its own cluster.
        iii. Else, *Switch* C to the closest cluster.
        iv. Recompute the centroid for every cluster that has changed membership.

The algorithm allows the construction of both backbone-dependent and backbone-independent libraries to custom sizes by using clustering factor *p* to define the desired degree of similarity between independent conformers. In this work, clustering factors of 0.3 Å and 1.0 Å were used for backbone-dependent and backbone-independent rotamer libraries, respectively.

For all calculation types, conformer libraries were smaller than the standard rotamer libraries. As an example, the number of sidechain conformations for the chorismate mutase calculations described in Table 2-3 were as follows: backbone-independent rotamer, 14229; backbone-independent conformer, 5955; backbone-dependent rotamer, 7945; and backbone-dependent conformer, 5539.

Calculation parameters

All non-Gly, non-Pro residues reasonably within the natural active sites were included in calculations. Residues with any atom within a 5 Å radius from any atom in the crystallographic ligands were included, less those residues separated from the natural ligand by backbone elements and plus a few adjacent residues not within the 5 Å cutoff. The positions designed were (all in chain A unless otherwise designated): chorismate mutase, 28, 32, 35, 39, 46, 47, 48, 51, 52, 55, 81, 84, 85, 88, 7B, 11B, 14B, 18B; streptavidin, 23, 24, 25, 27, 43, 45, 46, 47, 49, 50, 79, 86, 88, 90, 92, 108, 110, 112, 128, 130; and triosephosphate isomerase, 10, 12, 95, 97, 165, 170, 211, 230.

In ligand placement test cases, designed residues were restricted to ligand-contacting residues or alanine as follows: Arg, Lys, Gln, Glu, or Ala in chorismate mutase; Ser, Asn, Tyr, Asp, or Ala in streptavidin, and Glu, His, Lys, or Ala in triosephosphate isomerase. Four calculations on triosephosphate isomerase were run as smaller component calculations, as indicated in Table 2-2, because of prohibitive size as a single calculation.

Energy functions and optimization

Energy functions included scaled van der Waals,[21] hydrogen-bonding, and electrostatic terms.[22] A surface-area-based solvation potential[23] was used in sequence design calculations but not for ligand placement, where solvation energy would have been heavily outweighed by geometric considerations. Sequences were optimized with respect to the energy function using FASTER[25,26] or HERO.[27] On occasion, a top-ranked sequence contained more than one instance of a given specified geometric contact, owing to the energy benefit applied for these contacts. In these cases, Monte Carlo[38,39] was used

to sample around the global minimum energy sequence and the top-ranked sequence with a single instance of each geometric contact was reported.

References

1. Bolon, D. N.; Mayo, S. L., Enzyme-like proteins by computational design. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 14274-14279.

2. Dwyer, M.A.; Looger, L.L.; Hellinga, H.W., Computational design of a biologically active enzyme. *Science* **2004**, 304, 1967-1971. *This article has been retracted (Dwyer *et al. Science* **2008**, *319*, 569).

3. Mendes, J.; Guerois, R.; Serrano, L., Energy estimation in protein design. *Curr. Opin. Struct. Biol.* **2002**, *12*, 441-446.

4. Vizcarra, C.L.; Mayo, S.L., Electrostatics in computational protein design. *Curr. Opin. Chem. Biol.* **2005**, *9*, 622-626.

5. Dahiyat, B.I.; Mayo, S.L., De novo protein design: Fully automated sequence selection. *Science* **1997**, *278*, 82-87.

6. Malakauskas, S.M.; Mayo, S.L., Design, structure, and stability of a hyperthermophilic protein variant. *Nat. Struct. Biol.* **1998**, *5*, 470-475.

7. Shimaoka, M.; Shifman, J.M.; Jing, H.; Takagi, J.; Mayo, S.L.; Springer, T.A., Computational design of an integrin I domain stabilized in the open high affinity conformation. *Nat. Struct. Biol.* **2000**, *7*, 674-678.

8. Looger, L. L.; Dwyer, M. A.; Smith, J. J.; Hellinga, H. W., Computational design of receptor and sensor proteins with novel functions. *Nature* **2003**, *423*, 185-190.

9. Kuhlman, B.; Dantas, G.; Ireton, G.C.; Varani, G.; Stoddard, B.L.; Baker, D., Design of a novel globular protein fold with atomic-level accuracy. *Science* **2003**, *302*, 1364-1368.

10. Pierce, N.A.; Winfree, E., Protein Design is NP-hard. *Prot. Eng.* **2002**, *15*, 779-782.

11. Taylor, R.D.; Jewsbury, P.J.; Essex, J.W., A review of protein-small molecule docking methods. *J. Comput. Aided Mol. Des.* **2002**, *16*, 151-166.

12. Lilien, R.H.; Stevens, B.W.; Anderson, A.C.; Donald, B.R., A novel ensemble-based scoring and search algorithm for protein redesign and its application to modify the substrate specificity of the gramicidin synthetase A phenylalanine adenylation enzyme. *J. Comput. Biol.* **2005**, *12*, 740-761.

13. Chakrabarti, R.; Klibanov, A.M.; Friesner, R.A., Computational prediction of native protein ligand-binding and enzyme active site sequences. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 10153-10158.

14. Chakrabarti, R.; Klibanov, A.M.; Friesner, R.A., Sequence optimization and designability of enzyme active sites. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 12035-12040.

15. Hellinga, H.W.; Richards, F.M., Construction of new ligand binding sites in proteins of known structure I. Computer-aided modeling of sites with pre-defined geometry. *J. Mol. Biol.* **1991**, *222*, 763-785.

16. Ponder, J.W.; Richards, F.M., Tertiary templates for proteins: Use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol.* **1987**, *193*, 775-791.

17. Dunbrack, R.L, Jr.; Cohen, F.E., Bayesian statistical analysis of protein sidechain rotamer preferences. *Prot. Sci.* **1997**, *6*, 1661-1681.

18. Lovell, S.C.; Word, J.M.; Richardson, J.S.; Richardson, D.C., The penultimate rotamer library. *Proteins* **2000**, *40*, 389-408.

19. Xiang, Z.; Honig, B., Extending the accuracy limits of prediction for sidechain conformations. *J. Mol. Biol.* **2001**, *311*, 421-430.

20. Shetty, R.P.; de Bakker, P.I.W.; DePristo, M.A.; Blundell, T.L., Advantages of fine-grained side chain conformer libraries. *Prot. Eng.* **2003**, *16*, 963-969.

21. Dahiyat, B.I.; Mayo, S.L., Probing the role of packing specificity in protein design. *Proc. Natl. Acad. Sci. USA* **1997**, *94*, 10172-10177.

22. Dahiyat, B.I.; Gordon, D.B.; Mayo, S.L., Automated design of the surface positions of protein helices. *Protein Sci.* **1997**, *6,* 1333-1337.

23. Street, A.G.; Mayo, S.L., Pairwise calculation of protein solvent-accessible surface areas. *Fold. Des.* **1998**, *3*, 253-258.

24. Lazaridis, T.; Karplus, M., Effective energy functions for proteins in solution. *Prot. Struct. Funct. Genet.* **1999**, *35*, 133-152.

25. Desmet, J.; Spriet, J.; Lasters, I., Fast and accurate sidechain topology and energy refinement (FASTER) as a new method for protein structure optimization. *Prot. Struct. Funct. Genet.* **2002**, *48*, 31-43.

26. Allen, B.D.; Mayo, S.L., Dramatic performance enhancements for the FASTER optimization algorithm. *J. Comput. Chem.* **2006**, *27*, 1071-1075.

27. Gordon, D.B.; Hom, G.K.; Mayo, S.L.; Pierce, N.A., Exact rotamer optimization for protein design. *J. Comput. Chem.* **2003**, *24*, 232-243.

28. Lassila, J.K., Keeffe, J. R., Oeschlaeger, P.; Mayo, S.L., Computationally designed variants of *Escherichia coli* chorismate mutase show altered catalytic activity. *Protein Eng. Des. Sel.* **2005**, *18*, 161-163.

29. Kraut, D.A.; Carroll, K.S.; Herschlag, D., Challenges in enzyme mechanism and energetics. *Annu. Rev. Biochem.* **2003**, *72*, 517-571.

30. Benkovic, S.J.; Hammes-Schiffer, S., A perspective on enzyme catalysis. *Science* **2003**, *301*, 1196-1202.

31. Bolon, D.N.; Voigt, C.A.; Mayo, S.L., *De novo* design of biocatalysts. *Curr. Opin. Chem. Biol.* **2002**, *6*, 125-129.

32. Lee, A.Y.; Karplus, P.A.; Ganem, B.; Clardy, J., Atomic structure of the buried catalytic pocket of *Escherichia coli* chorismate mutase. *J. Am. Chem. Soc.* **1995**, *117*, 3627-3628.

33. Hyre, D.E.; Le Trong, I.; Merritt, E.A.; Eccleston, J.F.; Green, N.M.; Stenkamp, R.E.; Stayton, P.S., Cooperative hydrogen bond interactions in the streptavidin-biotin system. *Protein Sci.* **2006**, *15*, 459-467.

34. Jogl, G.; Rozovsky, S.; McDermott, A.E.; Tong, L., Optimal alignment for enzymatic proton transfer: Structure of the Michaelis complex of triosephosphate isomerase at 1.2-Ångstrom resolution. *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 50-55.

35. Word, J.M.; Lovell, S.C.; Richardson, J.S.; Richardson, D.C., Asparagine and glutamine: Using hydrogen atom contacts in the choice of sidechain amide orientation. *J. Mol. Biol.* **1999**, *285*, 1735-1747.

36. Wiest, O.; Houk, K. N., On the transition state of the chorismate-prephenate rearrangement. *J. Org. Chem.* **1994**, *59*, 7582-7584.

37. Mayo, S. L.; Olafson, B. D.; Goddard, W. A., DREIDING: A generic force field for molecular simulations. *J. Phys. Chem.* **1990**, *94*, 8897-8909.

38. Metropolis, N.; Rosenbluth, A.W.; Rosenbluth, M.N.; Teller, A.H.; Teller, E. Equation of state calculations by fast computing machines. *J. Chem. Phys.* **1953**, *21*, 1087-1092.

39. Voigt, C.A.; Gordon, D.B.; Mayo, S.L., Trading accuracy for speed: A quantitative comparison of search algorithms in protein sequence design. *J. Mol. Biol.* **2000**, *299*, 789-803.

40. Dwyer, M.A.; Looger, L.L.; Hellinga, H.W., Retraction of Dwyer *et al., Science* **2008**, *304* (5679) 1967-1971. *Science 319*, 569.

Table 2-1. RMSD and number of wild-type contacts as a function of rotational step size and rotamer library[a,b]

Chorismate Mutase

	Rotational step size				
Rotamer Library[c]	30°	20°	15°	10°	5°
Conformer: bb-ind	-	-	0.61 ± 0.03 (4.0)	0.55 ± 0.05 (4.0)	0.47 ± 0.04 (4.7)
with xtal rotamers	-	-	0.61 ± 0.03 (4.0)	0.55 ± 0.05 (4.0)	0.47 ± 0.04 (4.7)
Rotamer: bb-ind	-	-	3.88 ± 0.37 (0.0)	2.88 ± 1.44 (0.0)	3.01 ± 1.61 (0.0)
with xtal rotamers	-	-	1.57 ± 1.70 (2.7)	0.51 ± 0.00 (4.0)	0.52 ± 0.01 (4.0)
Conformer: bb-dep	-	-	3.66 ± 0.11 (1.0)	3.59 ± 0.08 (1.0)	3.60 ± 0.09 (1.0)
with xtal rotamers	-	1.67 ± 1.78 (3.3)	1.57 ± 1.83 (3.7)	0.60 ± 0.08 (4.3)	0.54 ± 0.06 (5.0)
Rotamer: bb-dep	-	-	-	-	-
with xtal rotamers	-	-	-	0.49 ± 0.04 (4.3)	0.52 ± 0.01 (4.0)

Streptavidin-Biotin

	Rotational step size				
Rotamer Library[c]	30°	20°	15°	10°	5°
Conformer: bb-ind	-	-	-	-	0.27 ± 0.09 (5.0)
with xtal rotamers	-	0.24 ± 0.09 (5.0)	0.24 ± 0.07 (5.0)	0.26 ± 0.06 (5.0)	0.20 ± 0.13 (5.0)
Rotamer: bb-ind	-	-	0.77 ± 0.42 (2.3)	0.60 ± 0.14 (3.0)	0.60 ± 0.05 (2.7)
with xtal rotamers	0.37 ± 0.17 (5.0)	0.24 ± 0.09 (5.0)	0.24 ± 0.07 (5.0)	0.26 ± 0.06 (5.0)	0.30 ± 0.17 (5.0)
Conformer: bb-dep	-	-	-	0.25 ± 0.12 (5.0)	0.20 ± 0.07 (5.0)
with xtal rotamers	-	0.24 ± 0.09 (5.0)	0.24 ± 0.07 (5.0)	0.22 ± 0.03 (5.0)	0.29 ± 0.09 (4.0)
Rotamer: bb-dep	-	-	-	0.82 ± 0.28 (2.3)	0.66 ± 0.02 (3.0)
with xtal rotamers	-	0.24 ± 0.09 (5.0)	0.24 ± 0.07 (5.0)	0.26 ± 0.06 (5.0)	0.16 ± 0.06 (5.0)

Triosephosphate Isomerase

	Rotational step size				
Rotamer Library[c]	30°	20°	15°	10°	5°
Conformer: bb-ind	-	1.87 ± 1.07 (0.7)	3.59 ± 2.28 (1.0)	0.28 ± 0.07 (3.0)	0.24 ± 0.05 (3.0)
with xtal rotamers	-	1.31 ± 0.29 (1.0)	1.95 ± 2.28 (1.3)	0.27 ± 0.06 (3.0)	0.15 ± 0.02 (3.0)
Rotamer: bb-ind	5.09 ± 0.05 (0.3)	0.60 ± 0.12 (1.7)	0.55 ± 0.25 (2.3)	0.34 ± 0.04 (2.3)	0.25 ± 0.08 (3.0)
with xtal rotamers	5.06 ± 0.05 (0.3)	0.60 ± 0.12 (2.0)	0.37 ± 0.04 (3.0)	0.25 ± 0.04 (3.0)	0.15 ± 0.02 (3.0)
Conformer: bb-dep	-	-	-	-	-
with xtal rotamers	-	-	-	-	0.15 ± 0.02 (3.0)
Rotamer: bb-dep	3.28 ± 0.73 (1.7)	0.60 ± 0.12 (1.7)	0.37 ± 0.05 (2.3)	0.31 ± 0.04 (2.3)	0.25 ± 0.08 (3.0)
with xtal rotamers	3.28 ± 0.73 (2.3)	0.60 ± 0.12 (2.3)	0.37 ± 0.05 (3.0)	0.29 ± 0.03 (3.0)	0.15 ± 0.02 (3.0)

[a] Dashes indicate that required contacts were not satisfied in at least one of three trials.
[b] Values are non-hydrogen-atom RMSD in Ångstroms relative to crystallographic ligands or bicyclic ring atom RMSD relative to crystallographic ligand for biotin (i.e., the pentanoic acid moiety was not considered in biotin RMSDs). Averages and standard deviations from three random initial positions are reported. Numbers in parentheses are the number of contacts where the amino acid position was the same as in the wild-type structure, averaged over the three trials. Maximum possible number of wild-type contacts: chorismate mutase, 5; streptavidin, 5; triosephosphate isomerase, 3.
[c] bb-ind: backbone-independent, bb-dep: backbone-dependent.

Table 2-2. RMSD and number of wild-type contacts as a function of rotational and translational step sizes[a,b]

Chorismate mutase

Translational step size (Å)	Rotational step size 30°	20°	15°	10°	5°	Time (10°, hours)[c]
0.6	1.69 ± 1.54 (2.3)	2.61 ± 1.67 (1.3)	0.77 ± 0.10 (4.3)	0.73 ± 0.02 (4.0)	0.61 ± 0.06 (4.7)	3
0.5	0.91 ± 0.20 (3.7)	0.72 ± 0.07 (4.0)	0.83 ± 0.06 (3.3)	0.74 ± 0.05 (4.0)	0.60 ± 0.13 (4.3)	10
0.4	2.02 ± 1.99 (2.3)	0.60 ± 0.04 (4.7)	0.59 ± 0.13 (4.0)	0.57 ± 0.12 (4.3)	0.53 ± 0.13 (4.3)	11
0.3	1.73 ± 1.51 (2.3)	0.61 ± 0.07 (4.3)	0.62 ± 0.15 (4.3)	0.58 ± 0.07 (4.0)	0.65 ± 0.04 (4.0)	12
0.2	1.71 ± 1.53 (2.3)	0.62 ± 0.10 (4.0)	0.60 ± 0.09 (4.0)	0.54 ± 0.07 (4.0)	0.56 ± 0.05 (4.0)	33

Streptavidin-biotin

Translational step size (Å)	Rotational step size 30°	20°	15°	10°	5°	Time (10°, hours)[c]
0.6	-	1.16 ± 0.60 (3.7)	1.67 ± 1.02 (3.7)	0.88 ± 0.44 (4.3)	0.84 ± 0.48 (4.3)	5
0.5	2.05 ± 0.59 (1.7)	0.91 ± 0.44 (5.0)	0.84 ± 0.61 (5.0)	0.99 ± 0.91 (3.7)	-	18
0.4	1.32 ± 1.39 (3.7)	0.80 ± 0.09 (5.0)	0.67 ± 0.28 (5.0)	0.96 ± 0.72 (3.7)	-	19
0.3	0.63 ± 0.16 (5.0)	1.08 ± 0.49 (5.0)	0.57 ± 0.21 (5.0)	1.03 ± 0.48 (4.3)	-	18
0.2	0.60 ± 0.32 (5.0)	0.70 ± 0.34 (5.0)	0.80 ± 0.24 (5.0)	-	-	-

Triosephosphate isomerase

Translational step size (Å)	Rotational step size 30°	20°	15°	10°	5°	Time (10°, hours)[c]
0.6	3.80 ± 2.14 (0.3)	5.22 ± 0.32 (0.0)	1.29 ± 0.91 (1.3)	2.39 ± 2.54 (1.7)	2.40 ± 2.58 (2.0)	0.4
0.5	3.92 ± 1.94 (0.0)	5.64 ± 0.45 (0.3)	4.47 ± 1.45 (0.0)	1.33 ± 1.01 (1.7)	-	2
0.4	3.13 ± 1.77 (0.3)	1.96 ± 1.05 (2.0)	0.47 ± 0.24 (1.7)	0.78 ± 0.66 (3.0)	-	2
0.3	3.44 ± 1.96 (0.3)	0.59 ± 0.18 (2.0)	0.60 ± 0.29 (2.3)	0.46 ± 0.11 (3.0)	-	2
0.2	2.33 ± 1.80 (0.7)	0.68 ± 0.10 (2.3)	0.49 ± 0.12 (3.0)	0.44 ± 0.11 (3.0)	-	5

[a] Dashes indicate that required contacts were not satisfied in at least one of three trials or that the calculation was too large to complete.

[b] Values are non-hydrogen atom RMSD in Ångstroms relative to crystallographic ligands or bicyclic atom RMSD relative to crystallographic ligand for biotin (i.e. the pentanoic acid moiety was not considered in biotin RMSDs). Averages and standard deviations from three random initial positions are reported. Numbers in parentheses are the number of contacts where the amino acid position was the same as in the wild-type structure, averaged over the three trials. Maximum possible number of wild-type contacts: chorismate mutase, 5; streptavidin, 5; triosephosphate isomerase, 3.

[c] Wall clock time; calculations performed on a 16-processor cluster.

Table 2-3. Results from targeted placement procedure as a function of rotamer library.

Chorismate mutase

Rotamer library[a]	log(initial ligand variations)	RMSD (Å)[b] (WT contacts)	Time (hours)[c]
Conformer: bb-ind	7.88	0.60 (5)	16
with xtal rotamers	7.88	0.68 (3)	18
Rotamer: bb-ind	8.18	3.61 (0)	51
with xtal rotamers	8.18	0.66 (4)	62
Conformer: bb-dep	7.64	3.62 (1)	8
with xtal rotamers	7.64	0.68 (4)	9
Rotamer: bb-dep	7.76	2.31 (1)	14
with xtal rotamers	7.76	0.66 (4)	16

Streptavidin-biotin

Rotamer library[a]	log(initial ligand variations)	RMSD (Å)[b] (WT contacts)	Time (hours)[c]
Conformer: bb-ind	7.07	0.64 (5)	1.4
with xtal rotamers	7.07	0.64 (5)	1.4
Rotamer: bb-ind	7.20	0.54 (4)	3.5
with xtal rotamers	7.20	0.34 (4)	3.4
Conformer: bb-dep	6.35	0.37 (5)	0.2
with xtal rotamers	6.35	0.54 (4)	0.2
Rotamer: bb-dep	7.17	3.50 (0)	2.6
with xtal rotamers	7.17	0.19 (5)	2.8

Triosephosphate isomerase

Rotamer library[a]	log(initial ligand variations)	RMSD (Å)[b] (WT contacts)	Time (hours)[c]
Conformer: bb-ind	7.31	0.49 (3)	1.3
with xtal rotamers	7.31	0.49 (3)	1.3
Rotamer: bb-ind	7.78	0.46 (3)	8.7[d]
with xtal rotamers	7.78	0.46 (3)	8.7[d]
Conformer: bb-dep	6.82	7.51 (0)	0.3
with xtal rotamers	6.82	0.78 (3)	0.3
Rotamer: bb-dep	7.58	0.51 (3)	4.3[d]
with xtal rotamers	7.58	0.51 (3)	4.9[d]

[a] bb-ind, backbone-independent; bb-dep, backbone-dependent.
[b] RMSDs calculated as described in Table 2-1. Maximum possible number of wild-type contacts: chorismate mutase, 5; streptavidin, 5; triosephosphate isomerase, 3.
[c] Wall clock time; calculations performed on a 16-processor cluster.
[d] Calculation was performed as a series of smaller calculations.

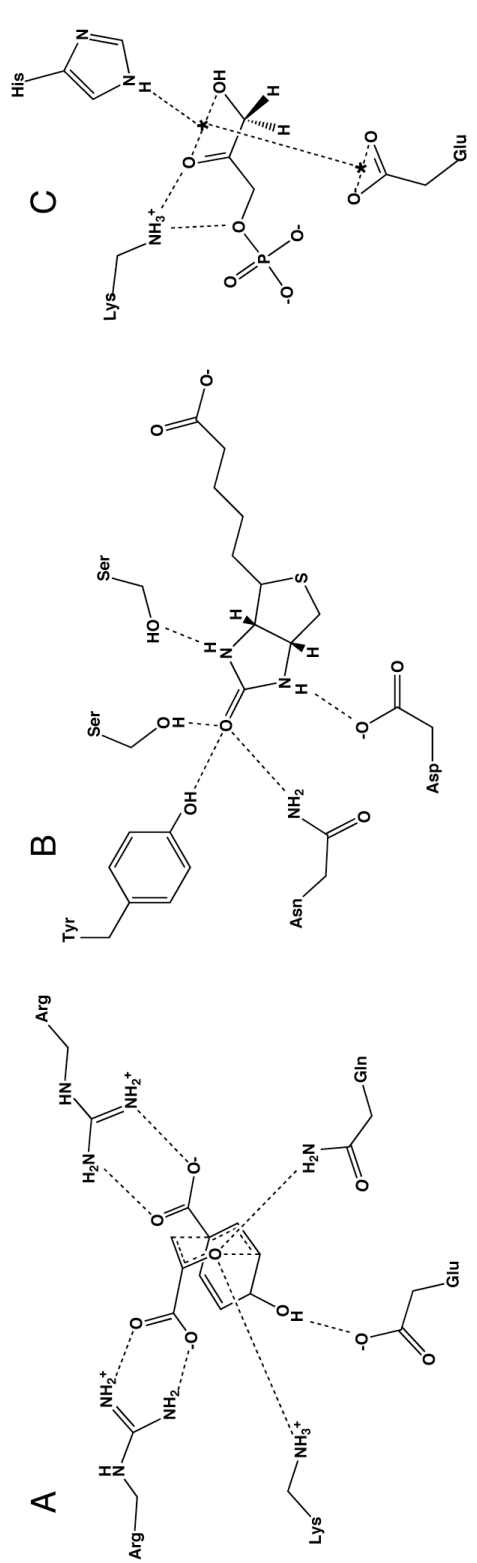


Figure 2-1. Contact geometries specified in small molecule pruning step. Ranges of distances, angles, and torsions were allowed that included the crystallographic geometries. (*A*) Chorismate mutase. (*B*) Biotin in streptavidin. (*C*) Triosephosphate isomerase Michaelis complex, modeled using an approach similar to that of Reference 2. Asterisks indicate pseudoatoms used in geometry definitions.

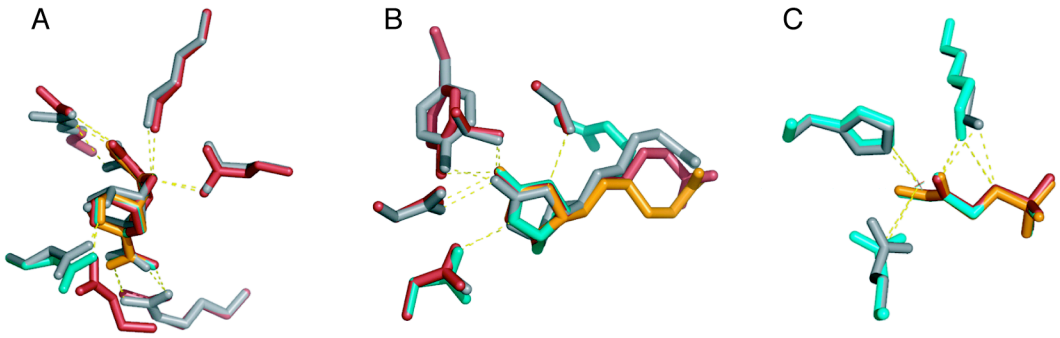


Figure 2-2. Sample results from test calculations presented in Table 2-1. Crystallographic side chains and ligands are shown in gray. Results from three trials using different initial random rotational positions are shown in red, teal, and orange. In cases where three colors are not visible, the selected rotamers from two or more calculations were identical. Results are shown from calculations with 5° rotation and the backbone-independent conformer library. (*A*) Chorismate mutase. An alternate backbone position was chosen for a glutamate-hydroxyl contact in one trial (red side chain, lower left). (*B*) Biotin in streptavidin. Note that the biotin pentanoic acid moiety samples different conformations in the calculation and the surrounding side chains were not designed. (*C*) Triosephosphate isomerase.

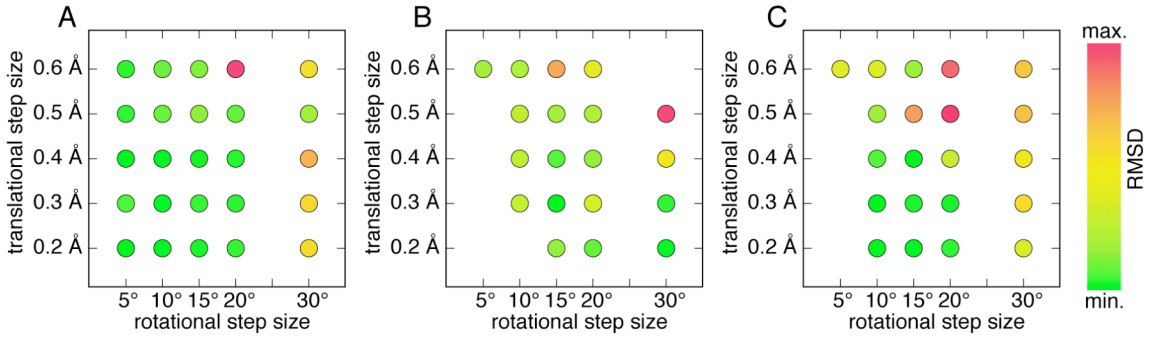


Figure 2-3. Effect of rotational and translational step sizes. Each spot represents the average of three trials with initial random starting positions. Missing points indicate that one or more trials could not identify wild-type-like contacts or else that the calculation was prohibitively large; no calculations were performed using a 25° rotational step size. Colors indicate non-hydrogen atom RMSD as described in the tables. (*A*) Chorismate mutase (min., 0.53 Å; max., 2.61 Å), (*B*) Streptavidin-biotin (min., 0.57 Å; max., 2.05 Å), (*C*) triosephosphate isomerase (min., 0.44 Å; max., 5.64 Å).

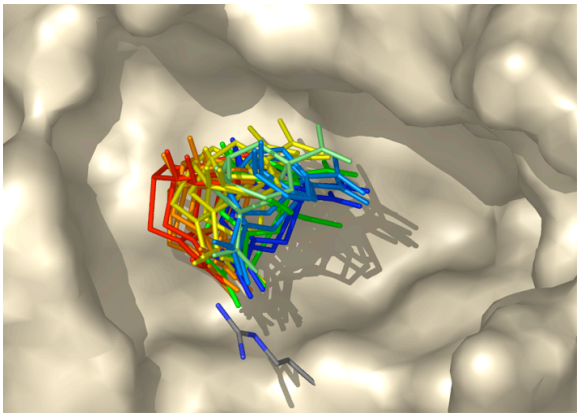


Figure 2-4. Targeted placement procedure. For a given side chain rotamer, small molecule ligands are placed such that they are able to meet specified geometric criteria. This is repeated for every possible conformation of the amino acid at every designed position. Shown is a subset of orientations of a chorismate mutase transition-state structure in contact with one conformation of arginine.

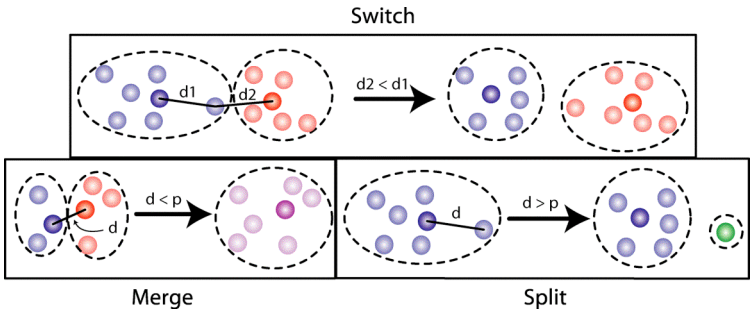


Figure 2-5. **The three clustering moves are illustrated by showing the state of a sample system before and after the move is performed.** Each dot represents a single sidechain conformation taken from the PDB. Distances represent sidechain RMSDs between pairs of conformers. Dots sequestered together by a dashed line and colored the same are members of the same cluster. Darker-colored dots denote cluster representatives.

Chapter III

Towards the computational design of a Kemp elimination enzyme and investigation of an inactive design

A majority of the experimental work described in this chapter was performed in the Mayo laboratory. Dr. Leonard Thomas of the Caltech Molecular Observatory solved and refined the crystal structure of HG-1 and Gert Kiss of the Houk lab (UCLA) carried out the molecular dynamics simulations.

Abstract

A general method for the *de novo* design of enzymatic activity has long been a major goal in computational protein design. Previous *de novo* enzyme design efforts placed little, if any, emphasis on the analysis of inactive designs, and the highly active designed enzymes that were recently reported were identified through the synthesis and screening of a large number of designs. Here, we describe an iterative method for the computational design of enzymes that combines computational enzyme design techniques with molecular dynamics simulations and crystallographic analysis of inactive designs. We used these methods to design and analyze an inactive design for the catalysis of the Kemp elimination, which helped us to understand the causes of our design's inactivity, including excessive solvent exposure and active site flexibility. We then applied this information to adjust our design procedure for future rounds of design.

Introduction

In early incarnations of computational protein design methodology, the strategy for software development was put forth in terms of the so called "protein design cycle" (Figure 1-1) where the experimental evaluation of an initial design is used to inform the adjustment of the design process for further rounds of designs.[1,2] These steps would ideally be continued iteratively until the protein sequences predicted by the algorithm have the desired characteristics. However, there is little evidence that this strategy has actually been used outside of force-field parameterization purposes,[1,3-5] and most inactive proteins from subsequent design efforts were likely discarded without comment or further investigation into the cause of the inactivity. This is unfortunate because little information can be gained from such an approach to protein design. Without detailed analysis of failed designs, flaws in the design procedure cannot be identified and adjusted to produce designed proteins with the desired characteristics.

Despite the lack of the type of systematic approach described above, the field of computational enzyme design has seen some success in the past ten years. The first successful computational introduction of *de novo* enzymatic activity into an inert scaffold was a protozyme, which catalyzed the hydrolysis of an activated ester with a very small rate enhancement (k_{cat}/k_{uncat} of about 10^2).[6] Additional incremental advances were made with the design of metalloenzymes and computationally designed changes in enzyme specificity.[7-9] In 2008, a breakthrough in *de novo* computational enzyme design came when the Rosetta protein design software was used to introduce catalytic activity for two distinct chemical reactions into a variety of inert scaffolds.[10,11] The reaction rates of these designed enzymes are impressive, but still well below those of natural enzymes.

Directed evolution was subsequently used to substantially increase the k_{cat}/K_m and the rate enhancement of one of the active designs.[11]

The synthesis and experimental evaluation of each potential enzyme is expensive both in terms of time and money. In the case of the enzymes designed by Röthlisberger *et al.*, over 70 individual designs, which were predicted to be active by their protein design methodology, were screened in order to identify 8 active enzymes.[11] While their methodology was successful in terms of producing active enzymes, the necessity of a "shotgun" approach suggests an incomplete understanding of some details of the enzymatic system and/or inaccurate modeling by the protein design algorithm.

Here, we focused on the development of a single design to rigorously test our understanding of enzymatic catalysis and the applicability of the protein design cycle to *de novo* enzyme design problems. To this end, we used our protein design software, ORBIT, in an attempt to design a catalyst for a model chemical system, the Kemp elimination (Figure 1-2). These methods resulted in a design called HG-1, which showed no Kemp elimination activity. In collaboration with the Houk lab at UCLA and the Molecular Observatory at Caltech, molecular dynamics simulations and crystallographic techniques were used to identify the possible sources of the inactivity, and a plan for future redesigns of the protein to create an active Kemp elimination enzyme was outlined.

Materials and Methods

Transitions state (TS) structure

Our TS structure was previously calculated *ab initio*[12] using the partial atomic charges used listed in Table 3-1. To assist with defining geometric constraints in the calculations, a pseudoatom named PSA with no charge and a negligible radius was added 0.01 Å from H3 in the plane of the TS. The final TS structure is shown in Figure 3-1 along with the atom names.

Scaffold selection

For the initial design, the xylanase from the thermophilic fungus *T. aurantiacus* (TAX) was chosen as the scaffold. TAX has a $(\alpha/\beta)_8$-barrel fold with a large, solvent-exposed active site. The crystal structure (PDB code: 1GOR) is 1.7 Å resolution with xylobiose bound in the active site.[13] The structure from the PDB was used without minimization and hydrogens were added with Molprobity.[14]

Active site search

All non-Pro positions within 5 Å of xylobiose in 1GOR were designated as design positions (residues 50, 83, 84, 87, 90, 130, 172, 207, 209, 267, and 275). Catalytic positions were defined as all of the carboxylate residues within 5 Å of xylobiose in 1GOR (E46, E131, and E237). All design positions were allowed to sample all conformations of Gly, His, Phe, Trp, Ser, Thr, and Tyr. Catalytic positions were allowed to sample all conformations of Glu and Asp. A backbone-independent conformer library was used to approximate sidechain flexibility.[15] As described in Chapter II, a library of

TS poses was generated in the active site by targeted ligand placement in which a defined set of ligand structures is built with respect to each conformation of the general base (Asp/Glu) at each catalytic position. The contact geometries that represent this set of ligand poses are listed in Table 3-2. After the generation of ligand poses, interaction energies were calculated, and TS-sidechain interaction energies were then biased to favor interactions that satisfy the geometric constraints in Table 3-3. An additional energy-biasing step was carried out to favor the hydrogen bond interactions to the base that were required (Table 3-4). The energy calculation and energy biasing steps are described in detail in Chapter II.

Active site repacking

For the repacking calculation, the initial TS position was taken from the results of the active site search. From this initial position, the TS structure was translated ± 0.2 Å in x, y, and z in 0.1 Å steps and rotated 5° in each direction in 2.5° steps. The geometric constraints from Tables 3-3 and 3-4 were applied to enforce the contacts between the TS and each of the three catalytic residues identified in the active site search (Y90, E237, and W275) as well as between E237, and H209. Positions 83 and 239, which are positioned such that they could be members of a potential hydrogen bond network with E237 were allowed to sample all conformations of all residues except for Cys and Pro. All other positions that clashed with the TS or catalytic residues including 21, 46, 87, 89, and 267 were allowed to sample all conformations of any non-polar residue.

Lazaridis-Karplus occlusion-based solvation[16] was applied with a scale factor of 1.0 for nonpolar burial and nonpolar exposure and a scale factor of 0.6 for polar burial.

Other standard ORBIT parameters were applied as in Lassila *et al.,* and a backbone-independent conformer library was used to represent sidechain flexibility.[15] As in the active site search, sidechain-TS interaction energies were biased to favor those contacts that satisfy the geometries in Tables 3-3 and 3-4. Sequence optimization was carried out with FASTER,[17,18] and a Monte Carlo-based algorithm[19,20] was used to sample sequences around the minimum energy conformation identified by FASTER (FMEC).

Hydrophobic active site repacking

This repacking calculation was similar to the previous active site repacking described above. All polar residues within the active site that were not directly contacting the base were designed and restricted to sampling conformations of hydrophobic residues only (Ala, Val, Leu, Ile, Phe, Tyr, Trp, and Met). Residues 46, 47, 84, 87, 89, 130, 131, 207, and 267 were allowed to sample all conformations of Ala, Val, Leu, Ile, Phe, Tyr, and Trp. In addition, residues 83, 21, and 50 were also allowed to sample conformations of Met.

Gene synthesis and cloning

The gene for TAX was back-translated from the protein sequence using the codon usage bias of *E. coli* in DNA 2.0.[21] The DNA sequence for a Factor Xa cleavage site and six-histidine purification tag were added to the 3′ end of the gene. Overlapping oligonucleotides spanning the gene sequence and flanking primers were designed using the Assembly PCR Oligo Maker web server.[22] The basic melting temperature calculation method was used along with a 50 mM cation concentration, a DNA concentration of 0.5

μM, maximum oligo length of 40, a 55°C annealing temperature, and a 40°C acceptable overlap T_m. The resulting oligonucleotides were 31-41 base pairs in length and the full-length gene was constructed by recursive PCR of the overlapping oligonucleotides using a method based on the one described by Stemmer *et al.*[23]

A primer mix was made by combining 2 μL each of the oligonucleotide primers, which had been diluted to 10 μM in H_2O. The 50 μL gene assembly reactions contained 5 μL 10x KOD buffer (Applied Biosystems), 1 mM $MgCl_2$ (Applied Biosystems), 0.2 mM dNTP mix (Novagen), 2.5 U KOD hot start (Applied Biosystems), and 10 μL of the primer mix. Gene amplification was carried out in 50 μL reactions containing 1 μL of the assembly reaction, 5 μL 10x KOD buffer (Applied Biosystems), 1 mM $MgCl_2$ (Applied Biosystems), 0.2 mM dNTP mix (Novagen), 2.5 U KOD hot start (Applied Biosystems), and 1 μM of the forward and reverse flanking primers. The reactions were carried out on Mastercycler Personal Thermocycler (Eppendorf) using the temperature program described in Table 3-5.

After amplification, the reactions were purified with a QIAquik PCR Purification Kit (Qiagen) then digested with *Bam*HI and *Nde*I (New England Biolabs). The digested genes were then run on a 1% agarose gel, extracted and purified using a QIAquik Gel Extraction Kit (Qiagen), and ligated into similarly digested pET11a vector (Novagen).

Site-directed mutagenesis

Mutagenesis primers were designed using the rules outlined in the Site-Directed Mutagenesis Kit (Stratagene). Primer melting temperatures were calculated using Equation 3-1:

$$T_m = 81.5 + 0.41(\%GC) - \frac{675}{N} - \%mismatch \quad . \qquad \text{[Equation 3-1]}$$

Mutagenesis primers were phosphorylated in 50 μL reactions containing 5 μL 10x T4-ligase buffer (New England Biolabs), 4 μg primer, and 10 units T4-polynucleotide kinase (New England Biolabs). Phosphorylation reactions were incubated at 37°C for 2 hours. Mutants were created using a protocol modified from the QuickChange Multi Site-Directed Mutagenesis Kit (Stratagene). Each 50 μL site-directed mutagenesis reactions contained 50 ng template, 120 ng of each phosphorylated mutagenesis primer, 3 μL each 10x Pfu Turbo buffer, 3 μL 10 x Taq ligase buffer (New England Biolabs), 2 mM dNTP mix (Stratagene), 2.5 U Pfu Turbo (Stratagene), and 40 U Taq DNA ligase (New England Biolabs). PCR was carried out on Mastercycler Personal Thermocycler (Eppendorf) using the temperature program described in Table 3-5.

Protein expression and purification

TAX and HG-1 were expressed in BL-21 (DE3) *E. coli* cells in LB with ampicillin. 30 mL starter cultures were grown overnight at 37°C and then used to innoculate 1L cultures. The cells were grown with shaking at 37°C until OD_{600} ~ 0.3 and were then grown at 25°C until OD_{600} ~ 0.6. Expression was induced with 1 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) and carried out at 25°C for 12-18 hours. Cells were harvested by centrifugation and were resuspended in 30 mL buffer 1 (20 mM Tris pH 7.4, 300 mM NaCl, 10 mM imidazole). 10 mM $MgCl_2$, ribonuclease A, and deoxyribonuclease I (Stratagene) and the cells were lysed mechanically with an Emulsiflex-C5 (Avestin). The soluble fraction was incubated for ~ 1 hour with 3 mL Ni-

NTA resin (Qiagen) that had been equilibrated with 10 bed volumes of buffer 1. The resin was washed in a gravity column with 10 bed volumes of buffer 2 (20 mM Tris pH 7.4, 300 mM NaCl, 20 mM imidazole) and HG-1 was eluted with 3 bed volumes of buffer 3 (20 mM Tris pH 7.4, 300 mM NaCl, 250 mM imidazole). The eluate was concentrated using Amicon 10,000 MWCO centrifugal concentrators (Millipore) and the buffer was exchanged to 50 mM sodium citrate, 150 mM NaCl pH 5.5 using PD-10 desalting columns (GE).

Protein characterization

The molecular weights of all proteins were confirmed by electrospray mass spectrometry at the Protein/Peptide MicroAnalytical Laboratory (PPMAL) at Caltech. The secondary structure and stability of the proteins were characterized by circular dichrosim (CD) spectroscopy with an Aviv 62DS spectrometer equipped with a thermoelectric temperature controller. All CD experiments were carried out in a 1 mm cuvette with 10 μM protein in 50 mM MES pH 5.5, 100 mM NaCl, or 50 mM potassium phosphate, 100 mM NaCl pH 7.5. Wavelength scans were performed at 25°C with 3 sec of averaging at each point. Thermal denaturation was monitored at 222 nm from 4°C to 98°C with 1°C temperature steps, 2 min of temperature equilibration time, and 30 sec of averaging. Apparent midpoints of thermal denaturation (T_m) were obtained using the equation of Minor and Kim.[24] T_m values should be considered approximate and not actual thermodynamic parameters, as the denaturation was not reversible in any case tested.

Substrate synthesis and purification

5-nitrobenzisoxazole was synthesized as described.[25] Briefly, 3.5 g benzisoxazole (Alfa Aesar) was added to 15 mL concentrated H_2SO_4 cooled in an ice water bath. 2 mL of concentrated HNO_3 was added dropwise to the mixture. Nitrification was carried out for 30 min with stirring at 0°C. The reaction mixture was poured into 100 mL ice water and the yellow precipitate was filtered. The solid was recrystalized three times from 60 mL of warm absolute ethanol yielding long, colorless needles. m.p.: 126 °C (lit.[25] m.p. 126-127°C).

IR ($CDCl_3$): ν =1620 (arom);1520, 1352 (NO_2) cm^{-1}

1H-NMR ($CDCl_3$): δ = 7.75 (d, 1H); 8.5 (dd, 1H); 8.75 (d, 1H); 8.9 (d, 1H) ppm

MS: m/z = 163.1 (100%, M^-), 133.0 (47%, M^- - NO), 117.9 (7%. M^- - NO_2)

Nuclear magnetic resonance (NMR) analysis was performed on a Varian Mercury 300 MHz machine and IR spectra were recorded using a Perkin-Elmer Spectrum BX spectrometer. Electrospray mass spectrometry was performed at PPMAL (Caltech).

Kinetic measurements

The production of the phenolate product (Figure 1-2) was monitored at 405 nm using a Shimadzu UV-1601 spectrophotometer equipped with a temperature-controlled cell holder. A 40 mM stock of 5-nitrobensoxazole was made in acetonitrile and stored at -20°C. Assays were carried out with 250 μM protein in 40 mM buffer, 100 mM NaCl at 20°C and 37°C.

Crystallization conditions

Crystallization conditions were set up by hand in 24-well hanging drop plates using 2.0 × 2.0 μL drops and Index, Crystal 1, and Crystal 2 screens (Hampton Research) as well as a variety of custom screens using conditions similar to the crystallization conditions for 1GOR.[13] The protein sample contained 10 mg/mL HG-1 in 50 mM sodium citrate pH 5.5, 100 mM NaCl. A single large crystal (~ 0.5 mm) with multiple protrusions was found in buffer containing 0.8 M posassium-sodium tartrate, 0.1 M Tris pH 8.5, 0.5% PEG-MME 5000. A smaller single crystal (~ 0.1 mm) was identified in 100 mM sodium phosphate pH 6.5, 12% PEG-600.

X-ray diffraction data collection and refinement

The spikes projecting off the 0.5 mm crystal were broken off for diffraction screening. The best diffraction was obtained from a crystal cryoprotected in the reservoir solution (0.8 M posassium-sodium tartrate, 0.1 M Tris pH 8.5, 0.5% PEG-MME 5000) with 15% ethylene glycol.

Data were collected using a MicroMax-007HF X-ray generator with a RAXIS IV++ detector (Rigaku Corp.) All data were processed using CrystalClear (Rigaku Corp.) and Mosflm.[26] The indexed and scaled data were further evaluated using CCP4.[27] Molecular replacement was carried out using Phaser version 1.3.3[28] and 1GOR as the starting model.[13] Further refinement was performed with Coot[29] and Refmac.[30]

Molecular dynamics (MD) simulations

MD simulations were carried out for 20 ns for each enzyme-substrate complex at

NPT conditions (constant number of particles, pressure, and temperature) with a pressure of 1 bar and temperature of 300 K. The TIP3P explicit solvent model[31] was used to solvate the protein in an octahedral-shaped volume, ensuring a solvent layer of at least 10 Å from any point of the protein surface. Each face of the octahedral box is connected to a mirror image of itself, which allows the system to be treated with periodic boundary conditions: when a water molecule "escapes" the original octahedral box, it re-enters from the opposite face of that same box. This ensures equilibration through diffusion, while the number of particles is kept constant. The temperature is regulated and evenly distributed through a Langevin equilibration scheme. The benzisoxazole substrate was parameterized from quantum mechanical calculations in order to be treated correctly by the AMBER force field.[32] Prior to production MD, the geometry of the X-ray-based structure was optimized and slowly heated to 300 K over a time period of 300 ps at NVT conditions, and then equilibrated at NPT conditions for 2 ns.

Results and Discussion

Design procedure

TS theory suggests that natural enzymes achieve their large rate accelerations through preferential stabilization of the TS.[33] Thus, our computational enzyme design strategy is based on a TS or high-energy intermediate structure and the active site residues are chosen for their ability to stabilize the TS as well as participate in the reaction chemistry.[15]

As mentioned in Chapter I, our computational protein design software was optimized for stabilizing small proteins and required adaptation to the unique requirements of enzymatic catalysis.[15] Specifically, the interactions that are predicted to stabilize the TS were to be enforced through geometric requirements and the positioning of the TS was varied systematically within the binding pocket during the sequence optimization. The details of this design methodology are described in Chapter II.

The catalytic contacts that were predicted to stabilize the Kemp elimination (KE) TS are shown in Figure 3-2. These contacts were inspired by the catalytic residues observed in the catalytic antibody 34E4 (Figure 1-3). A carboxylate-containing residue was chosen as a general base in analogy to the Glu that was determined to be the general base in the catalytic antibody 34E4.[34] In addition, a carboxylate base avoids problems with the protonation state variations of His and the large number of rotameric states of Lys. A π-stacking residue was required to facilitate electron delocalization within the TS. A hydroxyl group was added to stabilize the negative charge that develops on the benzisoxazole oxygen in the TS.[12] No analogous hydrogen bond donor is observed in the catalytic antibody, although water is postulated to fill this role.[34] During the course of this study, a similar active site arrangement was used by Röthlisberger *et al.* to obtain active KE enzymes.[11]

In addition to the direct contacts to the TS, a supporting contact to the general base was also required to stabilize its conformation. As shown in Figure 3-3, this contact is a hydrogen bond to the general base from a His (Hie and Hid are His residues, which are protonated at Nε or Nδ, respectively), Glu, or Asp residue. This contact was required

through a second, separate geometry-biasing step. The geometric definitions for the three catalytic contacts and the additional base contact are shown in Table 3-2.

Active site search

The active site search resulted in multiple possible active site configurations that satisfied all four of the required contacts. The active sites were assessed visually and active site 1-2 was chosen for repacking (Figure 3-4A). In 1-2, the general base is a wild-type Glu at position 237 that is in a conformation similar to that seen in the crystal structure. The base is held in place by a wild-type His at position 209 that preserves a contact seen in the crystal structure of the wild-type TAX. A wild-type Trp at position 275 makes a stacking contact to the TS and is in its crystallographic conformation. Finally, the hydrogen bond contact is fulfilled by a Tyr at position 90.

Active site repacking

The active site repacking calculation resulted in a seven-fold mutant called HG-1 (Figure 3-4B). This design features an extensive hydrogen bond network supporting the conformation of the base, E237, which is similar to the hydrogen bond network surrounding E237 in the wild-type protein. In the crystal structure of TAX, His209 makes a bridging contact between Glu237 and Asp239, indicating that it is in its fully protonated state which would also serve to stabilize the negative charge on both carboxylates. As a result, the pKa of Glu237 is probably quite low. This bridging contact is part of a larger hydrogen bond network that includes the indole of Trp275. One difference between the hydrogen bond network in the design and the one in TAX is

the D127N mutation, which should help to ensure that H209 is not positively charged, thus stabilizing the protonated form of E237 and increasing its pKa. The pKa of E237 was predicted to be 7.1 by PROPKA,[35] which may be sufficient for proton abstraction from the substrate.

Other mutations in the active site include W267A, W87G, A21M, E46M, and S89F (Figure 3-4B), which were introduced to make room in the pocket for substrate binding and to pack tightly around the substrate once it is bound.

HG-1 activity assays

HG-1 showed no KE activity over background under any of the conditions tested (pH 5.0-9.0, 20-37°C) (Figure 3-5). Analysis by circular dichroism (CD) spectroscopy showed that the protein had similar secondary structure as the wild-type scaffold, indicating that the inactivity was not due to global unfolding or misfolding of the protein (Figure 3-6A). The T_m for HG-1, also determined by CD, was about 10°C lower than the wild-type scaffold at pH 5.5 (Figure 3-6B). At pH 7.5, an additional 10°C decrease in melting temperature was seen compared to wild-type TAX at pH 5.5. However, the HG-1 was fully folded at all of the temperatures and pHs tested at which KE activity was assayed. Single and double mutants of HG-1 were also tested including N239D, N239D/M46A, and M46A. No rate acceleration was seen for any of these variants above the scaffold-catalyzed reaction.

HG-1 crystal structure

In an effort to understand the source of the inactivity of HG-1, we crystallized and solved the crystal structure of HG-1 to a resolution of 2.0 Å in collaboration with the Molecular Observatory at Caltech (Table 3-6). The overall RMSD of HG-1 with respect to the scaffold is 0.36 Å, confirming that the overall fold has been maintained (Figure 3-7A).

The structure fits well to the electron density as shown in Figure 3-7B. One concern with using molecular replacement (MR) to calculate the phases is potential bias from the template structure 1GOR. As shown in Figure 3-8, the electron density around the mutated positions is well-defined and shows no signs of the wild-type amino acid, indicating little bias.

The active site residues of HG-1, including the base, the hydrogen bond donor, and most of the hydrogen bond network, overlay well with the conformations predicted by ORBIT (Figure 3-9). The largest deviations of the active site residues from the design are in the π-stacking residue (W275), the associated Arg (R276) and N172, which acts as a hydrogen bond bridge between the indole of W275 and H209 (Figure 3-9B). W275 and R276 are both rotated out of the active site with respect to the design, removing potential binding interactions with the substrate and exposing the active site to solvent. In addition, N172 is rotated 90° from the predicted conformation, but it is unclear if this deviation is the cause for the conformation of W275.

The crystal structure revealed the presence of six ordered water molecules in the active site of HG-1 (Figure 3-10), five of which must be displaced before substrate binding could occur. The presence of these waters could result in a large desolvation

penalty, which could prevent substrate binding as well as a reduction the pKa of the base through stabilization of the anionic form.

Overall, the crystal structure indicates that the active site configuration predicted by ORBIT is correct with the exception of N172, W275, and R276. The flexibility of the active site, indicated by the deviation of W275 and the associated R276, could result in the loss of favorable binding interactions. The movement of these residues could also expose the active site to solvent, further decreasing the energetic benefit of substrate binding. Unfortunately, the crystal structure could not unambiguously identify the cause of the protein's inactivity, but these two factors could be significant impediments to activity.

MD simulations

The hypothesis of active site flexibility and solvent exposure being the cause of the inactivity was supported by MD simulations carried out by Gert Kiss in the Houk lab at UCLA. The design model of HG-1 with the substrate bound in the active site was used as the starting structure of the simulations. Figure 3-11 shows that the substrate-protein complex exists predominantly in two states. In state 1, the substrate is stably bound and maintains a substrate-base distance of about 3 Å. In state 2, a major reorientation of the substrate occurs and the substrate exits the active site. Closer analysis of the active site residues during the transition between state 1 and state 2 indicates that within 2 ns, R276 flips away from the active site, followed by W275 at 5 ns. The loss of these binding interactions seems to directly contribute to the reorientation and subsequent expulsion of the substrate from the active site around 7 ns. In addition, solvent molecules moving

rapidly into and out of the active site preferentially interact with the base and other polar residues causing significant destabilization of the bound substrate, actually pushing it from the active site. These results suggesting that the active site is very flexible near W275 and R276 and is occupied by a large number of solvent molecules agree with findings from the crystal structure of HG-1.

Design of a more hydrophobic active site

The more hydrophobic design of HG-1 resulted in the sequence HG-1h, an eight-fold mutant with respect to HG-1 (N47F, K50M, M64A, H83V, T84F, F89L, N130F, and E131A). Figure 3-12 shows that the overall conformation of the active site in HG-1h, including the TS and catalytic residues, is very similar to that in HG-1. The pKa of E239 in HG-1h was predicted to be about 1.0 pH unit higher than in HG-1. MD analysis of this design shows that as in HG-1, the active site is very flexible and allows the substrate to exit the active site during the simulation. The active site is still very solvent exposed, and solvent molecules still enter the active site over the course of the simulation to interact with the remaining polar residues and crowd the substrate out of the active site. Figure 3-13A shows representative configurations of HG-1h before and after the MD simulation. In the final configuration, the substrate-base distance is over 10 Å and the substrate-base distances shown in Figure 3-13B and 3-13C indicate that the contact is not maintained for any significant length of time. These results suggest that our redesign is not likely to be active for many of the same reasons as HG-1.

Conclusions

While we were unsuccessful in designing an active KE enzyme, we were able to solve the crystal structure of one of our inactive designs, allowing us to gain further insight into the cause of the inactivity. Unfortunately, neither the crystal structure nor the MD simulations of HG-1 were able to unambiguously determine the cause of HG-1's inactivity. However, both analyses indicated that the active site is very solvent exposed, flexible, and occupied by a large number of water molecules. While it is not certain that these factors are the cause of the inactivity, any one of them could have a negative effect on substrate binding as well as on the base pKa and all three issues could be addressed by choosing an active site that is more deeply buried within a protein scaffold. The natural active site of TAX, which was the focus of the designs presented here, was too solvent exposed and flexible to be amenable to this reaction. Future redesigns of this scaffold must locate the active site away from the natural binding pocket in a location that is more deeply buried, perhaps in the barrel of the scaffold. Other scaffolds with less solvent-exposed active sites should also be explored. A more deeply buried active site could effectively reduce the flexibility of the binding residues and shield the active site from solvent, though it remains to be seen if the resolution of these factors alone will result in an active KE enzyme.

Acknowledgements

We thank Paul Cheong from the Houk lab at UCLA for providing the coordinates and charges for the KE TS. Gert Kiss from the Houk lab at UCLA carried out the MD simulations and created Figure 3-11. Dr. Leonard Thomas of the Caltech Molecular

Observatory solved and refined the crystal structure of HG-1. We also thank Dr. Jonathan Lassila and Dr. Roberto Chica for discussions about enzymes and enzyme kinetics.

References

1. Dahiyat, B. I.; Mayo, S. L., Protein design automation. *Protein Sci.* **1996**, *5*, 895-903.

2. Dahiyat, B. I.; Mayo, S. L., De novo protein design: fully automated sequence selection. *Science* **1997**, *278*, 82-87.

3. Dahiyat, B. I.; Mayo, S. L., Probing the role of packing specificity in protein design. *Proc. Natl. Acad. Sci. USA* **1997**, *94*, 10172-10177.

4. Dahiyat, B. I.; Gordon, B.; Mayo, S. L., Automated design of the surface positions of protein helices. *Protein Sci.* **1997**, *6*, 1333-1337.

5. Zollars, E. S.; Marshall, S. A.; Mayo, S. L., Simple electrostatic model improves designed protein sequences. *Protein Sci.* **2006**, *15*, 2014-2018.

6. Bolon, D. N.; Mayo, S. L., Enzyme-like proteins by computational protein design. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 14274-14279.

7. Kaplan, J.; DeGrado, W. F., *De novo* design of catalytic proteins. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 11566-11570.

8. Ashworth, J.; Havranek, J. J.; Duarte, C. M.; Sussman, D.; Monnat, R. J.; Stoddard, B. L.; Baker, D., Computational redesign of endonuclease DNA binding and cleavage specificity. *Nature* **2006**, *441*, 656-659.

9. Chen, C. Y.; Georgiev, I.; Anderson, A. C.; Donald, B. R., Computational structure-based redesign of enzyme activity. *Proc. Natl. Acad. Sci. USA* **2009**.

10. Jiang, L.; Althoff, E. A.; Clemente, F. R.; Doyle, L.; Röthlisberger, D.; Zanghellini, A.; Gallaher, J. L.; Betker, J. L.; Tanaka, F.; Barbas, C. F.; Hilvert, D.; Houk, K. N.; Stoddard, B. L.; Baker, D., De novo computational design of retro-aldol enzymes. *Science* **2008**, *319*, 1387-1391.

11. Rothlisberger, D.; Khersonsky, O.; Wollacott, A.; Jiang, L.; Dechancie, J.; Betker, J.; Gallaher, J. L.; Althoff, E. A.; Zanghellini, A.; Dym, O.; Albeck, S.; Houk, K. N.; Tawfik, D.; Baker, D., Kemp elimination catalysts by computational enzyme design. *Nature* **2008**, *453*, 190-195.

12. Na, J.; Houk, K. N.; Hilvert, D., Transition state of the base-promoted ring-opening of isoxazoles. Theoretical prediction of catalytic functionalities and design of haptens for antibody production. *J. Am. Chem. Soc.* **1996**, *118*, 6462-6471.

13. Lo Leggio, L.; Kalogiannis, S.; Eckert, K.; Teixeira, S. C.; Bhat, M. K.; Andrei, C.; Pickersgill, R. W.; Larsen, S., Substrate specificity and subsite mobility in *T. aurantiacus* xylanase 10A. *FEBS Lett.* **2001**, *509*, 303-308.

14. Davis, I. W.; Murray, L. W.; Richardson, J. S.; Richardson, D. C., MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res.* **2007**, *35*, W375-W383.

15. Lassila, J. K.; Privett, H. K.; Allen, B. D.; Mayo, S. L., Combinatorial methods for small-molecule placement in computational enzyme design. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 16710-16715.

16. Lazaridis, T.; Karplus, M., Effective energy function for proteins in solution. *Proteins* **1999**, *35*, 133-152.

17. Desmet, J.; Spriet, J.; Lasters, I., Fast and accurate side-chain topology and energy refinement (FASTER) as a new method for protein structure optimization. *Proteins* **2002**, *48*, 31-43.

18. Allen, B. D.; Mayo, S. L., Dramatic performance enhancements for the FASTER optimization algorithm. *J. Comput. Chem.* **2006**, *27*, 1071-1075.

19. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., Equation of state calculations by fast computing machines. *J. Chem. Phys.* **1953**, *21*, 1087-1092.

20. Voigt, C. A.; Gordon, D. B.; Mayo, S. L., Trading accuracy for speed: A quantitative comparison of search algorithms in protein sequence design. *J. Mol. Biol.* **2000**, *299*, 789-803.

21. Villalobos, A.; Ness, J. E.; Gustafsson, C.; Minshull, J.; Govindarajan, S., Gene Designer: a synthetic biology tool for constructing artificial DNA segments. *BMC Bioinformatics* **2006**, *7*, 285-292.

22. Rydzanicz, R.; Zhao, X. S.; Johnson, P. E., Assembly PCR oligo maker: a tool for designing oligodeoxynucleotides for constructing long DNA molecules for RNA production. *Nucleic Acids Res.* **2005**, *33*, W521-525.

23. Stemmer, W. P.; Crameri, A.; Ha, K. D.; Brennan, T. M.; Heyneker, H. L., Single-step assembly of a gene and entire plasmid from large numbers of oligodeoxyribonucleotides. *Gene* **1995**, *164*, 49-53.

24. Minor, D. L.; Kim, P. S., Measurement of the bold β-sheet-forming propensities of amino acids. *Nature* **1994**, *367*, 660-663.

25. Lindemann, H.; Thiele, H., The chemistry of benzene-α, β-isoxazoles. *Justus Liebigs Annalen Der Chemie* **1926**, *76*, 63-81.

26. Leslie, A. G. W., Recent changes to the MOSFLM package for processing film and image plate data. *Joint CCP4 + ESF-EAMCB Newsletter on Prot. Crystallography* **1992**, *26*.

27. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr., Sect. D* **1994**, *50*, 760-763.

28. McCoy, A. J.; Grosse-Kunstleve, R. W.; Storoni, L. C.; Read, R. J., Likelihood-enhanced fast translation functions. *Acta Crystallogr., Sect. D* **2005**, *61*, 458-464.

29. Emsley, P.; Cowtan, K., Coot: model-building tools for molecular graphics. *Acta Crystallogr., Sect. D* **2004**, *61*, 458-464.

30. Murshudov, G. N.; Vagin, A. A.; Dodson, E. J., Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr., Sect. D* **1997**, *53*, 240-255.

31. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L., Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926-935.

32. Pearlman, D. A.; Case, D. A.; Caldwell, J. W.; Ross, W. S.; Cheatham, T. E.; DeBolt, S.; Ferguson, D.; Seibel, G.; Kollman, P., AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Comput. Phys. Commun.* **1995**, *91*, 1-41.

33. Lienhard, G. E., Enzymatic catalysis and transition-state theory. *Science* **1973**, *180*, 149-154.

34. Debler, E. W.; Ito, S.; Seebeck, F. P.; Heine, A.; Hilvert, D.; Wilson, I. A., Structural origins of efficient proton abstraction from carbon by a catalytic antibody. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 4984-4989.

35. Li, H.; Robertson, A. D.; Jensen, J. H., Very fast empirical prediction and interpretation of protein pKa values. *Proteins* **2005**, *61*, 704-721.

Table 3-1. Partial atomic charges for 5-nitrobenzisoxazole (5-NBZ).

atom name	charge
O1	-0.586
N2	-0.212
C3	0.041
C4	-0.252
C5	-0.148
C6	0.099
N6	0.517
O61	-0.484
O62	-0.506
C7	-0.168
C8	-0.269
C9	0.516
H3	0.490
H5	0.275
H7	0.251
H8	0.212

Table 3-2. Variation of contact geometry for targeted ligand placement. The TS was placed at every indicated geometry for every rotamer of Asp or Glu in the design calculation. Because Asp and Glu both have two equivalent, but distinctly named, oxygens in their sidechains, a separate set of geometries is described for each oxygen.

Placement: H3 from Aspartate (OD1)

type	atom1	atom2	atom3	atom4	min	max	step
DISTANCE	OD1	H3			1.1	1.5	0.2
ANGLE	OD2	OD1	H3		70.0	110.0	20.0
TORSION	CG	OD2	OD1	H3	160.0	200.0	20.0
ANGLE	OD1	H3	PSA		70.0	110.0	20.0
TORSION	OD2	OD1	H3	PSA	0.0	360.0	20.0
TORSION	OD1	H3	PSA	C3	160.0	200.0	20.0

Placement: H3 from Aspartate (OD2)

type	atom1	atom2	atom3	atom4	min	max	step
DISTANCE	OD2	H3			1.1	1.5	0.2
ANGLE	OD1	OD2	H3		70.0	110.0	20.0
TORSION	CG	OD1	OD2	H3	160.0	200.0	20.0
ANGLE	OD2	H3	PSA		70.0	110.0	20.0
TORSION	OD1	OD2	H3	PSA	0.0	360.0	20.0
TORSION	OD2	H3	PSA	C3	160.0	200.0	20.0

Placement: H3 from Glutamate (OE1)

type	atom1	atom2	atom3	atom4	min	max	step
DISTANCE	OE1	H3			1.1	1.5	0.2
ANGLE	OE2	OE1	H3		70.0	110.0	20.0
TORSION	CD	OE2	OE1	H3	160.0	200.0	20.0
ANGLE	OE1	H3	PSA		70.0	110.0	20.0
TORSION	OE2	OE1	H3	PSA	0.0	360.0	20.0
TORSION	OE1	H3	PSA	C3	160.0	200.0	20.0

Placement: H3 from Glutamate (OE2)

type	atom1	atom2	atom3	atom4	min	max	step
DISTANCE	OE2	H3			1.1	1.5	0.2
ANGLE	OE1	OE2	H3		70.0	110.0	20.0
TORSION	CD	OE1	OE2	H3	160.0	200.0	20.0
ANGLE	OE2	H3	PSA		70.0	110.0	20.0
TORSION	OE1	OE2	H3	PSA	0.0	360.0	20.0
TORSION	OE2	H3	PSA	C3	160.0	200.0	20.0

Table 3-3. Geometric constraints contacts between the active site residues and the transition state in the active site search. Three contacts are required: Asp/Glu, Phe/Trp and Ser/Thr/Tyr. Distance measurements are given in Ångströms. Angle, torsion, and plane measurements are given in degrees.

Contact: Asp/Glu to H3

residue	type	atom1	atom2	atom3	atom4	min	max
Asp	DISTANCE	OD1	H3			1.0	1.6
	ANGLE	OD2	OD1	H3		69.0	111.0
	TORSION	CG	OD2	OD1	H3	159.0	201.0
	ANGLE	OD1	H3	PSA		69.0	111.0
	TORSION	OD2	OD1	H3	PSA	0.0	360.0
	TORSION	OD1	H3	PSA	C3	159.0	201.0
Asp	DISTANCE	OD2	H3			1.0	1.6
	ANGLE	OD1	OD2	H3		69.0	111.0
	TORSION	CG	OD1	OD2	H3	159.0	201.0
	ANGLE	OD2	H3	PSA		69.0	111.0
	TORSION	OD1	OD2	H3	PSA	0.0	360.0
	TORSION	OD2	H3	PSA	C3	159.0	201.0
Glu	DISTANCE	OE1	H3			1.0	1.6
	ANGLE	OE2	OE1	H3		69.0	111.0
	TORSION	CD	OE2	OE1	H3	159.0	201.0
	ANGLE	OE1	H3	PSA		69.0	111.0
	TORSION	OE2	OE1	H3	PSA	0.0	360.0
	TORSION	OE1	H3	PSA	C3	159.0	201.0
Glu	DISTANCE	OE2	H3			1.0	1.6
	ANGLE	OE1	OE2	H3		69.0	111.0
	TORSION	CD	OE1	OE2	H3	159.0	201.0
	ANGLE	OE2	H3	PSA		69.0	111.0
	TORSION	OE1	OE2	H3	PSA	0.0	360.0
	TORSION	OE2	H3	PSA	C3	159.0	201.0

Contact: Ser/Thr/Tyr to O1

residue	type	atom1	atom2	atom3	atom4	min	max
Ser	DISTANCE	OG	O1			2.6	4.0
	ANGLE	OG	HG	O1		150.0	180.0
	ANGLE	HG	O1	N2		100.0	160.0
	TORSION	HG	O2	N2	C3	120.0	240.0
Thr	DISTANCE	OG1	O1			2.6	4.0
	ANGLE	OG1	HG1	O1		150.0	180.0
	ANGLE	HG1	O1	N2		100.0	160.0
	TORSION	HG1	O2	N2	C3	120.0	240.0
Tyr	DISTANCE	OH	O1			2.6	4.0
	ANGLE	OH	HH	O1		150.0	180.0
	ANGLE	HH	O1	N2		100.0	160.0
	TORSION	HH	O2	N2	C3	120.0	240.0

Contact: Phe/Trp to PS2

residue	type	atom1	atom2	atom3	atom4	atom5	atom6	min	max
Phe	PSEUDO_ATOM	PS1, equidistant between CE1 and CD2							
	PSEUDO_ATOM	PS2, equidistant between C4 and C9							
	DISTANCE	PS1	PS2					3.0	4.0
	PLANE	CG	CE1	CE2	C5	C8	N2	0.0	40.0
Trp	PSEUDO_ATOM	PS1, equidistant between CE2 and CD2							
	PSEUDO_ATOM	PS2, equidistant between C4 and C9							
	DISTANCE	PS1	PS2					3.0	4.0
	PLANE	CD1	CE3	CH2	C5	C8	N2	0.0	40.0

Table 3-4. Geometric constraints for additional base contact. A single hydrogen bond contact was required to one of the carboxylate oxygens of the general base. Distances are given in Ångströms. Angles are given in degrees.

Contact: His to OD1/OD2

residue	type	atom1	atom2	atom3	min	max
Hie	DISTANCE	NE2	OD1		2.6	4.0
	ANGLE	NE2	HE2	OD1	130.0	180.0
	ANGLE	HE2	OD1	CG	120.0	180.0
Hie	DISTANCE	NE2	OD2		2.6	4.0
	ANGLE	NE2	HE2	OD2	130.0	180.0
	ANGLE	HE2	OD2	CG	120.0	180.0
Hid	DISTANCE	ND1	OD1		2.6	4.0
	ANGLE	ND1	HD1	OD1	130.0	180.0
	ANGLE	HD1	OD1	CG	120.0	180.0
Hid	DISTANCE	ND1	OD2		2.6	4.0
	ANGLE	ND1	HD1	OD2	130.0	180.0
	ANGLE	HD1	OD2	CG	120.0	180.0

Contact: His to OE1/OE2

residue	type	atom1	atom2	atom3	min	max
Hie	DISTANCE	NE2	OE1		2.6	4.0
	ANGLE	NE2	HE2	OE1	130.0	180.0
	ANGLE	HE2	OE1	CD	120.0	180.0
Hie	DISTANCE	NE2	OE2		2.6	4.0
	ANGLE	NE2	HE2	OE2	130.0	180.0
	ANGLE	HE2	OE2	CD	120.0	180.0
Hid	DISTANCE	ND1	OE1		2.6	4.0
	ANGLE	ND1	HD1	OE1	130.0	180.0
	ANGLE	HD1	OD1	CD	120.0	180.0
Hid	DISTANCE	ND1	OE2		2.6	4.0
	ANGLE	ND1	HD1	OE2	130.0	180.0
	ANGLE	HD1	OE2	CD	120.0	180.0

Table 3-5. Thermocycler temperature programs for gene construction and mutagenesis reactions.

gene assembly		
Temp (°C)	**Time (min)**	**# cycles**
95	1	
95	1	25x
55	1	
68	14	
4	hold	

amplification		
Temp (°C)	**Time (min)**	**# cycles**
95	1	
95	1	23x
55	1	
68	14	
4	hold	

site-directed mutagenesis		
Temp (°C)	**Time (min)**	**# cycles**
95	1	
95	1	30x
55	1	
68	14	
4	hold	

Table 3-6. Crystallographic statistics for HG-1.

Data collection	
Space group	$P2_12_12_1$
Cell dimensions	
a,b,c (Å)	48.3, 72.5, 74.6
α,β,γ (deg)	90.0, 90.0, 90.0
Resolution (Å) *	2.0 (2.1 - 2.0)
Number of reflections	48807
unique reflections	18096
Rmerge (%) *	2.5 (4.4)
I/σI *	30.7 (21.1)
Completeness (%) *	99.2 (98.2)
Redundancy *	2.7 (2.6)

Refinement	
Resolution (Å)	29.5 - 2.0
Number of reflections	
working set	17163
test set	924
R_{work}/R_{free}	16.4 / 23.0 %
No. atoms	
protein	2326
solvent	187
B-factors	
protein	13.9
water	19.7
R.m.s deviations	
bond lengths (Å)	0.027
bond angles	1.938
Ramachandran plot	
favored (%)	91.0
allowed (%)	8.6
generously allowed (%)	0.4
disallowed (%)	0

* Last shell is shown in parentheses

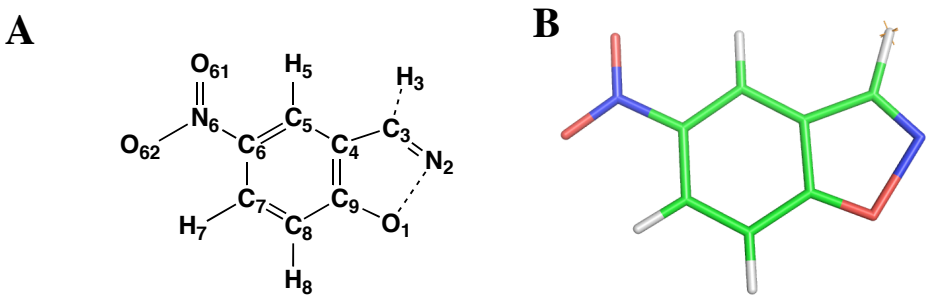


Figure 3-1. 5-NBX transition state. (A) 5-NBX atom names. The bonds being broken are shown as dotted lines. (B) Stick representation of the transition state structure. The pseudoatom is shown as an orange star near H3.

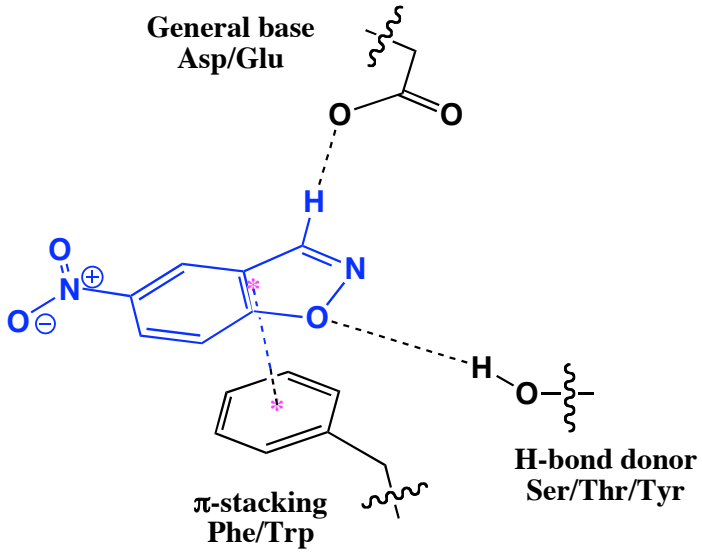


Figure 3-2. Kemp elimination ideal active site. The enzyme design calculations required either an Asp or Glu as a general base, a Ser, Thr, or Tyr as a H-bond donor to the phenolic oxygen, and a π-stacking residue above or below the plane of the TS structure (blue). The pseudoatoms that were used to define the π-stacking contact are shown in pink.

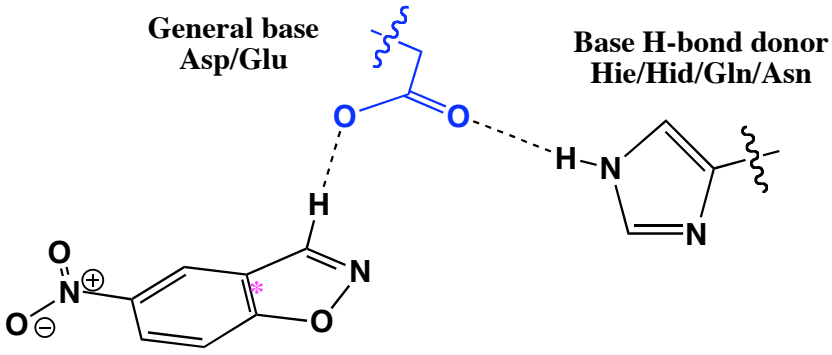


Figure 3-3. Additional hydrogen bond contact. An additional contact (Hie/Hid/Gln/Asn) was required to the base (shown in blue).

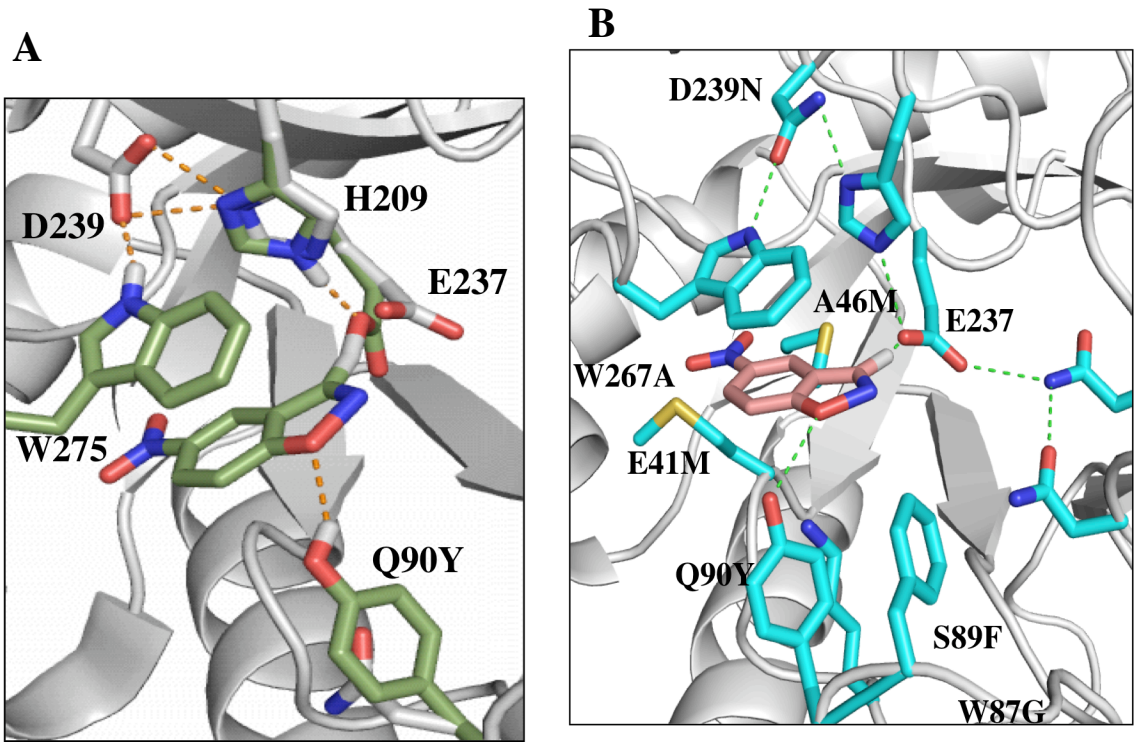


Figure 3-4. Predicted active site structure of HG-1. **(A)** Active site 2-1, identified through the active site search. The catalytic residues and TS are shown in green overlaid with the wild-type crystal structure in grey. **(B)** Repacked active site. The TS model is shown in pink and all seven mutations are identified. The hydrogen bond network supporting the conformation of the base is indicated with green dotted lines.

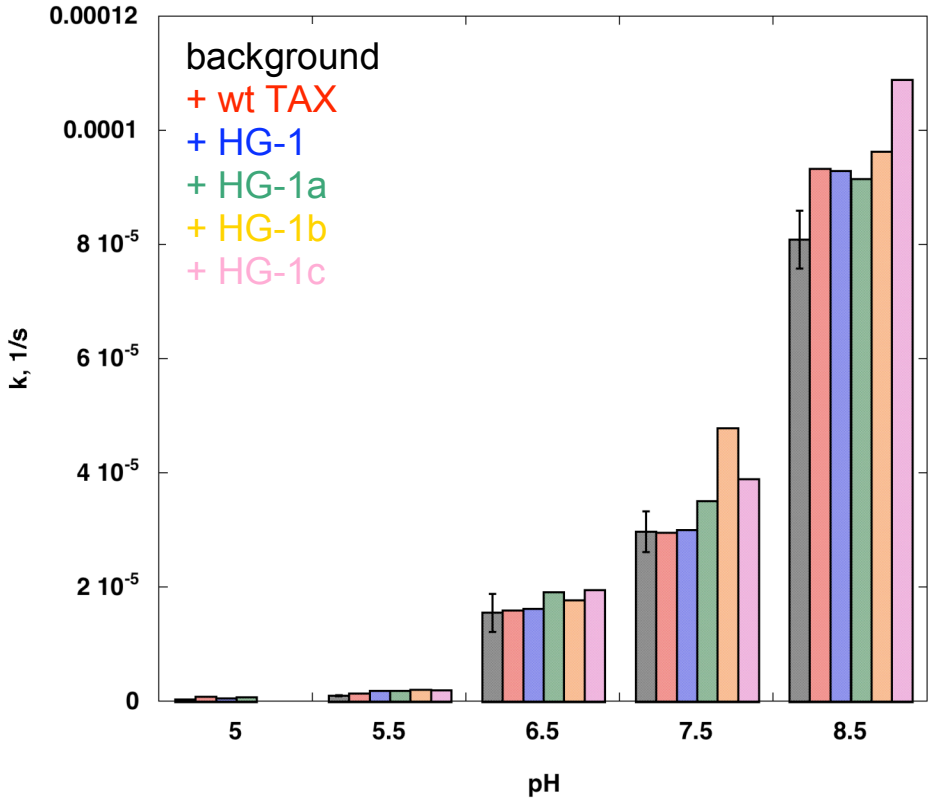


Figure 3-5. First-order rate constants for KE reaction. The rate constants for the reactions including 200 μM wild-type TAX, HG-1, HG-1a (HG-1/N239D), HG-1b (HG-1/N239D/ M46A), HG-1c (HG-1/M46A), and the buffer-catalyzed background reaction are given. No significant rate acceleration was seen above that for the scaffold (wild-type TAX)-catalyzed reaction for any of the variants assayed under any of the conditions tested. Sodium citrate was used for pH 5.0 and 5.5, sodium phosphate was used for pH 6.5 and 7.5, and sodium borate was used for pH 8.5.

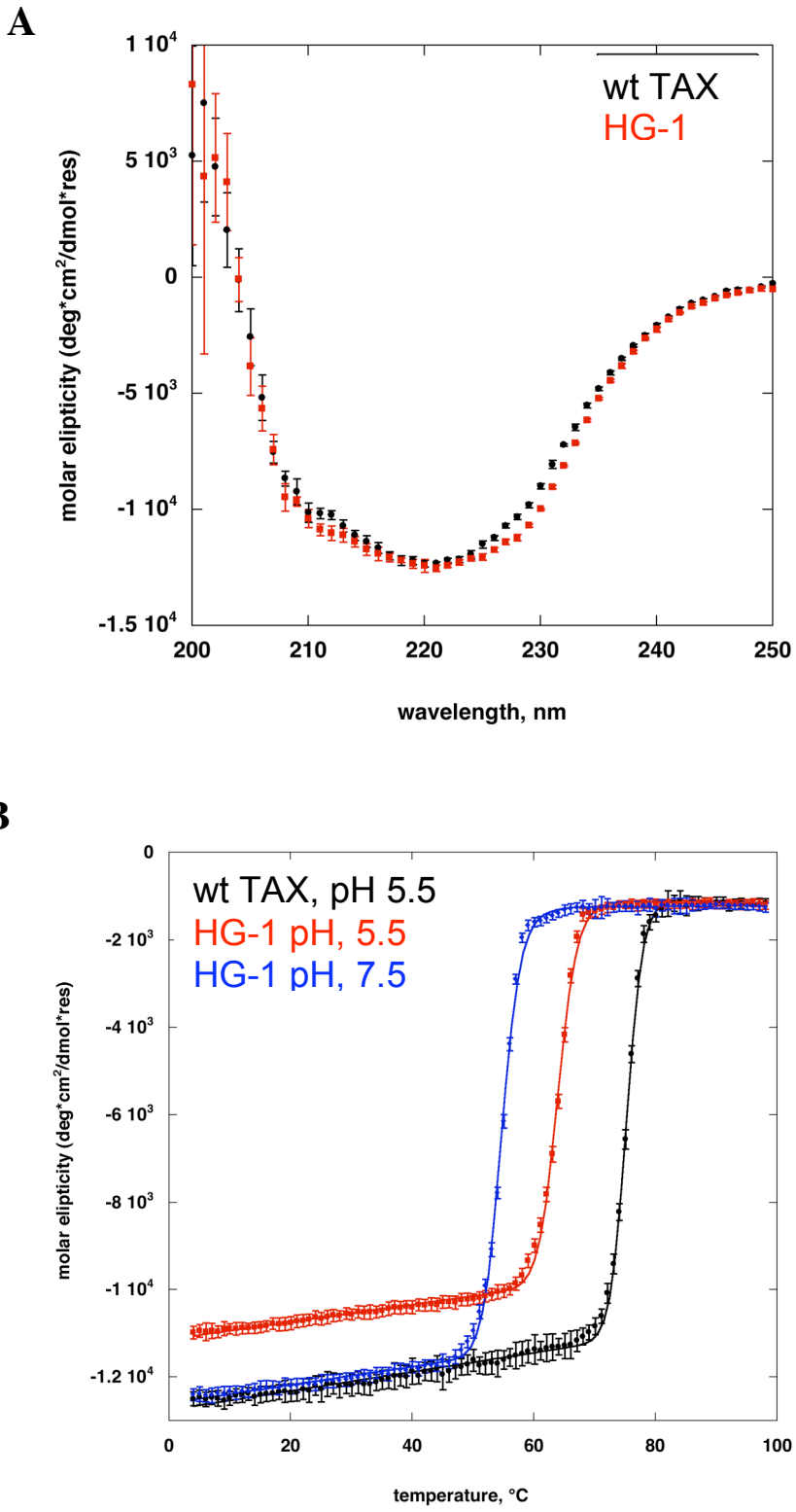


Figure 3-6. CD analysis of HG-1. **(A)** Far-UV wavelength scan indicating that HG-1 is folded. **(B)** Thermal denaturation curve indicating that HG-1 is significantly destabilized compared to TAX.

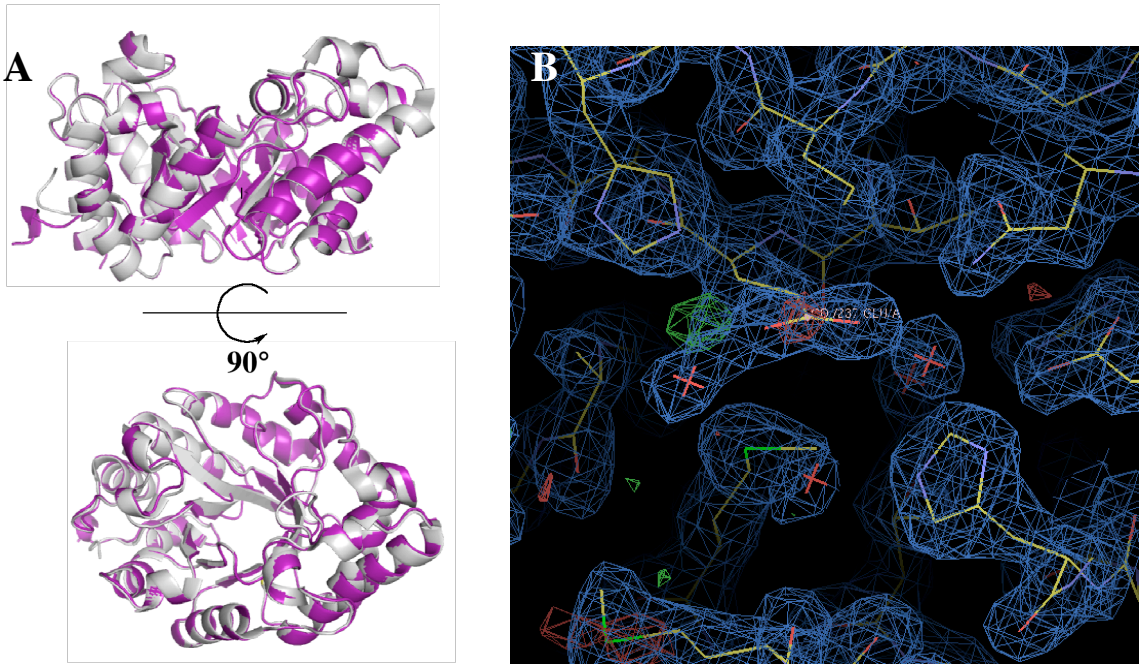


Figure 3-7. Crystal structure of HG-1. **(A)** Overlay of backbone structure of HG-1 (magenta) and wild-type TAX structure 1GOR (grey). **(B)** Electron density in active site of HG-1.

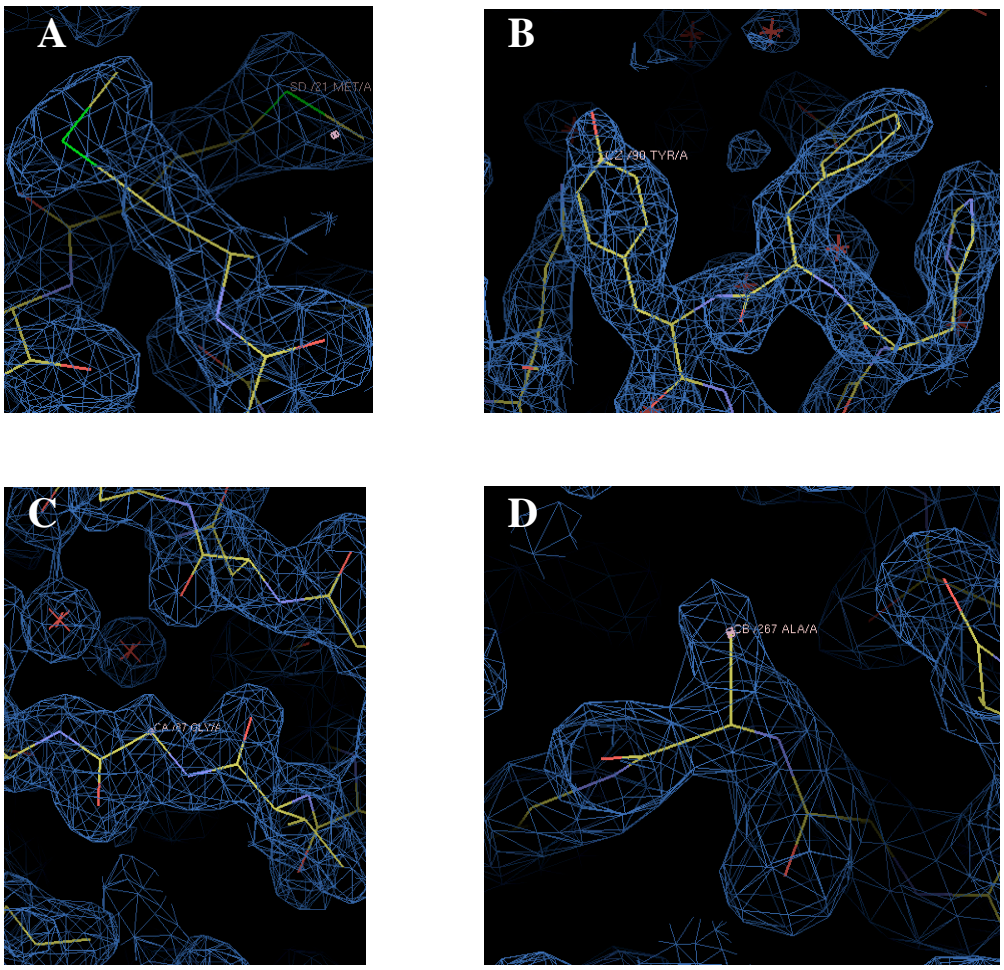


Figure 3-8. Electron density of HG-1 mutation sites. **(A)** A24M/E46M. **(B)** S89F/Q90Y. **(C)** W87G. **(D)** W267A.

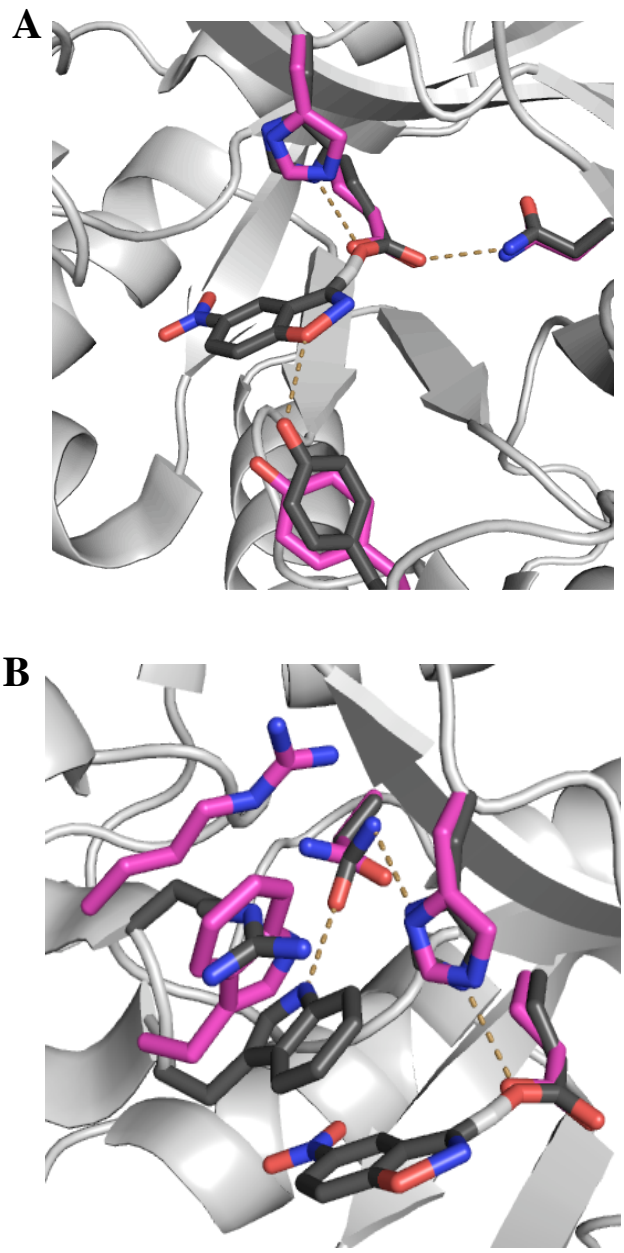


Figure 3-9. Overlay of HG-1 crystal structure active site with design model. The structure predicted by ORBIT (grey) is very similar to the actual conformations seen in the crystal structure (magenta). Predicted hydrogen bonds are represented as yellow dotted lines. **(A)** E237 and the supporting hydrogen bond network and Y90. **(B)** E237 and supporting hydrogen bond network including H209, N239, and W275.

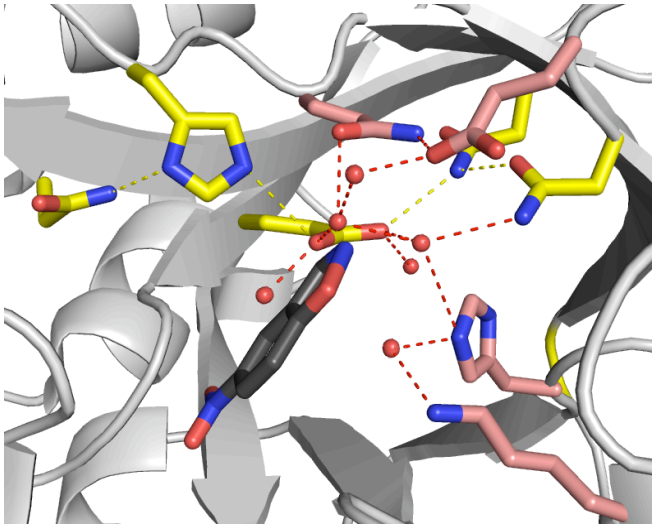


Figure 3-10. Ordered water molecules near E237 in the active site. Residues that participate in the hydrogen bond network to E237 are shown in yellow, additional polar residues in the active site are shown in pink, and water molecules are shown as red spheres. The modeled TS is shown in grey.

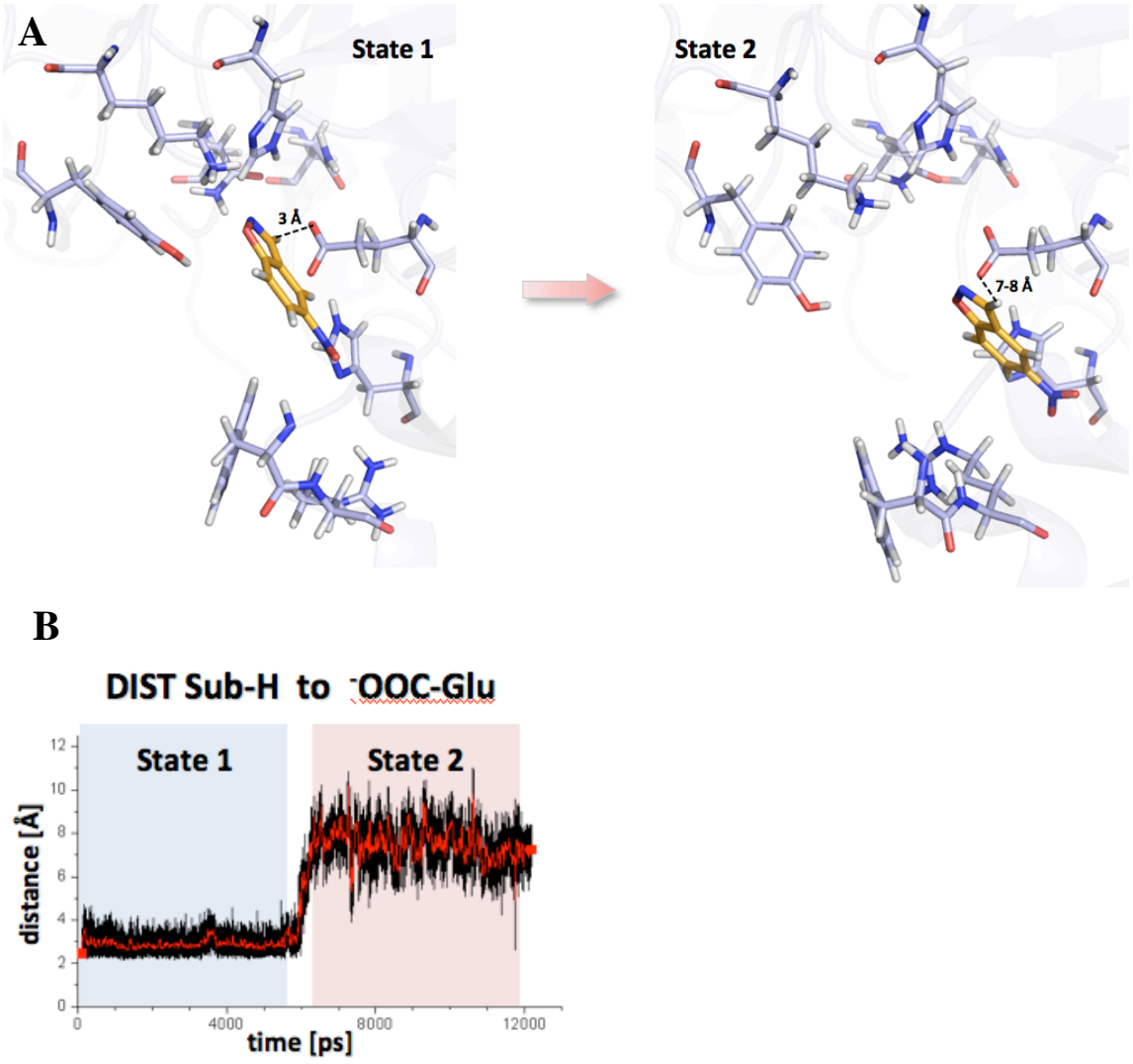


Figure 3-11. MD analysis of HG-1. **(A)** Representative structures of HG-1 in state 1 where the substrate is bound in the active site and in state 2 after the substrate has exited the active site. **(B)** The substrate-base distance over the course of the simulation. States 1 and 2 are indicated.

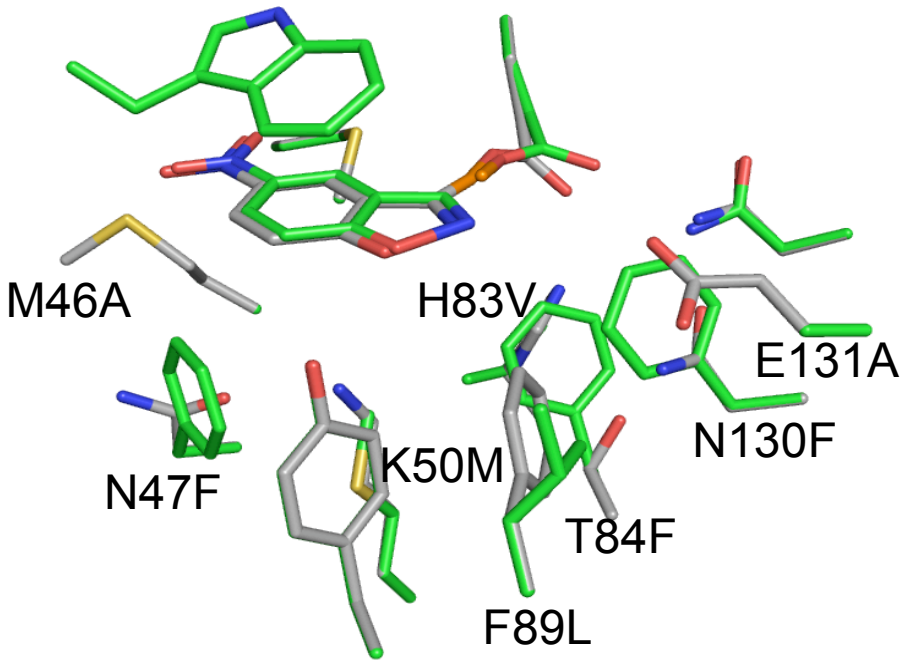


Figure 3-12. Active site of HG-1h. The active site residues of HG-1h (green) are shown overlaid with the HG-1 active site (grey). Mutations with respect to HG-1 are indicated.

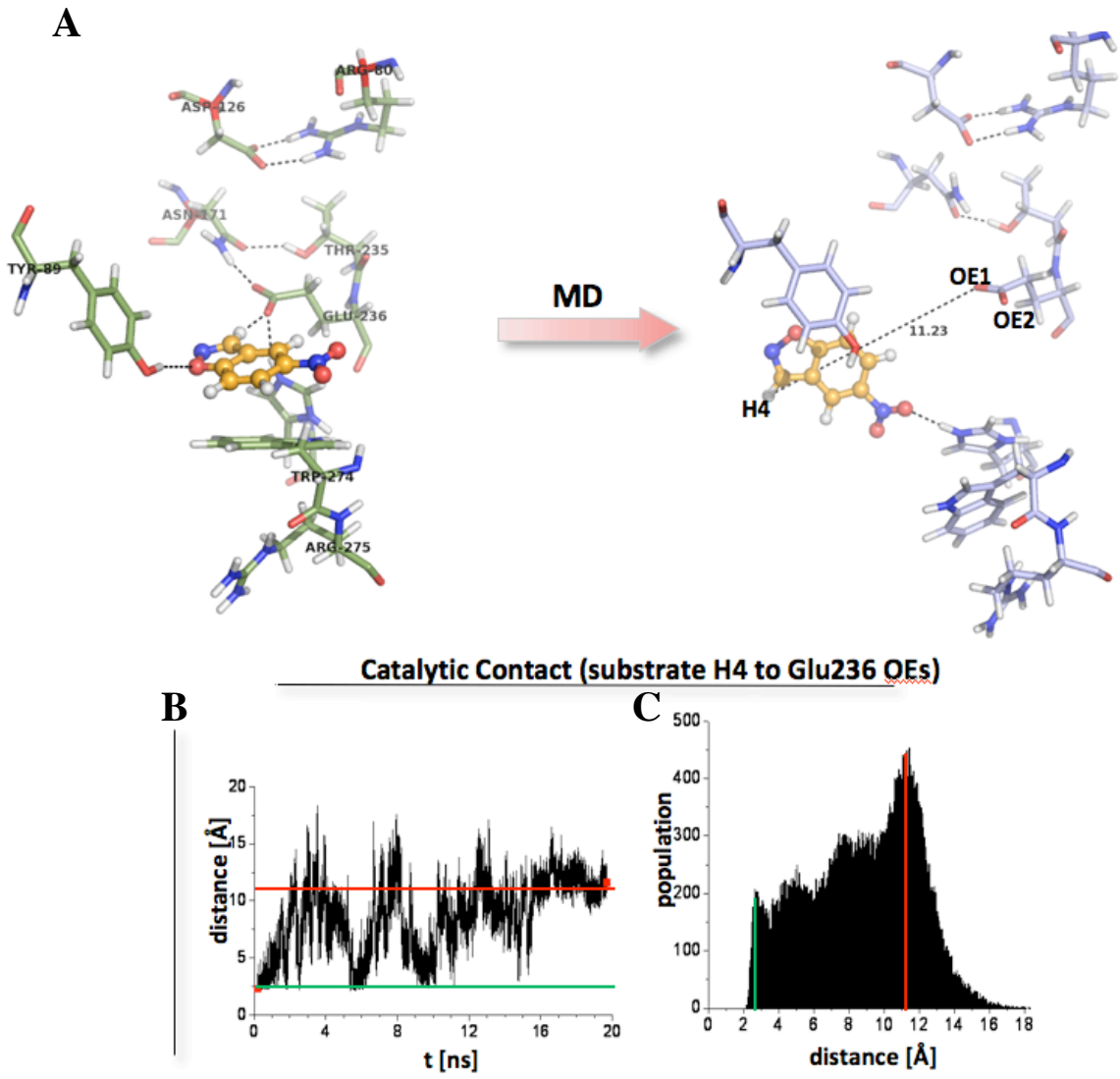


Figure 3-13. MD analysis of HG-1h. **(A)** Representative structures before (green) and after the MD simulation (blue). **(B)** The substrate-base distance over the course of the simulation. **(C)** The distribution of the substrate-base distances for the entire simulation. The substrate-base distance for the initial configuration is indicated by a green line and the substrate-base distance for the final configuration is indicated by a red line.

Chapter IV

The computational design and molecular dynamics simulation analysis of multiple Kemp elimination enzymes

All of the experimental work described in this chapter was performed in the Mayo laboratory. Gert Kiss of the Houk lab (UCLA) carried out the molecular dynamics simulations.

Abstract

Crystallographic and molecular dynamics (MD) simulation analysis of our inactive Kemp elimination (KE) design HG-1 suggested that possible reasons for the inactivity lay in the flexible, solvent-exposed active site. In an effort to construct a more hydrophobic, rigid active site in the same protein scaffold, we placed the catalytic residues farther into the barrel of the protein, resulting in an active KE enzyme. Using similar computational methods, three other active KE designs were also identified in two scaffolds that had been previously shown to be amenable to computational design procedures for this reaction. We also show that MD analysis of these designs is able to accurately differentiate between active and inactive designs, suggesting a novel *in silico* screening step for the computational enzyme design process.

Introduction

Most of the expense and time associated with computational enzyme design is in the experimental evaluation of designs predicted by the software. After a protein sequence is predicted to have activity for a given reaction, the corresponding DNA sequence must be ordered from a commercial source or constructed from scratch, the heterologous expression of the protein must be optimized and a purification scheme must be developed. Each of these steps has the potential for unforeseen problems unique to the system of interest, which may take weeks to overcome.[1] Indeed, much work has been dedicated to overcoming challenges associated with heterologous protein expression including optimizing gene sequences, using cell-free extracts, and employing eukaryotic expression systems.[1,2]

Before activity assays can be carried out, the substrate must be purchased or chemically synthesized and purified. Only then can the expressed protein can be assayed for activity and stability and characterized structurally. At a minimum, the entire process starting from the computationally predicted protein sequence can take a month and cost thousands of dollars in synthetic genes, oligonucleotides, and reagents.

Recent successes in *de novo* computational enzyme design have relied in part on the synthesis and screening of a large number of putative enzymes.[3,4] Although these methods did result in active enzymes, the time and costs associated with their synthesis and screening make the process extremely inefficient and point to inaccuracies in the design process. *De novo* enzyme design with our protein design software, ORBIT, and the recently developed software called Phoenix, has also been challenging (see Chapter

III and Appendix A), indicating that we also have undetermined problems in our design procedure.

Two strategies could help us address the inefficiencies in computational enzyme design: (1) Use the protein design cycle, mentioned in Chapter I and Chapter III, to analyze the inactive or less active designs and modify the design procedure accordingly. (2) Identify sequences from the design process that are likely to be inactive so that experimental time and money are focused on active enzymes.

Here, we attempt to use both strategies in the context of the Kemp elimination (KE). First, we use information from the analysis of a previous inactive KE design to inform the direction of new designs for the KE in the *Thermoascus aurantiacus* xylanase (TAX) scaffold. This cyclic design procedure resulted in a highly active KE enzyme that appears to accelerate the reaction with a 10^6-fold rate increase.

Second, we created a small set of KE enzymes, which were analyzed by a new MD simulation procedure to observe dynamics of the transition state (TS) structure, explicit solvent, and the catalytic residues within the active site. These analyses correctly predicted the presence or absence of activity in most of the designs. A method of this type could serve as an initial screen to help us eliminate false-positives from the design process.

The inability of our design procedure to reliably predict active sequences and the success of the MD simulations in making that distinction point to several limitations of our design procedure. Both ORBIT and Phoenix use fixed backbone structures to represent the scaffold and neither can currently model dynamics or explicit solvation. However, both of these factors are modeled in MD simulations. The combination of

fixed-backbone computational enzyme design and MD simulation could combine the best features of both procedures to improve the accuracy and efficiency of the computational enzyme design process.

Materials and methods

Active site placement

The protein design software Phoenix, which is similar to ORBIT, was used for these designs.[5,6,7]

For the active site search in design HG-2, all non-Pro positions within 5 Å of D127 in 1GOR that point into the barrel of TAX were designated design positions (residues 17, 42, 44, 81, 83, 129, 170, 172, 209, 234, 236, 237, 265, and 267). These positions were allowed to sample all conformations of Gly, Phe, Trp, Ser, Thr, and Tyr. Position 127 was defined as the only catalytic position and was allowed to sample all conformations of Asp. As in the design of Chapter III, a backbone-independent conformer library was used.[8] The library of TS poses was generated through the targeted ligand placement as discussed in Chapter II and Chapter III using the geometric variations in Table 3-2. The other catalytic contacts were enforced using the geometric constraints described in Chapter III (Table 3-3) except that the additional hydrogen bond contact to the base (Table 3-4) was not required. The catalytic contacts identified in this design were D127, W44, and S265.

The active site placement for the designs in scaffolds 1THF and 1A53 were carried out in a similar manner. The design positions and the required contacts for these

designs are shown in Table 4-1 along with the actual residues chosen as catalytic contacts in the active site search.

Active site repacking

For the repacking calculation, the initial TS position was taken from the active site chosen after the active site search. From this initial position, the TS structure was translated ± 0.4 Å in x, y, and z in 0.2 Å steps and rotated 10° in each direction in 5° steps. The geometric constraints from Tables 3-3 and 3-4 were applied to enforce the contacts between the TS and each of the three catalytic residues identified in the active site search.

For design HG-2, residues 42, 21, 81, 83, 84, 125, 130, 172, 234, 236, and 267 were designated design positions and were allowed to sample all conformations of all residues except Pro and Cys. A second shell of float positions was designated around the design positions (residues 16, 17, 46, 47, 50, 79, 87, 90, 170, 207, 209, 239, 275, and 276), which were allowed to sample all conformations of the wild-type residue at that position. Positions 44, 127, and 265 were allowed to sample all rotamers of Trp, Asp, and Ser, respectively. The residues that were designated as design and float residues for the calculations in scaffolds 1THF and 1A53 are shown in Table 4-1.

An occlusion-based solvation potential was applied with scale-factors of 0.05 for nonpolar burial, 2.5 for nonpolar exposure, and 1.0 for polar burial.[7] Other standard parameters were applied as in Lassila *et al.* and a backbone-independent conformer library was used to represent side chain flexibility.[8] As in the active site search, sidechain-TS interaction energies were biased to favor those contacts that satisfy the

geometries in Table 3-3 and Table 3-4. Sequence optimization was carried out with FASTER,[9,10] and a Monte Carlo-based algorithm[11,12] was used to sample sequences around the minimum energy conformation from FASTER (FMEC).

Gene synthesis and cloning

Genes for the designed proteins were designed and constructed as described in Chapter III. The DNA sequence for a factor Xa cleavage site and a His_6-tag added to the C-terminus of all genes. Closely related sequences (e.g., wild-type TAX, HG-1, and HG-2) were synthesized from a common oligonucleotide set with new oligonucleotides introduced as needed for differences in the gene sequence. Site-directed mutagenesis of the genes to create point mutations is also described in Chapter III.

Protein expression and purification

For initial activity screening, 200 μL LB/ampicillin starter cultures were inoculated with a single BL-21 (DE3) *E. coli* colony and grown at 37°C overnight with shaking. The entire starter culture was used to inoculate 5 mL of Overnight Express Instant TB media (Novagen) with ampicillin in 24-well culture plates (Whatman). The plates were fitted with Bugstopper Venting Capmats (Whatman) and incubated for 3 hours at 37°C with shaking. Expression was continued at 18°C overnight with shaking. The cells were harvested by centrifugation and then washed with 1 mL phosphate buffered saline. Cell pellets were frozen on dry ice for 30 min, then thawed at 4°C and resuspended in 400 μL lysis buffer (50 mM sodium phosphate, pH 8.0, 300 mM NaCl, 2.5 mM imidazole, 1x CelLytic B (Sigma-Aldrich), 1 mg/mL lysozyme (Sigma-Aldrich),

and 1 U Benzonase endonuclease (Merck)). The lysate was centrifuged at 5000 × g for 20 min at 4°C and crude protein samples were taken from the supernatant.

For further purification of 5 mL cultures, His-Select plates (Sigma-Aldrich) were equilibrated with 600 μL of equilibration buffer (50 mM $NaPO_4$ pH 8.0, 300 mM NaCl). The crude supernatant was applied to the plate wells and the columns were washed with 600 μL of wash buffer (50 mM $NaPO_4$ pH 8.0, 300 mM NaCl, 5 mM imidazole). The purified protein was eluted in 500 μL elution buffer (50 mM $NaPO_4$ pH 8.0, 300 mM NaCl, 250 mM imidazole).

For larger-scale production, proteins were expressed in 1 L LB/ampicillin cultures of BL-21 (DE3) *E. coli* as described in Chapter III except that the expression was induced at 18°C for up to 18 hours. The proteins were purified using 1 mL of Ni-NTA resin (Qiagen) and the elution buffer was exchanged with 50 mM sodium citrate, 150 mM NaCl pH 5.5 or 25 mM HEPES, 100 mM NaCl pH 7.25.

Protein concentration determination

Protein concentrations were measured by UV absorbance after protein denaturation in 8 M guanidinium hydrochloride for 10 min with a dilution of at least 10x. The extinction coefficient of each protein at 280 nm was calculated based on the number of Trp and Tyr residues in the sequence (Table 4-2).

Protein characterization

Circular dichroism (CD) and mass spectrometry (MS) were carried out as in Chapter III. CD spectra and thermal denaturation curves were obtained with 10 μM protein in 25 mM HEPES, 100 mM NaCl pH 7.25 or 25 mM MES, 100 mM NaCl pH 5.5.

KE activity screening

5-nitrobenzisoxazole (5-NBZ) was synthesized and purified as described in Chapter III. The reactions were carried out in a total volume of 200 μL in black 96-well microtiter plates with clear bottoms (Greiner) at 27°C. The reaction contained 20 μL of the lysate supernatant and 1 mM 5-NBZ diluted in 25 mM HEPES, 100 mM NaCl pH 7.25. The acetonitrile concentration was 2% for all reactions. The production of the phenolate product (Figure 1-2) was monitored by an increase in absorbance at 380 nm using a Safire^2 microplate reader (Tecan).

Kinetic measurements

Kinetic parameters were determined by monitoring phenolate production at 380 nm (ε = 15,800 $cm^{-1}M^{-1}$)[13] using a Shimadzu UV-1601 spectrophotometer equipped with a temperature-controlled cell holder. A 50 mM stock of 5-nitrobensoxazole was made in acetonitrile and stored at –20°C. Assays were carried out with 5 μM protein in 25 mM HEPES, 100 mM NaCl pH 7.25 at 27°C. Assays contained substrate concentrations between 16 μM and 1 mM and the acetonitrile concentration was kept constant at 2% of the 1 mL reaction volume. Initial reaction rates were determined from the linear portion

of the reaction progress curve and were corrected for the initial rate contribution for the buffer catalyzed (no protein) reaction.

Kaleidagraph was used to fit the data to the Michaelis-Menten equation (Equation-1) to determine the kinetic parameters k_{cat} and K_m:

$$v_0 = \frac{k_{cat} \cdot [E]_T \cdot [S]_0}{K_m + [S]_0} \qquad \text{Equation-1}$$

where v_0 is the initial reaction rate, $[E]_T$ is the total enzyme concentration and $[S]_0$ is the initial substrate concentration. In the cases where the enzyme was not saturated because of limited substrate solubility, the data were fit to a line to determine k_{cat}/K_m.

pH profiles

pH profiles were determined in 96-well microtitre plates by monitoring the absorbance increase at 380 nm with a Safire^2 microplate reader (Tecan). The final concentration of protein was 10 μM in a 200 μL total volume. Each reaction was carried out in 25 mM buffer (pH 5.4: MES, pH 6.6 and 7.0: potassium phosphate, pH 7.25: HEPES, pH 8.0 and 9.0: Tris) and 100 mM NaCl. A dilution series of 5-NBZ stocks was made in acetonitrile, resulting in substrate concentrations between 0.39 mM and 50 mM. 4 μL of each substrate stock was added to each reaction well. Initial rates were calculated from the first 400 sec of measurements. The initial rates of the buffer-catalyzed reaction (no protein) were subtracted from the protein-catalyzed initial rates. A pathlength of 0.67 cm was assumed for a volume of 200 μL in the 96-well plates.

Crystallography

Crystallization screens were set up by the Caltech Molecular Observatory. Crystals for HG-2 were obtained in three conditions at 20°C and a protein concentration of 9.5 mg/mL: (1) 100 mM MES, pH 6.5, 1.6 M magnesium sulfate (Figure 4-5C); (2) 100 mM HEPES, pH 7.5, 20 mM magnesium chloride, 22% (w/v) polycrylic acid 510 sodium salt; (3) 100 mM HEPES, pH 7.5, 1.26 M ammonium sulfate. Crystals for HG-2/K50A were obtained at 20°C in 200 mM ammonium sulfate, 30% (w/v) PEG 8000, 9.9 mg/mL protein.

Molecular dynamics (MD) simulations

MD simulations were carried out as described in Chapter III, Materials and Methods.

Results and Discussion

Second-generation designs

The design calculations for the second-generation designs were carried out using Phoenix, a computational protein design software package similar to ORBIT, which has also been developed in the Mayo lab.[5,6] Like ORBIT, Phoenix utilizes geometric-constraint-based methods to generate libraries of ligand poses including targeted ligand placement and rotation/translation;[8] stochastic search methods, including FASTER and Monte Carlo, are used for sequence optimization.[10]

The second-generation designs were carried out in the same scaffold as HG-1 (see Chapter III) with identical geometric constraints. A wild-type carboxylate residue in the TAX scaffold, D127, was identified as being in a promising position to serve as the catalytic base. This residue is deeply buried inside the core of the protein in an area that appeared to have enough room for the TS and other active site residues. The area surrounding D127 is adjacent to the wild-type binding pocket but is located farther into the barrel of the $(\alpha/\beta)_8$ scaffold (Figure 4-1). One major difficulty associated with using an active site that is not located in the natural binding pocket of a scaffold is potential destabilization of the protein by disruption of important interactions in the core to make room for the TS and catalytic residues.

The active site chosen in the active site search consisted of D127 as the general base, W44 as the π-stacking residue and S265 as the hydrogen bond donor (Figure 4-2A). This active site displays good stacking between W44 and the TS, and good hydrogen bond geometry between D127 and the TS H3. In addition, the oxazole ring is pointed into the back of the active site pocket and is well shielded from solvent. The nitro substituant on the TS is solvent-exposed and may be in a position to interact favorably with the wild-type residue K50.

Active site repacking resulted in a sequence that is 12 mutations away from wild-type TAX (Table 4-1, Figure 4-2B). Overall, the residues surrounding the TS and catalytic residues are hydrophobic and are predicted to pack well around the TS structure. In addition to the required polar catalytic contacts, the wild-type residue K50 is predicted to make a favorable interaction with the TS nitro group.

MD analysis of the design, which was done blind to the experimental results, indicated that the active site undergoes a conformational rearrangement that results in two distinct states with active site RMSD values of 1.85 and 2.28 Å from the initial configuration, respectively (Figure 4-3 and Figure 4-4A). This rearrangement results from changes in the secondary structure of the protein due to removal of stabilizing interactions (especially between D127 of β-1 and R81 of β-2) to make room for the active site. As a result of this change in secondary structure, D127 is able to make a favorable hydrogen bond interaction with the backbone carbonyl of G81 about 10 ns into the simulation, drawing it into an alternative conformation that coincides with the transition between the two states.

State 1 is more structurally similar to the original design than state 2 and many of the designed contacts are maintained. In state 2, the substrate has rotated 90° from its initial position, losing the stacking contact with W44 and the electrostatic contact with K50 (Figure 4-3). However, the base-substrate contact is maintained in both states as well as during the transition (Figure 4-4B), due in part to the flexibility of the methionine-rich active site. The designed hydrogen bond contact with S265 is not observed in either state, but the positioning of this residue may force the substrate closer to the base in state 1.

The simulation also shows that the binding site residues pack well around the substrate, preventing the invasion of any explicit solvent molecules into the active site to interact with the base. The lack of water in the active site suggests that the pKa of the base may be appropriately elevated to facilitate abstraction of the substrate proton. In

combination with the well-defined base-substrate contact, these analyses suggest that this design will be active.

A single Ni-NTA affinity chromatography purification step was sufficient to produce about 10 mg of pure HG-2 from 2 liters of culture (Figure 4-5A) and the mass was confirmed with electrospray MS (Figure 4-5B). CD analysis of HG-2 showed that the 12-fold mutant is folded with similar secondary structure to wild-type TAX (Figure 4-6A). However, HG-2 is significantly destabilized with respect to the wild-type protein (Figure 4-6B).

As predicted by the MD simulations, kinetic analysis of HG-2 indicates that the design is an active catalyst for the KE, far above the rate of the buffer catalyzed reaction (Figure 4-7A). The wild-type TAX scaffold was determined to have no activity for the KE. Substrate saturation was never reached for this design, so the kinetic constants k_{cat} and K_m could not be determined reliably. k_{cat}/K_m was determined to be 122 $M^{-1}s^{-1}$ from the slope of the initial rate versus initial substrate concentration plot, making the efficiency of this enzyme comparable to the best KE enzymes designed by Röthlisberger *et al.*[3] Knockout mutations of the base (D127N) and the hydrogen bond donor (S265A) show a significant decrease in activity compared to HG-2, with the base knockout losing almost all activity (Figure 4-7B). The loss of activity upon mutation of the putative catalytic residues indicates that these residues are, in fact, important for catalysis and support the proposed designed mechanism for this enzyme.

The pKa of the base was observed to be significantly elevated from the solvent-exposed pKa of aspartate (~ 7 versus ~ 4) (Figure 4-8A). The pH profile also showed a significant decrease in activity at high pHs, which may indicate the presence of a second

ionizable group in the active site with a high pKa or simply the unfolding of the protein at high pH. CD analysis ruled out the latter possibility, showing that the protein remains folded at pHs up to 9.0 under the conditions of the assays (25 mM HEPES, 100 mM NaCl pH 7.25, 27°C) (Figure 4-8B). Further investigation into the cause of the decrease in activity of HG-2 at high pH will be necessary.

Third-generation designs

In an attempt to increase activity of HG-2, point-mutants of HG-2 were selected based on the MD simulations. G81A was chosen to rigidify the local secondary structure to prevent the backbone hydrogen bond to D127 (Figure 4-3). However, MD analysis of HG2/G81A indicates that though the secondary structure of the loop may be stabilized somewhat, the base can adopt an alternate conformation, preventing interactions with the substrate. Subsequent experimental analysis of this mutant showed that it was less active than the original design HG-2 (Figure 4-9, Table 4-1).

MD analysis of HG-2 suggested that S265 was acting to push the substrate closer to the base and prevent water from entering the active site. The S265T mutation was chosen because of the slightly larger volume of Thr, which may provide better packing around the substrate. In addition, this mutation also reverts to the wild-type residue at this position. MD analysis of HG2/S265T predicted that this variant would be more active than the original design as the substrate-base interaction is maintained just as well as in HG-2, but no active site rearrangement occurs. Experimental evaluation showed that although this design has a slightly lower k_{cat} than HG-2, the K_m is also lower,

resulting in a k_{cat}/K_m that is about three fold higher than HG-2 (Figure 4-9, Table 4-1) or any of the KE enzymes from Röthlisberger *et al*.[3]

Additional designs

With one successful KE design in hand, we tested Phoenix's ability to recapitulate the active sites of enzymes with known activity that had been designed by Röthlisberger *et al*.[3] Starting with the scaffolds and base positions for KE07, KE10, and KE59, we applied our enzyme design methods to generate active sites within the binding pockets of *Thermotoga maritima* imidazoleglycerolphosphate synthase (PDB: 1THF)[14] and *Sulfolobus solfataricus* indole-3-glycerolphosphate synthase (PDB: 1A53)[15]. The methods used were the same as those described above for HG-2 except that lysine was allowed to be the hydrogen bond donor catalytic contact to fully access some of the designs.

Starting with the scaffold and catalytic base position from KE59 (1A52, E231), we used targeted ligand placement to generate ligand poses and required a π-stacking residue and a hydroxyl hydrogen bond donor in addition to the base contact as defined in Table 3-2. The sequence containing catalytic residues from KE59 (W110, S131, E231) was identified in the top 20 active site sequences from the design calculation. The catalytic residues for this design were predicted to overlay well with the catalytic residues of KE59 (Figure 4-10A). The repacking calculation introduced noncatalytic active site residues that differed from those in KE59, resulting in 1A53-1, an 8-fold mutant from KE59. However, this design showed good stacking interactions and an additional hydrogen bond donor contact to the base (S210) (Figure 4-10B).

MD analysis predicted that this design would be inactive because the catalytic base, E231, rearranges readily to form stable interactions with nearby S210 and S211 preventing productive interaction with the substrate (Figure 4-13A, B). In addition, the active site was not well packed around the substrate, allowing the substrate to adopt alternate binding modes (Figure 4-13C). The inactivity of this design was confirmed by experimental KE activity screening (Figure 4-11).

Starting with the scaffold and catalytic base position from KE10 (1A53, E178), targeted ligand placement was used to generate ligand poses. As seen in KE10, the only contacts required in the active site search were a π-stacking and base contact. Phoenix identified the catalytic residues of KE10 (E178 and W210) through the active site search. In addition, a second π-stacking contact was identified (W110), essentially sandwiching the substrate in the active site (Table 4-1, Figure 4-10C). The active site was repacked around the base and double π-stacking contacts, resulting in 1A53-2, a 12-fold mutant from the wild-type sequence and a 9-fold mutant from KE10 (Table 4-1, Figure 4-10D).

The MD analysis of this design showed that the binding pocket contributes to a significant reorientation of the substrate over the course of the simulation, at first maintaining the base-substrate contact (Figure 4-14A, B, C, E). Towards the end of the simulation, the flexible active site allows water molecules to enter and the substrate diffuse away from the base (Figure 4-14D, F). Because of the sharp distance distribution of the substrate-base contact (Figure 4-14E), the design was predicted to have some activity. However, it was not expected to be highly active because of the late intrusion of solvent and diffusion of the substrate. This design was determined to be active by experimental KE activity screening (Table 4-3, Figure 4-9).

In addition to the active site of 1A53-2 found in the KE10 recapitulation calculation, an alternate active site was identified with a base at E157 and π-stacking contacts at W110 and W210 (Table 4-1, Figure 4-10E). Active site repacking resulted in the design 1A53-3, which is an 11-fold mutant from the wild-type scaffold (Figure 4-10F).

MD analysis predicted 1A53-3 to be active because the active site configuration is maintained over the course of the simulation (Figure 4-15). Soon after the simulation begins, substrate rotates slightly from the initial configuration, but the base maintains contact with both the substrate and the π-stacking (W209) residue. In addition, only one water molecule is able to enter the active site and interact with the base (Figure 4-15B). Experimental evaluation of 1A53-3 confirmed the MD prediction of activity for this design (Table 4-3, Figure 4-9).

Two strategies were implemented in the recapitulation of KE07. In the first, a base, π-stacking residue, and a Lys hydrogen bond donor are required. However, Röthlisberger *et al.* found that the removal of the Lys hydrogen bond donor from the active site of KE07 actually increased activity because this residue may interact directly with the base, lowering the pKa.[3] Thus, in the second strategy, we required a hydroxyl hydrogen bond donor instead of the Lys. Starting with the base E101 in the scaffold 1THF, Phoenix identified W50 and K222 as the catalytic residues in the first calculation (Figure 4-12A) and W50 and S201 as the catalytic residues in the hydroxyl hydrogen bond donor case (Figure 4-12C). In both of these designs, W128 is identified as an additional π-stacking contact. Repacking of the E101/W50/W128/K222 active site resulted in the design 1THF-1, which is a 13-fold mutant from wild type and a 9-fold

mutant from KE07. In this design, the base is anchored by S78 and the substrate is held in place by the two Trps (Figure 4-12B). Repacking the E101/W50/W128/S201 active site resulted in the design 1THF-2, which is a 10-fold mutant from the wild type and a 10-fold mutant from KE07.

MD analysis of 1THF-1 indicated that the active site maintains its overall configuration with respect to the initial structure during the simulation (Figure 4-16A, B). The base-substrate contact is less well maintained than the previously described active designs (Figure 4-16C) because the base alternates its interactions between the substrate and solvent molecules in the pocket. As in the crystal structure of KE07, K222 appears to interact directly with the base, perhaps lowering the pKa.[3] Because the base-substrate distance seemed appropriate for catalysis, this design was predicted to be active, although less active than KE07. For 1THF-2, the overall configuration is the same as the initial structure (Figure 4-17A, B), and it also has an extremely sharp base-substrate distance distribution that is characteristic of previous active designs (Figure 4-17C). The simulation indicated that the substrate moved about 2 Å deeper into the pocket than was predicted in the design, preventing S201 from maintaining its contact to the substrate. However, in multiple KE designs, the hydrogen bond donor contact has been shown not to be critical for catalysis.[3] As a result, this design was predicted to be more active than 1THF-1. Experimental evaluation could not detect any activity in 1THF-1 but substantial activity was detected in 1THF-2 (Table 4-3, Figure 4-9), agreeing with the MD prediction.

We were able to recapitulate the placement of the active site residues in all three of the previously designed KE catalysts. However, the final sequences that were

generated by Phoenix were very different from the existing active designs. The sequences of the active enzymes found by Röthlisberger *et al.* are not necessarily the optimal sequences with respect to activity; each is just one of the many possible active sequences for a given set of catalytic residues. Despite our repacked sequences being at least eight mutations away from the active design, 3 out of 5 of our designs showed significant KE activity. In some cases, our sequences were more active than the corresponding sequence that we were trying to recapitulate and in other cases, our sequences were less active, or lacked activity completely. Starting with the initial sequence from computational enzyme design, it has been previously demonstrated that KE activity can be significantly improved through many rounds of directed evolution;[3] however, we have shown here that activity can also be optimized by additional rounds of design based on structural or MD analysis. In both cases, computational enzyme design served to generate an active starting point for further optimization by arranging critical catalytic residues in productive orientations in an environment amenable to the chemical reaction. This initial step is necessary for directed evolution, which generally cannot generate enzymatic activity *de novo*.[16]

MD analysis versus experiment

When compared with experimental results, MD analysis successfully predicted activity in HG-2, 1THF-2, 1A53-2, and 1A53-3, as well the inactivity of 1A53-1 (Table 4-4). In addition, the MD analysis was able to predict relative levels of activity in point mutants of HG-2. In the case of KE enzymes, the main criteria for prediction of activity versus inactivity were solvent inaccessibility of the active site, stable binding of the

substrate, and a stable substrate-base contact within hydrogen bonding distance. The success of these criteria in predicting activity agrees with the original work of Kemp *et al.* on the decomposition of benzisoxazoles, which shows that nonpolar solvents enhance the rate of the reaction significantly relative to water,[13,17] and more recent work, which shows that the nonpolar microenvironment of the binding pocket and the positioning of the base in the catalytic antibody and serum albumins are important.[18,19]

The criteria that are used to predict activity will likely vary for other reaction types and will help to elucidate the subtle requirements for a successful catalyst. However, determination of the best criteria for MD analysis may not be straightforward, especially for reactions that have been less well studied; MD evaluation of positive and negative controls for similar reactions will likely be necessary.

Conclusions

In this work, we used Phoenix to successfully design four active KE enzymes in three different inert scaffolds, demonstrating the applicability of Phoenix to *de novo* computational enzyme design. Of these enzymes, HG-2/S265T showed a catalytic efficiency that was 3 times higher than any other computationally designed KE enzyme. The success of the Phoenix KE designs is due, in large part, to iterative analysis and redesign of inactive enzymes as a part of the protein design cycle discussed in Chapters I and III. The analysis of inactive designs provides valuable information as to the deficiencies of our design procedure, allowing us to correct these problems to produce active enzymes.

We also show that MD simulations of these enzymes can predict the presence or absence of activity and, in the case of point mutants of HG-2, the relative level of activity can also be predicted. Once the MD methods are optimized and fully automated, these types of analyses could be used as a pre-screen to help eliminate any false-positives that may result from the design process, allowing us to focus our experimental efforts on designs that are most likely to be active. That the results of our computational protein design procedure can be assisted by the addition of MD analysis points to three major deficiencies in our design procedure: the reliance on a fixed backbone scaffold, discrete sidechain rotamers, and implicit solvation, all of which help make our design calculations tractable. The MD analysis of computationally designed enzymes can serve as an immediate solution to these problems until new methods that address these deficiencies are fully integrated.

While the activity of these Phoenix designs is encouraging, the enzymes described here still do not demonstrate rate accelerations or efficiencies close to those of natural enzymes. In addition, the K_ms of most of the computationally designed enzymes for this reaction so far have been above 1 mM, suggesting that the active sites are not optimized for substrate binding. Crystal structures of the designs described here are currently being pursued to identify possible improvements in the active site structure for better substrate binding in fourth-generation designs. In addition, directed evolution has previously been shown to be an effective supplemental strategy for significantly enhancing both the k_{cat} and K_m of designs with low-level activity.[3]

Even though the complete recapitulation of the active sites of previous KE designs was not possible, the redundancy of the active designs with respect to the base

position and scaffold demonstrate the ability of computational protein design to provide multiple correct answers to *de novo* enzyme design problems, yielding multiple starting points for further optimization by directed evolution. *In vitro* evolution studies are currently underway in the Hilvert lab at the Swiss Federal Institute of Technology to optimize the activity of some of these designs, and the results of these experiments should help us to better understand this model system and the limitations of computational enzyme design.

Acknowledgements

We would like to thank Gert Kiss from the Houk Lab at UCLA for carrying out the MD simulations and subsequent analysis, as well as providing the related figures. Thanks to Pavle Nikolovski at the Caltech Molecular Observatory for setting up crystallization screens and to Toni Lee for helping to express protein and optimize crystallization conditions, and for carrying out the crystallography of HG-2 and its variants. The gene for KE70 was generously provided by Daniela Röthlisberger of the Baker lab at the University of Washington. Thanks to Rebecca Bloomberg in Professor Donald Hilvert's lab at the Swiss Federal Institute of Technology in Zurich who is currently carrying out directed evolution of HG2/S265T.

References

1. Gustafsson, C.; Govindarajan, S.; Minshull, J., Codon bias and heterologous protein expression. *Trends Biotechnol.* **2004**, *22*, 346-353.

2. Higgins, S. J.; Hames, B. D., *Protein Expression: A Practical Approach*. Oxford University Press: 1999.

3. Rothlisberger, D.; Khersonsky, O.; Wollacott, A.; Jiang, L.; Dechancie, J.; Betker, J.; Gallaher, J. L.; Althoff, E. A.; Zanghellini, A.; Dym, O.; Albeck, S.; Houk, K. N.; Tawfik, D.; Baker, D., Kemp elimination catalysts by computational enzyme design. *Nature* **2008**, *453*, 190-195.

4. Jiang, L.; Althoff, E. A.; Clemente, F. R.; Doyle, L.; Röthlisberger, D.; Zanghellini, A.; Gallaher, J. L.; Betker, J. L.; Tanaka, F.; Barbas, C. F.; Hilvert, D.; Houk, K. N.; Stoddard, B. L.; Baker, D., De novo computational design of retro-aldol enzymes. *Science* **2008**, *319*, 1387-1391.

5. Allen, B. D. *Development and validation of optimization methods for the design of protein sequences and combinatorial libraries*. California Institute of Technology: Pasadena, CA, **2009**.

6. Nisthal, A.; Allen, B. D., (*Manuscript in preparation*). **2009**.

7. Chica, R.; Allen, B. D., (*Manuscript in preparation*). **2009**.

8. Lassila, J. K.; Privett, H. K.; Allen, B. D.; Mayo, S. L., Combinatorial methods for small-molecule placement in computational enzyme design. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 16710-16715.

9. Desmet, J.; Spriet, J.; Lasters, I., Fast and accurate side-chain topology and energy refinement (FASTER) as a new method for protein structure optimization. *Proteins* **2002**, *48*, 31-43.

10. Allen, B. D.; Mayo, S. L., Dramatic performance enhancements for the FASTER optimization algorithm. *J. Comput. Chem.* **2006**, *27*, 1071-1075.

11. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., Equation of state calculations by fast computing machines. *J. Chem. Phys.* **1953**, *21*, 1087-1092.

12. Voigt, C. A.; Gordon, D. B.; Mayo, S. L., Trading accuracy for speed: A quantitative comparison of search algorithms in protein sequence design. *J. Mol. Biol.* **2000**, *299*, 789-803.

13. Casey, M.; Kemp, D.; Paul, K.; Cox, D., Physical organic chemistry of benzisoxazoles. I. Mechanism of the base-catalyzed decomposition of benzisoxazoles. *J. Org. Chem.* **1973**, *38*, 2294-2301.

14. Lang, D.; Thoma, R.; Henn-Sax, M.; Sterner, R.; Wilmanns, M., Structural evidence for evolution of the β/α barrel scaffold by gene duplication and fusion. *Science* **2000**, *289*, 1546-1550.

15. Hennig, M.; Darimont, B. D.; Jansonius, J. N.; Kirschner, K., The catalytic mechanism of indole-3-glycerol phosphate synthase: crystal structures of complexes of the enzyme from *Sulfolobus solfataricus* with substrate analogue, substrate, and product. *J. Mol. Biol.* **2002**, *319*, 757-766.

16. Shao, Z.; Arnold, F. H., Engineering new functions and altering existing functions. *Curr. Opin. Struct. Biol.* **1996**, *6*, 513-518.

17. Kemp, D.; Casey, M., Physical organic chemistry of benzisoxazoles. II. Linearity of the Broensted free energy relation for the base-catalyzed decomposition of benzisoxazoles. *J. Am. Chem. Soc.* **1973**, *95*, 6670-6680.

18. Hu, Y.; Houk, K. N.; Kikuchi, K.; Hotta, K.; Hilvert, D., Nonspecific medium effects versus specific group positioning in the antibody and albumin catalysis of the base-promoted ring-opening reactions of benzisoxazoles. *J. Am. Chem. Soc.* **2004**, *126*, 8197-8205.

19. Seebeck, F. P.; Hilvert, D., Positional ordering of reacting groups contributes significantly to the efficiency of proton transfer at an antibody active site. *J. Am. Chem. Soc.* **2005**, *127*, 1307-1312.

Table 4-1. Summary of design calculations for Kemp elimination enzymes. The residues allowed for each required catalytic contact are indicated along with the actual catalytic residue chosen from the active site search in parentheses. The positions allowed to change identity during the active site search and repacking calculations are indicated as design residues. Float residues are allowed to change conformation but not identity.

		required contacts, allowed residues (actual catalytic contact)				repacking residues		
design	scaffold	base	π-stacking	H-bond	active site search design residues	design	float	mutations
HG-2	1GOR	D (D126)	F/W (T44W)	S/T/Y (S265T)	17, 42, 44, 81, 83, 126, 129, 170, 172, 209, 234, 236, 237, 265, 267	42, 21, 81, 83, 84, 125, 130, 172, 234, 236, 267	16, 17, 46, 47, 50, 79, 87, 90, 170, 207, 209, 239, 275 276	Q22M, T44W, R81G, H83G, T84M, N130G, N172M, A234S, T236L, E237M, T265S, W267F
1THF-1	1THF	E (S101E)	F/W (L50W)	K (L222K)	7, 9, 11, 48, 50, 101, 126, 128, 130, 169, 171, 176, 199, 201, 222, 224	9, 48, 78, 101, 103, 126, 128, 130, 169, 171, 199, 201	7, 11, 176	C9G, V48G, L50W, T78S, S101E, N103M, Q128W, D130A, L169A, T171V, I199G, S201M, L222K
1THF-2	1THF	E (S101E)	F/W (L50W/A128W)	S/T/Y (S201)	"	9, 48, 78, 126, 128, 130, 169, 171, 222	7, 11, 176, 199, 224	C9G, V48G, L50W, T78S, S101E, A128W, D130A, L169A, T171L, L222V
1A53-1	1A53	E (L231E)	F/W (K110W)	S/T/Y (L131S)	51, 81, 83, 89, 108, 110, 112, 131, 133, 157, 159, 178, 180, 182, 184, 210, 211, 231	51, 81, 108, 157, 159, 180, 210	83, 89, 112, 133, 178, 182, 184, 211	E51I, S81G, K110W, L131S, L157G, E159W, N180G, E210S, L231E
1A53-2	1A53	E (G178E)	F/W (K110W/E210W)	none	"	51, 81, 83, 108, 131, 157, 159, 180, 211, 231	8, 53, 89, 108, 112, 133, 182, 184	E51A, S81A, L83A, K110W, L131A, K157A, E159V, G178E, N180A, E210E, S211Q, L231G
1A53-3	1A53	E (L157E)	F/W (K110W/E210W)	none	"	81, 83, 108, 131, 178, 180, 211, 231	8, 53, 89, 112, 129, 133, 182, 184	L51E, S81A, L83A, K110W, L131A, L157E, E159I, N180M, E210W, S211Q, L231G

Table 4-2. Physical characteristics of protein variants.

	molecular weight (g/mol)	ε ($M^{-1}cm^{-1}$)	formal charge	# amino acids
1THF	29185	10360	-8	266
KE07	29350	17120	-5	266
1THF-1	29245	21480	-7	266
1THF-2	29225	21480	-8	266
1A53	30070	16360	0	261
KE59	29969	21920	+1	261
1A53-1	29974	27480	+1	261
1A53-2	29969	27480	+1	261
1A53-3	30071	27480	+1	261
1GOR	34564	55280	-4	318
HG-2	34440	55280	-4	318

Table 4-3. Experimental characterization of designed Kemp elimination enzymes.

parent	mutation	scaffold	k_{cat} (s^{-1})	K_m (mM)	k_{cat}/K_m (s^{-1} M^{-1})	k_{cat}/k_{uncat} [a]	T_m (°C)
HG-2	-	1GOR	NA	NA	130	NA	46.6
HG-2	D127N	1GOR	NA	NA	2	NA	51.0
HG-2	S265A	1GOR	NA	NA	54	NA	45.8
HG-2	K50A	1GOR	NA	NA	37	NA	49.3
HG-2	G81A	1GOR	NA	NA	17	NA	46.7
HG-2	S265T	1GOR	0.68 ± 0.4	1.6 ± 0.1	430	5.9E+05	47.9
1A53-2	-	1A53	0.012 ± 0.002	0.74 ± 0.2	16	1.0E+04	ND
1A53-3	-	1A53	0.015 ± 0.003	0.90 ± 0.03	17	1.3E+04	ND
1THF-2	-	1THF	NA	NA	5	NA	73.7
KE07 [b, c]	-	1THF	0.018 ± 0.001	1.4 ± 0.1	13	1.6E+04	NA
KE07 [b, d]	-	1THF	0.0089 ± 0.006	0.79 ± 0.09	11	7.7E+04	ND
KE70 [b, c]	-	1JCL	0.16 ± 0.05	2.1 ± 2	78	1.4E+05	NA
KE70 [b, d]	-	1JCL	0.070 ± 0.003	0.50 ± 0.06	146	6.0E+04	ND

ND = not determined
NA = not applicable
[a] k_{uncat} under the assay conditions was determined to be 1.16×10^{-6} s^{-1} by Rothlisberger *et al*.
[b] These enzymes were designed by Rothlisberger *et al*.
[c] The kinetic constants were determined by Rothlisberger *et al*.
[d] The kinetic constants were determined as a part of this thesis work.

Table 4-4. Design summary of Kemp elimination enzymes. MD prediction and experimental detection of activity are indicated.

					# mutations		activity	
name	**scaffold**	**base**	**π-stack**	**H-bond**	**from wt**	**from active design**	**MD**	**experimental**
KE59[a]	1A53	L231E	K110W	L131S	10			
1A53-1	1A53	L231E	K110W/E159W	L131S	9	8 (KE59)	no	no
KE10[a]	1A53	G178E	E210W	-	11			
1A53-2	1A53	G178E	K110W/E210W	-	12	9 (KE10)	yes	yes
1A53-3	1A53	L157E	K110W/E210W	-	11	-	yes	yes
KE07[a]	1THF	S101E	L50W	L222K	13			
1THF-1	1THF	S101E	K50W/A128W	L222K	13	9 (KE07)	yes	no
1THF-2	1THF	S101E	L50W/A128W	S201	10	10 (KE07)	yes	yes
HG-2	1GOR	D126	T44W	T265S	12	-	yes	yes
HG2-S265T	1GOR	D126	T44W	T265	11	1 (2.2.0)	increase	higher k_{cat}/K_m
HG2-G81A	1GOR	D126	T44W	T265S	13	1 (2.2.0)	decrease	decrease
HG2-K50A	1GOR	D126	T44W	T265S	13	1 (2.2.0)	decrease	decrease

[a] These KE enzymes were designed by Rothlisberger *et al*.

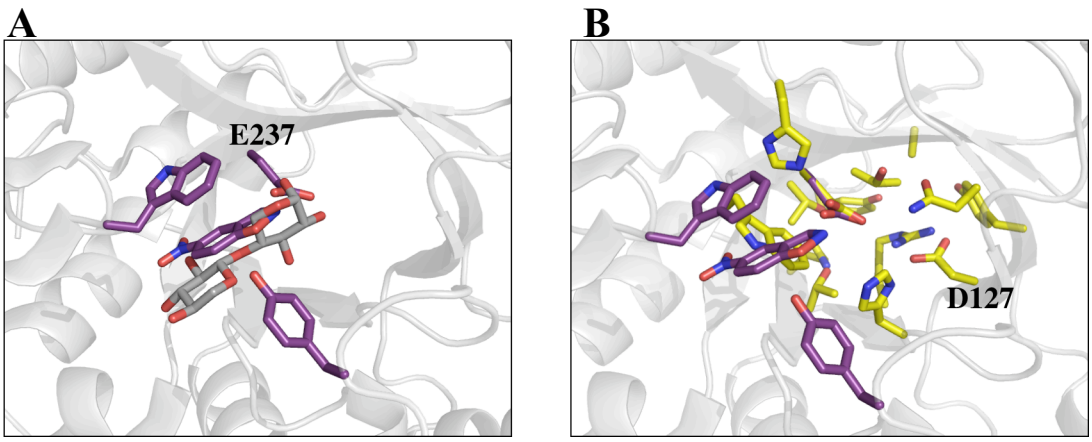


Figure 4-1. Active site locations of first- and second-generation designs in TAX. **(A)** The active site of first generation design 1.2.1 (purple) is in the native binding pocket for xylobiose (grey). **(B)** The active site of second-generation design HG-2 (yellow) is located deeper into the barrel of the scaffold.

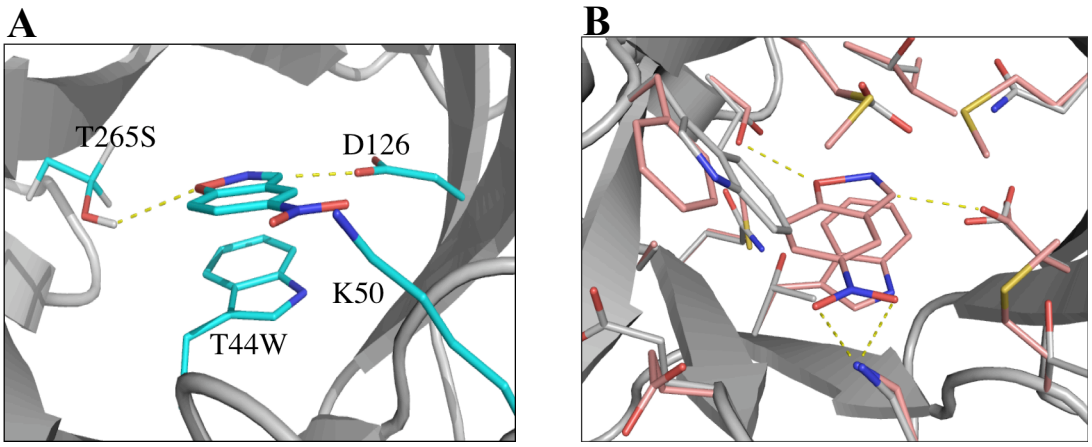


Figure 4-2. Active site of HG-2. (A) Catalytic residues identified through the active site search. **(B)** Repacked active site (pink) overlaid with the wild-type residues (grey).

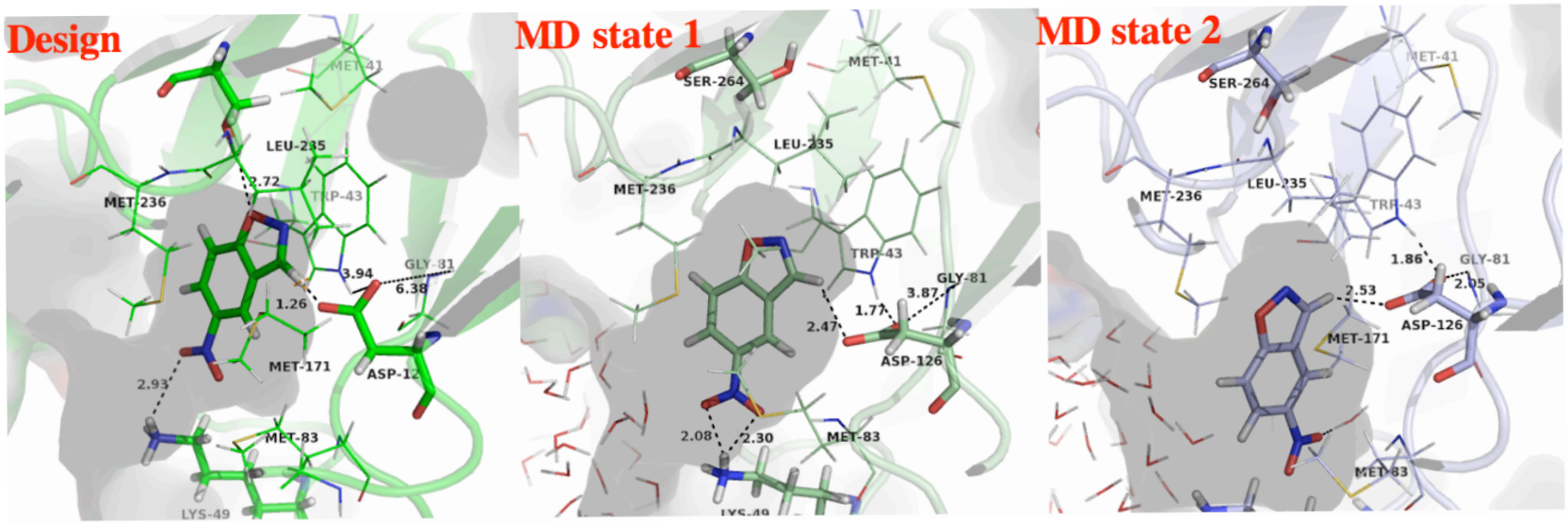


Figure 4-3. Active site structures of HG-2 during the MD simulation. Three active site configurations are shown: the initial configuration, a representative active site from state 1, and a representative active site from state 2.

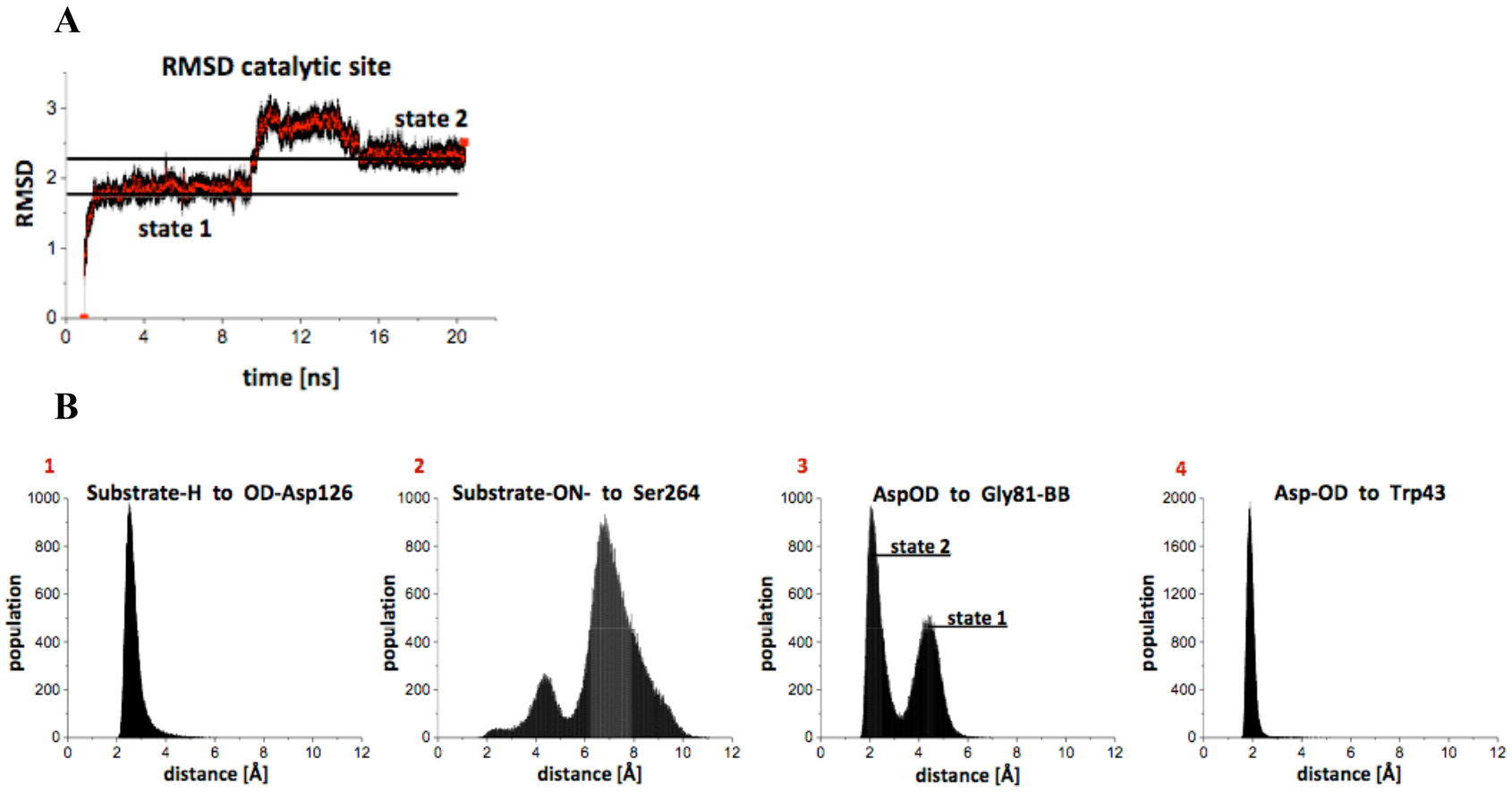
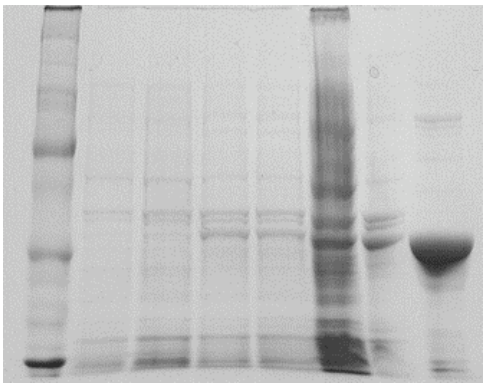
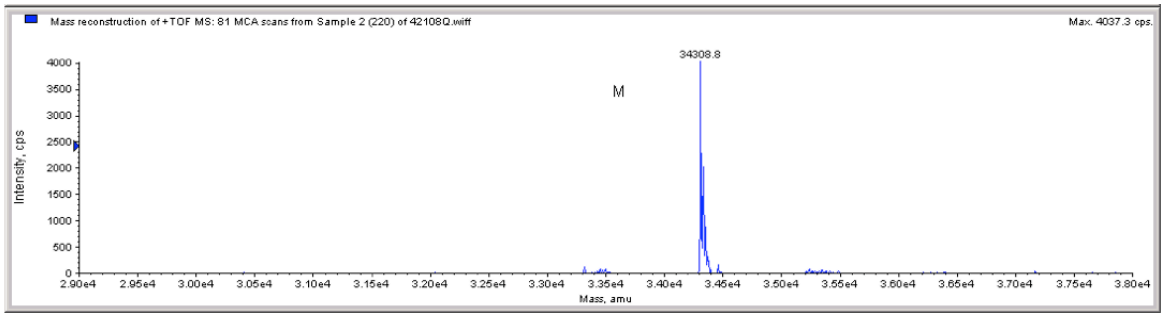


Figure 4-4. Distance distributions of HG-2. (A) Catalytic site RMSD with respect to the initial configuration as the simulation progresses. **(B)** Distance distributions for important catalytic contacts.

A



B



C

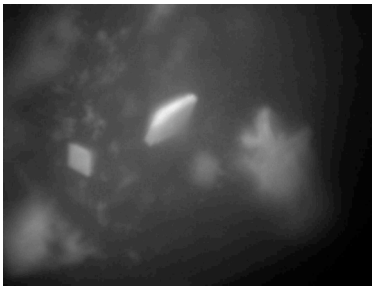


Figure 4-5. Expression and purification of HG-2. **(A)** SDS-PAGE gel of HG-2 expression and purification. Lane 1: MW marker. Lanes 2 and 3: uninduced cells. Lanes 4 and 5: cells 18 hours after IPTG induction. Lane 6: lysate supernatant. Lane 7: lysate pellet. Lane 8: Ni-NTA purification elution. **(B)** Electrospray mass spec. Expected mass: 34308.7, actual mass: 34308.8. **(C)** UV image of HG-2 crystals grown in 0.1 M MES, pH 6.5, 1.6 M Magnesium Sulfate with 9.5 mg/mL HG-2.

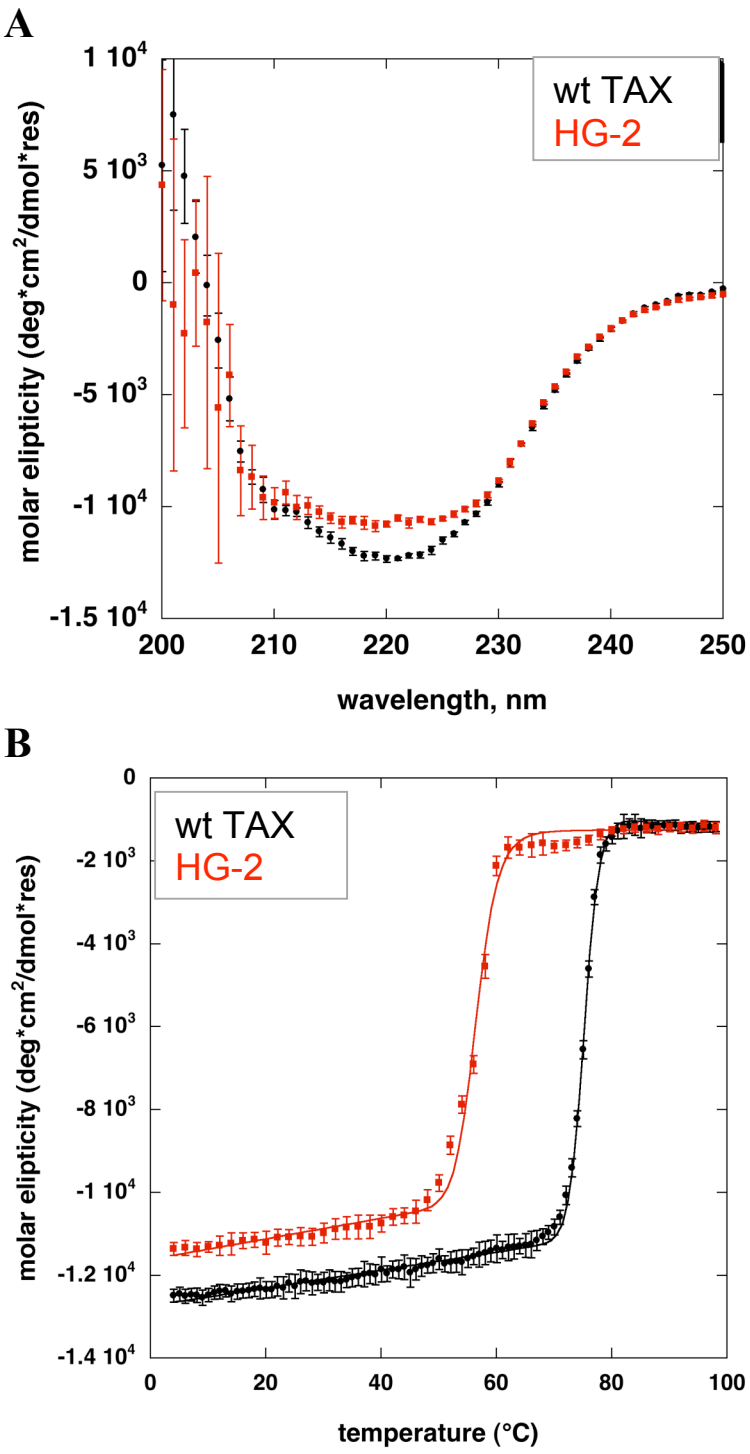


Figure 4-6. CD analysis of wild-type TAX and HG-2. (A) Far-UV wavelength scan. **(B)** Thermal denaturation. All experiments were carried out with 10 μM protein in 25 mM HEPES pH 7.25, 100 mM NaCl.

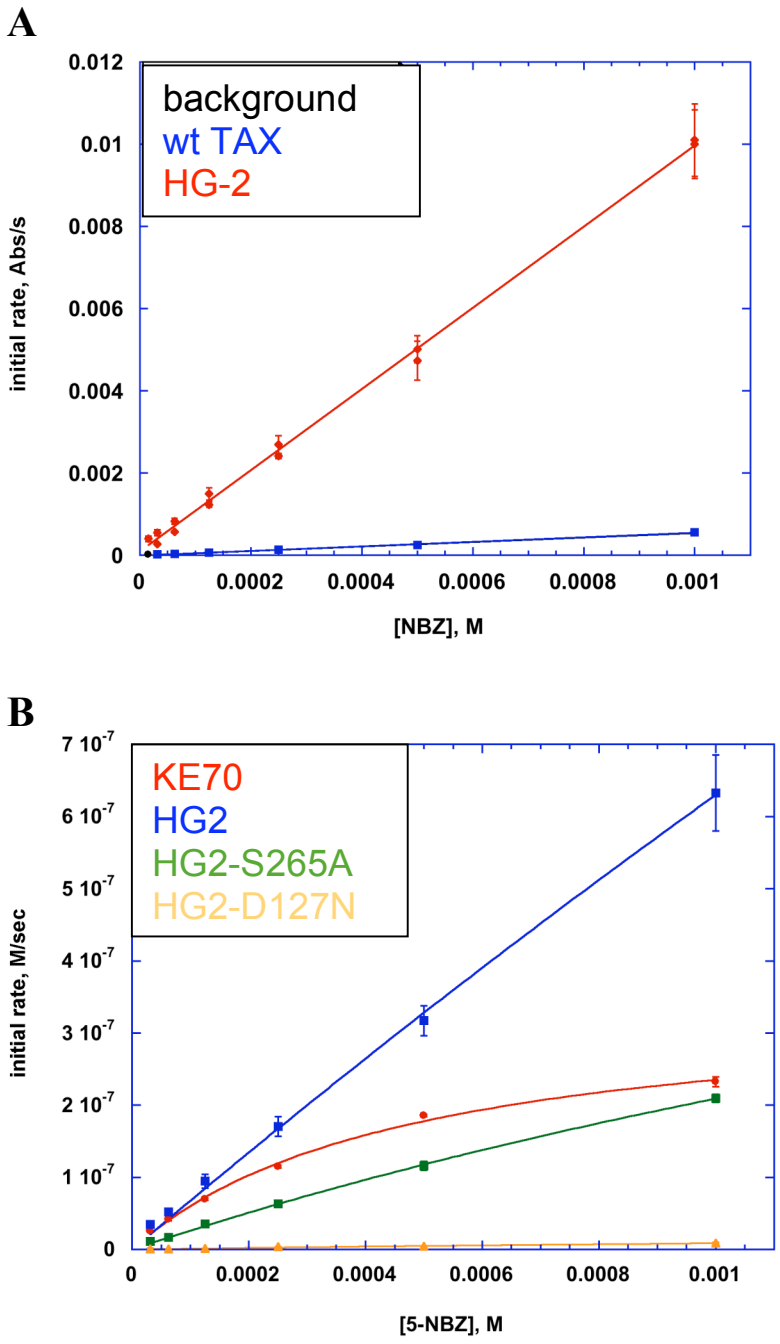


Figure 4-7. Kinetic characterization of second-generation enzymes. 5 μM protein is used in 25 mM HEPES pH 7.25, 100 mM NaCl, 27°C. **(A)** Michaelis-Menten plots of design HG-2 (red), wild-type TAX (blue) and background reaction (black). **(B)** Michaelis-Menten plots of KE70 (red), HG-2 (blue), and HG-2 knock out mutants HG-2-S265A (green) and HG-2-D127N (yellow).

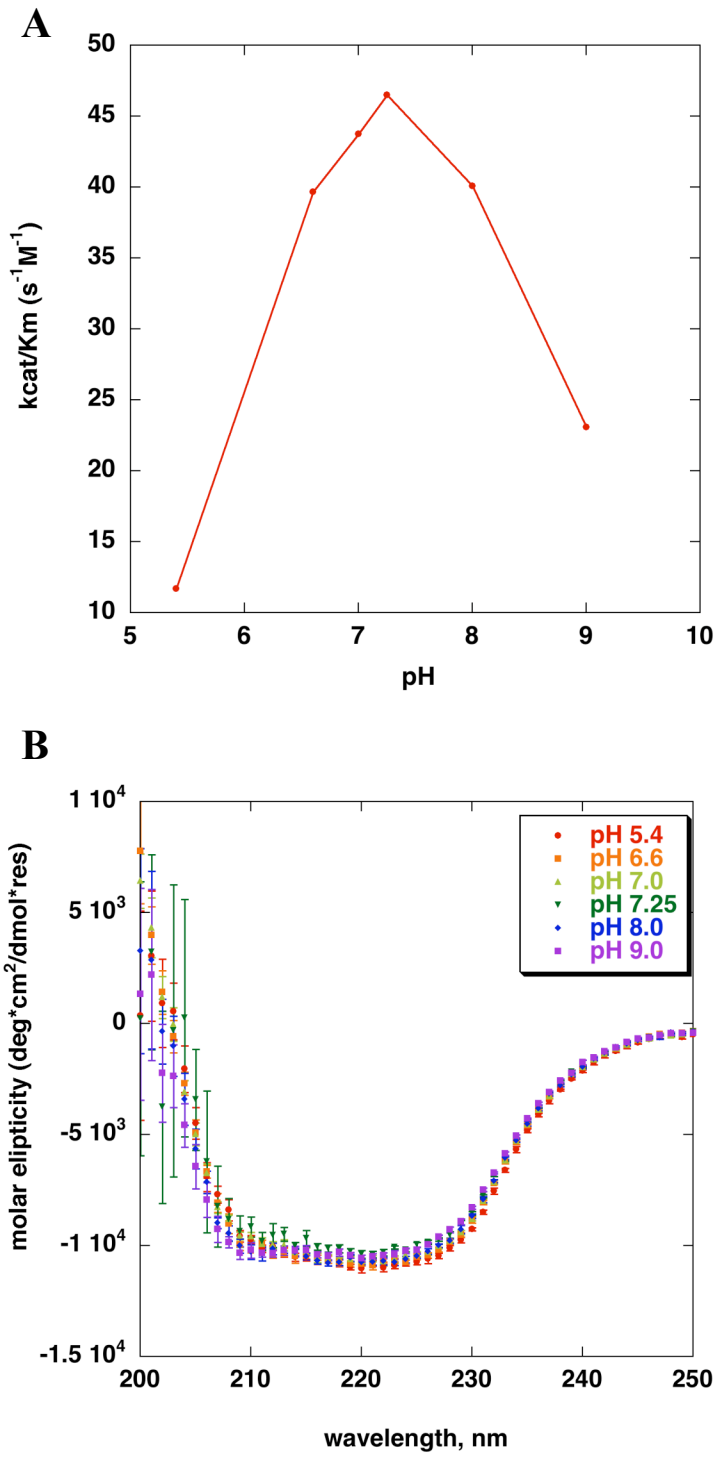


Figure 4-8. Effect of pH on the activity and structure of HG-2. (A) pH-rate profile of HG-2. **(B)** Far-UV CD wavelength scan of HG-2 at various pHs.

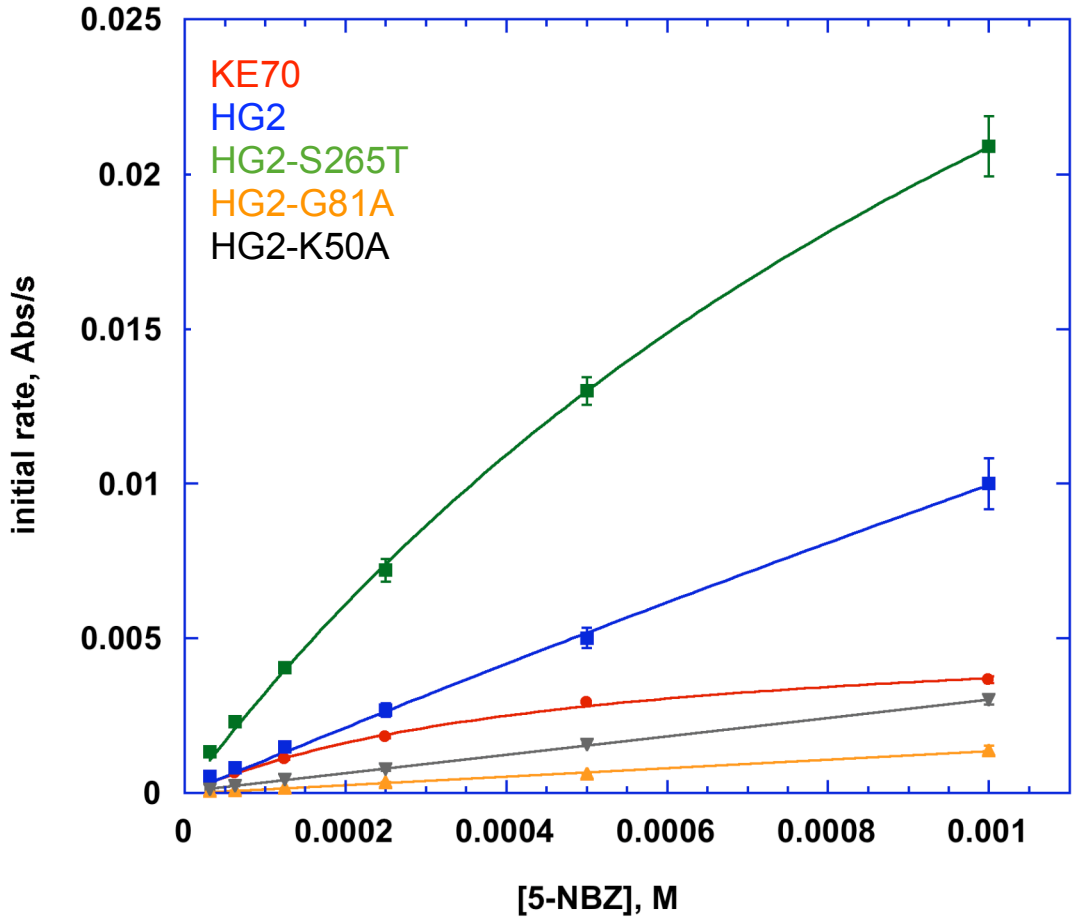


Figure 4-9. Kinetic characterization of third-generation enzymes. Michaelis-Menten plots of KE70 (red), design HG-2 (blue), 220-S265T (green), G81A (yellow), and 220-K50A (grey). 5 μM protein was used and the reactions were carried out in 25 mM HEPES pH 7.25, 100 mM NaCl, 27°C.

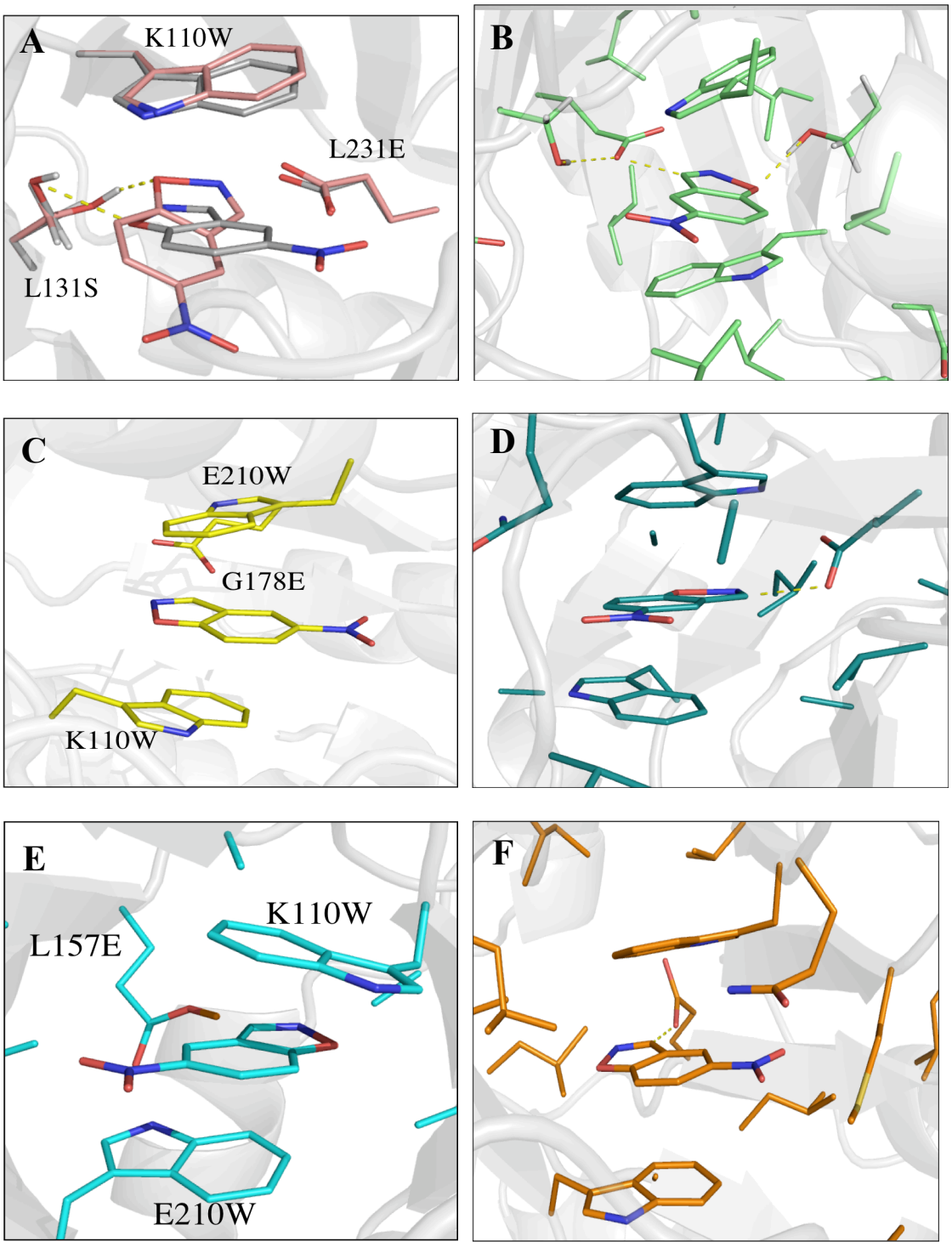


Figure 4-10. Active sites of designs in scaffold 1A53. **(A)** Catalytic residues of 1A53-1 (pink) overlaid with those from KE59 (grey). **(B)** Repacked active site of 1A53-1. **(C)** Catalytic residues of 1A53-2. **(D)** Repacked active site of 1A53-2. **(E)** Catalytic residues of 1A53-3. **(F)** Repacked active site of 1A53-3.

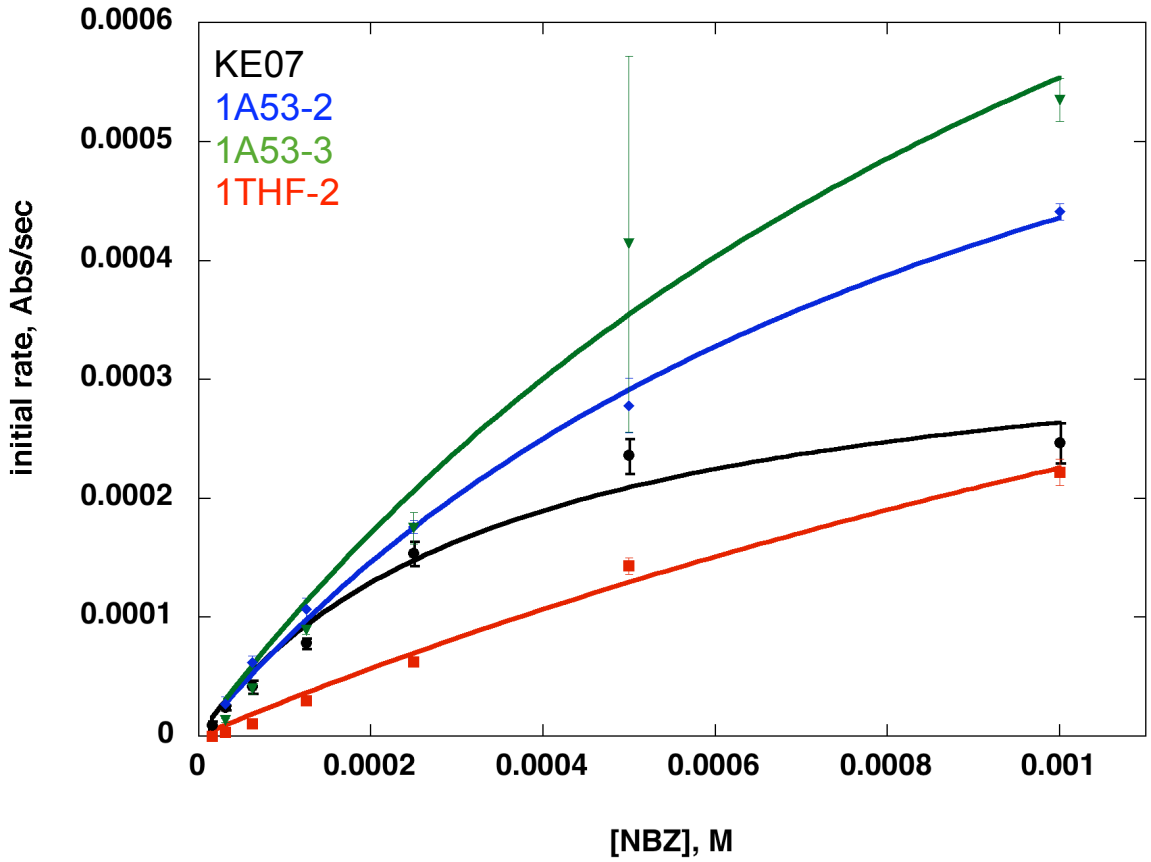


Figure 4-11. Kinetic characterization of designs in scaffolds 1A53 and 1THF. Standard errors are calculated from three measurements.

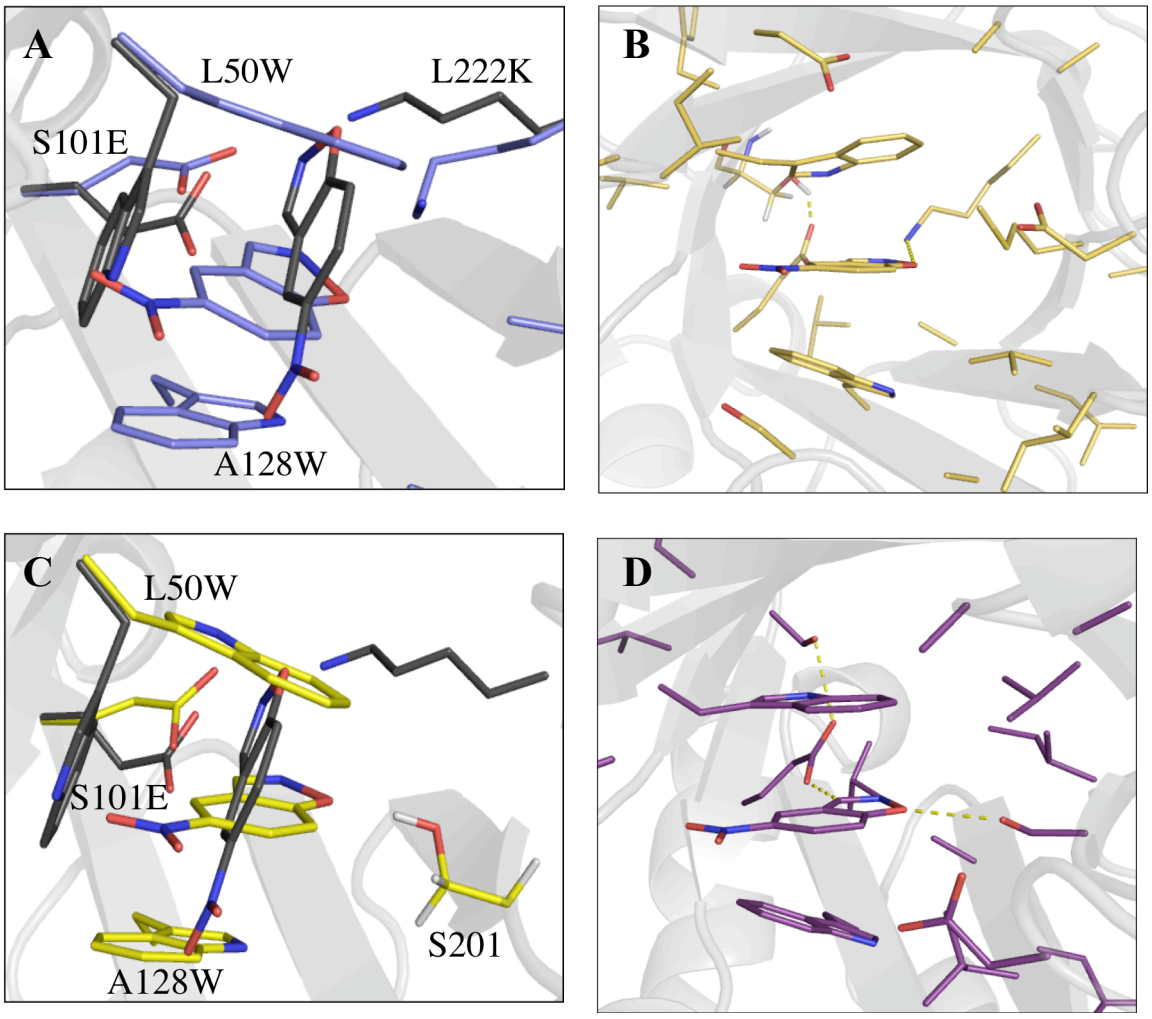


Figure 4-12. Active sites of designs in scaffold 1THF. (A) Catalytic residues of 1THF-1 (blue) overlaid with those from KE07 (grey). **(B)** Repacked active site 1THF-1. **(C)** Catalytic residues of 1THF-2 (yellow) overlaid with those from KE07 (grey). **(D)** Repacked active site of 1THF-2.

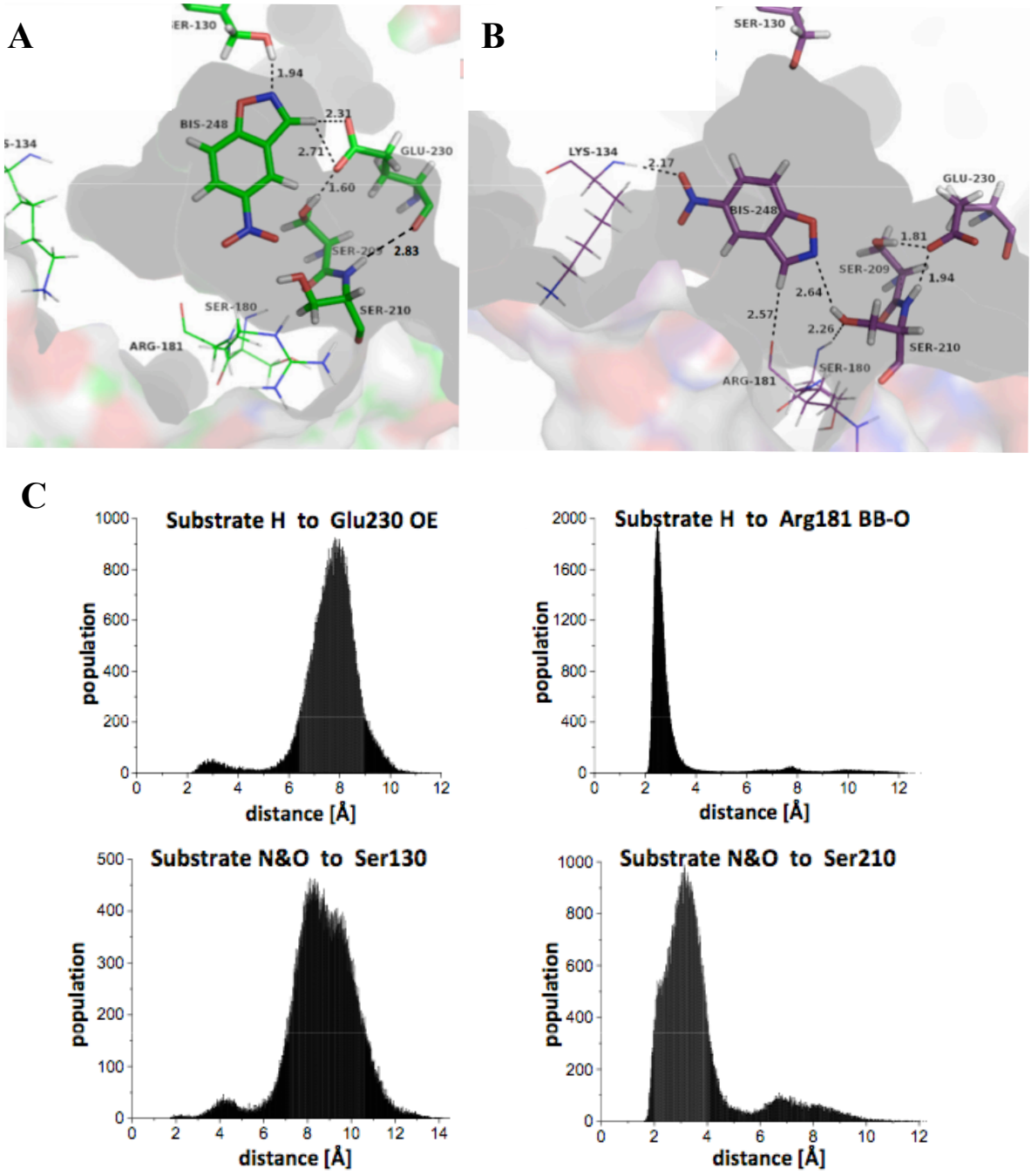


Figure 4-13. MD analysis of 1A53-1. **(A)** Initial active site configuration of 1A53-1. **(B)** Representative configuration after MD simulation. **(C)** Distance distributions of contacts between the substrate and catalytic residues.

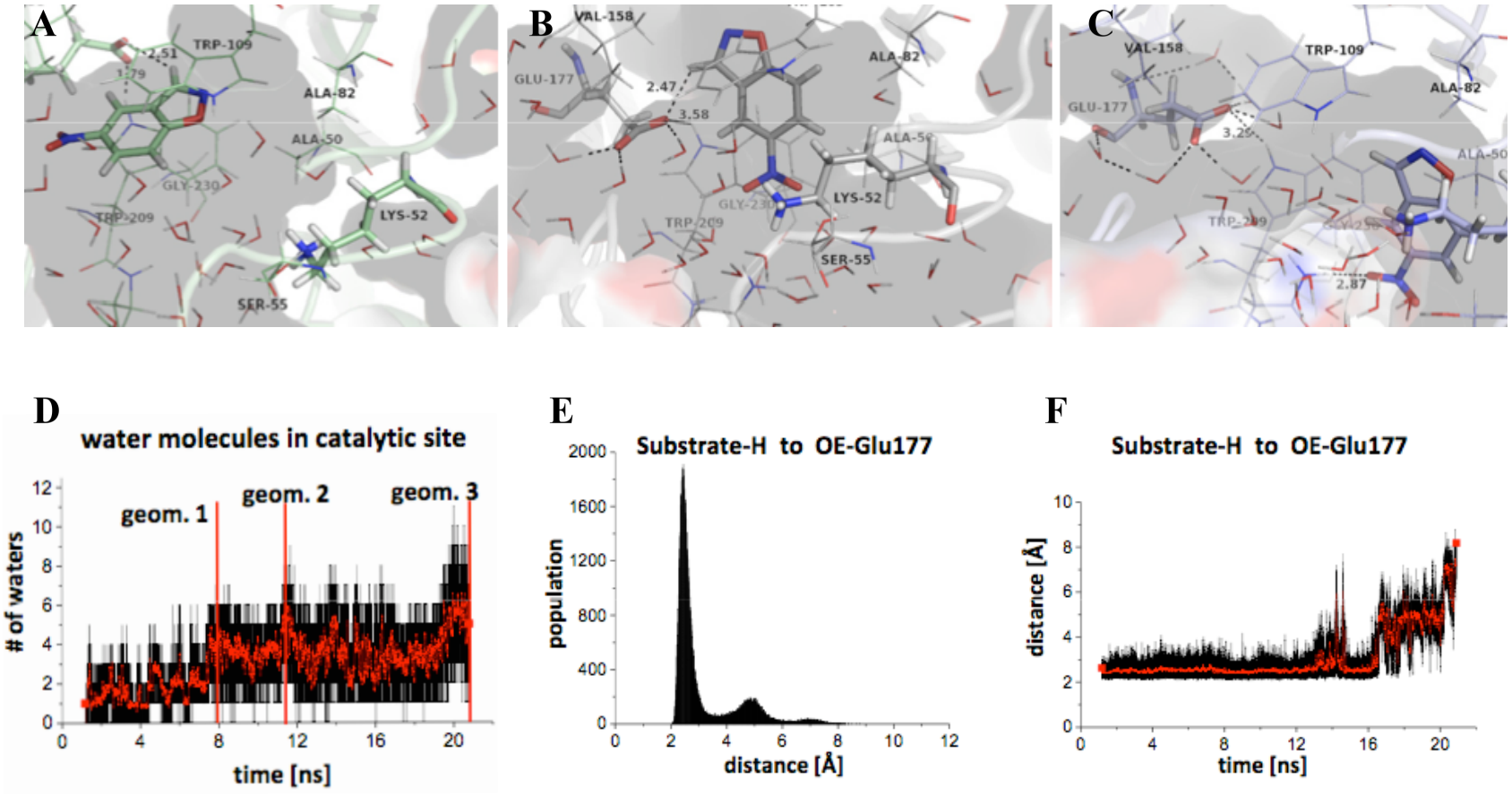


Figure 4-14. MD analysis of 1A53-2. **(A)** Initial active site configuration of 1THF-1. **(B)** Representative configuration midway through MD simulation (12 ns). **(C)** Representative geometry after MD simulation. **(D)** Number of water molecules that enter the active site over the course of the simulation. **(E)** Distance distribution of substrate-base contact. **(F)** Distance of the substrate-base contact over the course of the simulation.

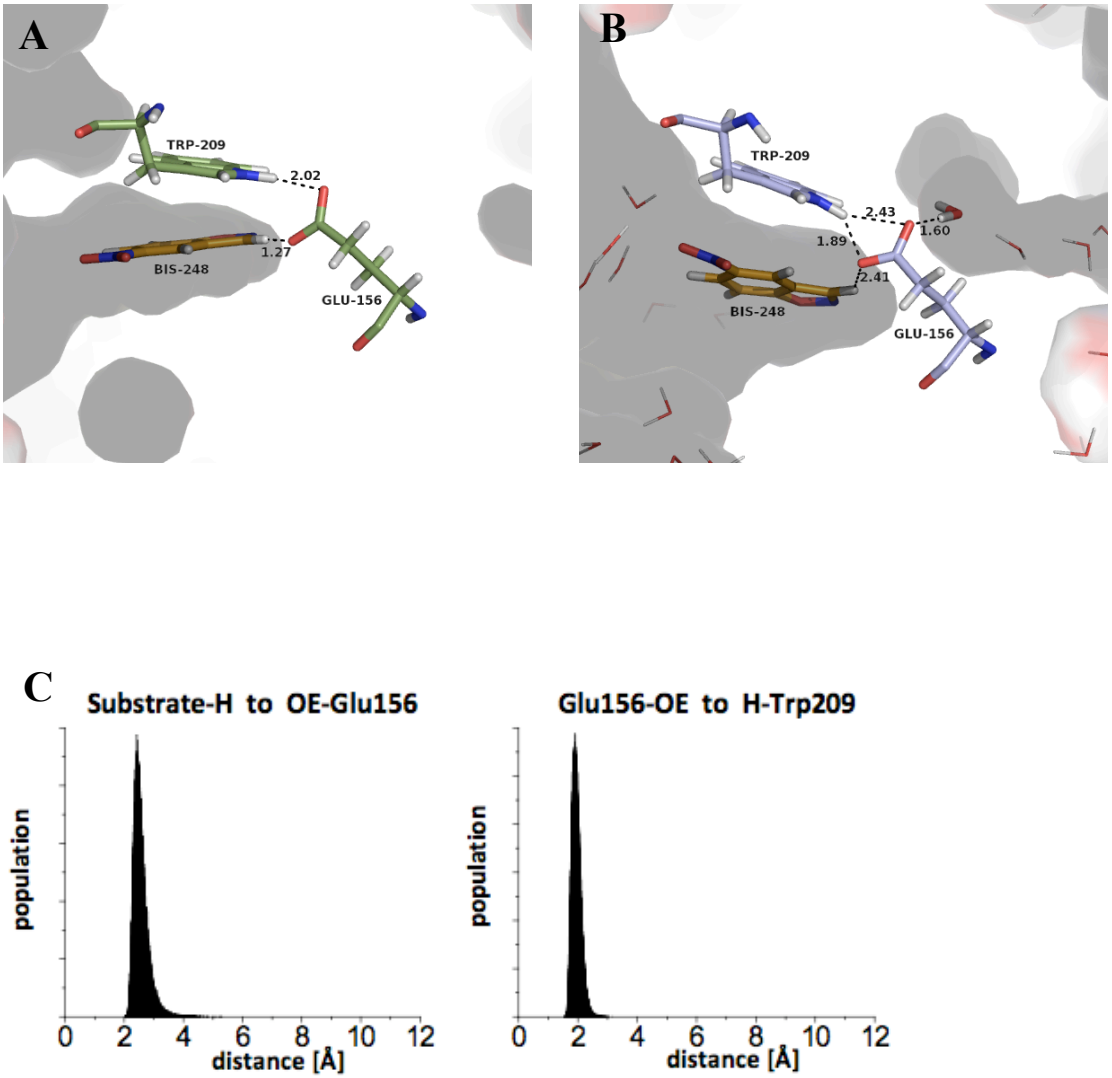


Figure 4-15. MD analysis of 1A53-3 **(A)** Initial active site configuration of 1A53-3. **(B)** Representative configuration after MD simulation. **(C)** Distance distributions of contacts between the substrate and catalytic residues.

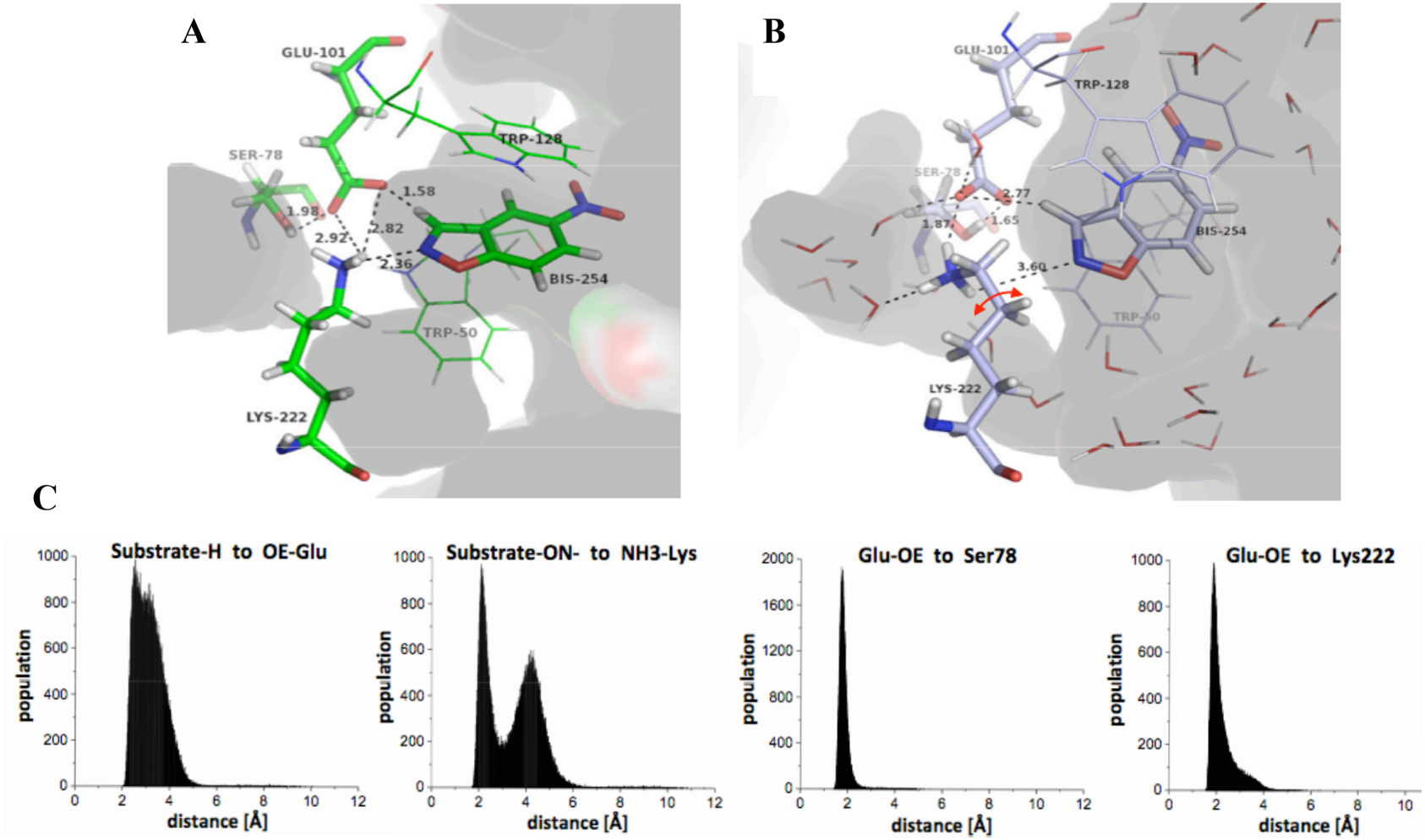


Figure 4-16. MD analysis of 1THF-1. **(A)** Initial active site configuration of 1THF-1. **(B)** Representative configuration after MD simulation. **(C)** Distance distributions of contacts between the substrate and catalytic residues.

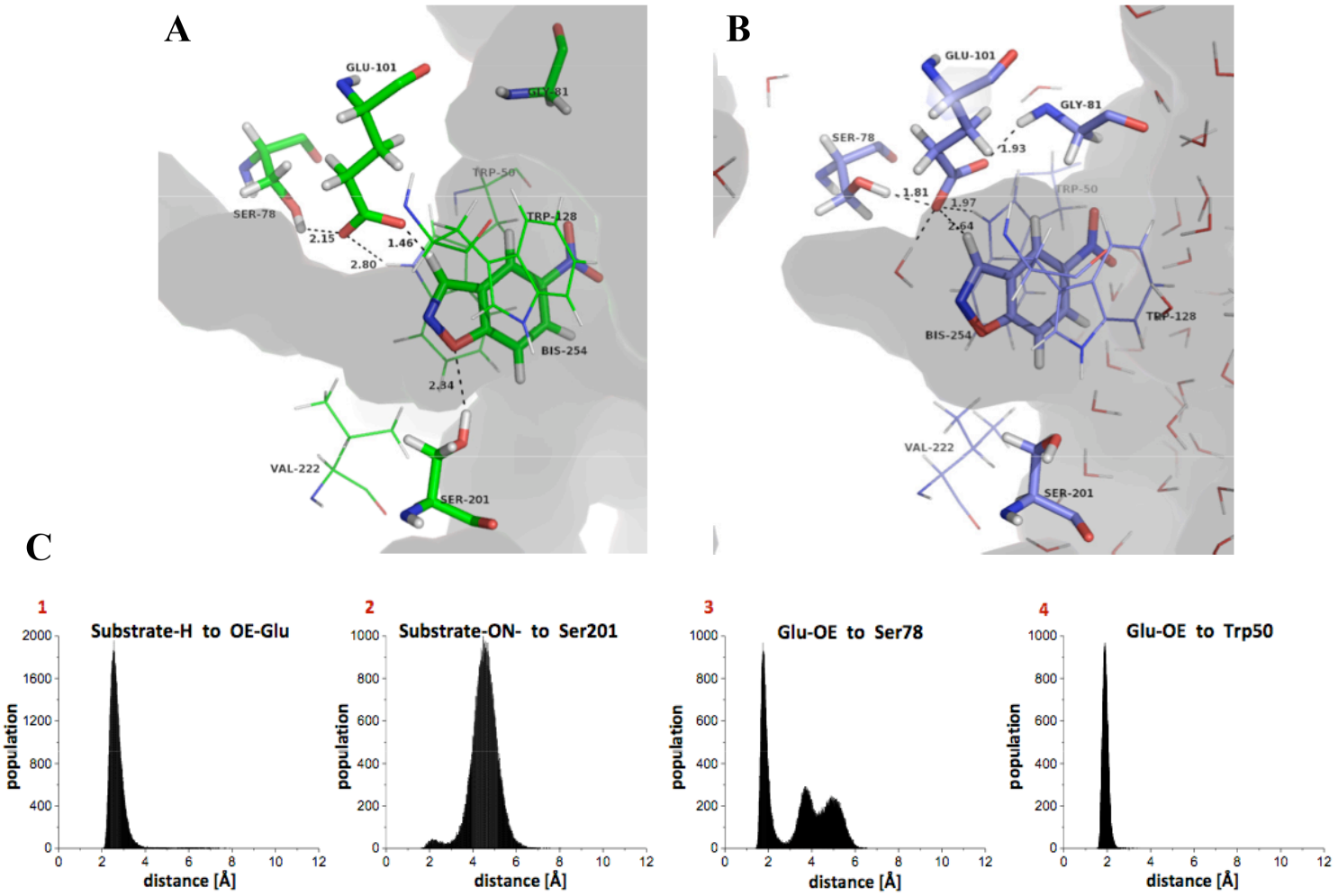


Figure 4-17. MD analysis of 1THF-2. **(A)** Initial active site configuration of 1THF-2. **(B)** Representative configuration after MD simulation**. (C)** Distance distributions of contacts between the substrate and catalytic residues.

Appendix A

Toward the computational design of a novel enantioselective hydrolase

Abstract

Enzymes are ideal catalysts of organic reactions because of their high efficiency and ability to perform a wide variety of chemical transformations with extreme chemo- and enantioselectivity under mild reaction conditions. The chemical repertoire of natural enzymes is not exhaustive, however, and certain reactions of industrial interest are not found in nature. A general method to design and optimize an enzyme for the catalysis of a specific reaction, especially difficult enantioselective reactions, would be a powerful tool in organic synthesis and would open up a range of reactions not currently accessible due to the lack of an appropriate catalyst. Towards this end, we used the ORBIT computational protein design software in an attempt to design an enzymatic catalyst for the kinetic resolution of N-benzoyl-L-phenylalanine through the selective hydrolysis of (S)-2-benzyl-4-phenyl-oxazolone-5-one (F-FOX) as a model system. This project did not result in an active enzyme. However, the methods developed here laid the groundwork for future active designs using other systems.

Introduction

Enantioselective enzymes

Enantiomerically pure drugs have become the new standard in the pharmaceutical industry. Chemists have long recognized the influence chiral molecules can have over a biological system. Indeed, many of the common drugs isolated from natural sources, such as morphine and quinine, are single enantiomers.[1] However, synthetically producing the single active enantiomer of a compound has always been challenging, if not impossible, and as a result, chiral drugs were historically produced as racemates.[2,3] Since the tragic effects of thalidomide were discovered in 1961, we have come to better understand the differing biological activity and pharmacokinetics of the individual enantiomers of a drug. These differences make enantiopure drugs preferable and even necessary in many cases. Fortunately, technology has advanced to make the synthesis of many single enantiomer drugs possible.[3]

Chemists use four basic methods to enantiomerically enrich a target compound: (1) isolation of chiral compounds from nature, (2) chemical asymmetric techniques, (3) chiral resolutions, and (4) enzymatic techniques. In spite of major advances in asymmetric synthetic techniques, nature is still arguably the richest source of chiral compounds and catalysts. However, even with this "chiral pool" of compounds, the selection, enantiopurity, and extractable quantities of these chiral compounds are limited.[4] Asymmetric strategies to access chiral compounds are far from being fully developed. Especially challenging reactions involve the formation of carbon-carbon bonds to a stereogenic center. Chiral resolutions of racemates into individual enantiomers are one of the oldest and most industrially important methods for enantio-enrichment. However,

this approach has the inherent drawback of a maximum theoretical yield of 50%, because by definition a racemate only has 50% of the desired enantiomer, except in special cases described below.[5,6]

With an ever-increasing demand for enantiomerically pure, complex, biologically active compounds (i.e., drugs), synthetic chemists are necessarily beginning to look beyond standard chemical strategies in favor of the high yields, efficiency, and strict selectivity of enzymatic reactions.[2] Although enzymes are widely used in industry, their use is often limited to reactions that are parts of natural metabolic processes. Exceptions include enzymes with natural limited selectivity such as some lipases.[7]

As previously mentioned in Chapter I, both directed evolution and catalytic antibodies have drawbacks that prevent their use as general tools for the introduction of new chemistries into enzymes. Here, I describe our initial strategy to introduce enantioselective hydrolysis activity into inert scaffolds using our computational protein design software, ORBIT.

Selected system: the dynamic kinetic resolution of F-FOX hydrolysis

Dynamic kinetic resolutions are a powerful method for transforming a racemic starting material into an enantiomerically pure product without the theoretical yield limitations imposed by standard chiral resolutions. Under the conditions of a dynamic kinetic resolution, the starting material racemizes and an equilibrium is established between the stereoisomers.[6] Consequently, by continuously removing one enantiomer from the mixture, 100% of the starting material can be converted into the desired enantiomer of the product compound. Our goal here was to design a stereoselective

enzyme to carry out a dynamic kinetic resolution. The enantioselective hydrolysis of (S)-2-benzyl-4-phenyl-oxazolin-5-one (F-FOX) to produce *N*-benzoyl-L-phenylalanine[8] was chosen as our model system (Figure A-1).

Initially, we found F-FOX hydrolysis to be an attractive system for this proof of principle because (1) F-FOX has reasonable solubility and stability in water;[9] (2) oxazolones have a characteristic absorption at around 245 nm that is reduced significantly upon hydrolysis, allowing the reaction to be easily monitored (see Materials and Methods); (3) F-FOX can be easily synthesized through a single-step reaction from commercially available materials;[10] (4) F-FOX racemizes quickly under basic conditions in aqueous solution (Figure A-3C);[11] (5) the benzyl sidechain of F-FOX gives a large hydrophobic handle with which to distinguish two enantiomers; (6) F-FOX is fairly rigid for its size, having only three degrees of freedom to model in the transition state; (7) F-FOX undergoes hydrolysis by a simple base-catalyzed mechanism;[9,11] (8) F-FOX is a strong fluorophore in aqueous solution providing a possible method for binding analysis fluorimetry (see Results); (9) F-FOX hydrolysis is not performed selectively by a natural enzyme. Some lipases have been shown to hydrolyze oxazolones, but these enzymes are nonselective hydrolases, the hydrolysis rates are modest, and the mechanism of oxazolone hydrolysis has not been established.[12,13]

In a method analogous to the generation of catalytic antibodies, our strategy was to first calculate a structure for the F-FOX transition state and then optimize the binding pocket of a protein around the structure using ORBIT to maximize the protein's affinity for the transition state.

Scaffold selection

In contrast to directed evolution, the protein design process requires a scaffold that is inert with respect to the reaction of interest. Several proteins were tested computationally as possible scaffolds to house the F-FOX hydrolase active site and transition state including *Homo sapien* retinol binding protein (PDB: 1BRP),[14] *E. coli* ribose binding protein (PDB: 2DRI),[15] *Thermus thermophilus* aspartate amino transferase (PDB: 1GCK),[16] and *E. coli* maltose binding protein (pdb: 1ANF).[17] As these proteins are all enzymes or ligand-binding proteins, they all have well-defined ligand-binding pockets, which is where the active site search was targeted. The periplasmic binding protein maltose binding protein (MBP) from *E. coli*, shown in Figure A-2, was chosen as the initial scaffold for the F-FOX hydrolase design because the size and shape of its binding pocket can accommodate the necessary catalytic residues and reaction transition state. Another attractive feature of this protein is its large size (41 kDa). In general, large proteins are convenient for enzyme design because the overall stability of the protein should not be significantly affected by mutations introduced to the binding pocket. Also, MBP not only binds to maltose with high affinity, but also binds to maltotriose and maltotetraose. The F-FOX transition state is of similar size and shape to maltose and maltotriose, suggesting that with some modifications to the residues within the binding pocket, MBP should be capable of binding to the F-FOX transition state as well. Our design calculations were based on the MBP crystal structure in complex with maltose at 1.67 Å resolution (PDB: 1ANF).[17]

Materials and Methods

Calculation of F-FOX transition state

Ab initio calculations of the first transition state of (S)-F-FOX hydrolysis (Figure A-1) were performed using the hybrid density functional B3LYP as implemented by the Jaguar 5.5 program package[18] and a 6-31G** basis set. The implicit solvation effects of methanol were calculated using a dielectric constant of 33.62 and a probe radius of 2.00 Å. Partial atomic charges were calculated with an electrostatic potential fit.

The system under investigation included the (S)-F-FOX structure with the phenyls removed to facilitate a faster calculation, a water molecule, and NH_3 to deprotonate the attacking water (Figure A-3A). A transition state structure for the hydrolysis of (S)-F-FOX was calculated using the following steps. First, a bond length scan was performed by varying the distance between the attacking water molecule and the carbonyl carbon of F-FOX until a maximum energy was found. The distance that produced the highest energy structure was used as the starting point for the transition state calculation. The final transition state structure has a water-oxygen to carbonyl-carbon distance of 1.79Å (Figure A-3B). We made 18 conformations of the transition state structure by varying the χ_1 and χ_2 angles of the benzyl group to correspond to the six phenylalanine rotamers in the Karplus and Dunbrack library[19] and by allowing the phenyl ring attached directly to the oxazolone to adopt positions that are ±15° out of the oxazolone plane as shown in Figure A-3C. The partial atomic charges for F-FOX were also calculated with Jaguar using an electrostatic potential fit.

Defining the design geometric constraints

To obtain a protein that exhibits high affinity and specificity towards the transition state structure, optimal contacts and ranges of contacts were defined between the transition state structure and three amino acid side chains based on ideal hydrogen bond configurations (Figure A-4). The geometric constraints shown in Figures 3-5 and 3-6 describe the activation of water by both the Nδ and Nε atoms of neutral histidine (in the Hie form of histidine, Nε is protonated; in Hid, Nδ is protonated) and the protonation of the oxazolone nitrogen by either nitrogen of protonated histidine (Hsp). Each of the constraints of these geometry definitions defines a range of allowable distances, angles or torsions among the atoms of the transition state and the side chain.

If the molecules adopt any orientation whose geometry falls within these ranges, they will be considered to be making a hydrogen bonding contact. There are four different ways the arginine guanidinium group can make a double hydrogen bonding contact to the two oxygens of the transition state (Figure A-6A) and the geometric constraints for arginine-transition state contacts are shown in Figure A-6B. To help define the geometries, pseudo-atoms with no volume, mass, or charge were created at the midpoint between each pair of nitrogens and between the two oxygens of the transition state (shown as pink stars in the figures). The distance between the pseudo-atom at the optimal orientation for the formation of two hydrogen bonds is 2.9 Å.

Design strategy and parameters

The overall strategy for the enzyme design was: (1) strip the side chains from the binding pocket, (2) find locations within the binding pocket that will allow all of the

defined contacts between the catalytic side chains and the transition state, (3) choose the optimal location for the active site, and (4) repack the rest of the binding pocket side chains around the catalytic residues and transition state. Steps 1-3 are referred to as the active site search and step 4 is called active site repacking.

As described previously, the scoring function is the sum of the van der Waals energy, hydrogen bond energy, electrostatics energy, and atomic solvation energy for the system. The atomic radii used to calculate the van der Waal's energies were scaled by 0.95.[20] Making the atoms appear smaller than their known van der Waal's radius softens all of the interactions and serves to compensate for the rigidity of the fixed backbone and discrete rotamers. We used a hydrogen bond potential function with a well depth of 8 kcal/mol.[21] Hydrogen bonds between two side chains and between a side chain and a remote backbone atom were treated equally, but hydrogen bonds between a side chain and its own backbone are not considered stabilizing and were scaled by 0.00. The surface area-based solvation terms were all weighted individually:[22] the benefit for non-polar surface area burial was 0.026 kcal/mol/Å^2, the penalty for nonpolar surface area exposure was scaled by a factor of 1.60, and the penalty for burial of polar atoms was 0.10 kcal/mol/Å^2.

The side chains were removed from all of the residues of MBP within 4 Å of maltose in the wild-type structure, creating a poly-glycine hole in the binding pocket. The residues within 8 Å of maltose in the wild-type structure were allowed to change conformation but not identity during the active site search. All other side chains and the backbone atoms were kept in the positions determined by the crystal structure. The geometry definitions for the Hie/Hid to transition state contact are special in that they

were also used to define rotamers for the transition state with respect to each neutral histidine rotamer. As these "substrate-target" residues sample positions/rotamers during the transition state search, the transition state structure samples all of its rotameric forms as it translates and rotates with respect to the neutral histidine rotamer. These translations and rotations were subject to the geometric constraints shown in Figure 4. Distances were sampled in 0.5 Å steps, and angles and torsions were sampled in 15 or 20° steps. The two additional catalytic residues also sampled positions within the poly-glycine region.

Active site search

In the active site search, single residue and pair-wise energies were calculated without application of solvation potentials. Any configuration meeting all the geometric constraints discussed above received an energy benefit of 100 kcal/mol, biasing the resulting configurations towards those that have a desired geometry with respect to the transition state. A penalty of 1000 kcal/mol was applied to any arginine or histidine located within the binding pocket that does not make a specified contact to the transition state, preventing extra catalytic residues from being selected based on favorable van der Waals interactions with the backbone or other residues. Optimization was performed using a modification of the FASTER algorithm,[23] resulting in a solution that specifies a single possible location of the active site.

A Monte Carlo search[24,25] was then performed to explore the "active site space" around the FASTER solution. To surmount the unrealistically deep local minima represented by a three-residue active site/transition state configuration within an empty

pocket, an increased high annealing temperature of 500,000 K was used. Even with this elevated temperature, only about 50% of the moves were successful. The Monte Carlo search provided a list of sequences that met all of the specified geometric and sequence requirements, ranked by energy; these structures were then evaluated individually. Evaluation criteria included transition state/catalytic residue geometry, favorable interactions of the catalytic residues with the backbone, wild-type residue identity at a catalytic residue position, positioning of the catalytic residues within the binding pocket and relative to one another, and overlap of the transition state structure with the position of the natural ligand within the binding pocket.

After the optimal active site location was chosen, the catalytic residue positons were fixed and FASTER was used to repack the rest of the binding pocket side chains around the catalytic residues and the transition state structure. Binding pocket residues capable of contacting the transition state or the catalytic residues could sample rotamers for all the amino acids except proline, methionine, and cysteine. The transition state was allowed to move in small steps according to its allowed geometries with respect to the neutral histidine. All other binding pocket residues and the catalytic residues were allowed to sample all conformations but were not allowed to change identity. Again, a Monte Carlo algorithm was used to sample sequence space around the solution provided by FASTER. A high annealing temperature of 4000 K was used. The top 20 Monte Carlo sequences were then subjected to a DEE-based algorithm that finds the lowest energy rotamer for each designed residue in the sequence.

F-FOX synthesis and characterization

F-FOX was synthesized as previously described[10] and purified via flash chromatography. (230-400 mesh silica gel, 60 Å purchased from Aldrich, 4:7 ethyl acetate:hexanes). *N*-benzoyl-L-phenylalanine and *N*-cyclohezyl-*N*'-2-(*N*-methyl morpholino)-ethylcarbodiimide p-toluenesulfonate were purchased from Aldrich and used without further purification. Evaporation under reduced pressure gave the oxazolone as a white solid; yield: 80.0 mg (50%) m.p. 70-72°C (Lit.[10] m.p. 70-71°C).

I.R. (CH_2Cl_2): ν = 1822; 1655; 1496; 1425; 1323; 1082; 1046; 965; 885 cm^{-1}.

1H-N.M.R. ($CDCl_3$): δ = 3.3 (q, 2H); 4.7 (t, 1H); 7.4 (t, 2H); 7.6 (d, 1H); 7.9 ppm (d, 2H).

NMR was performed on a Varian Mercury 300 MHz machine, and IR spectra were recorded using a Perkin-Elmer Spectrum BX spectrometer. UV-vis spectroscopy was performed using a Shimadzu UV-1601 spectrophotometer equipped with a temperature controlled cell holder. Fluorescence measurements were performed on a Photon Technology International fluorimeter equipped with a Model 180/814 photomultiplier detection system and a Ushio xenon short arc lamp. Fluorescence polarization experiments were carried out with 50 μM to 0.5 nM protein and 500 nM F-FOX. 500 nM BSA was used as a control.

Protein expression and purification

The gene for the F-FOX hydrolase (1ANF-FFH) was optimized for expression, synthesized with an N-terminal His_6 tag and Gly-Gly-Ser linker (Blue Heron, Bothell, WA), and cloned into a pETlla vector (Invitrogen). The plasmid was transformed into

BL-21(DE3) Gold cells (Strategene) by heat shock and protein expression was induced for 4 hours at 37°C with 1.0 mM IPTG. Cells were harvested by centrifugation at 5000 × g for 30 min and pellets were resuspended in 10 mM imidazole, 300 mM NaCl, 20 mM Tris pH 7.4. The cells were lysed mechanically and pelleted at 10,000 × g for 45 min. The soluble fraction was applied to a Ni-NTA Agarose column (QIAGEN) and washed with 10 column volumes of 10 mM imidazole, 300 mM NaCl, 20 mM Tris pH 7.4 and 10 column volumes of 20 mM imidazole, 300 mM NaCl, 20 mM Tris pH 7.4. The protein was eluted with 250 mM imidazole, 300 mM NaCl, 20 mM Tris pH 7.4 and the eluate was dialyzed exhaustively against 20 mM Tris pH 7.4, 50 mM NaCl and stored at 4°C until use. The overall purity of 1ANF-FFH was > 90% as determined by SDS-PAGE. The expected molecular weight of 1ANF-FFH was within 9 units of the weight determined by electrospray mass spectrometry. Expression yields for 1ANF-FFH were between 20 and 40 mg per liter of culture.

Circular dichroism

Circular dichroism (CD) data were recorded with an Aviv DS spectropolarimeter. The temperature was controlled with a thermoelectric unit. Experiments were performed on samples containing 16.6 μM protein in buffer containing 20 mM Tris pH 7.4 and 50 mM NaCl. Wavelength scans were performed in triplicate in 1nm steps from 250 nm to 190 nm with an averaging time of 1 sec. Thermal denaturation was performed in 1°C steps from 1 to 99°C with 2 min of equilibration at each new temperature and averaging over 30 sec. Thermal unfolding was monitored at 222 nm. Apparent melting temperatures were determined using the relation of Minor and Kim.[26]

Enzyme activity assays

Enzymatic activity was assayed using 3-5 μM 1ANF-FFH in 1 mL total volume and buffers of varying composition, salt concentration, pH and temperature. The assay was initiated by the addition of F-FOX in acetonitrile to a final concentration of 20–70 μM. If the assay temperature was above 25°C, the protein and the buffer (or the buffer and an equivalent amount of the protein storage buffer for controls) were equilibrated at that temperature for 10 min before the addition of substrate. The disappearance of the substrate with respect to time was observed by monitoring the absorbance of the sample at 245 nm for 10 min at each initial substrate concentration. Initial substrate concentrations were calculated using an extinction coefficient of 13000 $cm^{-1}M^{-1}$ (see Results section). The initial rates were calculated from the linear portion of the substrate concentration versus time curves. For preliminary enzyme activity assays, all rate constants were determined at three different initial substrate concentrations.

Fluorescence polarization

Fluorescence polarization measurements were performed on a Photon Technology International fluorimeter equipped with a Model 180/814 photomultiplier detection system, a Ushio xenon short arc lamp, and three polarizers from Photon Technology International. An excitation wavelength of 356 nm and an emission wavelength of 440 nm were used. Data were collected for 60 sec with a slit width of 1.25 mm. Experiments were carried out with 50 μM to 0.5 nM protein and 500 nM F-FOX. BSA was used as a positive control.

Chiral HPLC

A method for the separation of D-benzoyl-phenylalanine and L-benzoyl-phenylalanine (the products of F-FOX hydrolysis) was developed using chiral high-pressure liquid chromatography (HPLC). Good separation was found in 12% ethanol in hexanes with 0.1% trifluoroacetic acid on a 4.6 mm × 250 mm Chiracel AD column at a flow rate of 1 mL/min.

Site-directed mutagenesis

All mutagenesis reactions were carried out using site-directed mutagenesis as described in Chapter III, Materials and Methods.

Results and Discussion

F-FOX characterization

The UV-vis spectra of F-FOX and N-benzoyl-phenylalanine are shown in Figure A-7. The spectra are substantially different, but both the F-FOX reactant and benzoyl-phenylalanine product absorb at the same wavelengths. Because a large difference in the two spectra occurs at 245 nm, this wavelength was chosen to follow the reaction kinetics. F-FOX and N-benzoyl-phenylalanine were determined experimentally to have extinction coefficients of 13000 ± 311 $cm^{-1}M^{-1}$ and 6600 ± 141 $cm^{-1}M^{-1}$, Respectively, at 245 nm (data not shown).

The hydrolysis of F-FOX follows pseudo-first-order kinetics with a large dependence on pH. The background reaction occurs with a first-order rate constant of

0.00836 ± 0.000495 min^{-1} in 0.01 M KPi pH 6.6. As expected, the rate constant is positively correlated with pH (data not shown).

An excitation wavelength scan shows that F-FOX fluoresces upon excitation with 356 nm light with emission occurring at 440 nm. This fluorescence is strongly pH dependent (Figure A-8), suggesting that F-FOX fluoresces most strongly when it is in its deprotonated form. The loss of the acidic α-proton causes aromaticity to extend between the oxazolone and the phenyl side chain, facilitating fluorescence.

Due to the very different character of the two forms of F-FOX, its fluorescence is also strongly dependent on the nature of its environment. The fluorescence of F-FOX is completely quenched upon transfer into dichloromethane or acetonitrile, but it regains fluorescence with increasing water concentration in a mixture with acetonitrile. In contrast, most fluorescent molecules experience an increase in fluorescence in hydrophobic environments because these types of environments tend to shield the molecule from quenchers that are present in aqueous solutions.[27] Because the hydrophobic binding pockets of proteins can mimic the effects seen in nonpolar solvents, we attempted to take advantage of this unusual fluorescence property of F-FOX to monitor F-FOX-protein binding through a decrease in fluorescence. However, an assay of this type is challenging because many factors other than a change in the environment of a fluorophore can contribute to a reduction in fluorescence. In addition, measurements of a reduction in fluorescence are inherently less sensitive than measurements of an increase in fluorescence.

1*ANF-FFH active site*

Of the many possible active site locations, one was ultimately chosen by visual inspection. This site is shown in Figure A-9A. In this active site, all three catalytic residues make good hydrogen bonds to the transition state. Nε of Arg 66 forms an electrostatically favorable contact with the carbonyl oxygen of the transition state, which at 4.1 Å, is too long to call a true hydrogen bond. This particular contact may be important in stabilizing the oxyanion intermediate of the hydrolysis, and the long contact may give the transition state enough room to form the intermediate. In addition to the favorable geometry between the catalytic residues and the transition state, this particular active site was chosen because of the similarity of the rotameric forms of the designed residues and the wild-type residue (Figure A-9C). Hid 62 overlays perfectly with the indole of the wild-type Trp, and Arg 66 is not only a wild-type residue, but adopts a very similar rotamer to that in the crystal structure.

The similarity of the catalytic residues and wild-type residues is promising, suggesting that rotameric strain will not be a large destabilizing factor in the formation of this active site. Some of the catalytic residues of this active site also have additional stabilizing interactions. Hid 64 makes an additional hydrogen bond contact to a main chain carbonyl that was not required by the geometry definitions (Figure A-9B). This extra contact may stabilize the catalytic histidine in its appropriate orientation. The final design for the F-FOX hydrolase based on scaffold 1ANF (1ANF-FFH) is a seven-fold mutant including the two catalytic histidines. The catalytic arginine at position 66 is wild type. Figure A-10A shows the designed active site overlaid with the wild-type residues. Again, all of the designed residues adopt rotamers similar to those in the wild-type

protein. A space-filling representation of the active site and the binding pocket residues (Figure A-10B) shows that the benzyl side chain of the transition state is well packed. Favorable van der Waals contacts are made with designed residues such as Hid 64 and Phe 63 as well as with wild-type residues such as Tyr 155, Trp 230, and Trp 340.

CD analysis of 1ANF-FFH

CD analysis of 1ANF-FFH shows that the protein is well folded and has an overall secondary structure composition comparable to that of wild-type MBP (Figure A-11A). Difference in the curve magnitudes may be due to uncertainties in protein concentrations. The apparent melting temperature of 1ANF-FFH is very close to the melting temperature determined for MBP under identical conditions (58.3 ± 0.03°C for 1ANF-FFH versus 60.2 ± 0.05°C for MBP as determined by thermal denaturation) (Figure A-11B). The similarity in melting temperatures indicates that the seven mutations introduced into the binding pocket of MBP have little effect on the overall stability of the protein.

Enzyme assays

The apparent pseudo-first-order rate constants determined for F-FOX hydrolysis under various conditions are shown in Table A-1 and Figure A-12. The addition of micromolar quantities of 1ANF-FFH was found to increase the rate of F-FOX hydrolysis up to 10% above background at 25°C, which is similar to the rate enhancement seen upon addition of wild-type MBP. A rate increase of this size is not indicative of enzymatic activity and a similar effect upon addition of wild-type protein suggests that F-FOX

hydrolysis is being affected subtly by nonspecific interactions with the exterior of the protein. Pseudo-first-order rate constants were also established for the background hydrolysis of F-FOX in buffer and in buffer with 50 μM imidazole. The addition of imidazole did not enhance F-FOX hydrolysis significantly at any pH. An increase in buffer concentration from 10 mM to 50 mM results in an increase in F-FOX hydrolysis consistent with general base catalysis, but the addition of 1ANF-FFH and MBP do not increase the rate significantly in either case. Assays performed at 37°C show an increase in the rate of background hydrolysis compared to 25°C, but no significant rate enhancement was observed by the addition of 1ANF-FFH (data not shown).

FOX binding assay

1ANF-FFH did not show any FOX hydrolysis activity. However, FOX may have still been binding in the active site of 1ANF-FFH. Taking advantage of FOX fluorescence, fluorescence anisotropy was used to assay FOX binding. MBP has a large, irregular shape, so the effective size of FOX should increase dramatically upon being bound by the protein, resulting in an increase it its fluorescence anisotropy. As mentioned in the Methods, FOX fluorescence decreases in nonpolar solvents. The binding pocket of the protein is less polar than water, so FOX fluorescence is expected to decrease upon entering the binding pocket. The decrease in FOX fluorescence upon protein binding, along with the high background hydrolysis rate of FOX, make this measurement more difficult because both result in a decrease in signal. BSA was used as a control because it is known to have the ability to bind nonspecifically to many small molecules, especially hydrophobic small molecules.[28]

The titration of 1ANF-FFH at a constant FOX concentration is shown in Figure A-13. The titration curve shows that 1ANF-FFH binds to FOX much more weakly than BSA, which probably exhibits nonspecific binding. It is unclear if this weak 1ANF-FFH binding is specific or nonspecific.

1ANF-FFH modifications

In our analysis of the model of the inactive design, we identified four potentially beneficial mutations for 1ANF-FFH (Figure A-13). These include a double mutation of I239 and A96 to tryptophan at positions located in the hinge region distal to the ligand-binding site. Hellinga and coworkers reported that this double mutant induces a 60-fold increase in the affinity of wild-type MBP for maltose by shifting the equilibrium of MBP to its closed conformation.[29] We predicted that the N150D mutation would contribute to the stabilization of the protonated catalytic histidine in the desired configuration and protonation state. Two additional mutations, (Y155F) and another double mutant of polar residues to tryptophan (N12W and D14W) were not expected to directly influence the catalytic residues or transition state, but might contribute to the stabilization of the binding pocket itself (Figure A-13). These mutations were made using site-directed mutagenesis (see Materials and Methods) and the activity of the variants was assayed as described above. The N150D mutant and the A96W/I329W double mutant showed no significant F-FOX hydrolysis activity (Figure A-14). The N12W/D14W double mutant and the N150D/A96W/I239W triple mutant also showed no significant activity (Figure A-15).

Lessons learned from F-FOX hydrolase design

None of the F-FOX hydrolase designs have significantly catalyzed F-FOX hydrolysis under any of the conditions examined so far. In retrospect, some issues have been identified that will help inform our future selection of both the chemical system and the scaffold. First, the protonation state of the catalytic histidines (especially the base) is critical to the proposed hydrolysis mechanism. However, histidine has a pKa near neutral and can exist in three protonation states, making it difficult to control the histidine species present in the active site. Second, the F-FOX hydrolysis system has a background reaction rate that is relatively high. This is a benefit in the sense that the reaction is easy to carry out. However, any designed enzyme would have to perform better than this high background rate before activity could be detected. Finally, the flexibility of some of the scaffolds selected for these designs (especially MBP) may lead to active sites that are too exposed to solvent. As mentioned above, the catalytic histidines must be in their correct protonation states for the reaction to work and the presence of water in the active site may cause the protonation state of the histidines to shift unpredictably. In addition, the active site of the MBP designs is not fully formed until the hinge that connects the two lobes is closed. For future designs, it would be advantageous to choose scaffolds that always have a fully formed active site rather than one that spends at least part of its time in an open state where entropy and desolvation costs work against the formation of the active site.

Because we have no structural data for any of these designs, we can only speculate on reasons for their inactivity. However, the problems listed above combine to

form a reasonable argument for abandoning this system in favor of one that has fewer practical issues.

Acknowledgements

We thank Jonas Oxgaard from the Goddard group for help in determining the transition state structure for the F-FOX reaction and for general guidance in the use of the Jaguar software. We also thank Daniel Caspi and Doug Behenna from the Stoltz lab at Caltech for their help in developing the chiral HPLC assay and Ariele Hanek in the Dougherty lab for advice about F-FOX synthesis and characterization. Robert Dirks from the Pierce lab helped with the fluorescence polarization assays. This work was supported by the Howard Hughes Medical Institute and the Defense Advanced Research Project Agency.

References

1. Rouhi, A. M., Chiral business. *Chem. Eng. News* **2003**, *May*, 45-55.

2. Schoemaker, H. E.; Mink., D.; Wubbolts, M. G., Dispelling the myths- Biocatalysts in industrial synthesis. *Science* **2003**, *299*, 1694-1697.

3. Agranat, I.; Caner, H.; Caldwell, J., Putting chirality to work: The strategy of chiral switches. *Nat. Rev. Drug Discov.* **1992**, *1*, 753-768.

4. Blaser, H.-U., The chiral pool as a source of enantioselective catalysts and auxilaries. *Chem. Rev.* **1992**, *92*, 935-952.

5. Trost, B. M., Asymmetric catalysis: An enabling science. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 5348-5355.

6. El Gihani, M. T.; Williams, J. M. J., Dynamic kinetic resolutions. *Curr. Opin. Chem. Biol.* **1999**, *3*, 11-15.

7. Jaeger, K.-E.; Eggert, T., Lipases for biotechnology. *Curr. Opin. Biotechnol.* **2002**, *13*, 345-351.

8. Williams, A., The oxazolinone intermediate in the hdrolysis and aminolysis of N-benxoylglycine derivatives. *J. Chem. Soc., Perkin Trans. 1* **1975**, *2*, 947-953.

9. Goodman, M.; Levine, L., Peptide synthesis via active esters. IV. Racemization and ring opening reactions of optically active oxazolones. *J. Am. Chem. Soc.* **1964**, *86*, 2918-2922.

10. Hoyng, C.; McKena, M.; Walters, D., A convenient procedure for synthesis of derivatives of 2-oxazolin-5-one. *Synthesis-Stuttgart* **1982**, *3*, 191-193.

11. Goodman, M.; Stuben, K. C., Amino acid active esters III. Base-catalyzed racemization of peptide active esters. *J. Org. Chem.* **1962**, *27*, 3409-3416.

12. Brown, S. A.; Parker, M.-C.; Turner, N. J., Dynamic kinetic resolution: synthesis of optically active a-amino acid derivatives. *Tetrahedron: Asymmetry* **2000**, *11*, 1687-1690.

13. Crich, J. Z.; Brieva, R.; Marquart, P.; Gu, R.-L.; Flemming, S.; Sih, C. J., Enzymatic asymmetric synthesis of a-amino acids. Enantioselective cleavage of 4-subsitituted oxazolin-5-ones and thiazolin-5-ones. *J. Org. Chem.* **1993**, *58*, 3252-3258.

14. Zanotti, G.; Ottonello, S.; Berni, R.; Monaco, H. L., Crystal-structure of the trigonal form of human plasma retinol-binding protein at 2.5-Angstrom resolution. *Journal of Molecular Biology* **1993**, *230*, 613-624.

15. Bjorkman, A. J.; Binnie, R. A.; Zhang, H.; Cole, L. B.; Hermodoson, M. A.; Mowbray, S. L., Probing protein-protein interactions- The ribose-biniding protein in bacterial transport and chemotaxis. *Journal of Biological Chemistry* **1994**, *269*, 30206-30211.

16. Ura, H.; Nakai, T.; Kawaguchi, S.; Miyahara, I.; Hirotsu, K.; Kuramitsu, S., Substrate recognition mechanism of thermophilic dual-substrate enzyme. *J Biochem-Tokyo* **2001**, *130*, 89-98.

17. Quiocho, F. A.; Spurlino, J. C.; Rodseth, L. E., Extensive features of tight oligosaccharide binding revealed in high-resolution structures of the maltodextrin transport/chemosensory receptor. *Structure* **1997**, *5*, 997-1015.

18. Jaguar 5.5, Schrodinger, L.L.C., Portland, OR, **1991-2003**.

19. Dunbrak, R. L.; Karplus, M., Backbone-dependent rotamer library for proteins. *J. Mol. Biol.* **1993**, *230*, 543-574.

20. Dahiyat, B. I.; Mayo, S. L., Probing the role of packing specificity in protein design. *Proc. Natl. Acad. Sci. USA* **1997**, *94*, 10172-10177.

21. Dahiyat, B. I.; Gordon, B.; Mayo, S. L., Automated design of the surface positions of protein helices. *Protein Sci.* **1997**, *6*, 1333-1337.

22. Street, A. G.; Mayo, S. L., Pairwise calculation of protein solvent-accessible surface areas. *Folding & Design* **1998**, *3*, 253-258.

23. Desmet, J.; Spriet, J.; Lasters, I., Fast and accurate side-chain topology and energy refinement (FASTER) as a new method for protein structure optimization. *Proteins: Struct. Funct. Genet.* **2002**, *48*, 31-43.

24. Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H., Equation of state calculations by fast computing machines. *J. Chem. Phys.* **1953**, *21*, 1087-1092.

25. Kirkpatrick, S.; Gelatt, C. D.; Vecchi, M. P., Optimization by simulated annealing. *Science* **1983**, *220*, 671-675.

26. Minor, D. L., Jr.; Kim, P. S., Measurement of the beta-sheet-forming propensities of amino-acids. *Nature* **1994**, *367*, 660-663.

27. Lakowicz, J. R., *Principles of Fluroescence Spectroscopy.*, 2nd ed. Kluwer Academic/Plenum Publishers: New York, **1999**.

28. Kraghhansen, U., Molecular aspects of ligand binding to serum albumin. *Pharmacol. Rev.* **1981**, *33*, 17-53.

29. Marvin, J. S.; Hellinga, H. W., Manipulation of ligand binding affinity by exploitation of conformational coupling. *Nat. Struct. Biol.* **2001**, *8*, 795-798.

Table A-1. Apparent pseudo-first-order rate constants for F-FOX hydrolysis. Standard errors were calculated from assays at multiple initial substrate concentrations.

	k, min^{-1}			
pH	**50 mM KPi 50mM NaCl**	**+ 50 μM imidazole**	**+ 5 μM 1ANF-FFH**	**+ 5 μM MBP**
6.2	0.013 ± 0.0001	0.014 ± 0.001	0.014 ± 0.0006	0.017 ± 0.004
6.4	0.016 ± 0.0003	0.018 ± 0.001	0.018 ± 0.0003	0.021 ± 0.002
6.6	0.020 ± 0.0003	0.022 ± 0.0008	0.022 ± 0.0004	0.025 ± 0.003
6.8	0.024 ± 0.001	0.026 ± 0.002	0.029 ± 0.0006	0.030 ± 0.0003
7.0	0.031 ± 0.002	0.031 ± 0.003	0.034 ± 0.0011	0.035 ± 0.0005
7.2	0.035 ± 0.002	0.037 ± 0.001	0.040 ± 0.0004	0.042 ± 0.0018
7.4	0.041 ± 0.002	0.044 ± 0.001	0.047 ± 0.0018	0.050 ± 0.002
7.6	0.047 ± 0.002	0.05 ± 0.002	0.052 ± 0.0016	0.054 ± 0.002

	k, min^{-1}			
pH	**10 mM KPi 50mM NaCl**	**+ 50 μM imidazole**	**+ 5 μM 1ANF-FFH**	**+ 5 μM MBP**
6.6	0.010 ± 0.0009	0.010 ± 0.002	0.011 ± 0.001	0.012 ± 0.0007
6.8	0.012 ± 0.0009	0.012 ± 0.0009	0.015 ± 0.001	0.015 ± 0.001
7.0	0.015 ± 0.0012	0.016 ± 0.0006	0.016 ± 0.0006	0.017 ± 0.001
7.2	0.017 ± 0.0036	0.017 ± 0.0002	0.018 ± 0.001	0.020 ± 0.002
7.4	0.021 ± 0.0011	0.021 ± 0.001	0.023 ± 0.002	0.024 ± 0.001

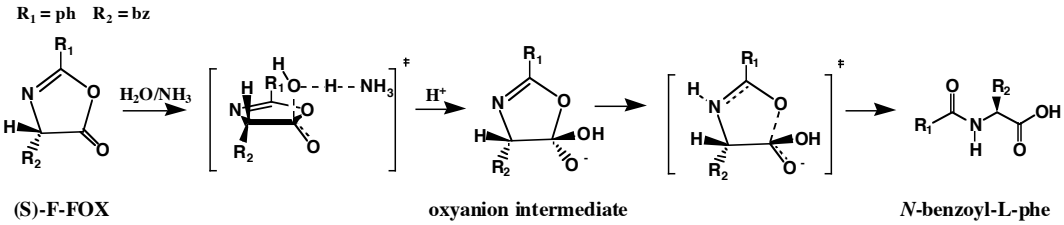


Figure A-1. Model system for dynamic kinetic resolution. Enantioselective hydrolysis of S-2-benzyl-4-phenyl-oxazolone-5-one ((S)-F-FOX).[8]

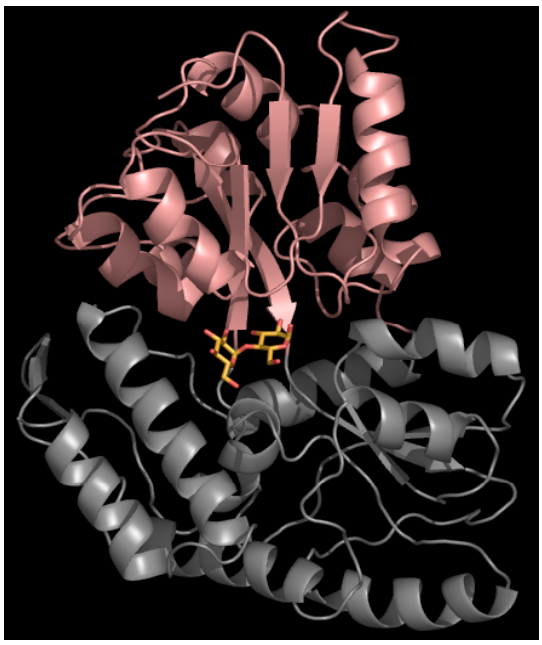


Figure A-2. Maltose binding protein structure. The scaffold for the design, maltose binding protein, in complex with maltose (pdb code: 1ANF).[17]

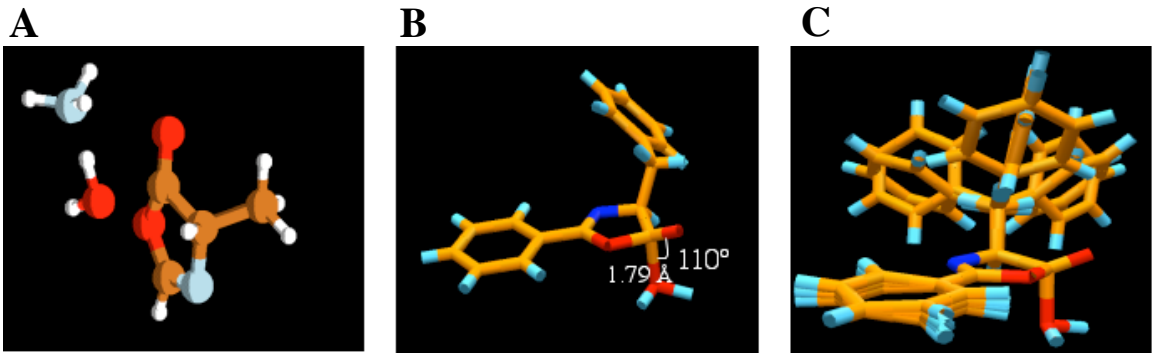


Figure A-3. (S)-F-FOX transition state structure. **(A)** The system for the *ab initio* calculation of the reaction transition state included a simplified (S)-F-FOX and a water molecule. NH_3 was added to activate the water. **(B)** Final transition state structure. **(C)** Rotamers of the transition state. The phenyl side chain can rotate 15° out of the plane of the oxazolone. The benzyl sidechain rotamers are based on the canonical backbone independent rotamers for phenylalanine.

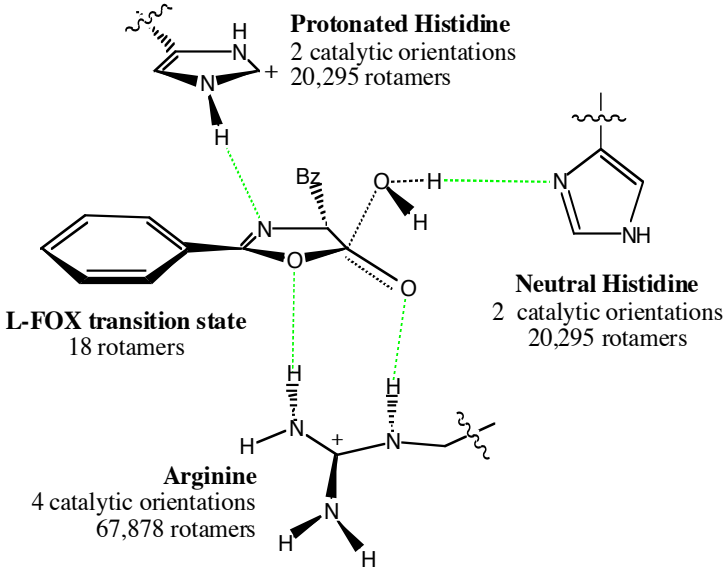


Figure A-4. Ideal active site contacts. Ideal hydrogen bond contacts between the desired catalytic residues and (S)-F-FOX hydrolysis transition state.

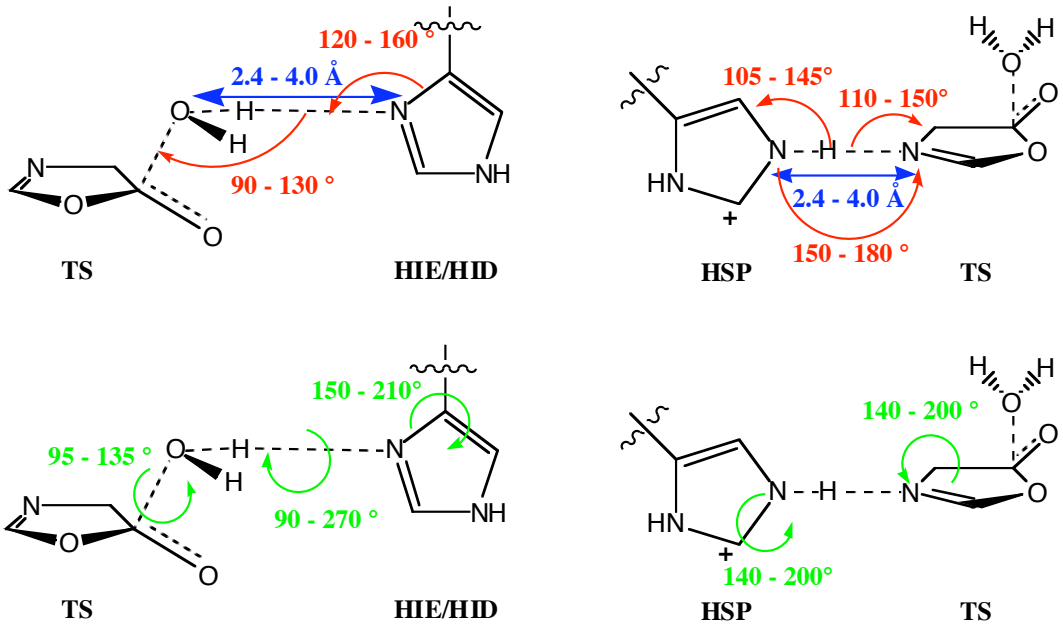


Figure A-5. Geometric constraints for the contacts between the catalytic residues and the (S)-F-FOX transition state (TS). Distance constraints are shown in blue, angle constraints are shown in red, and torsion constraints are shown in green. The transition state side chains were removed for clarity.

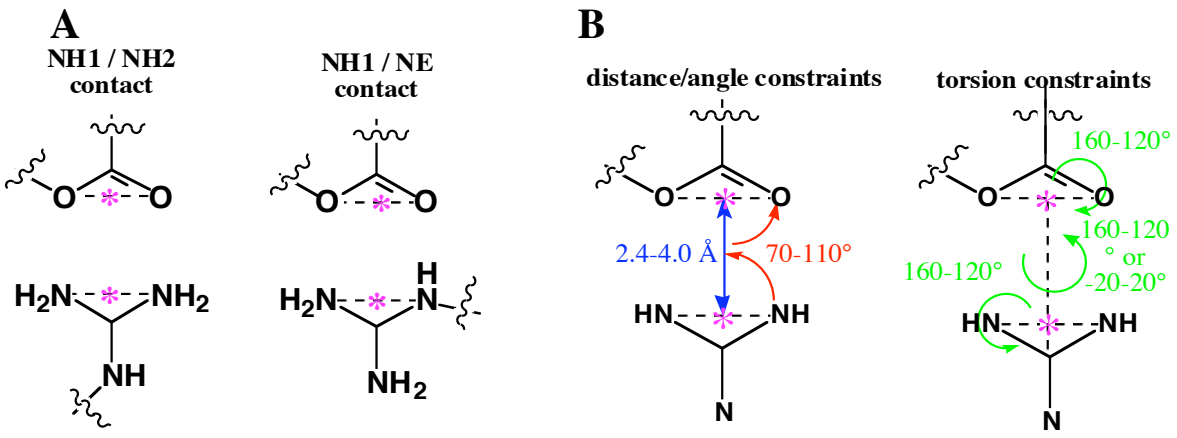


Figure A-6. Arginine–(S)-F-FOX geometric constraints. **(A)** Arginine can make four different double hydrogen bonding contacts to the transition state: NH1/oxazolone O and NH2/O; NH1/O and NH2/oxazolone O; NH1/oxazolone O and NE/O; NH1/O and NE/oxazolone O. Pseudo-atoms represented by pink stars. **(B)** Geometric constraints for the contacts between arginine and the transition state oxygens. Distance constraints are shown in blue, angle constraints are shown in red, and torsional constraints are shown in green. Pseudo-atoms are represented by pink stars.

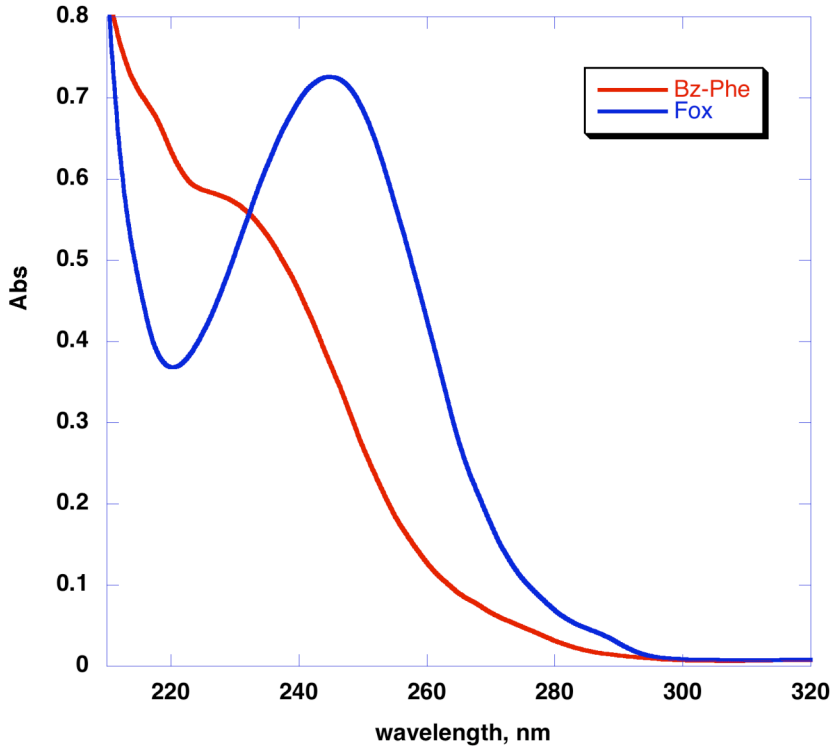


Figure A-7. UV-vis spectra of F-FOX and N-benzoyl-phenylalanine. λ_{max} for F-FOX is 244.5 nm.

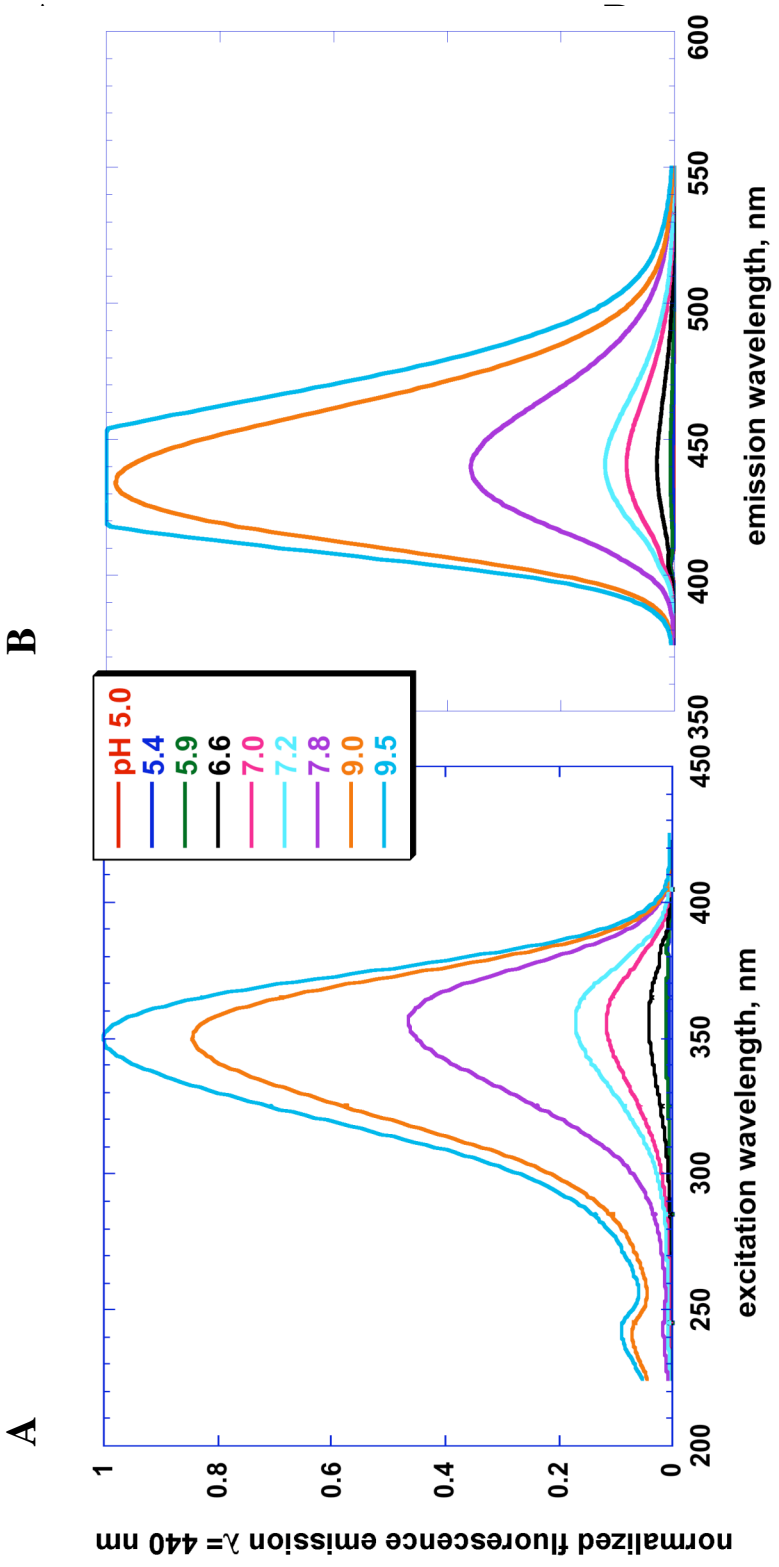


Figure A-8. Fluorescence of F-FOX. Wavelength scans were carried out using 4 μM F-FOX in 10 mM potassium phosphate buffer (**A**) pH dependence of F-FOX fluorescence λ_{em} = 440 nm. (**B**) pH dependence of F-FOX fluorescence emission at λ_{ex} = 356 nm.

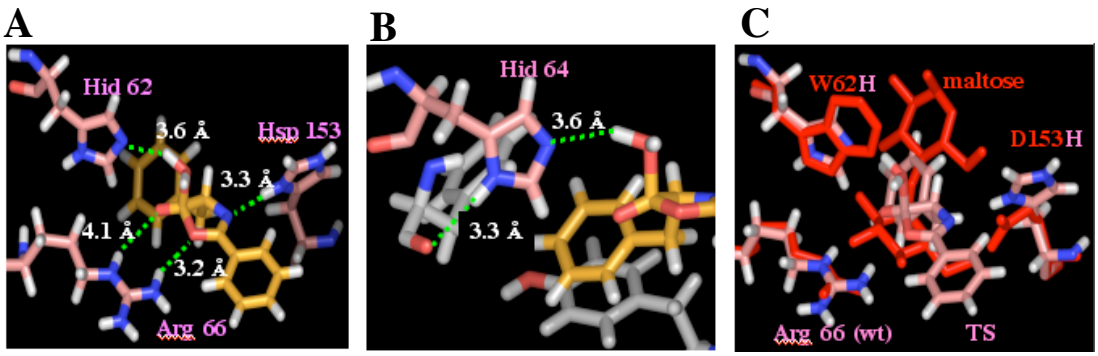


Figure A-9. Active site structure. **(A)** Chosen active site location and configuration. Hydrogen bonds are represented by dotted green lines. **(B)** Additional stabilization of Hid 64. **(C)** Designed active site residues (pink) overlaid with wild-type residues at those positions (red).

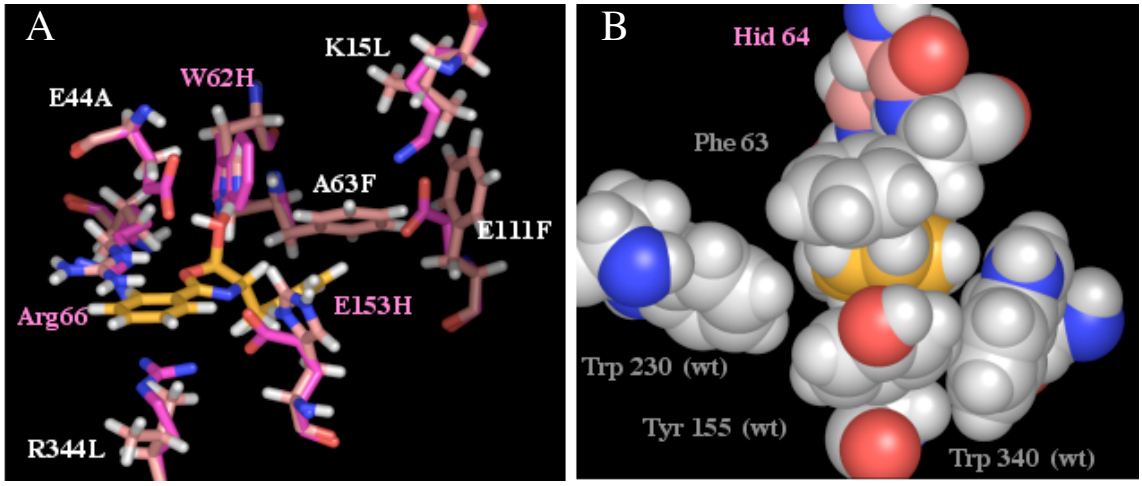


Figure A-10. Repacked active site. **(A)** The designed active site is shown in light pink and is overlaid with the wild-type binding pocket (magenta). The transition state structure is shown in yellow. **(B)** Space-filling representation of the active site. The benzyl side chain of the transition state is shown in yellow.

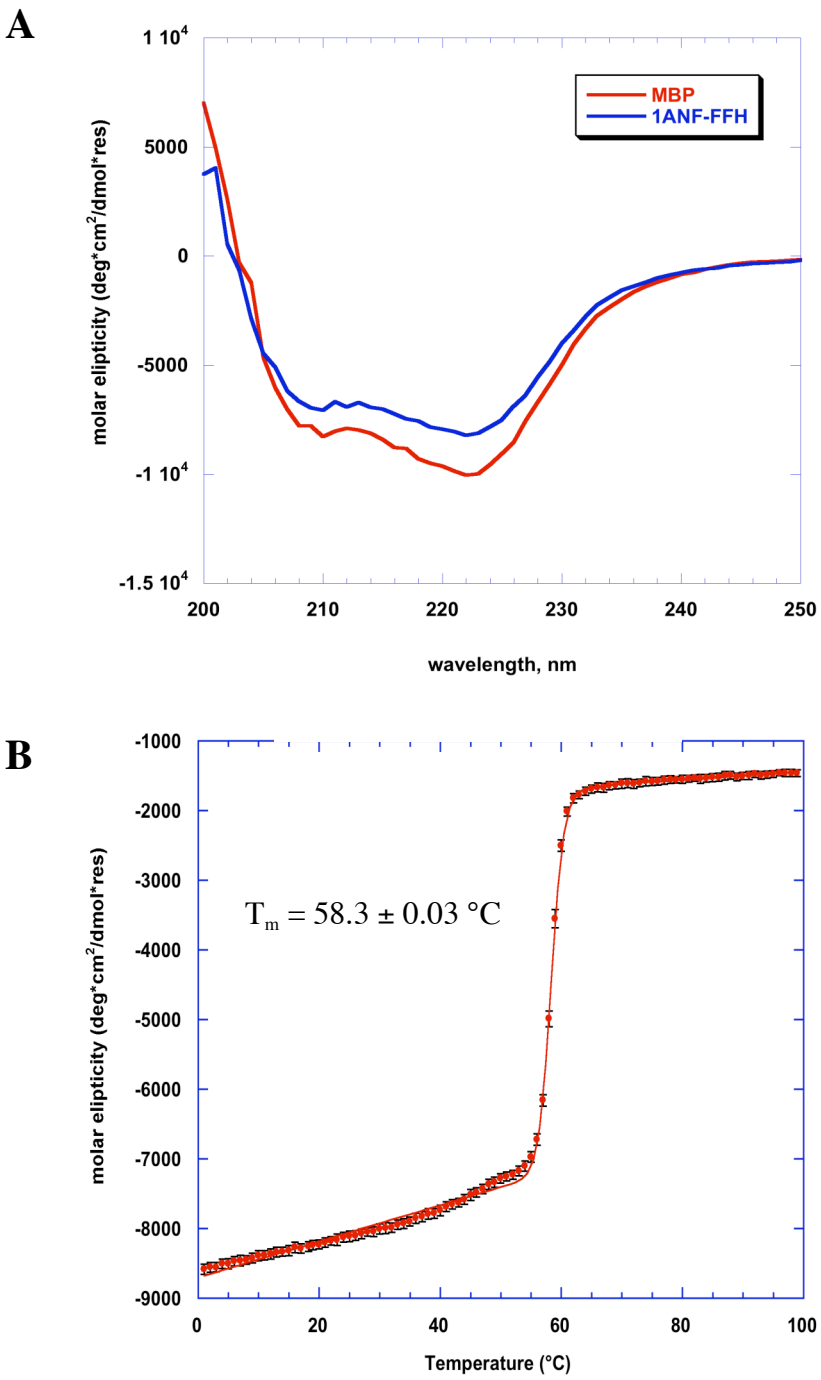


Figure A-11. 1ANF-FFH CD analysis. **(A)** CD wavelength scans of 1ANF-FFH compared to a scan for wild type MBP under identical conditions. **(B)** Thermal denaturation curve of 1ANF-FFH measured at 222 nm. All samples include 16.1μM 1ANF-FFH, 20 mM KPi 50 mM NaCl, pH 7.4.

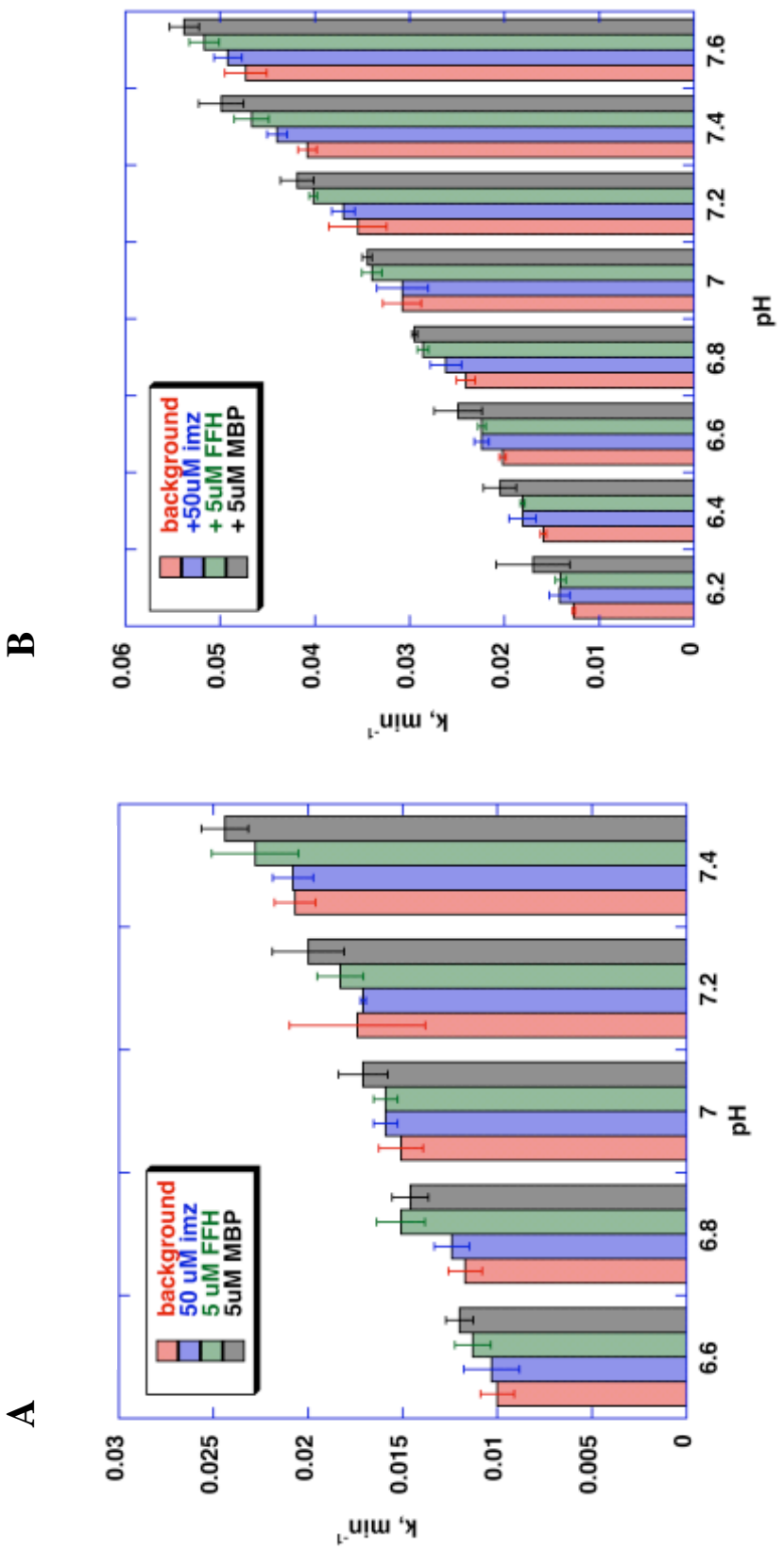
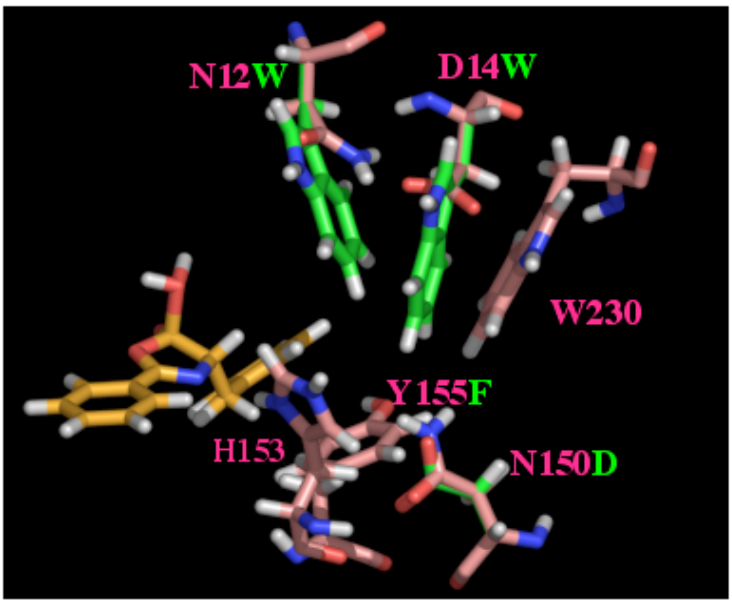


Figure A-12. F-FOX hydrolysis rate constants determined by UV-vis kinetics assays. The background rate constant was determined in the indicated buffer: (**A**) 10 mM KPi, 50mM NaCl or (**B**) 50 mM KPi, 50 mM NaCl. The rate constants for the hydrolysis of F-FOX in the indicated buffer in the presence of 50 μM imidazole, 5 μM 1ANF-FFH, and 5 μM MBP are shown in blue, green, and black, respectively.



Mutation	Purpose
N12W + D14W	overall stabilization of binding pocket
N150D	stabilization of protonated form of H153
Y155F	overall stabilization of active site
I239W + A96W	shift of MBP equilibrium to closed conformation

Figure A-13. Additional potential beneficial mutations to 1ANF-FFH. The transition state is shown in yellow. Wild-type residues are shown in pink and mutations are shown in green.

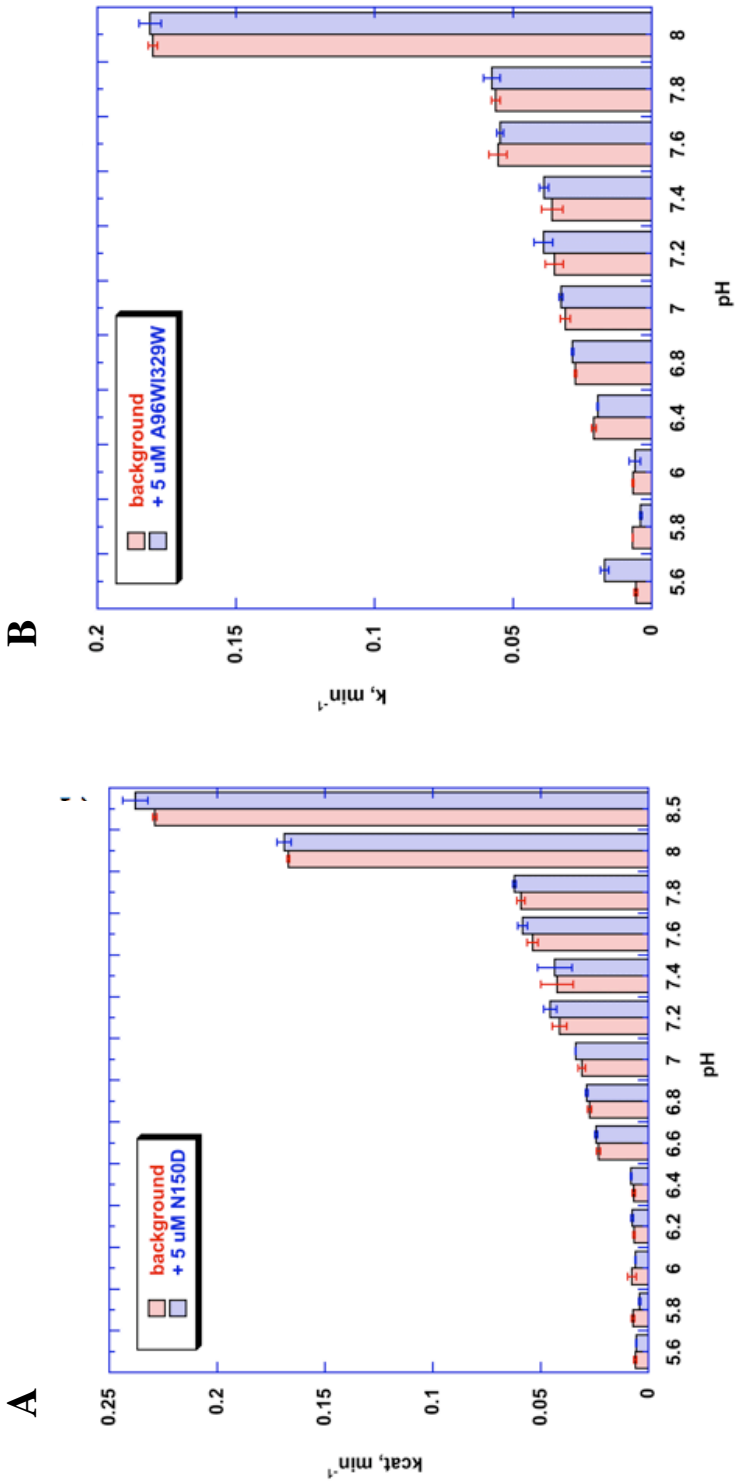


Figure A-14. 1ANF-FFH mutant apparent rate constants. The rate constants for **(A)** 1ANF-FFH N150D mutant and **(B)** 1ANF-FFH A96W/I329W double mutant were determined in 50 mM buffer, 50 mM NaCl at 25°C. No rate acceleration over background was seen for either variant under any condition tested.

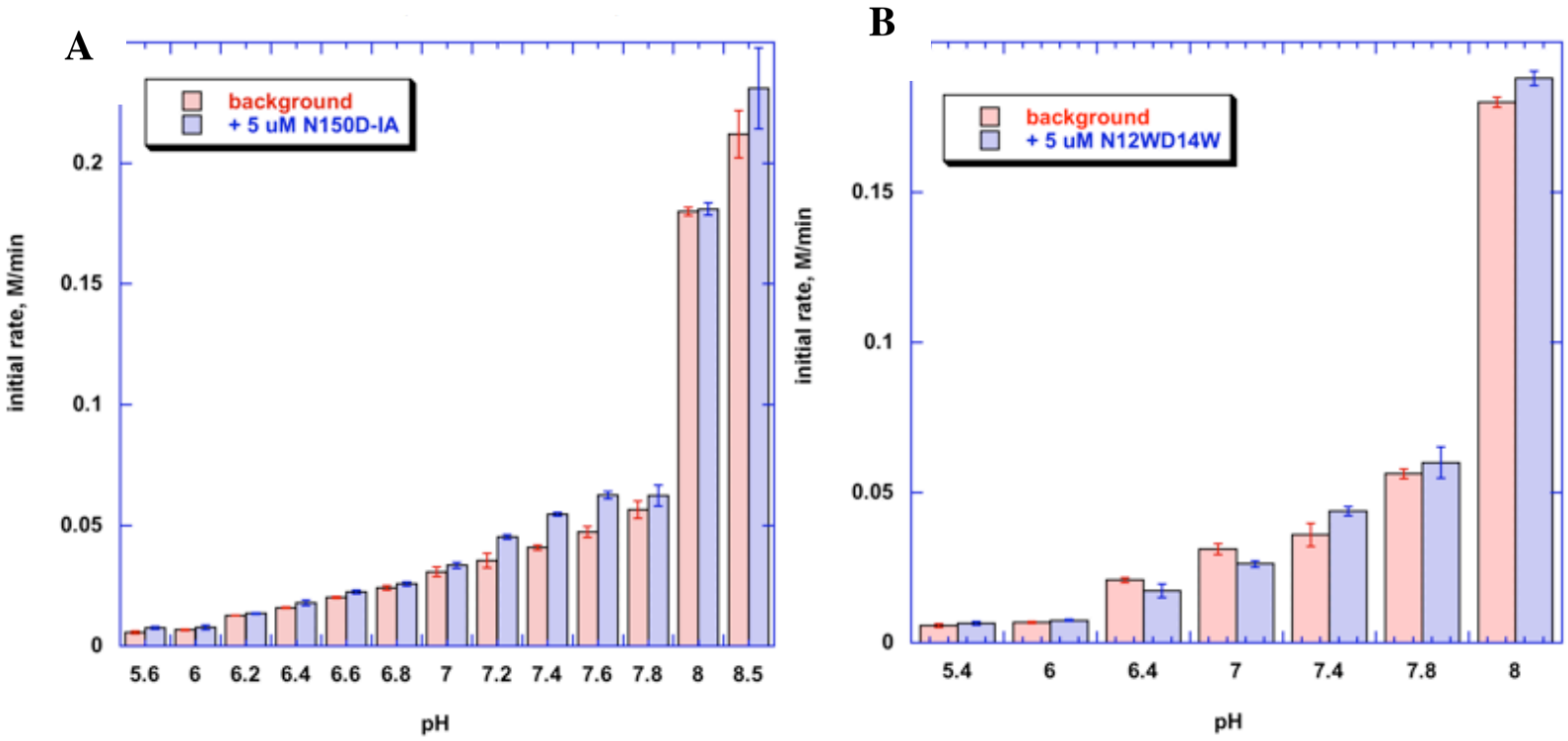


Figure A-15. 1ANF-FFH mutant apparent rate constants. The rate constants for: **(A)** 1ANF-FFH N150D/A96W/I329W triple mutant and **(B)** 1ANF-FFH N12W/D14W double mutant were determined in 50 mM buffer, 50 mM NaCl at 25°C. No rate acceleration over background was seen for either variant under any condition tested.

Appendix B

Using computational library design to alter the specificity of a xylanase

The work described here was carried out in the Mayo lab. All TAX activity screening and kinetic assays were carried out by Dr. Roberto Chica.

Abstract

Computational library design allows us to explore additional highly ranked sequences in the energy landscape near the minimum energy conformation that was predicted by a protein design calculation. Here, our goal was to use computational library design to adjust the specificity of a xylanase that exhibits a broad specificity for monosaccharide and disaccharide substrates. So far, none of our designs have exhibited an increase in specificity for the two substrates tested. However, this project is ongoing and calculations are currently being carried out on additional substrates.

Introduction

Computational library design is a special application of computational protein design discussed in Chapter I. Instead of identifying a single best sequence, the goal of computational library design is to specify a library of sequences that represents a group of top-scoring sequences from the computational design calculation and can be encoded by a single degenerate codon at each design position. This protein design procedure called Combinatorial Libraries Emphasizing and Reflecting Scored Sequences (CLEARSS) was developed in the Mayo lab and is advantageous over existing computational library design software because it ranks possible libraries via energy calculations of every member of each library and allows direct control over the size of the library.[1,2]

The relationship between protein structure and function is not completely understood. Consequently, many approximations are used in the protein design algorithm, including a fixed backbone structure, discrete sidechain rotamers, an energy function optimized for the stabilization of small proteins, and heuristic models of protein function that are necessary for incorporation into the force field. The screening of a large number of highly ranked protein sequences identified by computational library design can compensate for these limitations by allowing us to examine a larger portion of sequence space instead of a single sequence, improving our chances of finding a sequence that incorporates the function of interest into the protein.[3]

The xylanase from the thermophilic fungus *Thermoascus aurantiacus* (TAX) is a convenient system for testing computational library design procedures because it has been well studied with established mechanism, specificities, and kinetics (Table B-1, Figure B-1).[4,5] Glycosidase activity can be readily monitored using commercial

fluorogenic or chromogenic substrate analogs (Figure B-2). TAX has been recombinantly expressed in high yield in *Escherichia coli* and can be purified to homogeneity with a single affinity chromatography step (see Appendix D). Also, a high-resolution crystal structure of TAX bound to xylobiose is available at 1.7 Å resolution (PDB: 1GOR).[5]

The specificity of TAX is highly stringent in some respects and broad in others. TAX can hydrolyze *para*-nitrophenol (*p*NP)-derivatives of many pentose and hexose mono- and disaccharides, such as glucose and xylose; however it is inactive on others, such as mannose and maltose (Table B-1).[5] Upon modeling these inactive sugars into the active site of TAX, it is evident that steric clashes are the usual cause of the inactivity (Figure B-3). For example, glucopyranoside may be less active than xylopyranoside because the hydroxymethyl on C5 of glucose is expected to clash with W275. Mannopyranoside may be less active than glucopyranoside because the inverted stereochemistry at C2 causes the hydroxyl to clash with W84, N126, and E127. Maltose appears to be inactive because the $\alpha(1\rightarrow4)$ bond causes the second glucose to adopt a conformation that is almost perpendicular to the binding pocket. In general, disaccharides are more active than monosaccharides because they have an increased number of binding interactions to compensate for distortion in the transition state (TS).

In the work described here, we used CLEARSS to modify the specificity of TAX. In addition to evaluating the effectiveness of the library design calculation methods, these experiments were designed to elucidate the exact requirements for specificity in the TAX active site, especially the primary binding interactions that are critical for substrate

binding. None of the designs to increase the specificity for xylose or glucose have been successful so far; however, this project is ongoing.

Materials and Methods

Carbohydrate structures

The structures of the carbohydrate substrates were generated using the molecular modeling program BIOGRAF[6] to modify the xylobiose structure found in the crystal structure from Lo Leggio *et al*. (PDB: 1GOR).[5] The xylobiose structure was appropriately modified, hydrogen atoms were added, and the resulting structure was minimized for 50 steps. Charges were calculated using an electrostatic potential fit, methanol solvation including a dielectric constant of 33.62, and a hybrid density functional B3LYP as implemented by the Jaguar 5.5 software package and a 6-31G** basis set.[7] Rotamers were created using canonical torsions (60°, 180°, and 300°) for each rotatable bond in the region of carbohydrate that differs from xylobiose (e.g., the C5 hydroxymethyl group on glucose). The initial position of each carbohydrate structure within the TAX active site was determined by overlaying with the crystallographic xylobiose.

Phoenix calculations

Protein design calculations were carried out with Phoenix[1,2,8] using xylose- or glucose-based intermediates. For the xylose calculations, only Trp and Phe were allowed at position 275 to account for the results of the site-saturation mutagenesis, which showed

that only these two active residues at this position result in an active enzyme. Other design positions were 90 and 276. These positions were allowed to sample all rotamers of all amino acids except Pro. Residues in the immediate area of the design residues were allowed to change conformation but not identity (46, 47, 50, 83, 84, 87, 89, 130, 172, 207, 209, 239, and 267). The nucleophile of the reaction, E237, was required to be Ala to prevent steric clashes that could result from the close proximity of the nucleophile and the TS in the active site in the absence of a covalent bond.

An occlusion-based solvation potential was applied with scale factors of 0.05 for nonpolar burial, 2.5 for nonpolar exposure, and 1.0 for polar burial.[8] Other standard parameters were applied as in Lassila *et al.,* and a backbone-independent conformer library was used to represent side-chain flexibility.[9] The ligand was allowed to translate ± 0.2 Å in every direction in 0.2 Å steps and rotate ± 5° in every direction in 5° steps. As in the enzyme design calculations (Chapters II, III, IV, and Appendix A), geometric constraints were imposed to preserve important contacts between the intermediate and the active site sidechains. In this case, these contacts are wild-type ligand-binding contacts that are found in the crystal structure of TAX and are described in terms of distance only (Figure B-4).

To preserve the contact between xylose and K50, the Nζ of a Lys residue was required to be between 3.0 and 3.5 Å from O4 of xylose and between 2.5 and 3.0 Å from O3 of xylose. For the contact between H83 and xylose, the Nε of a Hid residue was required to be between 2.5 and 3.5 Å from O2 and O3 of xylose. Finally, the Nδ of Asn was required to be between 2.5 and 3.0 Å from O2 of xylose to preserve the contact between xylose and N172.

As in previous enzyme design calculations, sidechain-ligand interaction energies were biased to favor those contacts that satisfy the geometries. Sequence optimization was carried out with FASTER,[10,11] and a Monte Carlo-based algorithm[12,13] was used to sample sequences around the minimum energy conformation from FASTER (FMEC).

For the glucose calculations, hydrophobic residues (Ala, Ile, Leu, Val, Phe, Tyr, and Trp) were allowed at position 275 and nine internal rotamers of glucose were used to represent flexibility in the hydroxymethyl group. Otherwise, the glucose calculations were carried out in the same manner as the xylose calculations.

Library design calculations

Library design calculations were also carried out in Phoenix. The size of the libraries was set to 120 and the wild-type residue was required to be a member of the library at each design position.

Library construction

Construction of the libraries was carried out using splicing by overlap extension (SOE) mutagenesis to introduce degenerate codons at the design positions.[14] The site-saturation mutagenesis primers that were used are listed in Table B-2 (Integrated DNA Technologies). These primers were designed with the same rules as those for site-directed mutagenesis described in Chapter III, Materials and Methods. Mutagenesis primers for the xylose and glucose libraries were identical except for the replacement of the NNS site-saturation codon with the degenerate codon from the library design calculation (Table B-3). The mutagenesis fragments were constructed by combining 170

ng of template DNA (TAX-pET11a, described in Appendix D), 10x Thermopol buffer (New England Biolabs), 1 U Vent DNA polymerase (New England Biolabs), 0.5 mM dNTP mixture, 25 μmol forward or reverse mutagenesis primer, and 25 μmol forward or reverse flanking primer (TAX_NdeI_forward or TAX_BamHI_reverse) in a total reaction volume of 100 μL. The reactions were carried out on Mastercycler Personal Thermocycler (Eppendorf) using the temperature program described in Table B-4. The gene fragments were run on a 1% agarose gel and the bands were extracted and purified using a QIAquik Gel Extraction Kit (Qiagen).

Assembly reactions were carried out using an equimolar mixture of the forward and reverse fragments (200 ng total) and 10x Thermopol buffer (New England Biolabs), 0.5 mM dNTP mixture, 25 μmol forward and reverse flanking primers. and 1 U Vent DNA polmerase (New England Biolabs) in a total of 100 μL. The temperature program used for the PCR is described in Table B-4.

After amplification, the reactions were purified with a QIAquik PCR Purification Kit (Qiagen) then digested with *Bam*HI and *Nde*I (New England Biolabs). The digested genes were then run on a 1% agarose gel and the bands were extracted and purified using a QIAquik Gel Extraction Kit (Qiagen). The digested genes were then ligated into a similarly digested pET11a vector (Novagen). The libraries were transformed into *E. coli* BL-21 Gold (DE3) ultracompetent cells (Stratagene) and individual colonies were sequenced to confirm an adequate distribution of amino acids at each site (Agencourt).

TAX expression

Colonies were picked into 250 μL LB/ampicillin supplemented with 10% glycerol in 96-well plates and grown overnight at 37°C with shaking. These pre-cultures were used to inoculate 300 μL cultures in Overnight Express Instant TB media (Merck Biosciences), which were grown overnight at 37°C with shaking. The cells were pelleted, washed with phosphate buffered saline (PBS), and frozen on dry ice. The pellets were thawed at 30°C and resuspended with lysis buffer (1x CelLytic B (Sigma-Aldrich), PBS pH 7.4, 1 mg/mL hen egg white lysozyme (Sigma-Aldrich)). Lysis was carried out at 30°C for 30 min with shaking and the lysate was centrifuged for 15 min at 4°C at 3000 × g.

TAX activity assays

200 mM stocks of 4-methylumbelliferyl-β-D-glucopyranoside (MUG) and 4-methylumbelliferyl-β-D-xylopyranoside (MUX) were made in *N,N*-dimethylformamide (DMF). For the assays, the stocks were diluted to 10 mM MUG or MUX in citrate-phosphate buffer, pH 5.0 with a total DMF concentration of 5%. For end-point screening assays, 20 μL of cell lysate supernatant was added to 80 μL of the buffer with substrate in black 96-well microtiter plates with clear bottoms (Greiner). The plates were incubated for 30 to 60 min at 40°C and the fluorescence intensity was measured at λ_{em} = 445 nm (λ_{ex} = 360 nm) using a Safire^2 microplate reader (Tecan). Variants that exhibited end-point fluorescence intensity at least 1.5 times that of the median fluorescence value of the plate were considered to be active.

For kinetic analysis of selected variants, protein expression, purification, and concentration determination were carried out as for the TAX-based designs described in Chapter IV, Materials and Methods. Dilution series of MUX and MUG were made by serial two fold dilutions into DMF starting with a 250 mM stock solution in DMF. Final substrate concentrations ranged from 98 μM to 12.5 mM. In a clear bottom 96-well plate, 180 μL 50 mM citrate-phosphate buffer, pH 5.0 was combined with 10 μL substrate stock in DMF. The reaction was initiated by the addition of 10 μL TAX to a final concentration 30 μM. The release of 4-methylumbelliferone was monitored by an increase in fluorescence intensity at 445 nm with a λ_{ex} of 360 nm. The conversion factor for arbitrary fluorescence units (AFU) to concentration of 4-methylumbelliferone was determined to be 120 ± 23 AFU/μM by a standard curve.

Results and Discussion

The goal of our calculations here was to increase the specificity of TAX for both xylose and glucose individually. The design positions for these calculations were determined based on an overlay of the xylobiose and cellobiose-bound crystal structures of two xylanases (Figure B-5).[5,15] The positions with the most deviation in their conformations were identified as Q90, W275, and R276. To determine the tolerance of TAX for mutations in the active site, site-saturation mutagenesis was performed at these three sites. TAX activity assays on MUX indicated that in the case of xylose, only Phe and Trp are tolerated at position 275. At position 90, Gln (wild-type), Ser, and Arg are the residues that confer the most activity, but most other residues are also tolerated.

Position 276 shows a high tolerance for mutation with the greatest activity resulting from Arg (wild-type), Val, Leu, and Gly.

The results of the site-saturation mutagenesis suggest that our xylose libraries should be limited to Phe and Trp at position 275, and all residues should be sampled at the other two positions. Because no degenerate codon encodes just Phe and Trp, two separate calculations were carried out: one with Phe only at position 275 (Xyl1-Phe) and another with Trp only at 275 (Xyl2-Trp). In the library design calculations, the size of these two libraries was reduced to 60 so that the total size of the combined xylose libraries would be 120. Unfortunately, the signal for the assays of the MUG substrates was too low to get any meaningful site-saturation data. For the glucose calculations, we allowed only hydrophobic residues at position 275 to allow room for the C5 hydroxymethyl group and to help shield the active site from the solvent.

The results of the library design calculations are shown in Table B-3. In each of the libraries, a majority of the library members have calculated energies that are favorable. The two xylose libraries are identical except for the residue at position 275. Thus, the combination of these two libraries results in a full-sized library with 50% Trp and 50% Phe at position 275. It is evident from these results that because of the genetic code, the degenerate codon that is selected after the library design can encode for more amino acids than were represented in the initial design calculation. These cases test our assumptions about the restrictions on the identity of a given position and allow us to determine if, for example, it was reasonable to require only hydrophobic residues at position 275 in the glucose calculation.

The diversity in the xylose libraries at position 90 includes Arg, Asn, and Gln, which were found to be highly active from site-saturation mutagenesis. In all of the libraries, the high level of diversity at position 276 agrees with the tolerance of this site as seen during the site-saturation mutagenesis.

The designed combinatorial libraries were screened in crude cell lysate in 96-well plates, and the initial rates of the hydrolysis of MUG and MUX were compared. Some trends were observed from the results of this initial screening: (1) the W275F mutation eliminates MUG hydrolysis activity in variants with this mutation. This mutation also tends to decrease the activity for MUX somewhat, resulting in an increase in specificity for MUX. (2) Mutants with the R276V mutation (without the W275F mutation), show decreased activity for both substrates, but the overall specificity for MUG is increased. (3) The R276L mutation in the absence of W275F decreases the glucose hydrolysis rate while maintaining a high xylose hydrolysis rate, resulting in an increase in specificity for MUX.

Based on initial screening results, three point mutants were chosen for further kinetic characterization: W275F, R276V, and R276L. The kinetic constants resulting from these assays are shown in Table B-5. Specificity was determined by the ratio of k_{cat}/K_m of MUX and MUG. For wild-type TAX, the ratio of k_{cat}/K_m values of MUX to those for MUG is about 9. The R276V mutation decreased k_{cat}/K_m for both MUX and MUG. The R276V and R276L mutants were expected to increase the specificity of TAX for xylose over glucose. Instead, a slight decrease in MUX specificity was observed (2.9 and 2.2, respectively), but these changes in specificity are not large enough to be considered significant. In contrast, the W275F mutation, which was expected to decrease

the specificity for MUX over MUG, caused a slight increase in the specificity for xylose to about 10. The observed changes in specificity were not large enough to be considered significant. The flexibility of the TAX active site near positions 275 and 276 may have contributed to the inability to design a Kemp elimination enzyme in the natural binding pocket, as described in Chapter III. The inherent flexibility at these positions may allow the binding pocket to adjust to fit a variety of substrates regardless of the identity at these positions, making it impossible to change the specificity for glucose and xylose by solely manipulating the residues here.

Conclusions

So far, we have not been able to significantly alter the specificity of TAX using computational library design; however, xylose and glucose are the only substrates for which TAX has been redesigned. The variation in the structures of xylose and glucose may be too subtle and correspond to a region of the protein that is inherently flexible, allowing rearrangement upon binding an alternate substrate. Additional substrates with more substantial differences in structure from xylobiose (e.g., maltose or mannose) may prove to be better substrates for a specificity switch. Cellobiose may make a better substrate for a change in TAX specificity because it has an additional glucose molecule with a hydroxymethyl group that must also be accommodated. In addition, the higher specific activity that TAX shows for many disaccharides should make screening with these substrates more consistent than with monosaccharides, which exhibit low signal-to-noise ratios.

Other protein design strategies may also be employed to change the specificity of TAX. "Negative design" is one such strategy, where in addition to designing for increased binding to one specific substrate to shift specificity, a separate calculation is carried out with the goal of decreasing binding to another, alternate substrate. The best sequence in this type of calculation would have a low energy when bound to the target substrate and a high energy when bound to the alternate substrate.

Because of the limited number of substrates tested up to this point, we cannot draw any conclusions about our computational library design procedure. However, we are encouraged by the correlation between the site-saturation mutagenesis results and the sequences predicted in the library design calculation. As a larger number of substrates with more diverse structures are used to design combinatorial TAX libraries, we will be better able to assess the effectiveness of our library design methods.

References

1. Allen, B. D. *Development and validation of optimization methods for the design of protein sequences and combinatorial libraries*. California Institute of Technology: Pasadena, CA, **2009**.

2. Nisthal, A.; Allen, B. D., (*Manuscript in preparation*). **2009**.

3. Treynor, T. P.; Vizcarra, C.; Nedelcu, D.; Mayo, S. L., Computationally designed libraries of fluorescent proteins evaluated by preservation and diversity of function. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 48-53.

4. Lo Leggio, L.; Kalogiannis, S.; Bhat, M. K.; Pickersgill, R. W., High resolution structure and sequence of *T. aurantiacus* xylanase I: implications for the evolution of thermostability in family 10 xylanases and enzymes with (β)α-barrel architecture. *Proteins* **1999**, *36*, 295-306.

5. Lo Leggio, L.; Kalogiannis, S.; Eckert, K.; Teixeira, S. C.; Bhat, M. K.; Andrei, C.; Pickersgill, R. W.; Larsen, S., Substrate specificity and subsite mobility in *T. aurantiacus* xylanase 10A. *FEBS Lett.* **2001**, *509*, 303-308.

6. BIOGRAF version 3.21, Molecular Simulations, Inc., Burlington, MA, **1992**.

7. Jaguar 5.5, Schrodinger, L.L.C., Portland, OR, **1991-2003**.

8. Chica, R.; Allen, B. D., (*Manuscript in preparation*). **2009**.

9. Lassila, J. K.; Privett, H. K.; Allen, B. D.; Mayo, S. L., Combinatorial methods for small-molecule placement in computational enzyme design. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 16710-16715.

10. Desmet, J.; Spriet, J.; Lasters, I., Fast and accurate side-chain topology and energy refinement (FASTER) as a new method for protein structure optimization. *Proteins* **2002**, *48*, 31-43.

11. Allen, B. D.; Mayo, S. L., Dramatic performance enhancements for the FASTER optimization algorithm. *J. Comput. Chem.* **2006**, *27*, 1071-1075.

12. Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H., Equation of state calculations by fast computing machines. *J. Chem. Phys.* **1953**, *21*, 1087-1092.

13. Voigt, C. A.; Gordon, D. B.; Mayo, S. L., Trading accuracy for speed: A quantitative comparison of search algorithms in protein sequence design. *J. Mol. Biol.* **2000**, *299*, 789-803.

14. Ho, S. N.; Hunt, H. D.; Horton, R. M.; Pullen, J. K.; Pease, L. R., Site-directed mutagenesis by overlap extension using the polymerase chain reaction. *Gene* **1989**, *77*, 51-59.

15. Notenboom, V.; Birsan, C.; Warren, R. A.; Withers, S. G.; Rose, D. R., Exploring the cellulose/xylan specificity of the β-1,4-glycanase cex from *Cellulomonas fimi* through crystallography and mutation. *Biochemistry* **1998**, *37*, 4751-4758.

Table B-1. Specific activity of TAX on 2.5 mM *p*NP-glycosides. A dash indicates that no activity was detected. The data in this table were taken from Lo Leggio *et al*.[5]

substrate	activity (U/mg)
pNP-β-D-xylopyranoside	60
pNP-β-D-glucopyranoside	12.7
pNP-β-D-mannopyranoside	-
pNP-β-D-cellobioside	1290
pNP-β-D-lactoside	281
pNP-β-D-maltoside	-

Table B-2. Mutagenesis primers for site-saturation mutagenesis libraries. The degenerate codon is indicated in red. TAX_NdeI_forward and TAX_BamHI_reverse are the flanking primers used for the mutagenesis.

name	sequence
Q90X_forward	5'-CTGGTTTGGCACAGC**NNS**CTGCCGTCTTGGGTG-3'
Q90X_reverse	5'-CACCCAAGACGGCAG**SNN**GCTGTGCCAAACCAG-3'
W275X_forward	5'-GCCGATCCTGATTCT**NNS**CGCGCATCCACTACCC-3'
W275X_reverse	5'-GGGTAGTGGATGCGCG**SNN**AGAATCAGGATCGGC-3'
R276X_forward	5'- CGATCCTGATTCTTGG**NNS**GCATCCACTACCCCG-3'
R276X_reverse	5'- CGGGGTAGTGGATGC**SNN**CCAAGAATCAGGATCG-3'
TAX_NdeI_forward	5'-GAAGGAGATATACATATGGCAGAAGCG-3'
TAX_BamHI_reverse	5'- GTTAGCAGCCGGATCCCTAATGGTG-3'

Table B-3. Designed TAX libraries. The non-standard bases in the degenerate codons represent equimolar mixtures of bases: R = A, G; M = A, C; K = G, T; S = G, C; D = G, A, T; B = G, T, C; N = A, T, G, C.

library	residue #	wild-type	sampled amino acids	amino acids in library	degenerate codon
Xyl1-Phe	90	Q	ACDEFGHIKLMNQRSTVWY	HKNQRS	MRM
	275	W	F	F	TTC
	276	R	ACDEFGHIKLMNQRSTVWY	CFGILMRSVW	NKS
Xyl2-Trp	90	Q	ACDEFGHIKLMNQRSTVWY	HKNQRS	MRM
	275	W	W	W	TGG
	276	R	ACDEFGHIKLMNQRSTVWY	CFGILMRSVW	NKS
Glc	90	Q	ACDEFGHIKLMNQRSTVWY	HIKLMNQRS	MDS
	275	W	AFILVWVY	ACFGILMPSTVW	NBK
	276	R	ACDEFGHIKLMNQRSTWVY	ACFGILMPRSTVW	NBS

Table B-4. Thermocycler temperature programs for mutagenesis reactions.

SOE mutagenesis		
Temp (°C)	**Time (min)**	**# cycles**
94	5	
94	1	15x
55	1	
72	1	
72	2	
4	hold	

Table B-5. Kinetic constants for TAX variants with MUX or MUG.

	MUX			MUG				
TAX variant	k_{cat} (s^{-1})	K_m (mM)	k_{cat}/K_m ($s^{-1}M^{-1}$)	k_{cat} (s^{-1})	K_m (mM)	k_{cat}/K_m ($s^{-1}M^{-1}$)	MUX/MUG k_{cat}/K_m	MUG/MUX k_{cat}/K_m
wild-type	7.5E-03	3.3	2.26	1.4E-03	0.26	8.77	8.8	0.11
W275F	4.7E-03	2.6	1.8	1.6E-03	0.62	2.88	2.9	0.35
R276V	9.9E-03	2.6	3.82	9.1E-03	1.7	2.19	2.2	0.46
R276L	2.6E-03	2.0	1.28	3.5E-04	0.13	10.02	10.0	0.10

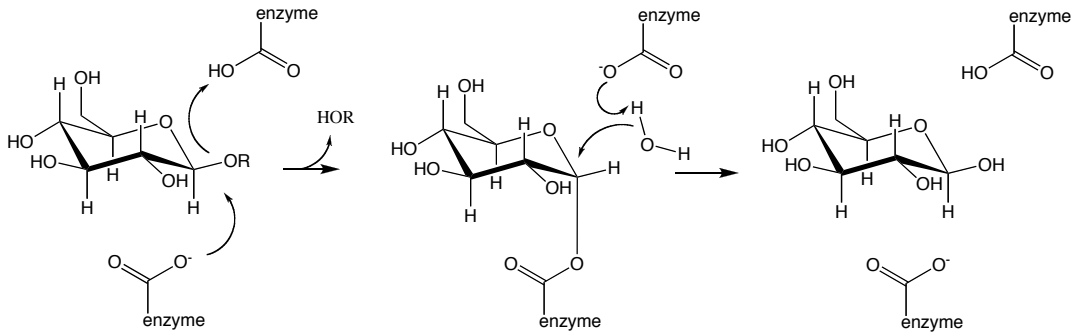


Figure B-1. Mechanism of retaining glycosidases.

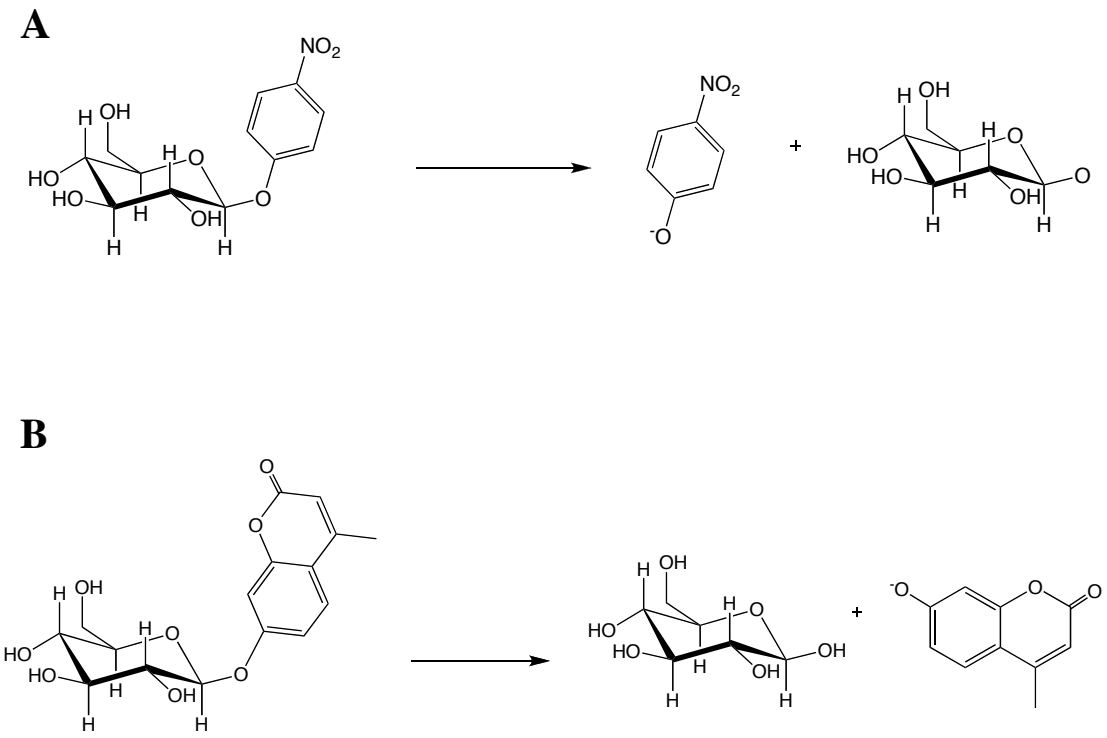


Figure B-2. Xylanase activity assays. **(A)** The hydrolysis of p-NP-glucopyranoside releases pNP, which absorbs with a λ_{max} of 410 nm. **(B)** The hydrolysis of 4-methylumbelliferyl-glucopyranoside releases 4-methylumbelliferyl, which is fluorescent with a λ_{ex} = 360 nm and a λ_{em} = 445 nm.

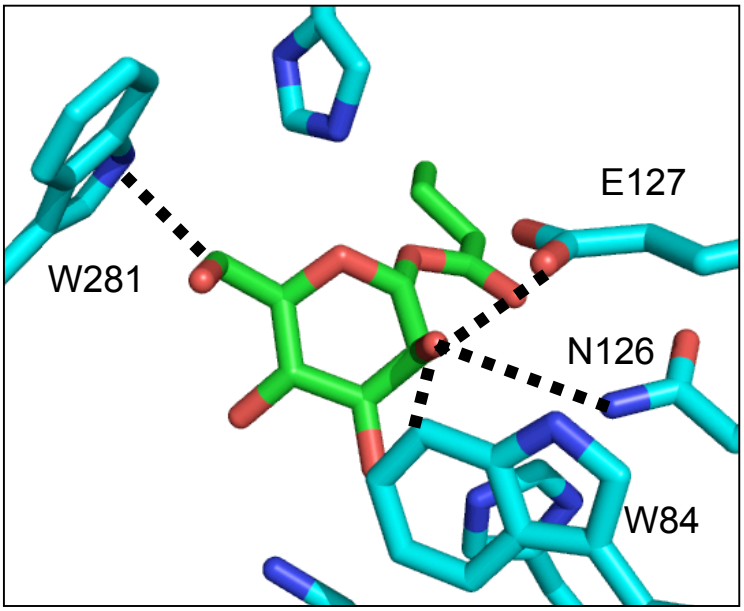


Figure B-3. Predicted clashes of mannose in the TAX active site. The modeled mannose-acyl intermediate structure is shown in green. Steric clashes are indicated with dotted lines.

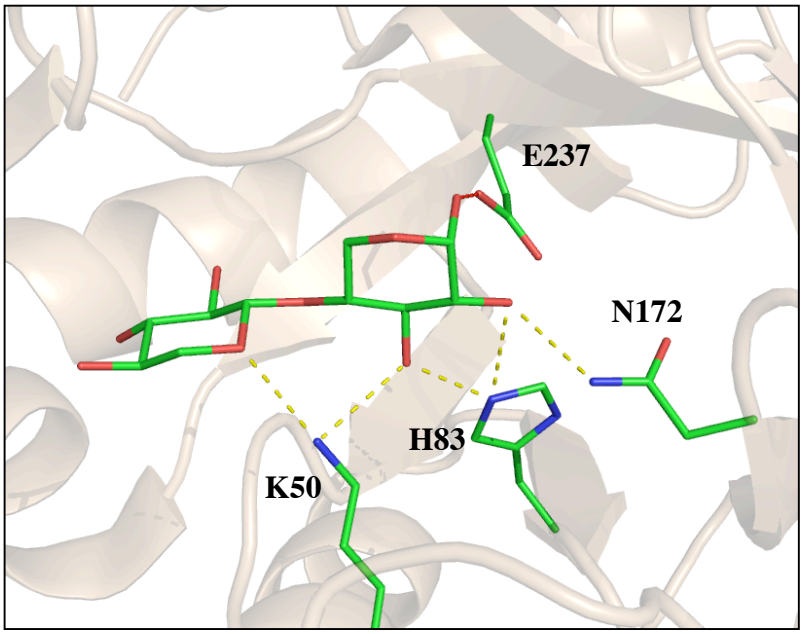


Figure B-4. Wild-type hydrogen bonds preserved in TAX calculations. The hydrogen bonds between K50, H83, N172, and the ligand are indicated with yellow dotted lines.

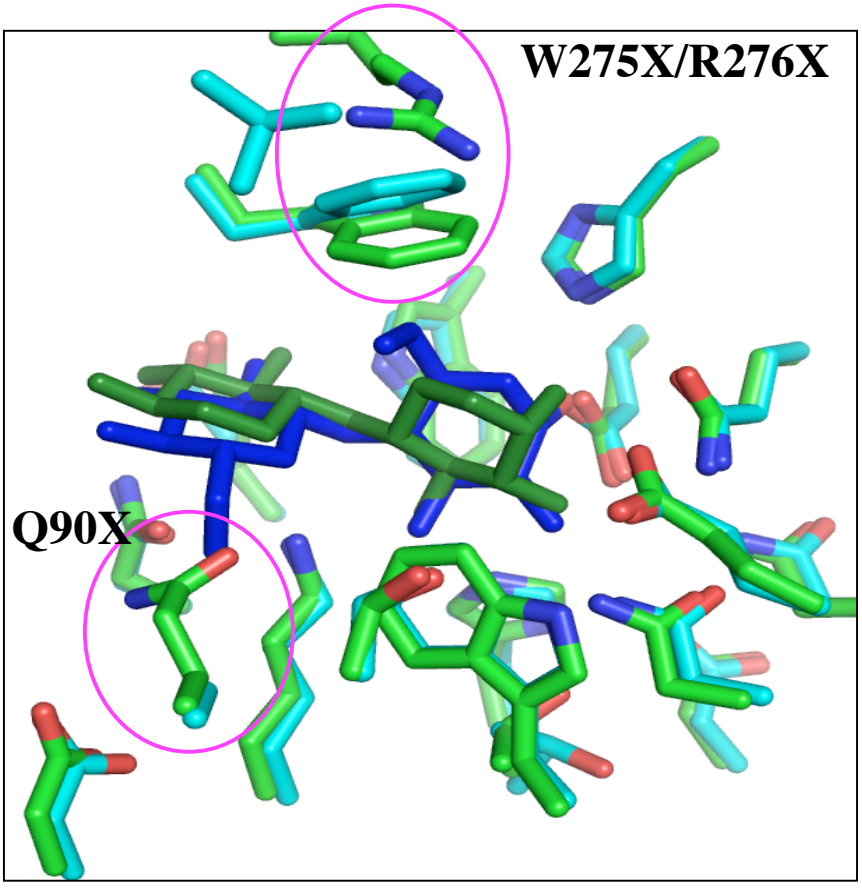


Figure B-5. Site-saturation mutagenesis positions in TAX. The *Cellulomonas fimi* xylanase is shown in cyan with the cellobiose shown in blue.[15] The TAX structure and xylobiose are shown in green.[5] Sites of site-saturation mutagenesis are indicated.

Appendix C

Altering the specificity of an androgen receptor

This project was carried out in collaboration with Prof. Robert Fletterick's lab at the University of California, San Francisco. The computational components of the project described here were carried out in the Mayo lab. All experimental work was carried out by Leslie Cruz in the Fletterick lab.

Abstract

Androgen receptors (AR) are nuclear hormone receptors that play a major role in sexual development. To study the role of AR on sexually stereotyped behavior, a mutant AR/ligand pair is needed that is orthogonal to the wild-type AR/androgen system. This variant AR could then be selectively activated in the presence of the wild-type AR and wild-type ligands. Here, we attempted to use our computational protein design methods to design an AR that can be activated by the nonnatural ligand 19PT. Our results so far indicate that none of our designed ARs are activated by 19PT.

Introduction

Some behavioral differences between the two sexes of a species are controlled on the molecular level. Male mice, for example, demonstrate aggressiveness towards other mice and are territorial, whereas female mice do not exhibit these behaviors.[1] These sexually dimorphic behaviors have prompted investigations into biochemical differences in the brains of male and female mice. However, few examples of molecular differences have been identified so far. One major difference has been observed in the distribution of androgen receptors (AR) in the hypothalamus of mice; male mice tend to have a larger number of highly localized neurons expressing AR, whereas females have fewer, more disperse neurons expressing AR.[1]

AR is a 110 kD nuclear hormone receptor, that is sequestered by a heat shock protein (HSP) in the cytoplasm. Upon binding dihydrotestosterone (DHT), a metabolic derivative of testosterone (TES), AR undergoes a conformational change that allows its translocation into the nucleus. Once in the nucleus, it homodimerizes and acts as a DNA transcription activator for genes that control sexual development and the maintenance of skeletal and muscular systems (Figure C-1).[2,3] While these roles of AR are widely accepted, less is known about the role of AR in the brain and its effect on sexually stereotypical behavior.[1]

Our goal here was to create a system that can facilitate the study of the role of AR in sex-differentiated behavior in a mouse model. To this end, we attempted to design an AR that can be activated by the nonnatural TES-analog 19PT (Figure C-2). The design of a novel AR-ligand pair is the first step to the design of a "magic pair" that is orthogonal to the wild-type AR-DHT system. The expression of this mutant AR in the

brains of mice would allow manipulation of mutant AR activity in a site-specific manner without affecting the endocrine system in the rest of the animal.

The strategy for redesigning the AR active site to accommodate 19PT is similar to that for designing enzyme active sites described in Chapters I-IV. As before, we used the computational protein design software ORBIT.[4] However, instead of designing around a reaction transition state (TS), we optimize the active site residues for binding to the ground-state ligand structure. Important binding contacts, such as hydrogen bonds, can be enforced using geometric constraints as in the enzyme design calculations. In addition, some rotational and translational freedom is given to the ligand during the design calculation and the internal flexibility of the ligand is modeled with canonical torsions of rotatable bonds, as in Appendix A.

Methods

Scaffold selection

Because it had been used for past analyses by the Fletterick lab, the 1.7 Å structure of AR from *Pan troglodytes* bound to DHT was chosen as the scaffold for the design.[5]

19PT structure and rotamer design

The DHT structure found in the crystal structure from Hur et al.[5] (PDB: 1T7T) was modified to produce the structure of 19PT using the molecular modeling program BIOGRAF.[6] Starting with the DHT structure, hydrogens and the 19-butyl group were added and the entire structure was minimized for 50 steps. Charges were calculated

using an electrostatic potential fit, methanol solvation, and the hybrid density functional B3LYP as implemented by the Jaguar 5.5 software package and a 6-31G** basis set.[7] Rotamers were created using canonical torsions (60°, 180°, and 300°) for each of the four rotatable bonds, resulting in 81 rotamers, which were then minimized for 250 steps using BIOGRAF (Figure C-3).[6]

Androgen receptor design for 19PT binding

The initial position of the 19PT ligand was determined by overlaying with the DHT structure in the 1T7T crystal structure.[5] The ligand was allowed to rotate ± 5° in x, y, and z in 5° steps and translate ± 0.2 Å in x, y, and z in 0.2 Å steps. Three positions (W741, M742 and M745) were chosen as design positions because of their proximity to the expected position of the 19PT butyl group based on its initial position (Figure C-4). All residues within 6 Å of DHT in the 1T7T structure (701, 704, 705, 707, 746, 749, 752, 746 768, 780, 784, 787, 873, 876, 877, 880, 889, 891, 895, 899) were allowed to change conformation but not identity.

As in Chapter III, Lazaridis-Karplus occlusion-based solvation was applied with scale factors of 1.0 for nonpolar burial and nonpolar exposure and a scale factor of 0.6 for polar burial (see Materials and Methods).[8] Other standard ORBIT parameters were applied as in Lassila *et al.* and a backbone-independent conformer library was used to represent sidechain flexibility[9]. Loose geometric constraints were applied to preserve hydrogen bonds that are present in the wild-type crystal structure. An Asp and Thr residue was required to be between 2.5 and 3.2 Å of one of the 19PT hydroxyls and an arginine contact was required between 4.0 and 5.0 Å of the other hydroxyl (Figure C-5).

Transcriptional activation assay

The activation of AR variants was monitored using a luciferase reporter assay in HeLa cells as described by Bohl *et al.*[10] In these experiments, the AR ligand binding domain is fused to the transcriptional activator Gal4 (Gal4-AR). Ligands were added to the culture after a day of expression and chemiluminescence was used to determine AR activation. Assays were carried out with 10 μM 19PT and 100 nM DHT. Percent activity was calculated based on DHT activity in Gal4-AR set at 100%.

Results and Discussion

The designs from the ORBIT calculations are summarized in Table C-1. In general, the sequences are hydrophobic and relocate the bulk in the active site from position 745 to position 715 to make room for the 19PT butyl group. The sequences resulting from the calculations fall into five groups, which can be separated based on the residue chosen at position 741. Groups 1-4 each contain 3 sequences that all have an alanine at position 745. All three sequences in each of these groups are structurally identical except for the residue at 715, which is either leucine, methionine, or isoleucine. These four groups are distinguished by the residue at position 741.

Group 1 has a tyrosine at position 741, whose hydrogen bond contact is unfulfilled. The 19PT structure in these designs does not have an extended butyl group. Group 2 designs are identical in structure to group 1, except that a phenylalanine is substituted for tyrosine at position 741 (Figure C-6A). Again, the butyl group of 19PT is

in a kinked conformation that may not be energetically favorable. Group 3 designs have a W715H mutation. In these designs, the histidine overlays well with the indole of the tryptophan in the wild-type crystal structure and the butyl of 19PT is in a fully extended conformation (Figure C-6B). Group 4 has a leucine at position 741, which causes the butyl group to adopt the same kinked conformation seen in group 2 (Figure C-6C). In contrast to the other sequences, group 5 has only a single sequence with a single W741A mutation and the 19PT butyl group is in a fully extended conformation (Figure C-6D).

These five groups of designs were submitted to the Fletterick lab for testing and a large number of variants were made. Only one of the variants (W741A) shows wild type-like DHT activation (Gal-AR) in the transcription activation assay and none of the AR variants that we designed so far shows significant activation by 19PT (Figure C-8). The loss of DHT activation in most of the designs indicates some kind of perturbation in the active site or a prevention of the conformational change necessary for activation. One of the limitations of computational protein design is that we can only design for ligand binding and not for AR activation. Unfortunately, in the case of AR, ligand binding is necessary but not sufficient for activation. To evaluate the success of the computational protein design process in this case, we would need to assay for binding directly. Currently, no binding assay exists for AR beyond crystallography, which is extremely low-throughput and not a reliable assay for binding due to the high potential for false negatives.

In addition, conformational changes of the backbone upon mutation of the binding pocket residues cannot be ruled out. In the crystal structure of AR/M745A, the backbone of the binding pocket is observed to shift slightly, changing the conformation of some of

the residues significantly, including W741 and N705 (Figure C-7).[11] Future design calculations could account for this backbone flexibility by using molecular dynamics or by including multiple backbone conformations during the design calculation.[12,13]

Conclusions

So far, we have not been able to use our computational protein design methods to design a mutant AR that can be selectively activated by the nonnatural ligand 19PT. However, without specific structural or ligand-binding data, it remains unclear if our methods failed or if this system is simply not amenable to computational design due to subtleties in the receptor structure-function relationship. This project is ongoing with experimental work in the Fletterick lab. Future computational studies of this system with molecular dynamics simulations may help us understand the effect of mutations on the activity of the receptor, and the incorporation of this information into future designs using multi-state design may aid us in creating active AR variants.

References

1. Shah, N. M.; Pisapia, D. J.; Maniatis, S.; Mendelsohn, M. M.; Nemes, A.; Axel, R., Visualizing sexual dimorphism in the brain. *Neuron* **2004**, *43*, 313-319.

2. Mooradian, A. D.; Morley, J. E.; Korenman, S. G., Biological actions of androgens. *Endocr. Rev.* **1987**, *8*, 1-28.

3. Keller, E. T.; Ershler, W. B.; Chang, C., The androgen receptor: a mediator of diverse responses. *Frontiers in Bioscience* **1996**, *1*, d59-71.

4. Street, A. G.; Mayo, S. L., Computational protein design. *Structure* **1999**, *7*, 105-109.

5. Hur, E.; Pfaff, S. J.; Payne, E. S.; Gron, H.; Buehrer, B. M.; Fletterick, R. J., Recognition and accommodation at the androgen receptor coactivator binding interface. *PLoS Biol.* **2004**, *2*, 1303-1312.

6. BIOGRAF version 3.21, Molecular Simulations, Inc., Burlington, MA, **1992**.

7. Jaguar 5.5, Schrodinger, L.L.C., Portland, OR, **1991-2003**.

8. Lazaridis, T.; Karplus, M., Effective energy function for proteins in solution. *Proteins* **1999**, *35*, 133-152.

9. Lassila, J. K.; Privett, H. K.; Allen, B. D.; Mayo, S. L., Combinatorial methods for small-molecule placement in computational enzyme design. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 16710-16715.

10. Bohl, C.; Miller, D. D.; Chen, J.; Bell, C. E.; Dalton, J. T., Structural basis for accommodation of nonsteroidal ligands in the androgen receptor. *J. Biol. Chem.* **2005**, *280*, 37747-37754.

11. Cruz, L.; Fletterick, R. J., (*Unpublished results*). **2007**.

12. Allen, B. D. *Development and validation of optimization methods for the design of protein sequences and combinatorial libraries*. California Institute of Technology: Pasadena, CA, **2009**.

13. Nisthal, A.; Allen, B. D., (*Manuscript in preparation*). **2009**.

Table C-1. AR-19PT design summary. The designs were separated into five groups based on their mutations at positions 741 and 745. Group 1 (blue) has W741Y and M745A mutations. Group 2 (orange) has W741F and M745A mutations. Group 3 (white) has W741H and M745A mutations. Group 4 (purple) has W741L and M745A mutations. Group 5 (yellow) has only a W741 mutation.

name	715	741	745	ORBIT energy (kcal/mol)	comments
wt	V	W	M		
1	L	Y	A	-216.9	Y742 makes no H-bond, butyl not fully extended
2	M	Y	A	-216.8	
3	I	Y	A	-216.8	
4	L	F	A	-216.0	butyl not fully extended, but takes up most of empty space
5	M	F	A	-215.9	
6	I	F	A	-215.9	
7	L	H	A	-215.5	His overlays well with indole of W741
8	M	H	A	-215.4	
9	I	H	A	-215.4	
11	L	L	A	-215.1	W241L has been made previously and does not preturb structure
12	M	L	A	-215.1	
13	I	L	A	-215.1	
14	V	A	M	-212.6	single mutation

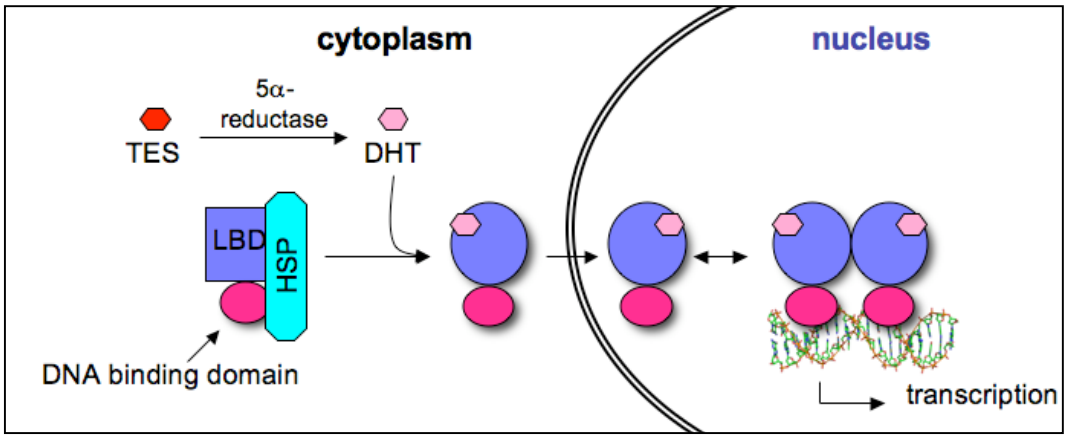
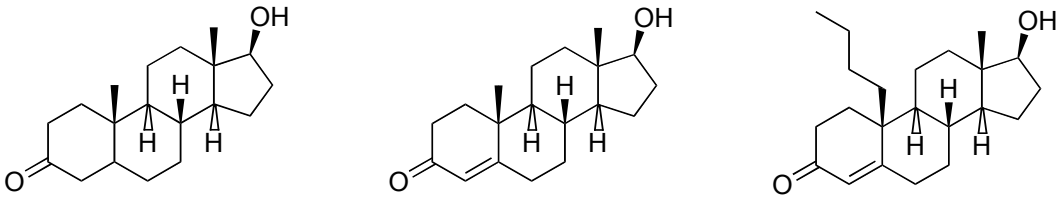


Figure C-1. Activation mechanism of AR.



dihydrotestosterone **testesterone** **19PT**

Figure C-2. Chemical structures of androgens of interest.

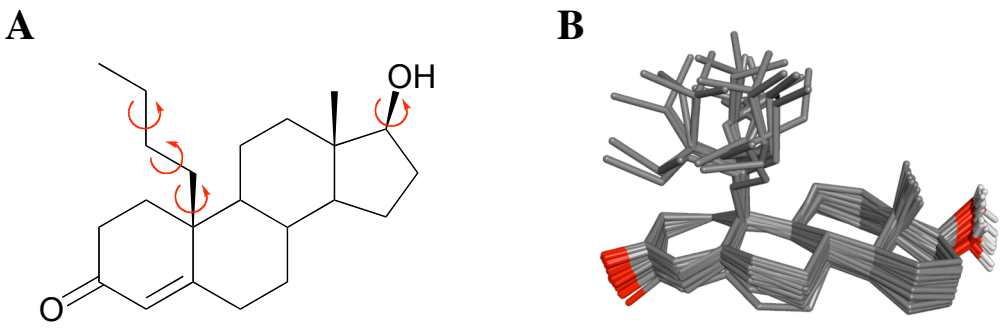


Figure C-3. Rotamers of 19PT. **(A)** Chemical structure of 19PT with red arrows indicating the positions of rotatable bonds. **(B)** Overlay of the 81 minimized 19PT rotamers used in the calculations.

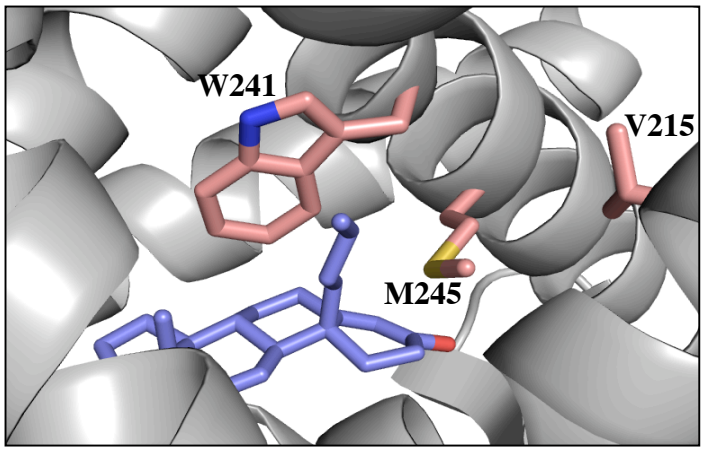


Figure C-4. Design positions in the active site of AR. Sidechains that were designed are indicated and a model of the 19PT ligand is shown in purple.

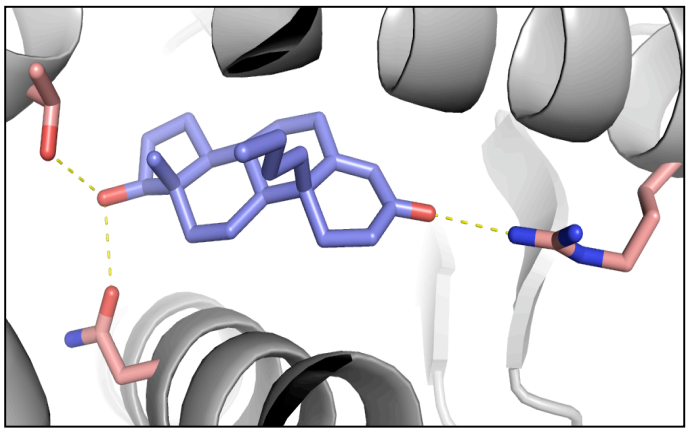


Figure C-5. Wild-type hydrogen bonds to 19PT. These hydrogen bond contacts were enforced through a simple geometry pruning step. The 19PT model is shown in purple.

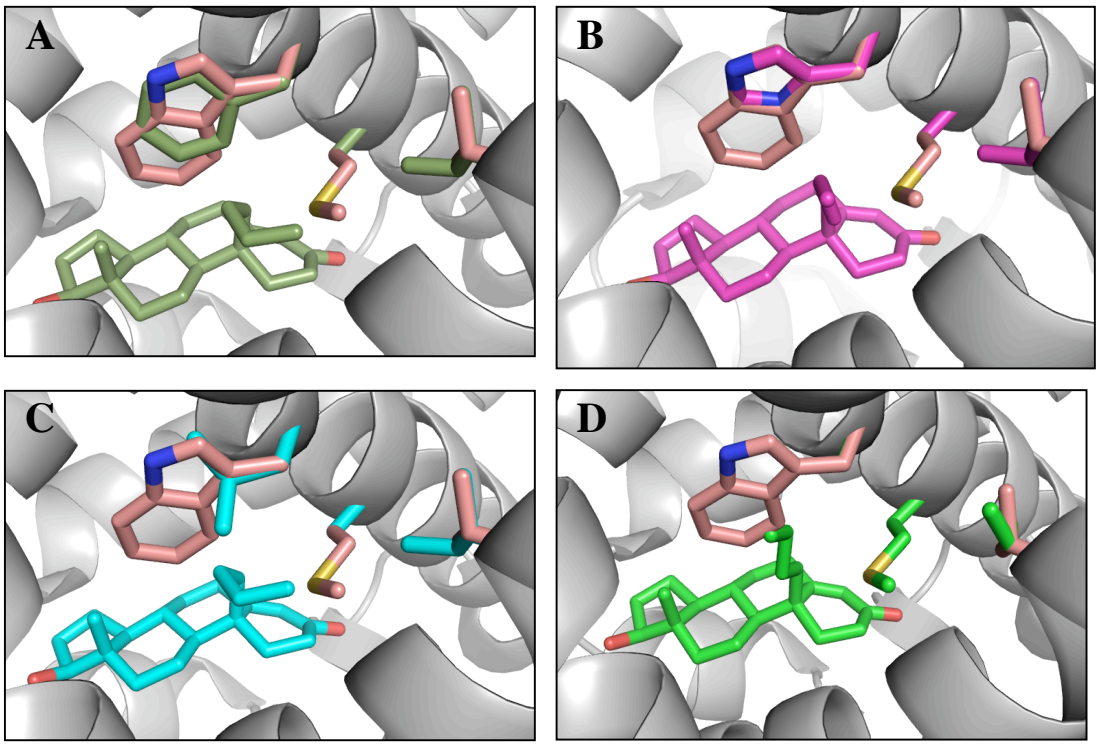


Figure C-6. ORBIT designs for AR binding of 19PT. Each design is shown overlaid with the wild-type AR crystal structure, shown in pink.[5] **(A)** AR/V215I/W241F/M245A design is shown in green. **(B)** AR/V215I/W241H/M245A is shown in magenta. **(C)** AR/V215I/W241L/M245A is shown in cyan. **(D)** W241A is shown in green.

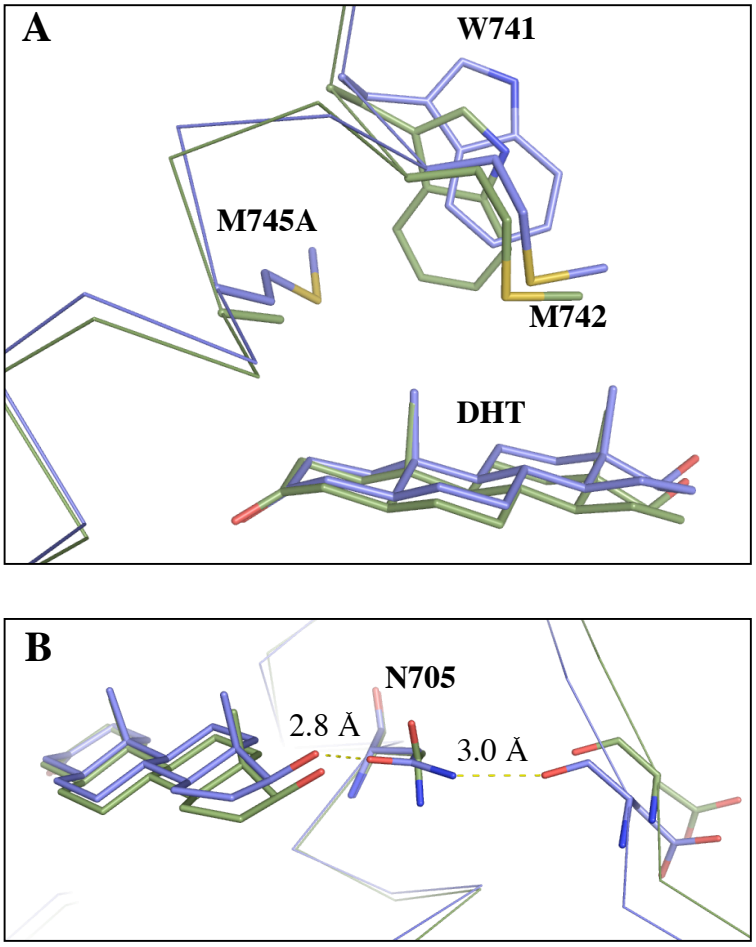


Figure C-7. Structural differences in the AR crystal structure caused by M745A. The structure of wild-type AR[5] is shown in blue and the structure of AR/M745A[11] is shown in green. DHT was bound in both structures. **(A)** Structural differences of W741 and M742. **(B)** Structural differences of N705.

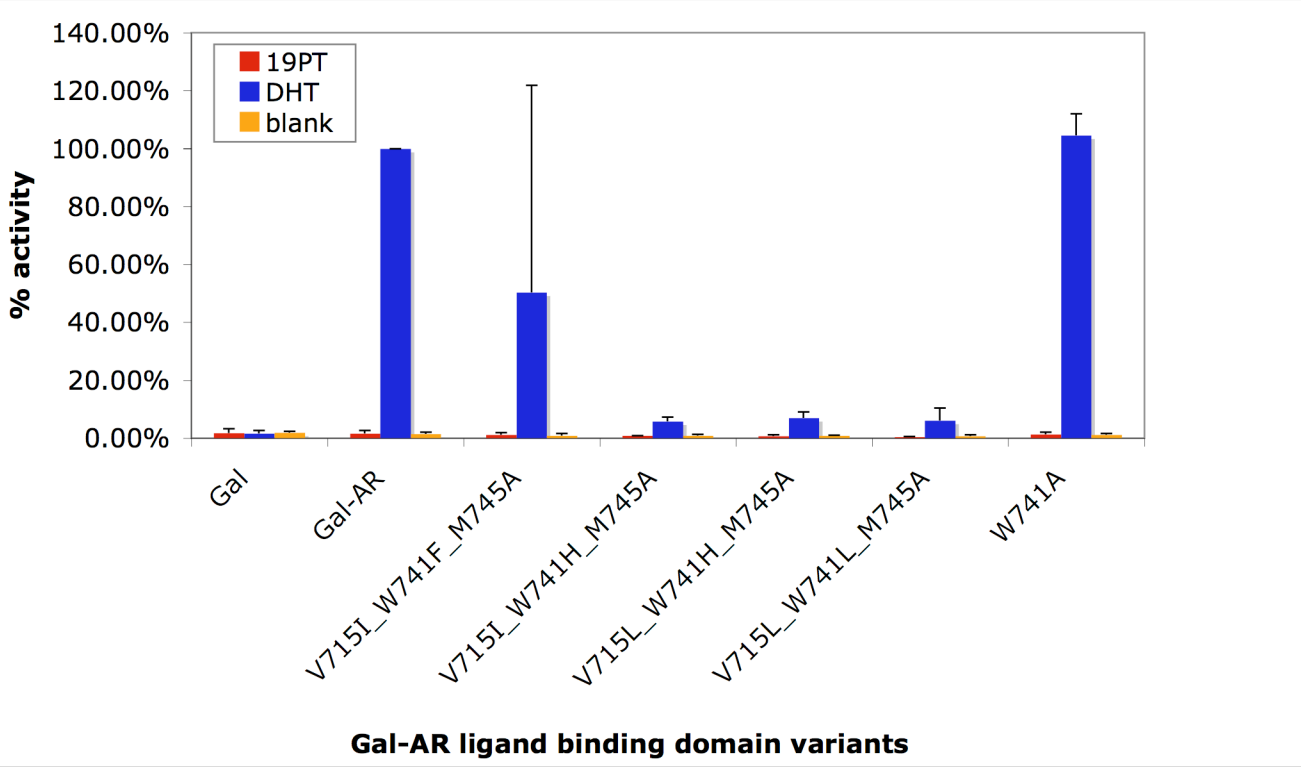


Figure C-8. Transcriptional activation assay. Percent activity is calculated based on the DHT activity in Gal-AR being 100%.

Appendix D

Recombinant expression and purification of a xylanase from the thermophilic fungus *T. aurantiacus*

Abstract

The *Thermoascus aurantiacus* xylanase 10A (TAX) is of interest for industrial and enzyme engineering purposes because of its ability to hydrolyze polymers of xylose at elevated temperatures. Previous expression of this xylanase has relied on natural production from fungal cell cultures, limiting the quantities that can be produced and hindering genetic manipulation that could be used to optimize existing activity or engineer novel activities. Here, we optimize the gene sequence of TAX for expression in *E. coli* and successfully express the protein in quantities up to 150 mg per liter of culture. In addition, a quick single-step purification protocol is developed that results in highly pure protein.

Introduction

Endo-β-1,4-xylanases are of industrial interest because of their hydrolytic activity against the internal glycosidic bonds of xylan, which is the major polysaccharide component of plant cell walls. The optimization of the stability and catalytic activity of such enzymes has potential applications in paper bleaching as well as in food and animal feed processing. Thus, thermostable xylanases such as xylanase 10A from the thermophilic fungus *Thermoascus aurantiacus* (TAX) are particularly well suited for industrial applications at elevated temperatures. Previously, TAX had only been obtained from the endogenous expression in *T. aurantiacus* found in samples of Indian soil.[1] Although natural expression gave sufficient quantities for crystallization studies, cultivation of fungal cells was necessary, requiring between 3 and 13 days of incubation to produce the protein.[2]

Our interest in TAX focused on its potential use as a scaffold for computational enzyme design. Three features of TAX make it attractive for these purposes: (1) The protein is thermostable, which generally indicates a robust scaffold more amenable to multiple mutations than its mesophilic homologs.[3] (2) TAX has a large natural binding pocket that could accommodate large substrates and multiple catalytic residues. (3) A high-resolution crystal structure is available at 1.7 Å with the natural ligand bound in the active site.[4]

The engineering of any protein requires an expression system amenable to genetic manipulation, thus the natural fungal system is not appropriate for the expression of designed TAX variants. Here, we present an efficient, recombinant *E. coli* expression

and purification protocol that yields up to 150 mg of active enzyme per liter of culture and is readily transferable to a high-throughput format.

Materials and Methods

Optimization of the TAX-His$_6$ gene

The protein sequence for TAX was taken from the FASTA sequence of the crystal structure of TAX (PDB:1GOR).[4] A hexahistidine purification tag was added to the C-terminus of the protein sequence preceded by a Factor Xa cleavage site resulting in the C-terminal sequence: GSIEGRGHHHHHH. The gene for TAX-His$_6$ was designed from the protein sequence using DNA 2.0 Gene Designer and optimized for expression in *E. coli*.[5] A unique *Nde*I restriction site was incorporated at the 5′ end of the gene, and a stop codon (TAG) along with a unique *Bam*HI restriction site were incorporated immediately following the 3′ end of the gene. The gene and protein sequences for TAX-His$_6$ are shown in Figure D-1.

Design of the oligonucleotides

Forty-eight overlapping oligonucleotides for the construction of the TAX-His$_6$ gene were designed using Assembly PCR Oligo Maker.[6] The overlapping oligonucleotides spanned the length of the gene and were less than 40 base pairs long with 16 to 20 base-pair overlapping regions (Table D-1). All oligonucleotides were synthesized by Integrated DNA Technologies at a 20 mmol scale with no additional purification.

Gene construction

The full-length TAX-His_6 gene was constructed by recursive PCR based on the method of Stemmer *et al*.[7] The protocol for gene construction is outlined in Chapter III, Materials and Methods. After construction, the gene was amplified using the primers TAX_forward 5′-(TAAGAAGGAGATATACATATGG)-3′ and TAX_reverse 5′-(AACTCAGCTTCCTTTCGGG)-3′, which were calculated to have melting temperatures of 63.9 and 67.6°C, respectively. The purified gene fragment was digested using *Bam*HI/*Nde*I (New England Biolabs) and was then ligated into a similarly digested pET11a plasmid (Novagen) yielding the TAX-His6-pET11a plasmid. The plasmid was then transformed into *E. coli* XL-1 Blue cells (Stratagene) and the gene sequence was confirmed by DNA sequencing.

Overexpression of TAX-His_6

TAX-His_6-pET11a was transformed into *E. coli* BL-21(DE3) cells for expression. A single colony was then used to inoculate 50 mL of LB containing 100 μg/mL ampicillin, which was then grown overnight at 37°C with shaking. 30 mL of the pre-culture was used to innoculate 1 L LB/ampicillin. The culture was grown at 37°C with shaking to an OD_{600} of approximately 0.3. The temperature was then reduced to 25°C and expression was induced at an OD_{600} of 0.6 with the addition of 1 mM isopropyl β-D-1-thiogalactopyranoside (IPTG). The culture was grown 18 hours and the cells were harvested by centrifugation.

Purification of TAX-His_6

The harvested cells were resuspended in lysis buffer (10 mM imidazole, 100 mM Tris pH 7.4, 300 mM NaCl), lysed mechanically with an Emulsiflex-C5 (Avestin), and then clarified by centrifugation at 10,000 × g for 40 min. The supernatant was then applied to 3 mL of Ni-NTA resin (Qiagen), which had been equilibrated with lysis buffer. The column was washed with 10 column volumes of lysis buffer and 10 column volumes of wash buffer (20 mM imidazole, 100 mM Tris pH 7.4, 300 mM NaCl). The protein was eluted with 3 column volumes of elution buffer (200 mM imidazole, 100 mM Tris pH 7.4, 300 mM NaCl). The eluate was dialyzed exhaustively against 50 mM sodium citrate, 50 mM NaCl, pH 5.5 and then concentrated using Amicon 10,000 MWCO centrifugal concentrators (Millipore).

Protein concentration determination

Protein concentrations were determined by UV absorbance after protein denaturation in 8 M guanidinium hydrochloride for 10 min with a dilution of at least 10×. An extinction coefficient at 280 nm of 55,280 $M^{-1}cm^{-1}$ was used.

Protein characterization

CD and MS were carried out as in Chapter III, Materials and Methods. CD spectra and thermal denaturation curves were obtained using 10 μM protein in 50 mM MES, pH 5.5, 100 mM NaCl.

Xylanase activity assays

The specific activity of TAX-His_6 was assayed using pNP-β-D-glucopyranoside as in Lo Leggio *et al*.[4] 2.5 mM substrate in 50 mM sodium citrate, pH 5.3 was incubated at 50°C for 5 min. The reaction was initiated with the addition of purified TAX-His_6 to a final volume of 300 μL. Protein concentrations between 170 μM and 1.7 μM were used. The reactions were mixed and incubated at 50°C for 15 min. The reaction was quenched with the addition of 400 uL of 20% Na_2CO_3 and the amount of liberated pNP was determined by measuring the absorbance at 410 nm. An extinction coefficient at 410 nm of 18,400 $M^{-1}cm^{-1}$ was used for pNP at pH 10.0.

Results and Discussion

We observed very high levels of TAX-His_6 expression in the BL-21 (DE3) *E. coli* cells (Figure D-2, lane 2) and were able to obtain 150 mg of highly pure protein per liter of culture after a single Ni-NTA affinity chromatography purification step (Figure D-2, lane 6). A majority of TAX-His_6 was expressed into the soluble fraction (Figure D-2, lane 3), but a significant amount of protein was found in the insoluble fraction, perhaps a result of the very high total amount of expressed protein. The mass was confirmed with electrospray mass spectrometry (Figure D-3), and is within 2 amu of the expected mass (34,433 kDa).

CD analysis shows that TAX-His_6 is folded and has a Tm of 75.2°C (Figure D-4). Unfortunately, no CD data is available for direct comparison with fungally expressed TAX. One source reports that TAX retains 40% of its activity at temperatures as high as

80°C for 1 hour. In the recombinant TAX-His_6, the protein is fully and irreversibly unfolded at this temperature. One subtle difference between TAX from fungal sources and from *E. coli* is the pyroglutamic acid modification of the N-terminal glutamate that is seen in the crystal structure of fungal TAX that is probably not present in the recombinant protein.[4] However, it is not clear that this small modification is enough to cause significant destabilization. The hexahistidine tag and Factor Xa cleavage site that were added to the protein sequence in TAX-His_6 may also contribute to the differences in stability, although we did not attempt to cleave the tag to determine if this was the cause of the destabilization. In addition, the strain of *T. aurantiacus* that was used to produce the xylanase in the crystal structure, which was the source of our protein sequence, is not explicitly stated. If the xylanase produced by the strain used in the temperature studies (C436) differed from the one that was used to produce the protein for the crystal structure, the protein sequences could be different, perhaps resulting in variations in thermal stability.

TAX-His_6 has a specific activity for pNP-β-D-glucopyranoside of 13.6 ± 3.9 U/mg. This is similar to the reported value from the natural expression system (12.7 U/mg), indicating that the enzyme is functionally unchanged when expressed in *E. coli*.[4]

Conclusions

We were able to express TAX-His_6 recombinantly, and because of our optimization of the gene for expression in *E. coli*, we obtained up to 150 mg of highly pure protein per liter of culture. The recombinantly expressed TAX-His_6 is folded, stable, and functionally indistinguishable from the protein obtained from the fungal

expression system. The high bacterial expression levels, ease of purification, and thermostability of TAX-His_6 make it an ideal scaffold for *de novo* computational enzyme design and other enzyme engineering studies. In addition, the method of constructing the TAX-His_6 gene through recursive PCR of overlapping 40-mer oligonucleotides facilitates future enzyme engineering projects on this scaffold. Multiple mutations can easily be introduced by simply substituting a subset of oligos that encode the mutations and carrying out the gene assembly and cloning steps to create the gene for the modified protein.

Acknowledgements

We would like to thank Leila Lo Leggio for discussion about the fungal TAX crystal structure and expression system.

References

1. Lo Leggio, L.; Kalogiannis, S.; Bhat, M. K.; Pickersgill, R. W., High resolution structure and sequence of T. aurantiacus xylanase I: implications for the evolution of thermostability in family 10 xylanases and enzymes with (β)α-barrel architecture. *Proteins* **1999**, *36*, 295-306.

2. Kalogiannis, S.; Owen, E.; Beever, D. E.; Bhat, M. K., Screening of ten strains of *Thermoascus aurantiacus* and characterization of a major xylanase. *Med. Fac. Landbouww. Univ. Gent.* **1995**, *60*, 1995-1998.

3. Besenmatter, W.; Kast, P.; Hilvert, D., Relative tolerance of mesostable and thermostable protein homologs to extensive mutation. *Proteins* **2007**, *66*, 500-506.

4. Lo Leggio, L.; Kalogiannis, S.; Eckert, K.; Teixeira, S. C.; Bhat, M. K.; Andrei, C.; Pickersgill, R. W.; Larsen, S., Substrate specificity and subsite mobility in *T. aurantiacus* xylanase 10A. *FEBS Lett* **2001**, *509*, 303-308.

5. Villalobos, A.; Ness, J.; Gustafsson, C.; Minshull, J.; Govindarajan, S., Gene Designer: a synthetic biology tool for constructing artificial DNA segments. *BMC Bioinformatics* **2006**, *7*, 285.

6. Rydzanicz, R.; Zhao, X. S.; Johnson, P. E., Assembly PCR oligo maker: a tool for designing oligodeoxynucleotides for constructing long DNA molecules for RNA production. *Nucleic Acids Res* **2005**, *33*, W521-525.

7. Stemmer, W. P.; Crameri, A.; Ha, K. D.; Brennan, T. M.; Heyneker, H. L., Single-step assembly of a gene and entire plasmid from large numbers of oligodeoxyribonucleotides. *Gene* **1995**, *164*, 49-53.

Table D-1. Assembly oligonucleotides for the construction of the TAX-His$_6$ gene. Overlapping regions are indicated with bold text and same-colored sections of adjacent oligonucleotides are complementary. All oligonucleotides are written from 5′ to 3′.

name	sequence	length
TAX-1t	TAAGAAGGAGATATACATATGGCAGAAGCGG	31
TAX-1b	TCAGTTGGTCAACGGATTGAGCCGCTTCTGCCATATGTAT	40
TAX-2t	CAATCCGTTGACCAACTGATTAAAGCTCGTGGTAAAGTGT	40
TAX-2b	GTCGGTTGCAACACCGAAATACACTTTACCACGAGCTTTA	40
TAX-3t	TTCGGTGTTGCAACCGACCAGAACCGCCTGACCACTGGC	39
TAX-3b	CTGAATGATCGCCGCGTTCTTGCCAGTGGTCAGGCGGT	38
TAX-4t	AACGCGGCGATCATTCAGGCAGATTTCGGTCAGGTTACCC	40
TAX-4b	CCCATTTCATGGAGTTCTCTGGGGTAACCTGACCGAAATC	40
TAX-5t	GAGAACTCCATGAAATGGGACGCGACCGAGCCTTCTCAA	39
TAX-5b	CGCCAGCGAAGTTGAAGTTGCCTTGAGAAGGCTCGGTCGC	40
TAX-6t	AACTTCAACTTCGCTGGCGCAGACTACCTGGTGAACTGG	39
TAX-6b	CAGCTTACCGTTCTGCTGCGCCCAGTTCACCAGGTAGTC	39
TAX-7t	CAGCAGAACGGTAAGCTGATCCGCGGTCATACGCTGGTT	39
TAX-7b	AGACGGCAGCTGGCTGTGCCAAACCAGCGTATGACCGCG	39
TAX-8t	ACAGCCAGCTGCCGTCTTGGGTGTCTTCCATCACCGATA	39
TAX-8b	TCACGTTGGTCAGAGTGTTTTTATCGGTGATGGAAGACAC	40
TAX-9t	AACACTCTGACCAACGTGATGAAGAACCATATCACTACCC	40
TAX-9b	TGCCTTTATAACGGGTCATCAGGGTAGTGATATGGTTCTTC	41
TAX-10t	GATGACCCGTTATAAAGGCAAAATCCGCGCGTGGGATGTT	40
TAX-10b	CCTCGTTGAATGCTTCATTCACAACATCCCACGCGCGGA	39
TAX-11t	GAATGAAGCATTCAACGAGGACGGCAGCCTGCGCCAGAC	39
TAX-11b	CACCGATCACATTCAGAAAAACGGTCTGGCGCAGGCTGC	39
TAX-12t	TTTTCTGAATGTGATCGGTGAAGATTACATCCCGATCGCAT	41
TAX-12b	GCAGCACGGGCGGTCTGGAATGCGATCGGGATGTAATC	38
TAX-13t	AGACCGCCCGTGCTGCAGATCCAAACGCTAAGCTGTACA	39
TAX-13b	AGTCCAGGTTGTAATCGTTAATGTACAGCTTAGCGTTTGG	40
TAX-14t	TAACGATTACAACCTGGACTCTGCGTCTTATCCGAAAACCC	41
TAX-14b	AACACGGTTCACGATGGCCTGGGTTTTCGGATAAGACGC	39
TAX-15t	GCCATCGTGAACCGTGTTAAACAGTGGCGTGCGGCTGG	38
TAX-15b	GCCGATGCCGTCGATCGGAACGCCAGCCGCACGCCACT	38
TAX-16t	CGATCGACGGCATCGGCTCCCAGACGCATCTGTCTGCA	38
TAX-16b	AGCACGCCAGCGCCCTGGCCTGCAGACAGATGCGTCTG	38
TAX-17t	AGGGCGCTGGCGTGCTGCAGGCCCTGCCGCTGCTGGC	37
TAX-17b	ACCTCCGGAGTGCCGGCGCTTGCCAGCAGCGGCAGGG	37
TAX-18t	CCGGCACTCCGGAGGTTGCAATCACCGAGCTGGATGTAG	39
TAX-18b	ATCAGTCGGGCTTGCGCCCGCTACATCCAGCTCGGTGAT	39
TAX-19t	GCGCAAGCCCGACTGATTATGTCAACGTCGTGAACGCG	38
TAX-19b	CACAAGACTGCACATTCAGGCACGCGTTCACGACGTTGAC	40
TAX-20t	CTGAATGTGCAGTCTTGTGTGGGCATTACCGTATGGGGT	39
TAX-20b	CCAAGAATCAGGATCGGCAACACCCCATACGGTAATGCC	39
TAX-21t	GCCGATCCTGATTCTTGGCGCGCATCCACTACCCCGCTG	39
TAX-21b	GGTTGAAATTACCGTCGAACAGCAGCGGGGTAGTGGATG	39
TAX-22t	GTTCGACGGTAATTTCAACCCGAAACCAGCTTACAACGCTA	41
TAX-22b	CTGTTGCAGGTCCTGAACGATAGCGTTGTAAGCTGGTTTC	40
TAX-23t	GTTCAGGACCTGCAACAGGGCAGCATCGAGGGTCGTGGT	39
TAX-23b	CCTAATGGTGGTGGTGATGGTGACCACGACCCTCGATGC	39
TAX-24t	CATCACCACCACCATTAGGGATCCGGCTGCTAACAAAGC	39
TAX-24b	AACTCAGCTTCCTTTCGGGCTTTGTTAGCAGCCGGA	36

T. aruantiacus xylanase (TAX-His$_6$)

```
    TAA GAA GGA GAT ATA CAT ATG GCA
                        NdeI    A

  1 GAA GCG GCT CAA TCC GTT GAC CAA CTG ATT AAA GCT CGT GGT AAA GTG TAT TTC GGT GTT
  1  E   A   A   Q   S   V   D   Q   L   I   K   A   R   G   K   V   Y   F   G   V

 61 GCA ACC GAC CAG AAC CGC CTG ACC ACT GGC AAG AAC GCG GCG ATC ATT CAG GCA GAT TTC
 21  A   T   D   Q   N   R   L   T   T   G   K   N   A   A   I   I   Q   A   D   F

121 GGT CAG GTT ACC CCA GAG AAC TCC ATG AAA TGG GAC GCG ACC GAG CCT TCT CAA GGC AAC
 41  G   Q   V   T   P   E   N   S   M   K   W   D   A   T   E   P   S   Q   G   N

181 TTC AAC TTC GCT GGC GCA GAC TAC CTG GTG AAC TGG GCG CAG CAG AAC GGT AAG CTG ATC
 61  F   N   F   A   G   A   D   Y   L   V   N   W   A   Q   Q   N   G   K   L   I

241 CGC GGT CAT ACG CTG GTT TGG CAC AGC CAG CTG CCG TCT TGG GTG TCT TCC ATC ACC GAT
 81  R   G   H   T   L   V   W   H   S   Q   L   P   S   W   V   S   S   I   T   D

301 AAA AAC ACT CTG ACC AAC GTG ATG AAG AAC CAT ATC ACT ACC CTG ATG ACC CGT TAT AAA
101  K   N   T   L   T   N   V   M   K   N   H   I   T   T   L   M   T   R   Y   K

361 GGC AAA ATC CGC GCG TGG GAT GTT GTG AAT GAA GCA TTC AAC GAG GAC GGC AGC CTG CGC
121  G   K   I   R   A   W   D   V   V   N   E   A   F   N   E   D   G   S   L   R

421 CAG ACC GTT TTT CTG AAT GTG ATC GGT GAA GAT TAC ATC CCG ATC GCA TTC CAG ACC GCC
141  Q   T   V   F   L   N   V   I   G   E   D   Y   I   P   I   A   F   Q   T   A

481 CGT GCT GCA GAT CCA AAC GCT AAG CTG TAC ATT AAC GAT TAC AAC CTG GAC TCT GCG TCT
161  R   A   A   D   P   N   A   K   L   Y   I   N   D   Y   N   L   D   S   A   S

541 TAT CCG AAA ACC CAG GCC ATC GTG AAC CGT GTT AAA CAG TGG CGT GCG GCT GGC GTT CCG
181  Y   P   K   T   Q   A   I   V   N   R   V   K   Q   W   R   A   A   G   V   P

601 ATC GAC GGC ATC GGC TCC CAG ACG CAT CTG TCT GCA GGC CAG GGC GCT GGC GTG CTG CAG
201  I   D   G   I   G   S   Q   T   H   L   S   A   G   Q   G   A   G   V   L   Q

661 GCC CTG CCG CTG CTG GCA AGC GCC GGC ACT CCG GAG GTT GCA ATC ACC GAG CTG GAT GTA
221  A   L   P   L   L   A   S   A   G   T   P   E   V   A   I   T   E   L   D   V

721 GCG GGC GCA AGC CCG ACT GAT TAT GTC AAC GTC GTG AAC GCG TGC CTG AAT GTG CAG TCT
241  A   G   A   S   P   T   D   Y   V   N   V   V   N   A   C   L   N   V   Q   S

781 TGT GTG GGC ATT ACC GTA TGG GGT GTT GCC GAT CCT GAT TCT TGG CGC GCA TCC ACT ACC
261  C   V   G   I   T   V   W   G   V   A   D   P   D   S   W   R   A   S   T   T

841 CCG CTG CTG TTC GAC GGT AAT TTC AAC CCG AAA CCA GCT TAC AAC GCT ATC GTT CAG GAC
281  P   L   L   F   D   G   N   F   N   P   K   P   A   Y   N   A   I   V   Q   D

901 CTG CAA CAG GGC AGC ATC GAG GGT CGT GGT CAC CAT CAC CAC CAC CAC CAT TAG
301  L   Q   Q   G   S   I   E   G   R   G   H   H   H   H   H   H

    GGA TCC GGC TGC TAA CAA AGC CCG AAA GGA AGC TGA GTT
     BamHI
```

Figure D-1. Protein and DNA sequences for TAX-His$_6$. The hexahistidine tag is shown in pink, the Factor Xa cleavage site is in red, and the restriction endonuclease cleavage sites are in green.

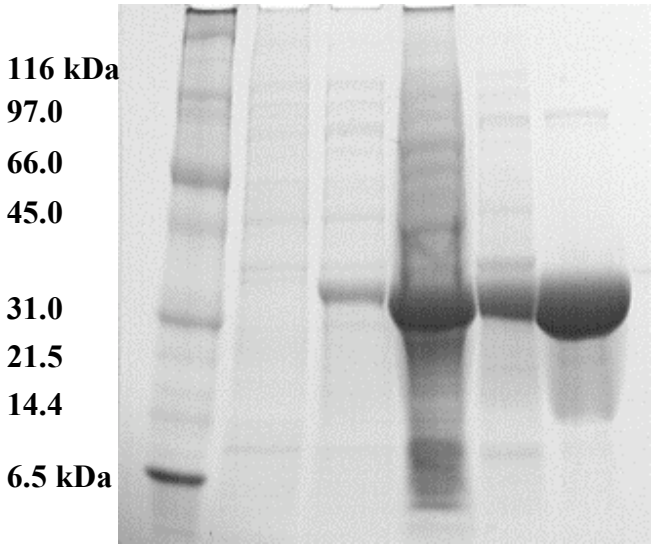


Figure D-2. SDS-PAGE analysis of TAX-His$_6$ expression and purification. Approximate molecular weights are indicated. Lane 1: Molecular weight marker. Lane 2: uninduced cells. Lane 3: cells after 18 hours of induction. Lane 4: lysate supernatant. Lane 5: lysate pellet. Lane 6: Ni-NTA affinity column eluate.

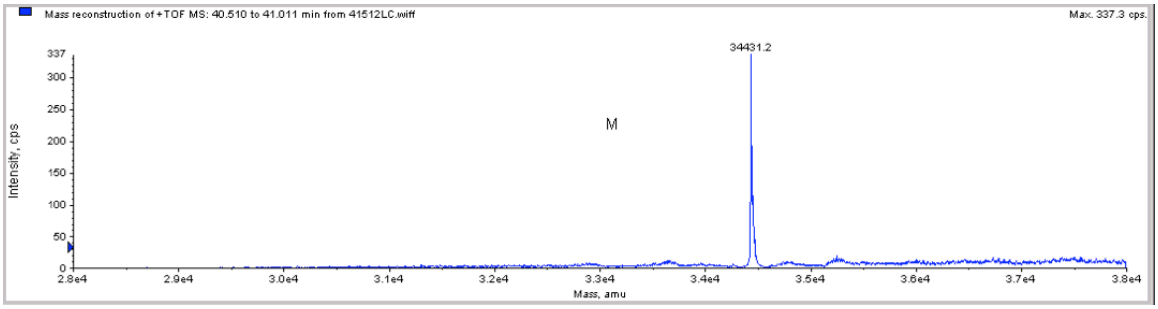


Figure D-3. Mass spectrometry analysis of TAX-His$_6$. The actual mass (34431.2) is within 2 amu of the expected mass (34433 amu).

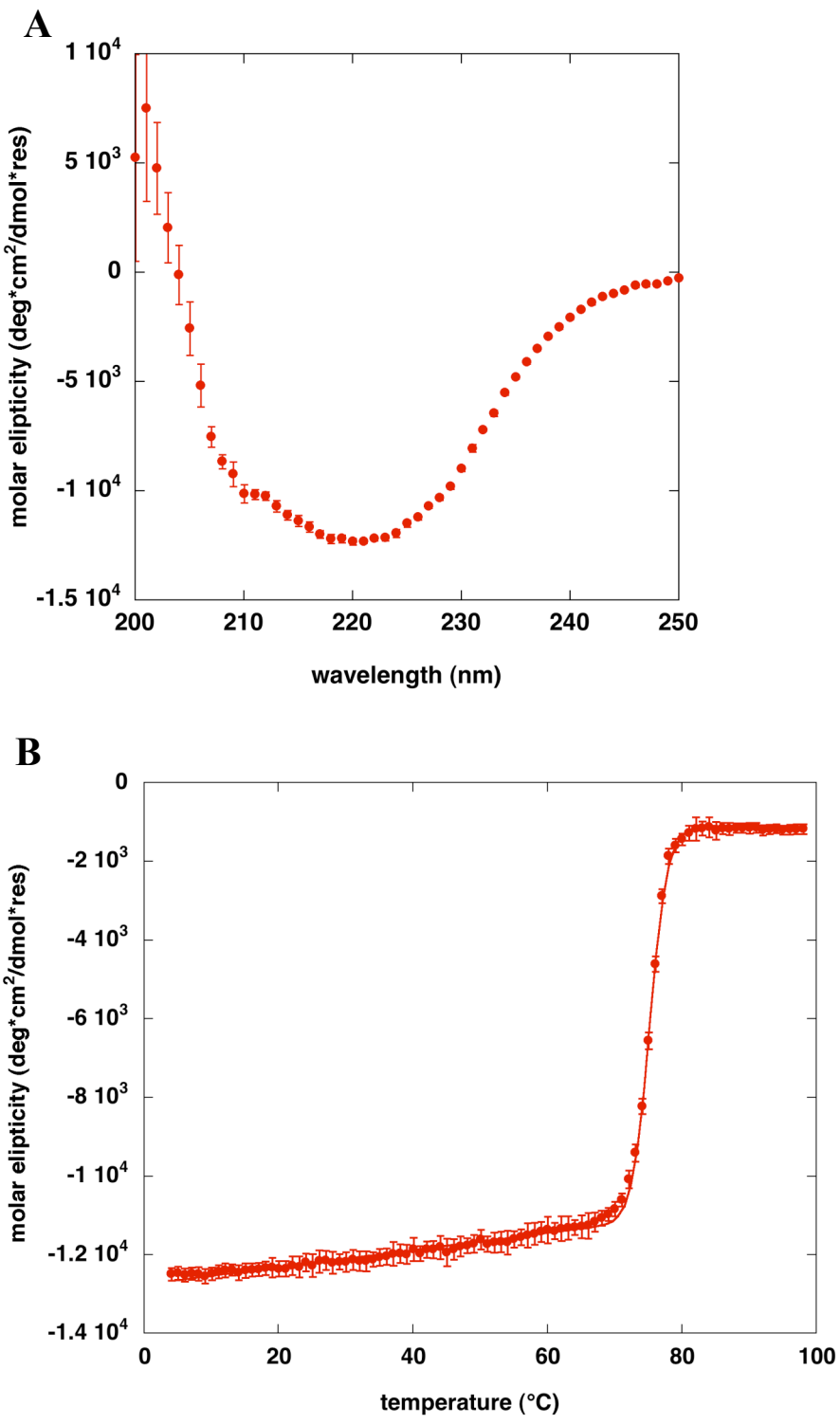


Figure D-4. CD analysis of TAX-His$_6$. **(A)** Far-UV wavelength scan, 25°C. **(B)** Thermal denaturation monitored at 222 nm. All experiments were carried out with 10 μM protein in 50 mM MES, pH 5.5, 50 mM NaCl.