Title
Understanding and Improving Designed Enzymes by Computer Simulations

Author
Bhowmick, Asmit

Publication Date
2016

Peer reviewed|Thesis/dissertation

Understanding and Improving Designed Enzymes by Computer Simulations

By
Asmit Bhowmick

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Chemical Engineering

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:
Professor Teresa Head-Gordon, Chair
Professor Berend Smit
Professor Susan Marqusee

Fall 2016

Abstract

Understanding and Improving Designed Enzymes by Computer Simulations

By

Asmit Bhowmick

Doctor of Philosophy in Chemical Engineering

University of California, Berkeley

Professor Teresa Head-Gordon, Chair

The ability to control for protein structure, electrostatics and dynamical motions is a fundamental problem that limits our ability to rationally design catalysts for new chemical reactions not known to have a natural biocatalyst. Current computational approaches for *de novo* enzyme design seek to engineer a small catalytic construct into an accommodating protein scaffold as exemplified by the Rosetta strategy. Here we consider 3 designed enzymes for the Kemp elimination reaction (KE07, KE70 and KE15) that showed minimal catalytic activity. KE07 and KE70 were subsequently improved by 2 orders of magnitude in catalytic efficiency by directed evolution and highlighted the shortcomings of the design process. This work studies two keys issues plaguing the designs – side chain conformational variability and electrostatics.

For the first part, a new Monte Carlo sampling method was developed that uses a physical forcefield and coupled with backbone variability and a backbone dependent rotamer library. Using transition state theory with energies/entropies calculated from Monte Carlo simulations, it is shown that in both KE07 and KE70, the initial design was over-optimized to stabilize the enzyme-substrate complex. Mutations introduced by directed evolutions led to destabilization of the enzyme-substrate complex and stabilization of the transition state. Furthermore, analysis of residue correlations via mutual information yielded hotspots, several of which were mutations during directed evolution. Laboratory mutations of these hotspots in the best variant of KE07 led to a drop in catalytic performance, demonstrating their importance. The metrics identified in KE07/KE70 studies were used to predict mutations to improve enzyme KE15 that had not been improved prior to this study. Several mutants, all predicted through computer simulations have now yielded better catalytic activity in the laboratory with the best one 10-fold better than the starting enzyme.

In order to quantify the role of electrostatics, a new method was developed using the AMOEBA polarizable forcefield that allowed splitting the contribution of electric field at the substrate by residues and solvent. The improvement in KE07 series could be tracked directly through changes in electric field at the substrate. In comparison, KE70 did not show a significant shift in electrostatic field, suggesting other factors like substrate binding may have been the reason for enhancement of activity. However, the common theme in both enzymes was the lack of participation (and in fact detrimental role) of the scaffold in the reaction. Future design efforts would benefit from an expanded theozyme and careful selection of scaffold based on electrostatic properties.

Generating efficient biocatalysts without using laboratory directed evolution would be an inflection point in the field of enzyme design. This work is a step in that direction.

Acknowledgements

There are many people I would like to thank and perhaps too many to mention here.

Firstly, I would like to thank my parents, Dr. Kundakali Bhowmick and Prof. Anil K. Bhowmick for their constant encouragement, support and reminding me of the other things in life beyond research.

A special thanks to my research lab mates for the hours spent together in lab as well as outside. A shout out to Saurabh Belsare, Alex Albaugh, Eugene Yedvabny, Sudhir C. Sharma, Omar Demerdash and Marielle Soniat.

I cannot thank Sukanya Sasmal enough. My lab mate and very close friend from my undergraduate years in IIT, life in graduate school was so much enjoyable because of you.

Thanks to Michael Mills, Carl Schreck and Jason Forster for tennis and dinner sessions, Nirupam Chakraborty for tennis, entire 260 squash peeps for the weekly matches and Maroof Adil for soccer.

Big thanks to the 2011 incoming Chemical Engineering graduate class at Cal for all the good times – I couldn't have asked for better classmates.

Studying at Berkeley, I have had the opportunity to interact with many brilliant professors through coursework and seminars - I would like to take this opportunity to thank them for their great work and impact on my academic and personal life.

From my undergraduate years, I would like to thank Prof. Sirshendu De at IIT Kharagpur for his inspiring teaching and research, Prof. Ramanan Krishnamoorti at University of Houston and Prof. Charles Cooney at MIT for their support.

Life in 2016 is unthinkable without the Internet. This thesis would not have been possible without online resources on Stack Overflow, Wolfram Alpha and Google/Google Scholar to name a few.

Chapter 1

Introduction

Recent developments in the field of computational chemistry have led to rationally engineered enzymes that can catalyze reactions for which no natural enzymes exist. In this chapter, I review the field of enzyme design, with a focus on a particular class of enzymes called Kemp eliminases that have been extensively studied in this thesis. I also set up the case for considering side chain conformational variability and electrostatics in these enzymes to yield better designs

1.1 INTRODUCTION

Enzymes are biocatalysts[1] that are capable of accelerating reactions up to 10^{20} fold under ambient conditions[2,3]– a remarkable feat considering most industrial catalysts function at elevated temperature and pressure[4, 5]. Human beings have been using enzymes since time immemorial for various daily activities like making bread, alcohol and cheese. Within our body there are thousands of different types enzymes that are working every second to keep our system functioning. In modern industries, enzymes have found additional use in food processing, starch and paper as well as biofuel production. Two attractive features of enzymes are – (a) they increase the reaction rates by decreasing the activation free barrier of the reaction and (b) they are highly specific. With an increased push towards energy efficient and sustainable technology[6], enzymes are an attractive alternative to energy-guzzling industrial catalysts. However, too few (or no) enzymes exist that can accelerate current industry favorite reactions. This situation can be remedied if we can design enzymes for those reactions. However before going into the current status of enzyme design, it is educational to know more about the various steps that have catalyzed the various stages of enzyme design.

The 1st and arguably most important aspect of designing enzymes is to understand how they work so well. Explaining how enzymes achieve such efficiency has been a topic of scientific research for the last 100 years or more. Well before the identity of enzymes was established, Emil Fischer proposed his seminal 'Lock and Key' model for enzymes in 1894 to explain why enzymes are so specific. That enzymes were proteins was only confirmed 32 years later by James Sumner. The next big development was in 1946 when Linus Pauling stated that enzymes were complementary in structure to the activated complex for the reaction catalyzed, thus leading to the observed speedup. This was the first attempt in explaining the tremendous kinetic rates seen for enzyme catalyzed reactions, almost a decade before the first protein crystal structure was published.

That happened in 1958 when the crystal structure of Myoglobin was reported[7] by Kendrew et. al. leading to the development of the field of structural biology. The dawn of the age of structural biology was an inflection point in studying enzymes. The availability of 3D enzyme structures aided in experimental and theoretical studies of these nanomachines. Pioneering work by Warshel, Jencks, Fersht and others laid the groundwork[8 9-15] for most of the currently accepted explanations in the field. These include but not limited to transition-state stabilization (by electrostatics), ground-state destabilization (by strain, entropy, desolvation), dynamical motions, covalent bonding etc. Although there is no consensus yet on which phenomena is more important

than others, there is a growing appreciation that enzymes use many of these physics together to maximize speedup.

In the 1980s, a parallel line of research started that attempted to engineer proteins to catalyze specific reactions[16]. Most of the research centered around very simple yet clever design ideas that used very little of the various catalytic proposals that were floated around that time to explain enzyme proficiency. The most popular one was developing catalytic antibodies[17, 18]. This was a direct application of Pauling's idea that enzymes stabilize the activated complex's structure and charge distribution. By cleverly using a hapten that is similar to the transition state of the reaction of interest, one can trick an antigen to produce catalytic antibodies that are capable of catalyzing the reaction. This method was used to catalyze various reactions including Diers Alders, acyl transfer, kemp elimination to name a few.

Another parallel field that took off in the early 90s was that of directed evolution[19-23]. The idea was to mimic natural evolution by introducing a bunch of mutations in the form of a library, usually randomly and selecting mutants that show the highest fitness[24]. The best mutants are mutated again and the process is repeated. By repeating this iteratively, one can dramatically alter a protein's fitness. Here fitness could mean catalytic ability, thermostability or selectivity to name a few. Popular strategies for constructing such mutant libraries are error-prone PCR, random shuffling, site-directed mutagenesis etc. Usually directed evolution requires very little knowledge of the structure and innards of an enzyme, making it a popular option to engineer enzymes. As we will discuss at length later, directed evolution has been crucial in improving activity of designed enzymes.

The 4th and final piece of the puzzle is the area of protein-folding prediction and using it in designing functional proteins[25-27]. Although the field is still evolving, current approaches ranging from a more knowledge-based approach like Rosetta[28] to a more physics-based approach like Folding@Home[29] are capable of folding a reasonable sized protein (less than 100 residues) starting from an extended peptide. Combining protein folding with enzyme design roughly corresponds to the following 3 steps -

(a) Identify a theoretical active site (theozyme) that can catalyze the reaction
(b) Identify protein scaffolds that can 'house' these theozymes.
(c) Make adjustments to the scaffolds to accommodate the active site and complement geometric and electronic properties.

The implementation of the above protocol has been done differently by different groups. The reader is referred to Ref [[30]] for a more comprehensive review of the different stages listed. The two most popular rational enzyme design techniques are RosettaMatch[28] and SABER[31]. The difference between these two protocols is in step 2 where one (RosettaMatch) tries to identify folds where the theozyme can be grafted into the fold while the other (SABER) tries to identify sequences that already have the functionality of the theozyme. Regardless of the differences, both protocols have had success in designing enzymes with minimal competence, providing room for optimism. Representative successes for this field include enzymes for Diels Alders[32], Kemp elimination[33], Retro-Aldol reaction[34] etc.

This is the stage where advances in directed evolution and catalytic antibodies play a pivotal role. The norm in the field currently is to use directed evolution to further enhance kinetic performance of the minimally competent enzymes[35-39]. The minimum desired activity is that of catalytic antibodies that surprisingly still outperform many improved variants for designed

enzymes[40]. Admittedly, the field is still in a nascent stage and most designs have almost certainly not optimized the various physics natural enzymes have mastered in using. It would thus be useful to reassess the various features the designs and improved variants possess to see how they were improved, using concepts developed over the last 50 years in understanding enzyme proficiency. The end goal is to incorporate these features into the design process, thus minimizing or completely doing away with laboratory directed evolution.

In the following section, I will introduce the specific system of interest in this thesis. A good model reaction that the community has worked on to develop designed enzymes is the Kemp elimination reaction. This base-catalyzed reaction proceeds in a single step and has been studied extensively both by simulations and experiments. The substrate 5-nitro benzisoxazole reacts with a base to form cyanophenol[41].

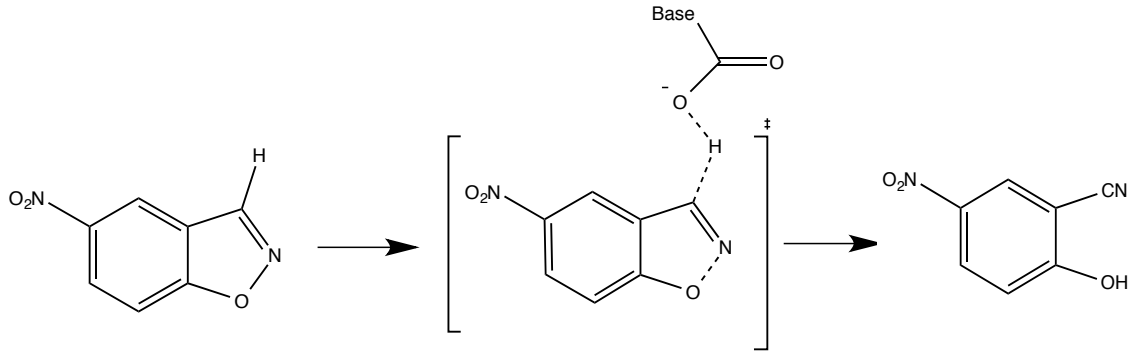


Figure 1.1: *The Kemp elimination reaction*. The one-step reaction scheme involving the abstraction of hydrogen from 5-nitro benzisoxazole by a catalytic base, leading to breaking of the C-H bond. Shown is the transition state that has a partial negative charge on the substrate oxygen with cleavage of the O-N bond and nascent formation of a C-N triple bond.

The attractive feature from an experimental point of view is the ease in studying the kinetics in a lab setting due to the unique absorption spectra of the substrate. From a simulation point of view, the simplicity of the reaction and complementary experimental work has led to it being studied widely. A variety of 'catalysts' have been shown to speed up this reaction. This includes catalytic antibodies (notably 34E4) [40], serum albumins, micelles and even charcoals[42]. Among designed enzymes, enzymes KE07, KE70 and KE59 developed using RosettaMatch and the HG series (by SABER) have received a lot of attention. The best rationally designed Kemp enzyme to date is HG-3 with a k_{cat}/K_M of 460 $M^{-1}s^{-1}$.

In this thesis I will be studying the following three designs - KE07, KE70 and KE15. KE stands for Kemp Eliminase. All the three enzymes were designed and their active site are illustrated in Figure 1.2. These enzymes were built using the Rosetta software and further information about the design process can be found in these references [30, 33]. When the enzymes were expressed and tested for kinetics, they performed very poorly. Table 1.1 tabulates the kinetic constants for these three specific enzymes.

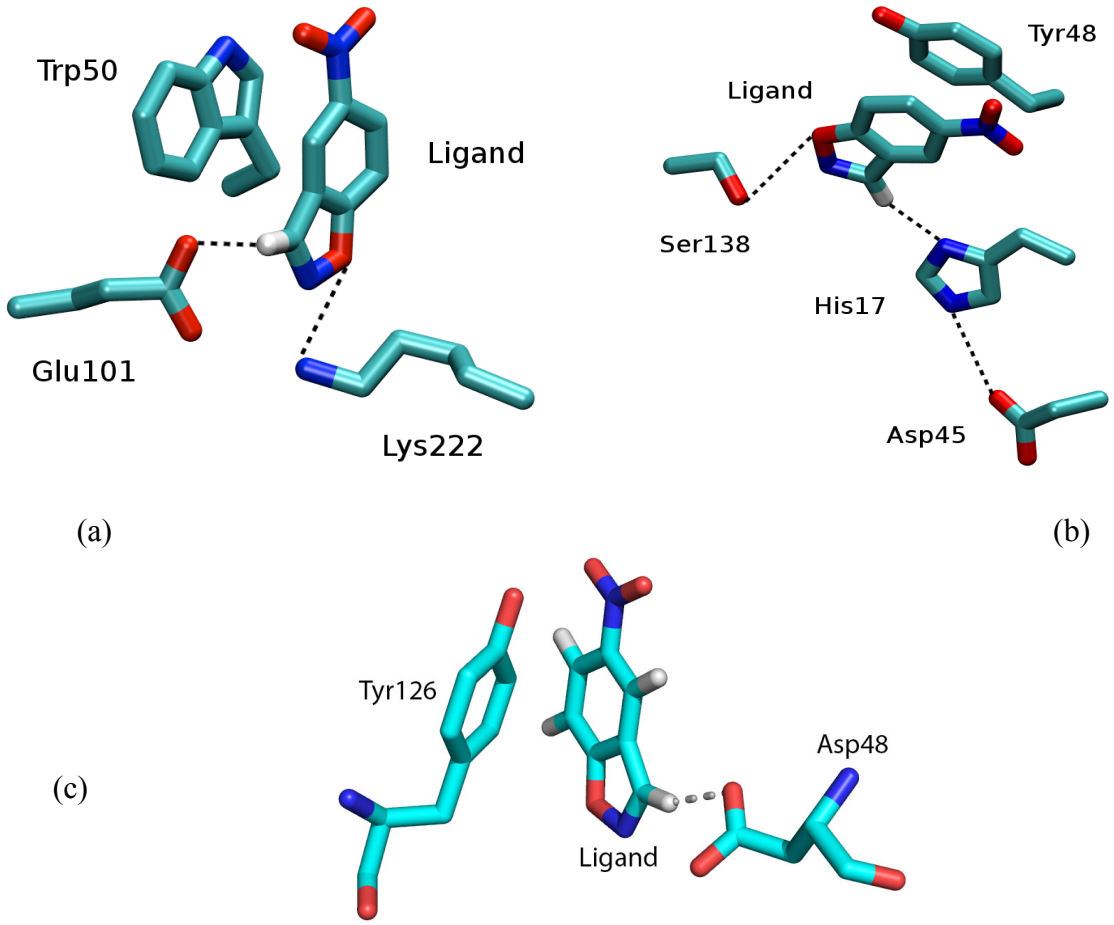


Figure 1.2: *The Kemp Eliminases*. (a) KE07 design, where the base is Glu-101. Additional stabilization is provided by Trp-50 (π-stacking) and Lys-222 (charge stabilizing) (b) KE70 design where the base is a His-Asp dyad. Ser-138 does the charge stabilization and Tyr-48 does π-stacking (c) KE15 design where is the base is Asp-48. No charge stabilization is present in this design.

Table 1.1: Kinetic constants k_{cat}/K_M for the 3 enzymes KE07, KE70 and KE15 for their designed state as well as the best variant after directed evolution. KE15 did not undergo any directed evolution. All values are in $M^{-1}s^{-1}$

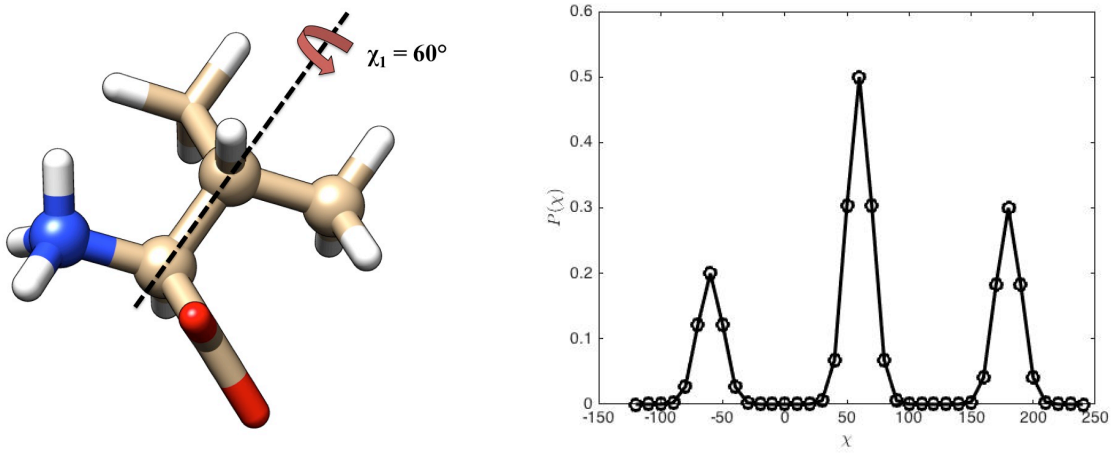
Enzyme	Design	Best Variant[2,3]
KE07	12	2600
KE70	126	57300
KE15	27	NA

The observation of a baseline activity validated the Rosetta methodology and was celebrated as an initial success. Unfortunately, most native enzymes operate in the diffusive limit i.e. their k_{cat}/K_M are of the order 10^6-10^8 $M^{-1}s^{-1}$, leaving a big room for improvement.

This work studies the effect of two important physical phenomena in designed enzymes – (a) side chain conformation variability and (b) electrostatic stabilization. These factors are usually unaccounted for during the design process despite substantial evidence to suggest that these two features are highly optimized by natural enzymes. Although experiments can give suggestive indications of these missing features, quantifying them through computer models will enable correcting this systematic error of the design protocols.

Side-Chain Conformational Motion

Proteins (and enzymes) are essentially polymers with the option of having up to 20 different types of monomers. Each monomer, called an amino acid (or residue, as will be referred to in this work) has a unique topology and chemical character that lends a characteristic feature to the protein. These characteristics are exhibited in the side chains of the amino acids. Beyond the chemistry of the side chains, their conformations can also be quite diverse. Extensive protein crystallography data from the last 50 years clearly highlight the extent to which these amino acids can switch conformations depending on the environment they are in. Often, their conformation motion is characterized by their dihedral angles. A representative dihedral angle for amino acid valine and its distribution found in crystallographic libraries[43, 44] is shown in figure 1.3.



(a) (b)

Figure 1.3: *Dihedral angles for valine*. (a) The side chain dihedral angle for valine (χ_1) defined by the 4 atoms shown in spheres (N-C_α-C_β-C_γ). (b) Side chain dihedral angles tend to cluster around a few well-separated values as illustrated here. For valine, these values are -60, 60 and 180°. The probabilities shown were obtained from an X-ray crystallographic database.

Clearly, the side chains dihedrals have a propensity to cluster around certain values that can be thought of as energy wells. The conformations can alternate between energy wells in solution[45] and have been shown to affect enzyme catalysis, binding and other important biological events.

Evidence from X-ray crystallography[46, 47] and NMR experiments[13, 48-50]that report on this kind of variability show convincing evidence of such motion that can potentially range from picoseconds to milliseconds (ps-ms). In Chapter 2, I will talk a bit more about these studies and how a new simulation method I developed was tested against the these experimental results.
The design process implemented in suites like Rosetta typically does not consider side-chain conformational motion in great detail. Crystal structures of the Kemp eliminases show virtually no backbone variability, hinting at a role of side-chain motion in facilitating catalysis. There is a decent body of work that has shown how side-chain motions can propagate over long distances to influence binding events, a phenomenon referred to as allostery[51-54]. Thus, although the design protocol is geared towards bringing the substrate in close proximity to a base in a confined environment, many residues can indirectly tune the efficiency of the enzyme, in many cases leading to complete loss of activity. In Chapter 3, I will discuss how we elucidated the role of side chain variability in improving the Kemp eliminases using transition-state theory.

Electrostatic Stabilization in Enzymes

Chapter 5 of this thesis considers the case of electrostatic stabilization provided to the transition state by designed Kemp eliminases. Recent experimental work using vibrational Stark effect spectroscopy (VSE) has shown that electrostatic stabilization is responsible for up to 10^5-fold improvement[55] in catalytic performance in ketosteroid isomerase enzyme. A further 10^3 fold is believed to come from chemical positioning of the catalytic residue Asp-40. These conclusions are in line with simulation studies by Warshel and others going back to the 1970s[11]. At that time, it was not obvious how enzymes can provide a better electrostatic environment compared to water. This can be understood by considering the solvation free energy of the transition state, ΔG_{sol}, which is a sum of 2 terms[56], a charge-dipole interaction ($\Delta G_{Q\mu}$) and a dipole-dipole interaction ($\Delta G_{\mu\mu}$) as seen in (1)

$$\Delta G_{sol} = \Delta G_{Q\mu} + \Delta G_{\mu\mu} \quad (1)$$

(1) is a linear response approximation and shows where enzymes win in terms of stabilizing a transition state. The 1st term is similar between solvent and enzymes. However, the 2nd term is costly for solvent as it has to undergo a large rearrangement to orient the dipoles in a favorable manner for the transition state. This is the so-called reorganization energy for stabilizing a transition state. A simple example is for water that has to break it's complicated network of interaction among itself to stabilize the substrate (Fig 1.4). In enzymes, these interactions are already anticipated, requiring minimal rearrangement of protein dipoles. It is believed that the protein pays this cost of orienting the dipoles appropriately for reaction during the folding process. Readers interested in more details about this mechanism should consult these references [56, 57]

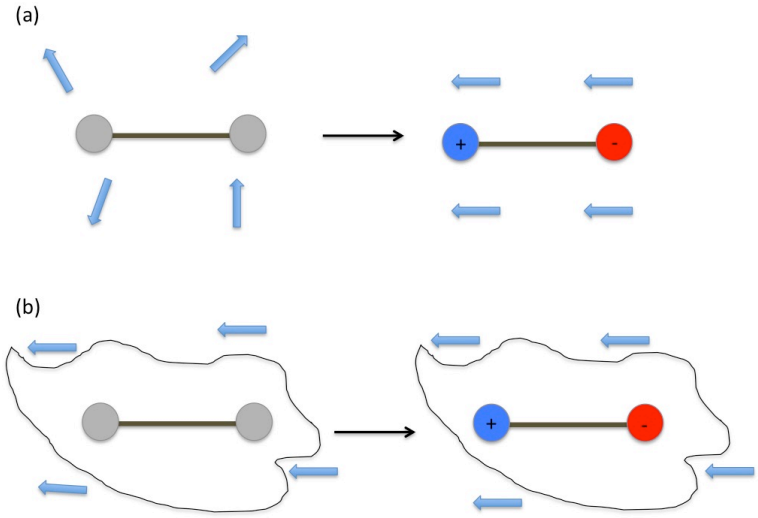


Figure 1.4: *Electrostatic stabilization in enzymes*. (a) In solvent, there is extensive rearrangement of dipoles adding to the cost of stabilizing transition state. (b) Enzymes on the other hand have a highly preorganized dipolar environment, much of which is included into the folding energy, giving a significant electrostatic advantage.

Given such overwhelming evidence for natural enzymes utilizing electrostatics to preferentially stabilize the transition state, it would be interesting to see if the designed enzymes have harnessed any electrostatic features. Since most designs activities are woeful, it is reasonable to suspect poor electrostatics in the active site. In chapter 5 we study these features in more details and lay out some suggestions that can be incorporated in future design protocols.

Improving enzymes rationally

The culmination of such detailed studies into the Kemp eliminase would be to implement some of the features known to be lacking in the original design but engineered in by laboratory directed evolution. Chapter 4 reports this effort by considering the enzyme KE15 that was also designed by the Rosetta protocol but did not undergo any subsequent directed evolution. The initial activity was 27 $M^{-1}s^{-1}$, similar to other designs. By considering metrics like mutual information and electrostatic stabilization, the goal would be to improve the enzyme by a significant amount.

1.2 REFERENCES

1. Fersht, A., *Enzyme Structure and Mechanism*. WH Freeman: New York, New York, 1985.
2. Radzicka, A.; Wolfenden, R., A Proficient Enzyme. *Science* **1995,** *267* (5194), 90.

3. Wolfenden, R.; Snider, M. J., The Depth of Chemical Time and the Power of Enzymes as Catalysts. *Accounts of Chemical Research* **2001,** *34* (12), 938-945.
4. Zamost, B. L.; Nielsen, H. K.; Starnes, R. L., Thermostable Enzymes for Industrial Applications. *Journal of Industrial Microbiology* **1991,** *8* (2), 71-81.
5. Cheetham, P., Principles of Industrial Enzymology: Basis of Utilization of Soluble and Immobilized Enzymes in Industrial Processes. *Handbook of enzyme biotechnology/editor, Alan Wiseman* **1985**.
6. Chu, S.; Majumdar, A., Opportunities and Challenges for a Sustainable Energy Future. *nature* **2012,** *488* (7411), 294-303.
7. Kendrew, J.; Dickerson, R.; Strandberg, B.; Hart, R.; Davies, D.; Phillips, D.; Shore, V., Structure of Myoglobin: A Three-Dimensional Fourier Synthesis at 2 A. Resolution. *Nature* **1960,** *185* (4711), 422-427.
8. Jencks, W. P., *Catalysis in Chemistry and Enzymology*. Courier Corporation: 1987.
9. Kirby, A. J., Enzyme Mechanisms, Models, and Mimics. *Angewandte Chemie International Edition in English* **1996,** *35* (7), 706-724.
10. Warshel, A.; Naray-Szabo, G.; Sussman, F.; Hwang, J., How Do Serine Proteases Really Work? *Biochemistry* **1989,** *28* (9), 3629-3637.
11. Warshel, A., Energetics of Enzyme Catalysis. *Proceedings of the National Academy of Sciences* **1978,** *75* (11), 5250-5254.
12. Russell, A. J.; Fersht, A. R., Rational Modification of Enzyme Catalysis by Engineering Surface Charge. *Nature* **1986,** *328* (6130), 496-500.
13. Henzler-Wildman, K.; Kern, D., Dynamic Personalities of Proteins. *Nature* **2007,** *450*, 964-72.
14. Henzler-Wildman, K. a.; Lei, M.; Thai, V.; Kerns, S. J.; Karplus, M.; Kern, D., A Hierarchy of Timescales in Protein Dynamics Is Linked to Enzyme Catalysis. *Nature* **2007,** *450*, 913-6.
15. Benkovic, S. J.; Hammes-Schiffer, S., A Perspective on Enzyme Catalysis. *Science* **2003,** *301* (5637), 1196-1202.
16. O'Neil, K. T.; DeGrado, W. F., How Calmodulin Binds Its Targets: Sequence Independent Recognition of Amphiphilic A-Helices. *Trends in biochemical sciences* **1990,** *15* (2), 59-64.
17. Pollack, S. J.; Jacobs, J. W.; Schultz, P. G., Selective Chemical Catalysis by an Antibody. *Science* **1986,** *234* (4783), 1570-1573.
18. Tramontano, A.; Janda, K.; Napper, A.; Benkovic, S.; Lerner, R. In *Catalytic Antibodies*, Cold Spring Harbor symposia on quantitative biology, Cold Spring Harbor Laboratory Press: 1987; pp 91-96.
19. Crameri, A.; Raillard, S.-A.; Bermudez, E.; Stemmer, W. P., DNA Shuffling of a Family of Genes from Diverse Species Accelerates Directed Evolution. *Nature* **1998,** *391* (6664), 288-291.
20. Moore, J. C.; Arnold, F. H., Directed Evolution of a Para-Nitrobenzyl Esterase for Aqueous-Organic Solvents. *Nature biotechnology* **1996,** *14* (4), 458-467.
21. Chen, K.; Arnold, F. H., Tuning the Activity of an Enzyme for Unusual Environments: Sequential Random Mutagenesis of Subtilisin E for Catalysis in Dimethylformamide. *Proceedings of the National Academy of Sciences* **1993,** *90* (12), 5618-5622.
22. Roberts, B. L.; Markland, W.; Ley, A. C.; Kent, R. B.; White, D. W.; Guterman, S. K.; Ladner, R. C., Directed Evolution of a Protein: Selection of Potent Neutrophil Elastase

Inhibitors Displayed on M13 Fusion Phage. *Proceedings of the National Academy of Sciences* **1992,** *89* (6), 2429-2433.
23. Giver, L.; Gershenson, A.; Freskgard, P.-O.; Arnold, F. H., Directed Evolution of a Thermostable Esterase. *Proceedings of the National Academy of Sciences* **1998,** *95* (22), 12809-12813.
24. Brustad, E. M.; Arnold, F. H., Optimizing Non-Natural Protein Function with Directed Evolution. *Current opinion in chemical biology* **2011,** *15* (2), 201-210.
25. Street, A. G.; Mayo, S. L., Computational Protein Design. *Structure* **1999,** *7* (5), R105-R109.
26. 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* (5649), 1364-1368.
27. Lippow, S. M.; Tidor, B., Progress in Computational Protein Design. *Current opinion in biotechnology* **2007,** *18* (4), 305-311.
28. Rohl, C. A.; Strauss, C. E.; Misura, K. M.; Baker, D., Protein Structure Prediction Using Rosetta. *Methods in enzymology* **2004,** *383*, 66-93.
29. Shirts, M.; Pande, V. S., Screen Savers of the World Unite! *Science* **2000,** *290* (5498), 1903-1904.
30. Kiss, G.; Çelebi‐Ölçüm, N.; Moretti, R.; Baker, D.; Houk, K., Computational Enzyme Design. *Angewandte Chemie International Edition* **2013,** *52* (22), 5700-5725.
31. Nosrati, G. R.; Houk, K., Saber: A Computational Method for Identifying Active Sites for New Reactions. *Protein Science* **2012,** *21* (5), 697-706.
32. Siegel, J. B.; Zanghellini, A.; Lovick, H. M.; Kiss, G.; Lambert, A. R.; Clair, J. L. S.; Gallaher, J. L.; Hilvert, D.; Gelb, M. H.; Stoddard, B. L., Computational Design of an Enzyme Catalyst for a Stereoselective Bimolecular Diels-Alder Reaction. *Science* **2010,** *329* (5989), 309-313.
33. Röthlisberger, D.; Khersonsky, O.; Wollacott, A. M.; Jiang, L.; DeChancie, J.; Betker, J.; Gallaher, J. L.; Althoff, E. a.; Zanghellini, A.; Dym, O.; Albeck, S.; Houk, K. N.; Tawfik, D. S.; Baker, D., Kemp Elimination Catalysts by Computational Enzyme Design. *Nature* **2008,** *453*, 190-5.
34. 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*, 1387-91.
35. Giger, L.; Caner, S.; Obexer, R.; Kast, P.; Baker, D.; Ban, N.; Hilvert, D., Evolution of a Designed Retro-Aldolase Leads to Complete Active Site Remodeling. *Nat Chem Biol* **2013,** *9* (8), 494-498.
36. Khersonsky, O.; Kiss, G.; Röthlisberger, D.; Dym, O.; Albeck, S.; Houk, K. N.; Baker, D.; Tawfik, D. S., Bridging the Gaps in Design Methodologies by Evolutionary Optimization of the Stability and Proficiency of Designed Kemp Eliminase Ke59. *Proc Natl Acad Sci U S A* **2012,** *109*, 10358-63.
37. Khersonsky, O.; Röthlisberger, D.; Dym, O.; Albeck, S.; Jackson, C. J.; Baker, D.; Tawfik, D. S., Evolutionary Optimization of Computationally Designed Enzymes: Kemp Eliminases of the Ke07 Series. *J. Mol. Bio.* **2010,** *396*, 1025-42.

38. Khersonsky, O.; Röthlisberger, D.; Wollacott, A. M.; Murphy, P.; Dym, O.; Albeck, S.; Kiss, G.; Houk, K. N.; Baker, D.; Tawfik, D. S., Optimization of the in-Silico-Designed Kemp Eliminase Ke70 by Computational Design and Directed Evolution. *J. Mol. Bio.* **2011,** *407*, 391-412.
39. Blomberg, R.; Kries, H.; Pinkas, D. M.; Mittl, P. R. E.; Grutter, M. G.; Privett, H. K.; Mayo, S. L.; Hilvert, D., Precision Is Essential for Efficient Catalysis in an Evolved Kemp Eliminase. *Nature* **2013,** *503* (7476), 418-421.
40. Thorn, S.; Daniels, R.; Auditor, M.; Hilvert, D., Large Rate Accelerations in Antibody Catalysis by Strategic Use of Haptenic Charge. *Nature* **1995**, 228.
41. Casey, M.; Kemp, D., Physical Organic Chemistry of Benzisoxazoles. I. Mechanism of the Base-Catalyzed Decomposition of Benzisoxazoles. *J. Org. Chem.* **1973,** *58*, 33-34.
42. Seebeck, F. P.; Hilvert, D., Positional Ordering of Reacting Groups Contributes Significantly to the Efficiency of Proton Transfer at an Antibody Active Site. *Journal of the American Chemical Society* **2005,** *127* (4), 1307-1312.
43. Jr, R. L. D., Rotamer Libraries in the 21 St Century. **2002**, 431-440.
44. Lovell, S. C.; Word, J. M.; Richardson, J. S.; Richardson, D. C., The Penultimate Rotamer Library. *Proteins: Structure, Function, and Bioinformatics* **2000,** *40* (3), 389-408.
45. Scouras, A. D.; Daggett, V., The Dynameomics Rotamer Library: Amino Acid Side Chain Conformations and Dynamics from Comprehensive Molecular Dynamics Simulations in Water. *Protein Science* **2011,** *20* (2), 341-352.
46. Shapovalov, M. V.; Dunbrack, R. L., A Smoothed Backbone-Dependent Rotamer Library for Proteins Derived from Adaptive Kernel Density Estimates and Regressions. *Structure* **2011,** *19*, 844-58.
47. Bower, M. J.; Cohen, F. E.; Dunbrack, R. L., Prediction of Protein Side-Chain Rotamers from a Backbone-Dependent Rotamer Library: A New Homology Modeling Tool. *Journal of molecular biology* **1997,** *267*, 1268-82.
48. Tuttle, L. M.; Dyson, H. J.; Wright, P. E., Side-Chain Conformational Heterogeneity of Intermediates in the Escherichia Coli Dihydrofolate Reductase Catalytic Cycle. *Biochemistry* **2013,** *52*, 3464-77.
49. Boehr, D. D.; Nussinov, R.; Wright, P. E., The Role of Dynamic Conformational Ensembles in Biomolecular Recognition. *Nature Chem. Bio.* **2009,** *5*, 789-96.
50. Fraser, J. S.; Clarkson, M. W.; Degnan, S. C.; Erion, R.; Kern, D.; Alber, T., Hidden Alternative Structures of Proline Isomerase Essential for Catalysis. *Nature* **2009,** *462*, 669-73.
51. Dubay, K. H.; Bothma, J. P.; Geissler, P. L., Long-Range Intra-Protein Communication Can Be Transmitted by Correlated Side-Chain Fluctuations Alone. *PLoS Comp. Bio.* **2011,** *7*, e1002168.
52. Einav, T.; Mazutis, L.; Phillips, R., Statistical Mechanics of Allosteric Enzymes. *The Journal of Physical Chemistry B* **2016,** *120* (26), 6021-6037.
53. Motlagh, H. N.; Wrabl, J. O.; Li, J.; Hilser, V. J., The Ensemble Nature of Allostery. *Nature* **2014,** *508* (7496), 331-339.
54. Wrabl, J. O.; Gu, J.; Liu, T.; Schrank, T. P.; Whitten, S. T.; Hilser, V. J., The Role of Protein Conformational Fluctuations in Allostery, Function, and Evolution. *Biophysical chemistry* **2011,** *159* (1), 129-141.
55. Fried, S. D.; Bagchi, S.; Boxer, S. G., Extreme Electric Fields Power Catalysis in the Active Site of Ketosteroid Isomerase. *Science* **2014,** *346* (6216), 1510-1514.

56. Warshel, A., Electrostatic Origin of the Catalytic Power of Enzymes and the Role of Preorganized Active Sites. *J. Biol. Chem.* **1998,** *273* (42), 27035-27038.
57. Warshel, A.; Sharma, P. K.; Kato, M.; Xiang, Y.; Liu, H. B.; Olsson, M. H. M., Electrostatic Basis for Enzyme Catalysis. *Chemical Reviews* **2006,** *106* (8), 3210-3235.

Chapter 2

A Monte Carlo method for generating side chain structural ensembles

In this chapter, I present a new Monte Carlo side chain entropy (MC-SCE) method that uses a physical energy function inclusive of long-range electrostatics and hydrophobic potential of mean force, coupled with both backbone variations and a backbone dependent side chain rotamer library, to describe protein conformational ensembles. Using the MC-SCE method in conjunction with backbone variability, one can reliably determine the side chain rotamer populations derived from both room temperature and cryogenically cooled X-ray crystallographic structures for CypA and H-Ras and NMR J-coupling constants for CypA, Eglin-C, and the DHFR product binary complexes E:THF and E:FOL. Furthermore, near perfect discrimination between a protein's native state ensemble and ensembles of misfolded structures for 55 different proteins was obtained, thereby generating far more competitive side chain packings for all of these proteins and their misfolded states. This chapter us based on the following publication

A. Bhowmick and T. Head-Gordon (2015). A Monte Carlo method for generating side chain structural ensembles. *Structure*.23(1):44-55

2.1 INTRODUCTION

Anfinsen's thermodynamic hypothesis[1] states that the native protein ensemble resides in a global minimum free energy basin that defines its functional state whether it be binding, catalysis, or signaling. This has been traditionally interpreted as a free energy basin dominated by O(~1) unique conformations, an interpretation heavily influenced by X-ray crystallographic protein structures that have proven to be invaluable for providing functional insight. Nonetheless, the perspective of considering just one native conformation opposes evidence that proteins are highly flexible[2], especially at the level of backbone displacements[3] that aid side chain packing rearrangements[4-7]. For example, new analysis of weak electron density features in X-ray crystallographic data has shown that a large percentage of PDB structures have alternate rotameric side chains[8, 9]. Furthermore, X-ray crystallographic structures that are cryogenically cooled also tend to overemphasize a level of uniqueness in native state structures that are too small and overpacked, and miss important catalytic side chain conformers that are present in room temperature crystallographic data [10, 11].

The thermodynamic manifestation of conformational flexibility is encompassed in entropic effects[12], with statistical fluctuations of side chain packing arrangements playing a dominant role. NMR groups have made quantitative progress on equating Lipari-Szabo order parameters, S^2, to conformational entropy for both the backbone and side chains[7, 13-15]. For example, NMR experiments on calmodulin[16] and CAP[6] proteins have shed light on this 'residual' free energy arising from the alternate conformations a side chain can take. A good percentage of side chains were found to have the side chain order parameter in the range $0.3<S^2<0.7$ which

indicates that these side chains may be populating alternate rotameric wells on the nanosecond-microsecond timescale, although the fast motions measured by S^2 are not always probing side chain rotamer transitions[17]. Instead, three bond J-coupling constants $^3J_{C\gamma N}$ and $^3J_{C\gamma CO}$ that report on χ_1 dihedral angle fluctuations in the broad picosecond-millisecond timescale have enabled quantitative estimation of different rotamer populations in solution[17]. In addition, recent work using relaxation experiments have also highlighted the dynamic nature of side chains up to the millisecond, and longer, timescale [18, 19].

Although it is true that the conformational flexibility of an unfolded protein compared to a folded protein is increased, numerical studies have shown that the number of possible ways of packing side chains on the backbone of a folded protein is by no means small or unique. Zhang and Liu reported that the total number of self-avoiding (i.e. with just hard sphere interactions) side chain conformations for the 17-residue protein 1ebx is of the order of 10^{11} [20], and this number would be expected to be larger for larger proteins. However theoretical approaches for sampling the low energy alternative side chain arrangements of a protein is a difficult problem, and while molecular dynamics (MD) simulations give a good description of side chain conformational change on the nanosecond to sub-microsecond level[21], the experimental estimates indicate that the timescales are much longer. While it is true that distributed computing paradigms such as Folding@Home [22] and special purpose hardware like the Anton computer [23] can reach the millisecond timescale for MD, we assert that computing the side chain populations and the thermodynamic entropy for tens to hundreds of native proteins and hundreds of their misfolded states, ad we have done in this study, is well beyond a comfortable scale for MD even using these two powerhouse computing platforms.

Therefore to circumvent the sampling issues imposed by MD, many groups have resorted to advanced Monte Carlo (MC) schemes[3, 20, 24] which are designed to more exhaustively sample the Boltzmann weighted populations of side chain conformations of the protein on the NMR timescale of ns to ms or even longer. In this work we develop a new MC approach for calculating side chain entropy (SCE) by introducing several new features that make our MC-SCE method more quantitative compared to past efforts, including a better convergent Rosenbluth sampling scheme[25], the use of an augmented Dunbrack library[26], a very robust physics-based energy function[27-29], and side chain rotamer sampling on an ensemble of backbone structures.

Here we use our MC-SCE algorithm to generate ~20,000 different side chain packings for native X-ray crystal backbones, and the same number for perturbations to the backbone using short MD simulations and so-called "backrub motions" by Friedland et al.[3], for 60 different proteins. As a first test of our MC-SCE algorithm, we use it to quantify the side chain rotamer populations on backbones derived from cryogenically cooled (CC) and room temperature (RT) X-ray crystallographic structures for CypA, and the Ser99Thr mutant [10] and for H-Ras[11]. We also compare directly to NMR J-coupling data for CypA[10], Eglin-C[30], and the DHFR binary complexes of E:THF and E:FOL [17]. We find overall excellent agreement across the full range of X-ray and NMR data. Finally we consider alternative rotamer packings for 55 native proteins and each of the hundreds of misfolded structures from a difficult Rosetta set that exhibit near-native features in their backbone fold[31]. We use our MC-SCE approach to provide the thermodynamic functions of energy (enthalpy), side chain entropy, and free energy to discriminate the native state of a protein from its misfolded states. This large validation suite shows that we can nearly perfectly discriminate between a protein's native state ensemble and ensembles of misfolded structures, and provide for an even more competitive decoy set with better optimized side chain packings.

2.2 MATERIALS AND METHODS

We introduce a new and more robust MC chain growth method to evaluate side chain entropy, MC-SCE, to estimate structural ensemble properties of proteins. We use an augmented Rosenbluth chain growth algorithm[20, 25, 43] to generate an ensemble of side chain packings for a given (and fixed) protein backbone. The algorithm starts with a PDB file of the enzyme, and all the side chain atoms, except the C_β atom, and any existing water molecules are eliminated. Backbone mobility is provided by a decoy library, backrub motions, or captured during a MD simulation. The side chain ensemble that can populate a provided bare backbone is then realized by growing side chains of each residue in a sequential manner with dihedral angle inputs from a backbone dependent rotamer library[26] to approximate the continuous nature of side chain dihedrals. We have augmented the rotamer library selection based on probabilities of occurrence in the PDB and by allowing for dihedral angle variations that are Gaussian distributed about a given rotamer value. All of the χ_1 and χ_2 torsional angles of all residues, except for arginine and lysine, were expanded by including a standard deviation, resulting in 3 values, χ_i and $\chi_i \pm \sigma$. After expansion, all the rotamers were further perturbed by about 0.5° to place them optimally with respect to the backbone. This is necessary because of the sensitivity of the energy function to slight changes in the protein that could distort statistics and increase the number of dead end chain growths. In our model, alanine and glycine have no dihedral degrees of freedom and hence no side chain entropy, and all residues are grown with ideal bond lengths and angles.

From the initial condition (step 0) of a bare backbone conformation *m*, for subsequent steps *i*, we develop a MC scheme whereby the residue *k* that has the lowest side chain partition function

$$Q_k = \sum_{\{\nu_k\}} e^{-\beta E_k^{(m,\nu_k)}} \tag{2}$$

is considered for the next side chain growth. For residue *k*, a side chain conformation ν_k is defined by the resulting set of dihedral angles selected from the rotamer library, i.e. (χ_1, χ_2, ….). Each side chain rotamer r_k is selected according to the following probability

$$\rho_k^{(m,r_k)} = \frac{P_{r_k}^{(pdb)} e^{-\beta E_k^{(m,r_k)}}}{\sum_{\{\nu_k\}} P_{\nu_k}^{(pdb)} e^{-\beta E_k^{(m,\nu_k)}}} \tag{3}$$

where $\{\nu_k\}$ are the possible side chain conformations for residue *k*, $E_k^{(m,r_k)}$ is the energy of interaction of side chain *k* with the backbone and all protein side chains grown so far using Eq. (3) only, and $P_{r_k}^{(pdb)}$ is the probability of the side chain conformation calculated using the values reported in the recent backbone-dependent Dunbrack library[26]. The reason for including this knowledge based $P_{r_k}^{(pdb)}$ is to guide the growth process especially early on when very few side chains have been placed and to minimize picking rotamers which are known to occur infrequently in the PDB database; conformations with probability less than 0.001 in the library were ignored. Once the side chain of a residue is placed, the process is repeated until all the side chains are grown, thereby creating one complete protein structure. This complete chain growth procedure for one N-residue enzyme structure is then repeated ~20,000 times to give an

ensemble of structures. Each structure *m* is then assigned a weight *W(m)* in order to get correct statistics in the canonical ensemble.

$$W(m) = e^{-\beta F_{solv}} \prod_{k=1}^{N} \frac{\sum_{\{\nu_k\}} P_{\nu_k}^{(pdb)} e^{-\beta E_k^{(m,\nu_k)}}}{P_{r_k}^{(pdb)}} \tag{4}$$

This is defined on the basis of our chain growth probabilities as well as now including the Boltzmann factor using the GB-HPMF implicit solvent model[27-29]. When the chain growth is unsuccessful because of unresolvable clashes, the partially grown structure is considered dead and its weight is set to zero.

The side chain entropy of a given residue *k* is evaluated using the Gibbs probabilistic definition of entropy.

$$S^{(k)} = -k_B \sum_{\{\nu_k\}} p_{\nu_k}^{(k)} \log p_{\nu_k}^{(k)} \tag{5}$$

where the probability $p_{\nu_k}^{(k)}$ of a conformational state ν_k of residue *k* is calculated using the weights of the structures in the ensemble

$$p_{\nu_k}^{(k)} = \frac{\sum_{m=1}^{M} W(m) \delta_{r_k,\nu_k}^{(m)}}{\sum_{m=1}^{M} W(m)} \tag{6}$$

The sum in Eq. (6) is over all of the successful structures grown by the Rosenbluth procedure. The Kronecker delta is 1 if the side chain conformation r_k that was picked for the residue *k* in the m-th structure is ν_k and 0 otherwise. The weights of each structure ensure that the probabilities are Boltzmann weighted. The total side chain entropy of a protein is calculated by summing over the individual entropy values

$$S_{SC} = \sum_{k}^{\#\,residues} S^{(k)} \tag{7}$$

NMR J-coupling calculations: Three-bond J-coupling values between the C_γ atom and the backbone carbonyl carbon ($^3J_{C\gamma CO}$) and amide nitrogen ($^3J_{C\gamma N}$) of the same residue can be calculated using

$$J_{XY} = A\cos^2(\theta+\delta) + B\cos(\theta+\delta) + C \tag{8}$$

where θ represents the dihedral angle between atoms (Y-C_α-C_β-X). The Karplus parameters (A,B,C,δ) are amino-acid specific and were taken from the original experimental sources. For Valine, ^3J values for only $C_{\gamma 1}$ have been reported in this paper.

The J-coupling value, $J_{XY}^{(k)}(m)$ for residue k in the m-th structure of our side chain ensemble was calculated from Eq. (8). These values were then used to calculate the average J-coupling value with

$$\bar{J}_{XY}^{(k)} = \frac{\sum_{m=1}^{M} W(m) J_{XY}^{(k)}(m)}{\sum_{m=1}^{M} W(m)} \tag{9}$$

where W(m) are the weights given in Eq. (4). We also calculated χ^2 values defined as

$$\chi_{XY}^2 = \frac{1}{N} \sum_{i \in \{k\}}^{N} \frac{(\bar{J}_{XY}^{(i)} - J_{XY}^{(i,\exp)})^2}{\sigma_i^2} \tag{10}$$

where N is the number of residue measurements taken. We have assumed that the primary source of experimental uncertainty is the Karplus parameters themselves; we assume an average uncertainty of σ=0.5 Hz given the differences found for these same scalar couplings for CypA.

Rosetta decoy set calculations. The single side chain native structure (the PDB) and the provided Rosetta decoy structures (with a given side chain arrangement) undergo local optimization, and are sorted in ascending order based on their energy in order to determine the E_{single} rankings. These minimized structures are then stripped of their side chains beyond the C_β position, and 20,000 alternate side chain packings with no steric clashes (which signals a failed chain growth) are generated on the native backbone and each Rosetta decoy backbone. The lowest energy structure for each ensemble is then minimized (to relax residual geometric artifacts arising from the fixed bond and bond angles assumed in the MC-SCE sampling using the rotamer library), and these minimized native and decoy structure for each protein are sorted in ascending order based on their energy in order to determine the E_{best} rankings. The side chain ensemble of structures generated for each backbone, native or decoy, shows a Gaussian distribution of energies, and we define the side chain entropy of the protein, S_{SC} in Eq. (7), based on Boltzmann weighted structures, Eq. (6), with energies below two standard deviations from the mean energy. We find that this subset of ~200 structures typically underestimates the entropy by ~5-10%, but since it is systematically applied across the protein and decoy sets, it suffices for this study.

2.3 RESULTS

Overview. We present results below based on a new and more robust MC side chain growth method to evaluate side chain entropy, MC-SCE, to estimate structural ensemble properties of proteins. Details are given in the Methods section, but the important points are highlighted here to better present the following results. Backbone structures are provided by either an X-ray crystal structure or a given backbone from a misfolded decoy library. Additional backbone variability on these starting structures is introduced in two independent ways: through so-called "backrub motions" [3], which lead to small backbone RMSD with respect to the crystal structure of ~0.1-0.7 Å, and from snapshots generated from a thermalized molecular dynamics simulation with explicit solvent that lead to slightly larger RMSD changes of ~0.6-1.3 Å.

Given these different backbones, the side chains atoms beyond the C_β position are stripped away, and then all are regrown using the MC-SCE algorithm to generate an *ensemble* of ~20,000 different side chain packings, allowing us to evaluate both the side chain entropy at each residue position and rotamer populations. Table S1 provides the definition of the side chain dihedral angles sampled. One of the key features of this work is the use of well-tested physics-based energy function based on Generalized-Born electrostatics and a hydrophobic potential of

mean force [27-29] to perform the Boltzmann weighting, and which is used to define the potential energy rank of all 20,000 structures. Here we demonstrate our ability to reliably reproduce and predict the side chain rotamer ensembles of the following class of problems: (1) cryo-cooled vs. room temperature X-ray crystallography for CypA and H-Ras, (2) both X-ray and NMR data taken on CypA, Eglin-C and the product binary complexes of DHFR, E:THF and E:FOL, and finally (3) native vs. misfolded state discrimination using a difficult Rosetta decoy set. All components of the MC-SCE approach (including the energy function) have been implemented into our in-house version of the TINKER[32] software package. As an example of the cost of the MC-SCE method, we can generate a side chain ensemble of CypA (164 residues) with 20,000 structures in ~12 hours using an MPI implementation that distributes work across 16 cores; this timing uses our in-house computing cluster with the AMD Opteron(TM) Processor 6274 (2.2 Ghz) cores.

Comparison with X-ray crystallography and NMR for CypA and H-Ras. Recently, Fraser et al. found population shifts in side chain rotamer states when comparing X-ray structures obtained under cyro-cooling vs. room temperature crystallization conditions for the proteins CypA [10, 11]. Given that the backbone differences between the CC and RT structures are negligible (RMSD ~ 0.1 Å), a good test of our MC-SCE algorithm would be to determine if we can predict the major and minor side chain rotamer populations that are reported in the CC and RT crystallographic data for CypA and H-Ras.

Experiments on CypA showed that alternate side chain conformations for Arg55 and Met61 were found with RINGER in the CC data, and additional side chain rotamer changes were evident for Leu98, Ser99 and Phe113 in the RT data, helping to explain the catalytically competent and incompetent conformations of the active site residues [10]. Table 2.1 reports the CC and RT X-ray experimental χ rotamers and their populations and the corresponding MC-SCE values for WT CypA and the Ser99Thr mutant. The MC-SCE calculations were done on the CC backbone, as well as an average over two RT backbones based on so-called major and minor conformers reported for the room temperature crystal structure (RT-M or RT-m). The 20,000 structures of the generated side chain packing ensemble for each backbone allow us to report MC-SCE population percentages. We also averaged over the 20,000 side chain ensembles generated for each backbone relevant to RT backbone variations: two backrub ensembles of 10 structures each based on the starting RT-M and RT-m backbones, and 3 backbones generated from MD snapshots at 0.2 ns, 2.0 ns and 4.0 ns. For side chain conformations predicted from the MC-SCE algorithm, the χ rotamers were binned as is done conventionally with bin centers on 60°, 180° and -60.

When performed on the CC X-ray backbone, our MC-SCE method predicts the same major conformer for residues Leu98(χ_1), Phe113(χ_1), Arg55(χ_3), and Met61(χ_2), as well as detecting the minor rotamer states for the latter two residues that was found from the RINGER analysis of weak electron density features. When performed on the RT X-ray backbone, our MC-SCE method also predicts the major and minor conformer for all four same residues. Furthermore, we determined an increase in SCE (using Eq. 7) when going from the CC to RT backbone as was observed in [11], in which the RT backbone allows for greater conformational flexibility of the side chains. Even better agreement with reported X-ray rotamer populations is found with a thermalized backbone (i.e. side chains grown on backrub and MD backbone ensembles) for these same residues as shown in Table 2.1. We also perform our MC-SCE calculations on the Ser99Thr mutant, which through active site interactions stabilizes the minor

rotamers for Phe113(χ_1), Arg55(χ_3), and Met61(χ_2) compared to the WT form, which is exactly what we observe in our simulations (Table 2.1).

Table 2.1. *X-ray crystallographic and MC-SCE generated side chain χ rotamers for active site residues of CypA.* Experimental rotamer populations [10] are the occupancies reported in the deposited PDB files for CypA and mutant (CC: 3k0m, RT: 3k0n, Ser99Thr: 3k0o). In certain cases, the minor rotamer was identified in the CC structure using the software Ringer [9]. MC-SCE calculations were done on the backbone of the cryo-cooled structure (CC) as well as an average over the backbone conformers M and m reported for the room temperature crystal structure (RT-M and RT-m). MC-SCE calculations were also performed on a RT backbone ensemble comprised of backrub motions and MD simulations (RT ensemble).

CypA		**X-ray Population**		**MC-SCE population using CC, RT, and Ensemble backbones**			**CypA Mutant Ser99Thr**	
Res	*χ Class*	*CC*	*RT*	*CC backbone*	*RT (M, m) backbone*	*RT ensemble*	*X-ray RT*	*MC-SCE RT*
Leu98 (χ_1)	60							
	180	**100.0**	**63.0**	100.0	50.0	57.5	**100.0**	19.0
	-60		**37.0**		50.0	42.5	**Ringer**	74.7
Ser99 (Thr99) (χ_1)	60		**37.0**		50.0	22.4	**Ringer**	44.3
	180	**100.0**	**63.0**				**100.0**	3.8
	-60			100.0	50.0	77.6		51.9
Phe113 (χ_1)	60	**100.0**	**63.0**	100.0	50.0	75.0		
	180							
	-60		**37.0**		50.0	25.0	**100.0**	100.0
Arg55 (χ_3)	60				3.5	17.3		
	180	**100.0**	**63.0**	78.6	45.2	53.0	**Ringer**	25.3
	-60	**Ringer**	**37.0**	21.4	51.3	29.7	**100.0**	74.7
Met61 (χ_2)	60	**40.0**	**37.0**	1.8	0.7	9.0	**100.0**	83.5
	180	**60.0**	**63.0**	98.2	99.3	91.0		1.3
	-60							15.2

In all cases, regardless of method for creating the backbone, we do not predict the 180° rotamer for Ser99(χ_1), and we do not find the same 180° dominant rotamer for the Thr99(χ_1) mutant (although we do predict it as a minor conformation). One possibility is that the energy function, and possibly the use of an implicit solvent model for water, accounts for this discrepancy, although our energy function with implicit solvent has been extensively validated[27-29]. When we perform MD with explicit solvent using the CC crystal as the start state, the 180° rotamer flips to the 60° rotamer and maintains that value for the entirety of the simulation run. Hence the very different energy functions and sampling methods (implicit vs. explicit solvent and MC vs. MD) favors an alternate rotamer to the major rotamer seen experimentally. Therefore we believe that overall the energy function used with MC-SCE is performing well. The fact that we are able to correctly predict the change in rotameric states for Phe113(χ_1), Arg55(χ_3), and Met61(χ_2) when going from WT to the SerThr99 mutant for CypA, indicates that the adoption of the 180° rotamer at position 99 for WT and mutant CypA is not necessary, suggesting that we

are seeing a similar cooperative network effect among residues that were analyzed in the NMR relaxation experiments[10].

To provide for better contact with NMR solution data for CypA and the Ser99Thr mutant, Table 2.2 reports $^3J_{C\gamma N}$ and $^3J_{C\gamma C}$ values evaluated from our MC-SCE ensemble populations and compared to the same experimentally measured values for various aromatic residues[10]. We evaluated our J-couplings using the Karplus equation parameterization values found in both [33] and [17], and they are also reported in Table 2.2. To put the comparison in some context, we also calculate the difference between the experimental J-coupling, $\overline{J}_{XY}^{(i,\exp)}$ and the average scalar coupling calculated from a given MC-SCE structural ensemble, $\overline{J}_{XY}^{(i)}$ for each residue (Eq. (9)), normalizing it by the uncertainty of the Karplus parameters and any experimental error, to generate χ_J^2 values (Eq. (10)). We use a conservative uncertainty value due to the Karplus equation of σ_J = 0.5 Hz for both types of scalar couplings, estimated from the difference in calculated J-couplings using the two Karplus equation parameterizations that use the same underlying structural ensemble. Any dominant error due to the underlying structural ensembles themselves would then correspond to values of $\chi_J^2 > 1$. Our calculated deviations are $\chi^2_{JC\gamma N}$ =0.53 and $\chi^2_{JC\gamma C}$ =0.99 indicating that the underlying structural ensembles are sound. As such, we also observe a change in J-coupling values for Phe113 which confirms a switch in rotameric state from 60° to -60° as per the experiment.

Table 2.2: *J-coupling data for CYPA calculated using MC-SCE on the cryo-cooled backbone of wild type CYPA (3k0m) and Ser99Thr mutant (3k0p*). The experimental values are taken from[10]. Two sets of Karplus parameters have been used to generate the MC-SCE scalar couplings: [33] CC(S), and [17], CC(T). Using either parameters, we observed a change in J-coupling values for Phe113 which confirms a switch in rotameric state from 60° to -60°.

	WT						**Ser99Thr**					
Residue	$^3J_{CC}$*(Hz)*			$^3J_{NC}$*(Hz)*			$^3J_{CC}$*(Hz)*			$^3J_{NC}$*(Hz)*		
	Expt	*CC (S)*	*CC (T)*	*Expt*	*CC (S)*	*CC (T)*	*Expt*	*CC (S)*	*CC (T)*	*Expt*	*CC (S)*	*CC (T)*
Phe25	3.6	3.91	4.36	0.8	0.36	0.37	3.7	3.91	4.37	0.7	0.36	0.37
Tyr79	3.3	3.78	4.39	1.0	0.44	0.42	3.4	3.79	4.40	0.7	0.41	0.41
Phe88	3.6	3.91	4.41	0.8	0.39	0.37	3.3	3.92	4.39	0.5	0.39	0.37
His92	3.6	4.22	5.06	1.0	0.57	0.47	3.9	4.22	5.06	1.2	0.57	0.51
Phe113	**1.1**	**0.43**	**0.38**	**0.9**	**0.51**	**0.41**	**2.8**	**3.9**	**4.41**	**0.8**	**0.46**	**0.39**
Phe145	3.3	3.91	4.39	0.9	0.36	0.37	3.5	3.91	4.38	0.2	0.44	0.37

Table 2.3 reports the CC and RT X-ray crystallographic and MC-SCE generated side chain χ rotamers and their populations for H-Ras. Again, the MC-SCE calculations were done on the CC backbone and its backrub variation, and the 20,000 structures of the generated side chain packing ensembles allow us to report MC-SCE population percentages. However the RT variations of the reported 9 individual side chains involved more than two rotameric states, and in combination would result in a large combinatorial number of RT crystal backbones that are inconvenient for performing the backrub motions. Instead we represent backbone variability using 3 MD snapshots at 0.2 ns, 2.0 ns and 4.0 ns to analyze the higher temperature data, given

its consistency with backrub motions for CypA and for the Rosetta data sets described further below.

Table 2.3. *X-ray crystallographic and MC-SCE generated side chain χ rotamers for active site residues of H-Ras.* Experimental rotamer populations are the occupancies reported in the deposited PDB files (CC: 1ctq, RT: 3TGP). In certain cases, the minor rotamer was identified using the software Ringer [8, 9]. MC-SCE calculations were done on the cryo-cooled backbone (CC), backrub ensemble of the cryo-cooled backbone (BR-CC) as well as on RT MD snapshots generated at 0.2, 2 and 4 ns time points to incorporate backbone flexibility.

H-Ras		**X-ray Populations**		**MC-SCE using CC backbone**		**MC-SCE using RT MD backbone**		
Res	*χ Class*	*CC χ*	*RT χ*	*CC*	*BR-CC*	*0.2ns*	*2.0 ns*	*4.0 ns*
Asp 30 (χ_1)	60		**55.0**					
	180	**100.0**	**45.0**	100.0	58.3	90.0	100.0	50.0
	-60				41.7	10.0		50.0
Glu 62 (χ_1)	60						1.5	
	180		**100.0**	7.7	41.6		24.2	66.7
	-60	**100.0**		92.3	58.4	100.0	74.3	33.3
Ser 65 (χ_1)	60	**100.0**		88.5	50.0	65.2	9.1	66.7
	180		**100.0**					
	-60			11.5	50.0	34.8	90.9	33.3
His 94 (χ_1)	60	**100.0**	**48.0**	61.6	50.0	11.6	7.6	100.0
	180		**52.0**	34.6	50.0	88.4	22.7	
	-60			3.8			69.7	
Val 103 (χ_1)	60	**Ringer**					1.5	
	180		**38.0**	38.5	58.3			
	-60	**100.0**	**62.0**	61.5	41.7	100.0	98.5	100.0
Gln 61 (χ_2)	60		**66.0**	23.1	25.0	1.5		
	180	**100.0**	**34.0**	3.8	25.0	30.4	90.9	66.7
	-60		**Ringer**	73.1	50.0	68.1	9.1	33.3
Arg 97 (χ_3)	60		**Ringer**	3.8		97.1	1.5	
	180	**100.0**	**100.0**	96.2	83.3	0.0	98.5	100.0
	-60				16.7	2.9		
Glu 98 (χ_2)	60				16.7	10.2	7.6	
	180	**100.0**		80.8	75.0	73.9	1.5	83.3
	-60		**100.0**	19.2	8.3	15.9	90.9	16.7
Gln99 (χ_2)	60		**Ringer**		8.3		39.4	
	180	**100.0**	**100.0**	84.6	83.3	92.7	54.6	66.7
	-60			15.4	8.3	7.3	6.1	33.3

When performed on the CC X-ray backbone, or its backrub variant, our MC-SCE method predicts the same major conformer for residues Asp30(χ1), Glu62(χ1), Ser65(χ1), His94(χ1), Val103(χ1), Arg97(χ3), Glu98(χ2), and Gln99(χ2), with the only exception being Gln61(χ2) in which the MC-SCE algorithm predicts it to be a minor (up to 25%) population. What is most interesting is that the MC-SCE method using the CC backbone can also determine the minor side

chain conformation detected in the RT crystal structure for Glu62(χ1), His94(χ1), Val103(χ1), Arg97(χ3), Glu98(χ2), as well as Gln61(χ2) which samples all three rotameric states. This suggests that the cryogenic backbones are not completely deficient for accommodating alternate rotamers, but apparently their electron density features are either far too weak to detect, or possibly that crystalline contact interactions favor certain rotamers. The MD results are also interesting, showing the time evolution of the rotamer populations for these residues as the backbone varies, flipping between the major and minor rotamer states.

However, although the MC-SCE does predict the major conformer, it does not predict the alternative rotamer preference observed in the RT X-ray data for either Asp30(χ1) or Ser65(χ1) on any backbone. Given that these residues are surface residues, they may be more prone to crystal packing artifacts that bias the populations of a particular rotamer class. Figure 2.1 shows that there are stabilizing interactions for these two residues with the surrounding lattice that favor the RT major rotamer that is experimentally observed; in particular Asp30 interacts with arginine and Ser65 shows very close approach to glutamic acid. Since we do not represent the crystal lattice, these favorable hydrogen-bonding interactions would not be present, and thus would not preferentially stabilize the experimentally observed RT major rotamer.

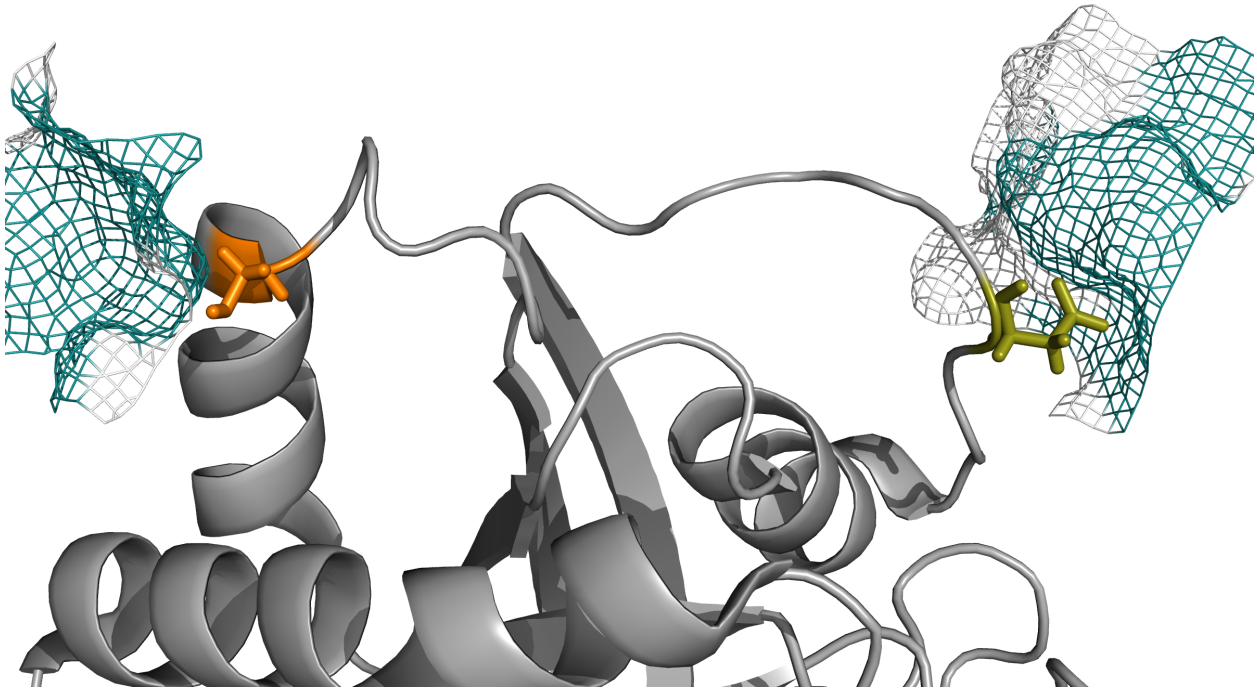


Figure 2.1. *The PDB backbone and crystallization conditions for H-Ras.* The residues represented are Asp30 (olive) and Ser65 (orange). They are important catalytic residues studied for H-Ras in which both the cryogenically cooled structure (1CTQ)[44] and the room temperature structure (3TGP)[11] were crystallized with a bound GTP ligand bound (purple). MC-SCE could not predict the major rotamer reported in the room temperature crystal structure for these 2 residues. The meshes represent the crystal elements nearby as reported in the room temperature crystal structure(3TGP). The figure was generated using the PyMOL Molecular Graphics System, Version 1.5.0.4 Schrödinger, LLC.

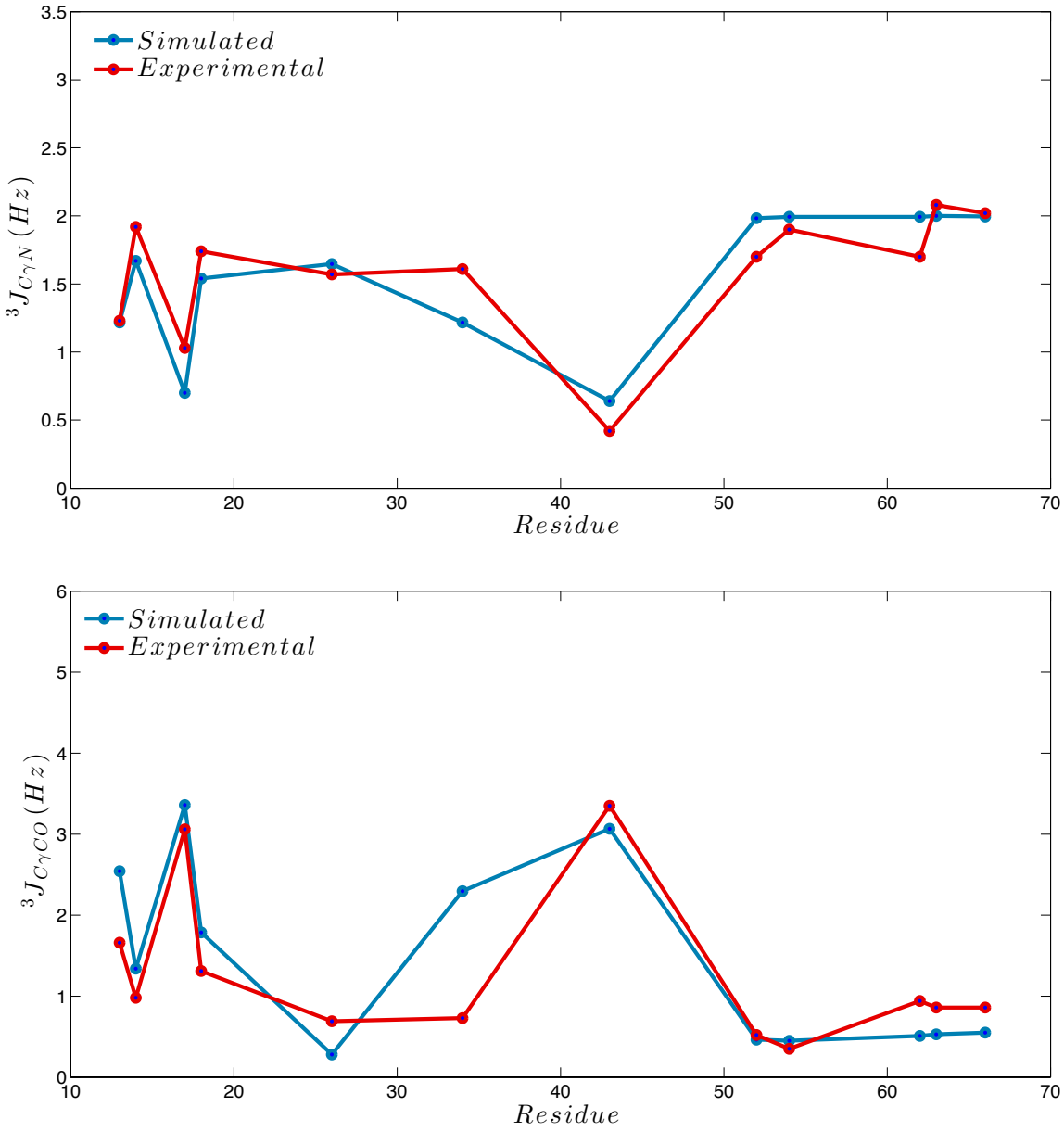


Figure 2.2. *J-coupling constants (a) $^3J_{C\gamma N}$ and (b) $^3J_{C\gamma CO}$ for Eglin-C.* The red symbols are the experimental data from [30]. The blue symbols are calculated from the MC-SCE ensemble using backbones from molecular dynamics and the Karplus parameterization from [17]

Comparison to NMR data for Eglin-C, E:THF, and E:FOL. We next analyze the MC-SCE approach against solution-based NMR scalar coupling constants $^3J_{C\gamma N}$ and $^3J_{C\gamma C}$ generated by Clarkson et al on Eglin-C [30] and by Tuttle and co-workers for the DHFR binary product complexes E:THF and E:FOL [17]. To calculate the scalar coupling constants, we again use the standard Karplus equation, Eq. (8), with Karplus parameters from [17]. Figures 2.2 shows the agreement between the experimental coupling values and the $\overline{J}_{XY}^{(i)}$ generated from the MC-SCE ensembles, taken on both the CC and MD backbones, for C-Eglin.

The overall χ_J^2 values for $^3J_{C\gamma N}$ is 0.36 and for $^3J_{C\gamma C}$ is 0.84 on the CC backbone, and these values change to $\chi^2_{JC\gamma N}$=0.20 and $\chi^2_{JC\gamma C}$=1.44 on the averaged molecular dynamics backbones, indicating that the structural ensembles are in overall good agreement with the rotamer populations for the 12 residues. Table S2 in the Supplementary materials provides a more detailed rotamer assignment for Eglin-C, and we note that although our J-coupling values for two of the residues, Thr 17 and Thr 26, are in excellent agreement with the experimental measurements, we do not agree with the experimental study in the assigned rotameric populations, suggesting that the experimental rotameric populations may be flawed.

We next consider the J-coupling constants for E:THF and E:FOL, requiring us to develop parameters for the bound ligand on which the NMR data was taken; the introduction of the ligand means that we can't generate backrub ensembles from the server[3], and hence we use MD data to provide for backbone variations. Figures 2.3 and 2.4 show the agreement between the experimental coupling values and the $\overline{J}_{XY}^{(i)}$ generated from the MC-SCE ensembles, taken on the averaged MD backbones, for the E:THF complex and E:FOL complex, respectively. The overall χ_J^2 values on the CC backbone is small ($\chi^2_{JC\gamma N}$ = 0.45 to 0.64) for both proteins, while the deviation in $^3J_{C\gamma C}$ is larger when measured on the MD generated backbones ($\chi^2_{JC\gamma C}$ = 2.71 to 3.06). The large χ_J^2 value for the $^3J_{C\gamma C}$ coupling for E:THF is due to genuine disagreement for what is the major rotamer for only three residues: Val40, His114, Thr123, although for His114 we find it to be a minor rotamer instead (Table S3 at the end of chapter). For the DHFR complex E:FOL we again find disagreement for the major rotamer for two residues: Val10, Val40, and Thr123. It is noteworthy that Val40 and Thr123 are among one of the few residues that have different major rotamers in the multiple DHFR complexes studied in [17].

For E:THF the MC-SCE structural ensembles show overall very good agreement across 46 of the 49 residue NMR measurements, with $\chi^2_{JC\gamma N}$ = 0.65 and $\chi^2_{JC\gamma C}$ = 1.62, in which the major rotamer is correctly selected for all of these residues. For E:FOL the MC-SCE structural ensembles show overall very good agreement across 20 of the 22 residue NMR measurements, with $\chi^2_{JC\gamma N}$ = 0.23 and $\chi^2_{JC\gamma C}$ = 1.33, in which the major rotamer is correctly selected for all of these residues. Problems in the structural ensembles that gives rise to disagreement with the $^3J_{C\gamma C}$ measurement for the two protein complexes are due to differences in the assignment of the minor rotamer for a smaller subset of residues, i.e. no minor rotamer detected, detected with a much smaller population, or assignment of a different minor rotamer state (Table S4).

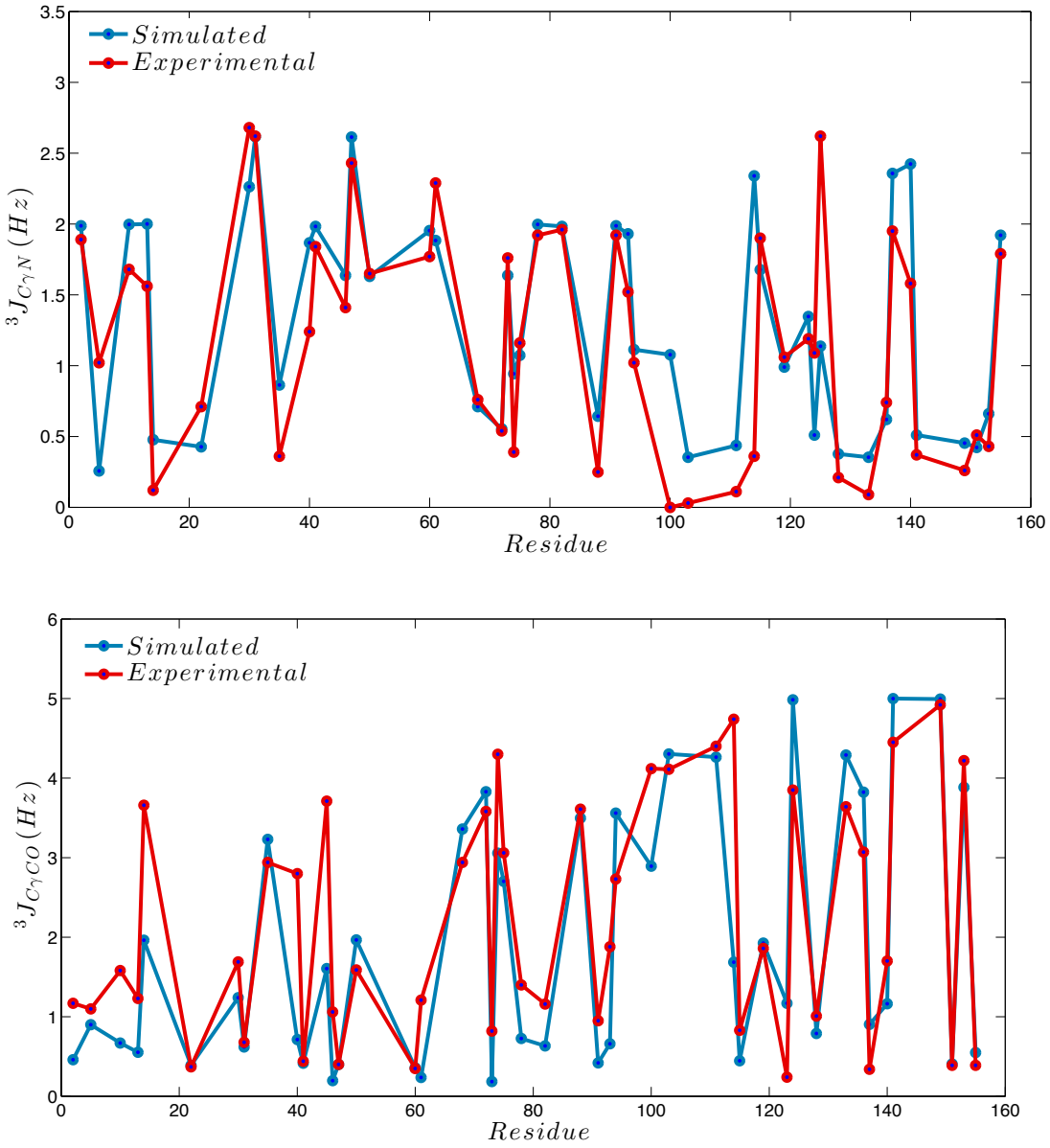


Figure 2.3. *J-coupling constants (a) $^3J_{C\gamma N}$ and (b) $^3J_{C\gamma CO}$ for the DHFR binary product complex E:THF.* The red symbols are the experimental data from [17]. The blue symbols are calculated from the MC-SCE ensemble using backbones from molecular dynamics and the Karplus parameterization from [17].

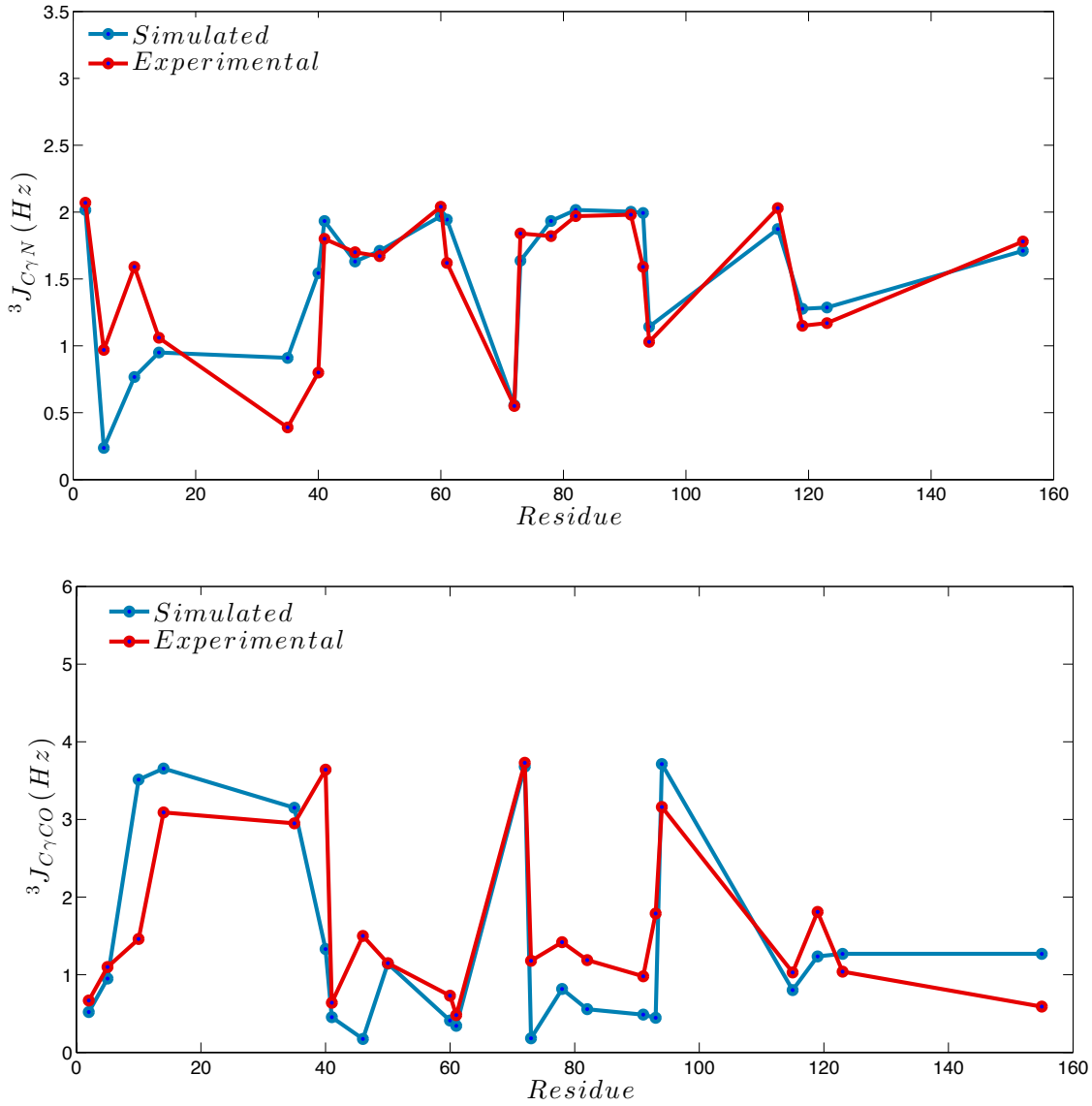


Figure 2.4. *J-coupling constants (a) $^3J_{C\gamma N}$ and (b) $^3J_{C\gamma CO}$ for the DHFR binary product complex E:FOL.* The red symbols are the experimental data from [17]. The blue symbols are calculated from the MC-SCE ensemble using backbones from molecular dynamics and the Karplus parameterization from [17].

Discrimination of native folded vs. misfolded states. Given the very good agreement between X-ray and NMR data and the MC-SCE ensembles, we next test our ability for selecting native state structures when compared to the 2007 Rosetta decoy set[34] generated by the popular fragment assembly folding program ROSETTA[35]. We define a traditional energy rank, E_{single}, as the energy of the *provided* side chain packing on a given backbone which is either the native X-ray PDB structure or the Rosetta decoy structures. Next we consider a free energy rank, F, based

on ensembles of alternative side chain packings for the given backbones for all 55 native proteins and misfolded structures using our MC-SCE method that evaluates the side chain entropy. This free energy function is defined as

$$F = E_{best} - TS_{SC} \quad (1)$$

where E_{best} is the structure whose side chain packing for a given backbone (native or decoy) is the lowest energy in the generated ensemble, and $-TS_{SC}$ is the temperature weighted side chain entropy (Eq. (7) in Methods). We also judge the quality of these thermodynamic metrics through calculation of a Z-score, the free energy (or E_{single} or E_{best}) difference between the native state quantity and the same quantity averaged over the misfolded states. A larger value of the Z-score signals better separation of the native structure from the misfolded conformers. All detailed data is reported in Table S4 at the end of the chapter in order for us to highlight the important points here.

The traditional rank based on our physics-based energy function, E_{single}, does a very good job of discrimination of the given native state from all of the members of a given decoy set, in which 40/55 proteins are ranked 1st with a Z-score of -3.76 for this subset (-2.95 over all proteins). Our energy function comfortably outperforms many recent popular statistical potentials like DFIRE[36](21/58), DOPE[37] (21/58), and EPAD[38] (34/58), and is competitive with other reported energy functions like EPAD2[38] (46/58) and PM6 [39] (49/49).

However, the free energy is the true thermodynamic quantity, and given that our MC-SCE algorithm can generate an ensemble of side chain conformer packings, we compare the native side chain ensembles and the respective decoy ensembles, based on the evaluation of the free energy, F, using Eq. (1). Using the free energy thermodynamic metric, the absolute native state discrimination improves modestly to 42/55 natives identified (Z-score for natives of -3.62), with the Z-score over all proteins improving slightly to -3.07. Even so, for 8 proteins whose native states were not selected, the ensemble F rank improved native state ranking, significantly in most cases, compared to using E_{single}: 1ail (rank 62 to 3), 1c8c (rank 47 to 2), 1enh (rank 81 to 13), 1hz6 (rank 7 to 3), 1rnb (rank 93 to 89), 1utg (rank 94 to 75), 1vcc (rank 4 to 2), 1ubi (rank 9 to 5), while 1pgx and 1dhn were 2nd ranked by either single or thermodynamic ensemble metrics (Table S4). A breakdown of the free energy shows that selection of the native conformation using the F rank is largely driven by E_{best}, since the Z-score based on the side chain ensemble best energy alone is lowered to -3.94 for all native states selected, and -3.27 over all proteins. This clearly indicates that the original native PDB structure and provided Rosetta misfolded structures have not optimized side chain arrangements for the given backbone. Furthermore, these lower energy side chain packings are providing sharper discrimination of folded vs. misfolded states. These results are consistent with a number of recent studies that have shown that weak features in the electron density maps from X-ray protein crystallography support alternate side chain packings that differ from the original reported side-chain rotamers [7-9, 11, 40].

In order to push toward better native state discrimination, we also considered additional ensemble characterizations involving the native state backbone, with the expectation that small perturbations to the backbone might allow for new side chain rotamer packings. These backbone changes may remove overly unique side chain rotamer states that arise from cryo-cooling [10, 11], as well as crystal contacts, oligomeric packing, or ligand-binding interactions[40]. For example, 1ail has been crystallized as a dimer, while 1c8c has a bound peptide, and thus are illustrative of perhaps why many of their decoys, generated independently from the original crystallization

conditions but with near-native features, are energetically better than the crystallized native state [40].

Therefore to test how the backbone perturbations influence the free energy ranking, we used backrub motions [3] that minimize repositioning of the backbone, but which can drastically affect side chain rotamer populations due to reorientation of the C_β atoms, for the 13 proteins in which the native state was not selected or very poorly predicted by the free energy function. We performed backrub motions on the X-ray backbone for each of these proteins, generating 10 different backrub structures. We then removed the side chains atoms beyond the C_β position for each, and then used our MC-SCE approach to generate side chain packing ensembles, in order to calculate thermodynamic rankings using E_{best}, and the free energy F, and their corresponding Z-scores (Table 4). In addition we also do the same MC-SCE procedure for backbones derived at the end of a short molecular dynamics simulation in explicit water at ambient temperature and pressure as an independent way to relax the crystalline constraints of the X-ray native structure. In both cases the native backbones were found to change by a little less than 1.0Å RMSD compared to their PDB structure, on average. The relative RMSD of the final thermalized native backbone with respect to the decoy set was unchanged on average, i.e. making the decoys no more or no less competitive for determining the native state ensemble.

Table 2.4. *Thermodynamic rankings and Z-scores of the native X-ray structure and MD and Backrub[3] relaxed backbones.*

Protein	RMSD (Å)	E_{best} Rank	E_{best} Z-score	F Rank	F Z-score	Protein	RMSD (Å)	E_{best} Rank	E_{best} Z-score	F Rank	F Z-score
1ail	**0.00**	**9**	**-1.22**	**3**	**-1.73**	**1pgx**	**0.00**	**3**	**-2.29**	**2**	**-2.23**
Backrub	0.26	1	-2.01	1	-2.33	Backrub	0.70	1	-3.08	1	-3.22
MD	1.13	1	-3.58	1	-3.36	MD	1.13	1	-2.72	1	-2.71
1c8c	**0.00**	**3**	**-2.01**	**2**	**-2.34**	**1rnb**	**0.00**	**90**	**1.18**	**89**	**1.17**
Backrub	0.33	3	-2.22	1	-2.59	Backrub	0.73	1	-3.87	1	-3.87
MD	0.56	3	-2.35	1	-2.96	MD	1.23	1	-4.07	1	-4.01
1dhn	**0.00**	**2**	**-2.40**	**2**	**-2.01**	**1ubi**	**0.00**	**10**	**-1.23**	**5**	**-1.38**
Backrub	0.41	1	-3.47	1	-2.82	Backrub	0.35	1	-2.66	1	-2.46
MD	0.94	1	-3.70	1	-3.34	MD	0.68	1	-2.92	1	-2.66
1enh	**0.00**	**14**	**-1.03**	**13**	**-1.02**	**1utg**	**0.00**	**81**	**0.94**	**75**	**0.73**
Backrub	0.31	1	-2.66	1	-2.47	Backrub	0.56	10	-1.29	10	-1.24
MD	0.60	1	-2.44	1	-2.32	MD	1.29	15	-0.99	26	-0.78
1gvp	**0.00**	**21**	**-0.81**	**18**	**-1.04**	**1vcc**	**0.00**	**3**	**-1.99**	**2**	**-2.06**
Backrub	0.65	1	-3.57	1	-3.67	Backrub	0.26	1	-3.38	1	-3.20
MD	1.14	1	-4.04	1	-3.86	MD	0.85	1	-3.77	1	-3.60
1hz6	**0.00**	**3**	**-2.03**	**3**	**-2.22**	**1vls**	**0.00**	**75**	**0.45**	**98**	**2.11**
Backrub	0.04	7	-1.76	6	-1.86	Backrub	0.93	1	-2.16	1	-1.81

MD	0.68	3	-2.11	3	-2.27	MD	1.02	1	-3.39	1	-2.58
256b	**0.00**	**1**	**-2.07**	**3**	**-1.65**						
Backrub	0.42	1	-2.81	1	-2.32						
MD	1.31	1	-4.02	1	-3.57						

The resulting drastic improvement in ranking– 53/55 proteins native states are now well distinguished from the misfolds– suggests that the initial failure of the free energy to identify the native state cannot be attributed primarily to the limitation of the energy function or MC-SCE sampling protocol. Instead, the small changes in backbone flexibility, consequences of which were also examined by Tyka and co-workers[40], highlights the sensitivity of SCE to subtle effects of the backbone configuration, which improved the discrimination for 11 of the 13 problematic proteins. Since the native state is selected for in ~96% proteins of the best available Rosetta decoy set, considered to be a challenging test of any new sampling method, statistical potential, physical force field or scoring function, MC-SCE appears to provide an excellent standard for native state prediction. Two proteins for which we did not discriminate for the native were 1hz6 (whose rank remained 3rd whether using the PDB or MD backbone) and 1utg (whose 75th native rank with the PDB structure rose to 10th with the backrub motions), and would require more careful consideration of available NMR data.

In order to check the similarity between the best energy native structure in our free energy ensemble with the deposited PDB crystal structure, the χ_1 torsional angles between the 2 structures were compared for each of the 55 proteins we analyzed. A residue was said to have had a change in torsional angle if the absolute value of their difference exceeded 40°, which is similar to the convention adopted by Bower and co-workers[41], and the fraction of the total residues that changed the χ_1 angle is listed in the final column of Table S4. On an average, our MC-SCE algorithm found an alternate χ_1 dihedral angle in the best free energy native structure compared to the crystal structure 25% of the time, consistent with the ~18% of alternate side chain rotamers on reexamination of electron density from 402 high resolution X-ray crystal structures[9]. Since Lang and co-workers only considered unbranched side chains in their electron density analysis, as well as ignoring density fitting with combinations of χ_1, χ_2, χ_3 etc., it would likely explain the quantitative discrepancy with what we have found since we considered all residues and the full rotameric set of χ values for any given amino acid. Figure 2.5 shows the typical distribution of side chain entropy on the 2CHF PDB backbone, where the side chain conformations that showed most variability did not exclusively select surface residues, but core positions as well.

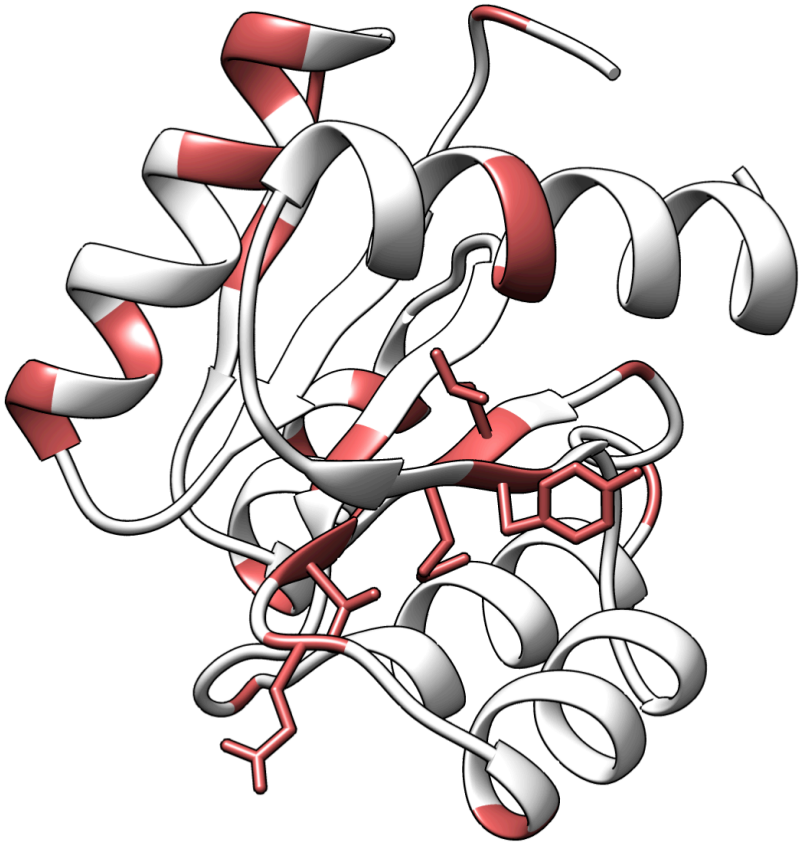


Figure 2.5. *The PDB native backbone of the Mg^{2+}-bound form of CheY[45] (2CHF).* The regions colored red are the side chain positions where alternate rotameric states was found by the MC-SCE algorithm compared to the PDB side chain packing. It is notable that side chain repacking occurs for both interior and surface residues. The Figure was generated using Chimera[46].

2.4 CONCLUSIONS

In summary, we have introduced a new MC-SCE algorithm for generating side chain packing ensembles, allowing us to predict side chain rotamer populations and side chain entropy, which we have compared to extensive data sets from both NMR and X-ray crystallographic experiments. We have validated our approach by making direct contact with X-ray crystallography and NMR data on side chain rotamer populations for CypA and its Ser99Thr mutant[10], HRas[11], Eglin-C[30], and the DHFR complexes E:THF and E:FOL [17]. For all proteins we find overall excellent agreement of rotamer values, their populations, and calculated J-couplings when compared to crystallographic data and with NMR experimental J-couplings.

We have shown that the side chain populations measured depend significantly on the given backbone structure, and hence our MC-SCE technique is aided by introducing small

deviations (~1.0 Å RMSD) from the crystallographic backbone structures using both backrub motions and thermalized explicit solvent molecular dynamics simulations. For the case of CypA and its Ser99Thr mutant we found all of the major and minor rotamers of all reported residues except for Ser(Thr)99. However, it had no discernable influence on our successful ability to predict the Phe113 catalytic rotameric state for WT as well as stabilizing the minor rotamers for Phe113(χ_1), Arg55(χ_3), and Met61(χ_2) in the active site of the mutant form[10].

For the protein H-Ras, we found that we can detect the minor or alternate rotamer state of a sidechain when the ensemble is generated on the CC backbone and its backrub variants, although the experimental density is only evident in the RT X-ray data[11]. In addition, we do not observe the same minor rotameric states that are experimentally found for Asp30 and Ser65. In both cases, i.e. our ability to detect new rotamers on the CC backbone or observing alternate minor rotamers to that found in the RT data, can be explained by the fact that the surrounding crystal lattice is not present in our approach. Previous work has shown that stabilizing packing interactions often arise from polar-polar interactions with the surrounding crystal lattice, and thus can influence the experimentally observed rotamer populations[42]. We found that such specific interactions with the surrounding lattice are present for Asp30 and Ser65, for example, and hence would not be predicted with our MC-SCE approach that instead represents aqueous solution conditions.

We have also compared our MC-SCE rotamer populations to those estimated from solution phase NMR data. Our calculated agreement with scalar coupling measurements for CypA, C-Elgin, and the two DHFR complexes E:THF and E:FOL were found to be overall excellent. The calculated scalar couplings using our MC-SCE method was well within experimental and Karplus parameter uncertainty for $^3J_{C\gamma N}$ for all four proteins, and for 85-100% of residues for the $^3J_{C\gamma C}$ measurement across the four data sets. The primary error for the E:THF and E:FOL complexes was the failure to predict the major rotamer for Val40 and Thr123, although these same residues were found to sample alternate rotameric states in the full series of DHFR complexes[17].

Finally, we have a highly reliable method for discrimination of native states from misfolded structures based on a difficult Rosetta decoy set. One consequence of our MC-SCE algorithm is that we find better side chain rotamer and packing representations of both the native state *and* the decoy set. This can be quantified for the decoy set by the Z-score between the PDB structure, i.e. the single backbone and side chain rotamers of the X-ray structure, and the E_{best} from the decoy set ensembles, which shrinks to –2.77. We have provided this new decoy set, Berkeley-SC-Ensemble, which we have made available at our web site http://thglab.berkeley.edu. It also includes the ensemble of new side chain packing arrangements on native PDB backbones that will be of interest to X-ray crystallographers and NMR groups.

2.5 REFERENCES

1. Anfinsen, C. B., Principles That Govern the Folding of Protein Chains. *Science* **1973,** *181* (4096), 223-30.
2. Kohn, J. E.; Afonine, P. V.; Ruscio, J. Z.; Adams, P. D.; Head-Gordon, T., Evidence of Functional Protein Dynamics from X-Ray Crystallographic Ensembles. *PLoS Comput Biol* **2010,** *6* (8).
3. Friedland, G. D.; Linares, A. J.; Smith, C. a.; Kortemme, T., A Simple Model of Backbone Flexibility Improves Modeling of Side-Chain Conformational Variability. *Journal of molecular biology* **2008,** *380*, 757-74.
4. Moorman, V. R.; Valentine, K. G.; Wand, a. J., The Dynamical Response of Hen Egg White Lysozyme to the Binding of a Carbohydrate Ligand. *Protein science : a publication of the Protein Society* **2012,** *21*, 1066-73.
5. Schnell, J. R.; Dyson, H. J.; Wright, P. E., Effect of Cofactor Binding and Loop Conformation on Side Chain Methyl Dynamics in Dihydrofolate Reductase. *Biochemistry* **2004,** *43*, 374-83.
6. Tzeng, S.-R.; Kalodimos, C. G., Protein Activity Regulation by Conformational Entropy. *Nature* **2012,** *488*, 236-40.
7. Fenwick, R. B.; van den Bedem, H.; Fraser, J. S.; Wright, P. E., Integrated Description of Protein Dynamics from Room-Temperature X-Ray Crystallography and Nmr. *Proc Natl Acad Sci U S A* **2014,** *111* (4), E445-54.
8. Lang, P. T.; Holton, J. M.; Fraser, J. S.; Alber, T., Protein Structural Ensembles Are Revealed by Redefining X-Ray Electron Density Noise. *Proc Natl Acad Sci U S A* **2014,** *111* (1), 237-42.
9. Lang, P. T.; Ng, H. L.; Fraser, J. S.; Corn, J. E.; Echols, N.; Sales, M.; Holton, J. M.; Alber, T., Automated Electron-Density Sampling Reveals Widespread Conformational Polymorphism in Proteins. *Protein Sci* **2010,** *19* (7), 1420-31.
10. Fraser, J. S.; Clarkson, M. W.; Degnan, S. C.; Erion, R.; Kern, D.; Alber, T., Hidden Alternative Structures of Proline Isomerase Essential for Catalysis. *Nature* **2009,** *462*, 669-73.
11. Fraser, J. S.; van den Bedem, H.; Samelson, A. J.; Lang, P. T.; Holton, J. M.; Echols, N.; Alber, T., Accessing Protein Conformational Ensembles Using Room-Temperature X-Ray Crystallography. *Proc Natl Acad Sci U S A* **2011,** *108* (39), 16247-52.
12. Baldwin, R. L.; Rose, G. D., Molten Globules, Entropy-Driven Conformational Change and Protein Folding. *Curr Opin Struct Biol* **2013,** *23* (1), 4-10.
13. Stone, M. J., Nmr Relaxation Studies of the Role of Conformational Entropy in Protein Stability and Ligand Binding. *Accounts of Chemical Research* **2001,** *34* (5), 379-388.
14. Lee, A. L.; Kinnear, S. A.; Wand, A. J., Redistribution and Loss of Side Chain Entropy Upon Formation of a Calmodulin-Peptide Complex. *Nat Struct Biol* **2000,** *7* (1), 72-7.
15. Mittermaier, a.; Kay, L. E.; Forman-Kay, J. D., Analysis of Deuterium Relaxation-Derived Methyl Axis Order Parameters and Correlation with Local Structure. *Journal of biomolecular NMR* **1999,** *13*, 181-5.
16. Frederick, K. K.; Marlow, M. S.; Valentine, K. G.; Wand, a. J., Conformational Entropy in Molecular Recognition by Proteins. *Nature* **2007,** *448*, 325-9.

17. Tuttle, L. M.; Dyson, H. J.; Wright, P. E., Side-Chain Conformational Heterogeneity of Intermediates in the Escherichia Coli Dihydrofolate Reductase Catalytic Cycle. *Biochemistry* **2013,** *52* (20), 3464-3477.
18. Farès, C.; Lakomek, N.-A.; Walter, K. F. a.; Frank, B. T. C.; Meiler, J.; Becker, S.; Griesinger, C., Accessing Ns-Micros Side Chain Dynamics in Ubiquitin with Methyl Rdcs. *Journal of biomolecular NMR* **2009,** *45*, 23-44.
19. Henzler-Wildman, K. a.; Lei, M.; Thai, V.; Kerns, S. J.; Karplus, M.; Kern, D., A Hierarchy of Timescales in Protein Dynamics Is Linked to Enzyme Catalysis. *Nature* **2007,** *450*, 913-6.
20. Zhang, J.; Liu, J. S., On Side-Chain Conformational Entropy of Proteins. *PLoS computational biology* **2006,** *2*, e168.
21. Li, D.-W.; Brüschweiler, R., A Dictionary for Protein Side-Chain Entropies from Nmr Order Parameters. *Journal of the American Chemical Society* **2009,** *131*, 7226-7.
22. Shirts, M.; Pande, V. S., Computing: Screen Savers of the World Unite! *Science* **2000,** *290* (5498), 1903-4.
23. Shaw, D. E.; Deneroff, M. M.; Dror, R. O.; Kuskin, J. S.; Larson, R. H.; Salmon, J. K.; Young, C.; Batson, B.; Bowers, K. J.; Chao, J. C.; Eastwood, M. P.; Gagliardo, J.; Grossman, J. P.; Ho, C. R.; Ierardi, D. J.; Kolossváry, I.; Klepeis, J. L.; Layman, T.; McLeavey, C.; Moraes, M. A.; Mueller, R.; Priest, E. C.; Shan, Y.; Spengler, J.; Theobald, M.; Towles, B.; Wang, S. C., Anton, a Special-Purpose Machine for Molecular Dynamics Simulation. *Comm. ACM* **2008,** *51* (7), 91-97.
24. DuBay, K. H.; Geissler, P. L., Calculation of Proteins' Total Side-Chain Torsional Entropy and Its Influence on Protein-Ligand Interactions. *Journal of molecular biology* **2009,** *391*, 484-97.
25. Rosenbluth, M. N.; Rosenbluth, A. W., Monte Carlo Calculation of the Average Extension of Molecular Chains. *The Journal of Chemical Physics* **1955,** *23*, 356.
26. Shapovalov, M. V.; Dunbrack, R. L., A Smoothed Backbone-Dependent Rotamer Library for Proteins Derived from Adaptive Kernel Density Estimates and Regressions. *Structure (London, England : 1993)* **2011,** *19*, 844-58.
27. Lin, M. S.; Fawzi, N. L.; Head-Gordon, T., Hydrophobic Potential of Mean Force as a Solvation Function for Protein Structure Prediction. *Structure* **2007,** *15* (6), 727-40.
28. Röthlisberger, D.; Khersonsky, O.; Wollacott, A. M.; Jiang, L.; DeChancie, J.; Betker, J.; Gallaher, J. L.; Althoff, E. a.; Zanghellini, A.; Dym, O.; Albeck, S.; Houk, K. N.; Tawfik, D. S.; Baker, D., Kemp Elimination Catalysts by Computational Enzyme Design. *Nature* **2008,** *453*, 190-5.
29. Lin, M. S.; Head-Gordon, T., Reliable Protein Structure Refinement Using a Physical Energy Function. *J Comput Chem* **2011,** *32* (4), 709-17.
30. Clarkson, M. W.; Gilmore, S. A.; Edgell, M. H.; Lee, A. L., Dynamic Coupling and Allosteric Behavior in a Non-Allosteric Protein. *Biochemistry* **2006,** *45* (25), 7693-7699.
31. Tsai, J.; Bonneau, R.; Morozov, A. V.; Kuhlman, B.; Rohl, C. A.; Baker, D., An Improved Protein Decoy Set for Testing Energy Functions for Protein Structure Prediction. *Proteins-Structure Function and Genetics* **2003,** *53* (1), 76-87.
32. Ponder, J. W. *Tinker: Software Tools for Molecular Design*, 5.0; Washington University School of Medicine: Saint Louis, 2009.
33. Schmidt , J. M., Asymmetric Karplus Curves for the Protein Side-Chain 3 J Couplings. *Journal of Biomolecular NMR* **2007,** *37* (4), 287-301.

34. Qian, B.; Raman, S.; Das, R.; Bradley, P.; McCoy, A. J.; Read, R. J.; Baker, D., High-Resolution Structure Prediction and the Crystallographic Phase Problem. *Nature* **2007,** *450*, 259-64.
35. Leaver-Fay, A.; Tyka, M.; Lewis, S. M.; Lange, O. F.; Thompson, J.; Jacak, R.; Kaufman, K.; Renfrew, P. D.; Smith, C. A.; Sheffler, W.; Davis, I. W.; Cooper, S.; Treuille, A.; Mandell, D. J.; Richter, F.; Ban, Y. E.; Fleishman, S. J.; Corn, J. E.; Kim, D. E.; Lyskov, S.; Berrondo, M.; Mentzer, S.; Popovic, Z.; Havranek, J. J.; Karanicolas, J.; Das, R.; Meiler, J.; Kortemme, T.; Gray, J. J.; Kuhlman, B.; Baker, D.; Bradley, P., Rosetta3: An Object-Oriented Software Suite for the Simulation and Design of Macromolecules. *Methods Enzymol* **2011,** *487*, 545-74.
36. Zhou, H.; Zhou, Y., Distance-Scaled, Finite Ideal-Gas Reference State Improves Structure-Derived Potentials of Mean Force for Structure Selection and Stability Prediction. *Protein Sci* **2002,** *11* (11), 2714-26.
37. Shen, M. Y.; Sali, A., Statistical Potential for Assessment and Prediction of Protein Structures. *Protein Sci* **2006,** *15* (11), 2507-24.
38. Zhao, F.; Xu, J., A Position-Specific Distance-Dependent Statistical Potential for Protein Structure and Functional Study. *Structure (London, England : 1993)* **2012,** *20*, 1118-26.
39. Faver, J. C.; Benson, M. L.; He, X.; Roberts, B. P.; Wang, B.; Marshall, M. S.; Sherrill, C. D.; Merz, K. M., The Energy Computation Paradox and Ab Initio Protein Folding. *PloS one* **2011,** *6*, e18868.
40. Tyka, M. D.; Keedy, D. a.; André, I.; Dimaio, F.; Song, Y.; Richardson, D. C.; Richardson, J. S.; Baker, D., Alternate States of Proteins Revealed by Detailed Energy Landscape Mapping. *Journal of molecular biology* **2011,** *405*, 607-18.
41. Bower, M. J.; Cohen, F. E.; Dunbrack, R. L., Prediction of Protein Side-Chain Rotamers from a Backbone-Dependent Rotamer Library: A New Homology Modeling Tool. *Journal of molecular biology* **1997,** *267*, 1268-82.
42. Dasgupta, S.; Iyer, G. H.; Bryant, S. H.; Lawrence, C. E.; Bell, J. A., Extent and Nature of Contacts between Protein Molecules in Crystal Lattices and between Subunits of Protein Oligomers. *Proteins: Structure, Function and Genetics* **1997,** *28*, 494-514.
43. Batoulis, J.; Kremer, K., Statistical Properties of Biased Sampling Methods for Long Polymer Chains. *Journal of Physics A: Mathematical and ...* **1988,** *21*, 127.
44. Scheidig, A. J.; Burmester, C.; Goody, R. S., The Pre-Hydrolysis State of P21(Ras) in Complex with Gtp: New Insights into the Role of Water Molecules in the Gtp Hydrolysis Reaction of Ras-Like Proteins. *Structure Fold. Des.* **1999,** *7*, 1311-1324.
45. Stock, A. M.; Martinez-Hackert, E.; Rasmussen, B. F.; West, A. H.; Stock, J. B.; Ringe, D.; Petsko, G. A., Structure of the Mg(2+)-Bound Form of Chey and Mechanism of Phosphoryl Transfer in Bacterial Chemotaxis. *Biochemistry* **1993,** *32*, 13375-13380.
46. Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E., Ucsf Chimera--a Visualization System for Exploratory Research and Analysis. *J. Comp. Chem.* **2004,** *25* (13), 1605-1612.

2.6 APPENDIX

Protein energy function. The protein energy function is based on the standard AMBER functional form, but for use with rotamer libraries it requires minor additional specification. The Van der Waals dispersion energy is defined as

$$E_{vdw} = \begin{cases} \infty, & r_{ij} < r_{ij}^{*} \\ 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right], & r_{ij} > r_{ij}^{*} \end{cases} \tag{S1}$$

where $r_{ij}^{*} = \alpha(r_i + r_j)$ and α is a clash check parameter that is set to 0.8 in this work. The Coulomb electrostatic energy is

$$E_{elec} = \sum_{\substack{i,j \\ i<j}} \frac{1}{\varepsilon_P}\frac{q_i q_j}{r_{ij}} \tag{S2}$$

where q_i and q_j are the partial charges of atoms i and j, respectively, and ε_P is a dielectric constant that is set to 4.

The GB-HPMF solvation model. The Generalized Born (GB) model(Onufriev et al., 2004) treats the electrostatic polarization energy as the interaction between the protein's charge distribution enclosed in a low dielectric region and the reaction potential it induces in the surrounding high dielectric solvent.

$$F_{GB} = -\frac{1}{2}\left(\frac{1}{\varepsilon_P} - \frac{1}{\varepsilon_W}\right)\sum_{i,j}\frac{q_i q_j}{f_{ij}^{GB}(r_{ij})} \tag{S3}$$

where f_{ij}^{GB} is a continuous function defined as

$$f_{ij}^{GB} = \left[r_{ij}^2 + R_i R_j \exp\left(\frac{-r_{ij}^2}{4R_i R_j}\right)\right]^{\frac{1}{2}} \tag{S4}$$

Here R_i and R_j are the born radii of atoms i and j, respectively, evaluated using the method proposed by Onufriev et. al and references therein(Onufriev et al., 2004).

The hydrophobic potential of mean force (HPMF) implicitly models the influence of water on the free energy of interaction between two small hydrophobic groups(Sorenson et al., 1999). The resulting potential of mean force exhibits two minima separated by a barrier: one for the hydrophobic molecules in contact and one for the hydrophobic groups separated by a water layer, which we calculate using a simple Gaussian functional form

$$F_{HPMF} = \sum_{i \in SA_i > A_c}^{N_c} \tanh(SA_i) \sum_{j \in SA_j > A_c}^{N_c} \tanh(SA_j) \sum_{k=1}^{3} h_k \exp\left(-\left[\frac{r_{ij} - c_k}{w_k}\right]^2\right) \tag{S5}$$

The parameters of this model and other details are available in Ref [(Lin et al., 2007)].

Local Minimization Procedure. The PDB conformations are given as Cartesian coordinates of heavy atoms, but real proteins and our atomistic energy function require that hydrogen positions to be specified as well. We use the Amber program(Case et al., 2010) to build the positions of the hydrogen atoms on the native PDB structures. Both the resulting native and all decoy structures are then optimized to their nearest local minimum using the L-BFGS (Broyden-Fletcher-Goldfarb-Shanno) limited memory quasi-Newton method(Liu and Nocedal, 1989; Press et al., 1992). Energies and derivatives are defined by the total energy, which is the sum of Eqs. (1) and (5) in the main paper.

Molecular dynamics relaxation. We ran MD simulations of several proteins using the Amber ff99SB force field(Hornak et al., 2006) and aqueous solvent represented by the TIP4P-Ew water model(Horn et al., 2004), which we chose because previous studies support its clear superiority relative to other biomolecular simulation force fields(Fawzi et al., 2008; Sgourakis et al., 2011; Wickstrom et al., 2009). We simulated each protein in a cubic box containing sufficient (multiple layers) water and counter ions to neutralize any protein net charge at 300K and 1atm. The pmemd module of AMBER(Case et al., 2010) was used to generate up to 4ns of NPT data for each protein; the short trajectories were intentional since the purpose of the simulation is to relax any residual effects of the X-ray crystal environment on the backbone structure(Tyka et al., 2011).

Table S1: *Dihedral Angle definitions used in MC-SCE to grow side chains of amino acids.* Standard bond lengths/angles were used to place side chain atoms. No attempt was made to regrow the C_β atom.

Residue	χ angle	Atoms	Residue	χ angle	Atoms
ARG	χ_1	N-C_α-C_β-C_γ	LEU	χ_1	N-C_α-C_β-C_γ
	χ_2	C_α-C_β-C_γ-C_δ		χ_2	C_α-C_β-C_γ-$C_{\delta1}$
	χ_3	C_β-C_γ-C_δ-N_ε	LYS	χ_1	N-C_α-C_β-C_γ
	χ_4	C_γ-C_δ-N_ε-C_ζ		χ_2	C_α-C_β-C_γ-C_δ
	χ_5	C_δ-N_ε-C_ζ-$N_{\omega1}$ (0°)		χ_3	C_β-C_γ-C_δ-C_ε
ASN	χ_1	N-C_α-C_β-C_γ		χ_4	C_γ-C_δ-C_ε-N_ζ
	χ_2	C_α-C_β-C_γ-$O_{\delta1}$	MET	χ_1	N-C_α-C_β-C_γ
ASP	χ_1	N-C_α-C_β-C_γ		χ_2	C_α-C_β-C_γ-S_δ
	χ_2	C_α-C_β-C_γ-$O_{\delta1}$		χ_3	C_β-C_γ-S_δ-C_ε
CYS	χ_1	N-C_α-C_β-S_γ	PHE	χ_1	N-C_α-C_β-C_γ
GLN	χ_1	N-C_α-C_β-C_γ		χ_2	C_α-C_β-C_γ-$C_{\delta1}$
	χ_2	C_α-C_β-C_γ-C_δ	PRO	χ_1	N-C_α-C_β-C_γ
	χ_3	C_β-C_γ-C_δ-$O_{\varepsilon1}$		χ_2	C_α-C_β-C_γ-C_δ
GLU	χ_1	N-C_α-C_β-C_γ	SER	χ_1	N-C_α-C_β-O_γ
	χ_2	C_α-C_β-C_γ-C_δ	THR	χ_1	N-C_α-C_β-$O_{\gamma1}$
	χ_3	C_β-C_γ-C_δ-$O_{\varepsilon1}$	TRP	χ_1	N-C_α-C_β-C_γ
HIS	χ_1	N-C_α-C_β-C_γ		χ_2	C_α-C_β-C_γ-$C_{\delta1}$
	χ_2	C_α-C_β-C_γ-$N_{\delta1}$	TYR	χ_1	N-C_α-C_β-C_γ
ILE	χ_1	N-C_α-C_β-$C_{\gamma1}$		χ_2	C_α-C_β-C_γ-$C_{\delta1}$
	χ_2	C_α-C_β-$C_{\gamma1}$-C_δ	VAL	χ_1	N-C_α-C_β-$C_{\gamma1}$

Table S2: χ_1 *rotamer population data for protein Eglin C as calculated by MC-SCE.* The experimental populations were taken from the reported crystal structure (PDB: 1CSE) and estimates from NMR J-coupling experiments (Clarkson et al., 2006). MC-SCE calculations were done on the deposited crystal structure backbone, a backrub ensemble of the crystal structure (BR-CC) and on backbones thermalized by MD at time points of 0.2, 2.0 and 4.0 ns.

Eglin C Residue		*Experimental Population*		*MC-SCE Population*	
	χ_1 Class	*CC χ_1*	*NMR χ_1*	*CC backbone*	*Thermal backbone*
Val13	60		**8.0**		<1.0
	180	**100.0**	**55.0**	98.0	54.0
	-60		**37.0**	2.0	45.0
Val14	60		**4.0**		<1.0
	180	**100.0**	**90.0**	100.0	82.0
	-60		**6.0**		17.0
Thr17	60	**100.0**	**36.0**	99.0	100.0
	180		**64.0**		
	-60			1.0	
Val18	60				
	180	**100.0**	**86.0**	100.0	74.0
	-60		**14.0**		26.0
Thr26	60		**81.0**		<1.0
	180		**9.0**		<1.0
	-60	**100.0**	**10.0**	100.0	98.0
Val34	60	**100.0**	**21.0**		<1.0
	180		**75.0**	98.0	58.0
	-60		**4.0**	2.0	41.0
Val43	60		**10.0**		15.0
	180		**10.0**		4.0
	-60	**100.0**	**80.0**	100.0	81.0
Val52	60		**17.0**		
	180	**100.0**	**81.0**	100.0	99.0
	-60		**2.0**		<1.0
Val54	60		**13.0**		
	180	**100.0**	**87.0**	100.0	100.0
	-60				
Val62	60		**9.0**		
	180	**100.0**	**83.0**	99.0	99.0
	-60		**8.0**	1.0	<1.0
Val63	60		**3.0**		
	180	**100.0**	**94.0**	100.0	100.0
	-60		**3.0**		
Val66	60				
	180	**100.0**	**91.0**	100.0	100.0
	-60		**9.0**		

Table S3: χ_1 *rotamer population data for certain residues in DHFR complex E:THF as calculated by MC-SCE.* The experimental populations were taken from the deposited crystal structure (PDB: 1RX5) and estimates from NMR J-coupling experiments(Tuttle et al., 2013). The MC-SCE calculations were done on the crystal structure backbone with the ligand docked in the crystallographic conformation (CC) as well as MD snapshots of enzyme-ligand complex taken at time points of 0.2, 2 and 4 ns.

Residue		*Experimental Population*		*MC-SCE Population*	
	χ_1 Class	*CC χ_1*	*NMR χ_1*	*CC backbone*	*Thermal backbone*
Ile2	60		**19.0**		
	180				
	-60	**100.0**	**81.0**	100.0	100.0
Ile5	60		**12.0**		
	180	**100.0**	**75.0**	86.0	100.0
	-60		**13.0**	14.0	
Ile14	60	**100.0**	**91.0**	55.8	34.0
	180		**9.0**	44.2	66.0
	-60				
Ile41	60		**10.0**		
	180			7.0	
	-60	**100.0**	**90.0**	93.0	100.0
Ile50	60		**29.0**	2.3	45.0
	180				
	-60	**100.0**	**71.0**	97.7	55.0
Ile60	60		**9.0**		
	180		**8.0**		
	-60	**100.0**	**83.0**	100.0	100.0
Ile61	60		**19.0**		
	180				
	-60	**100.0**	**81.0**	100.0	100.0
Ile82	60		**18.0**		4.0
	180				
	-60	**100.0**	**82.0**	100.0	96.0
Ile91	60		**14.0**		
	180				
	-60	**100.0**	**86.0**	100.0	100.0
Ile94	60	**100.0**	**59.0**	100.0	92.0
	180				
	-60		**41.0**		8.0
Ile115	60		**11.0**		
	180				17.0
	-60	**100.0**	**89.0**	100.0	83.0
Ile155	60		**9.0**		8.0
	180		**5.0**		<1.0
	-60	**100.0**	**85.0**	100.0	91.0

Table S3: *χ_1 rotamer population data for DHFR complex E:THF as calculated by MC-SCE (continued)*

Residue		*Experimental Population*		*MC-SCE Population*	
	χ_1 Class	*CC χ_1*	*NMR χ_1*	*CC backbone*	*Thermal backbone*
Val10	60				
	180	**100.0**	**74.0**	95.4	100.0
	-60		**26.0**	4.6	
Val13	60		**5.0**		
	180	**100.0**	**76.0**	100.0	100.0
	-60		**19.0**		
Val40	60			11.6	<1.0
	180		**20.0**	83.7	92.0
	-60	**100.0**	**80.0**	4.6	7.0
Val72	60		**4.0**		
	180		**6.0**		
	-60	**100.0**	**90.0**	100.0	100.0
Val75	60	**100.0**	**3.0**		
	180		**23.0**		33.0
	-60		**74.0**	100.0	67.0
Val78	60				
	180	**100.0**	**82.0**	100.0	100.0
	-60		**17.0**		
Val88	60		**3.0**		4.0
	180		**1.0**		6.0
	-60	**100.0**	**96.0**	100.0	90.0
Val93	60				
	180	**100.0**	**72.0**	100.0	94.0
	-60		**28.0**		6.0
Val99	60		**3.0**		
	180	**100.0**	**90.0**	100.0	100.0
	-60		**7.0**		
Val119	60		**29.0**	67.4	27.0
	180		**37.0**		33.0
	-60	**100.0**	**34.0**	32.6	40.0
Val136	60		**9.0**		
	180		**6.0**		<1.0
	-60	**100.0**	**86.0**	100.0	99.0
Tyr100	60		**8.0**		
	180				28.0
	-60	**100.0**	**92.0**	100.0	72.0
Tyr111	60		**1.0**		
	180				8.0
	-60	**100.0**	**99.0**	100.0	92.0
Tyr128	60	**100.0**	**88.0**	100.0	96.0

	180				
	-60		**12.0**		4.0
Tyr151	60	**100.0**	**94.0**	100.0	100.0
	180		**6.0**		
	-60				

Table S3: *χ_1 rotamer population data for DHFR complex E:THF as calculated by MC-SCE (continued)*

Residue type - THR		*Experimental Population*		*MC-SCE Populations*	
	χ_1 Class	*CC χ_1*	*NMR χ_1*	*CC backbone*	*Thermal backbone*
Thr35	60	**100.0**	**84.0**		100.0
	180		**11.0**		
	-60		**6.0**	100.0	
Thr46	60		**28.0**	32.6	
	180				
	-60	**100.0**	**72.0**	67.4	100.0
Thr68	60	**100.0**	**84.0**	100.0	100.0
	180		**9.0**		
	-60		**7.0**		
Thr73	60		**22.0**		
	180				
	-60	**100.0**	**78.0**	100.0	100.0
Thr113	60				1.0
	180				
	-60	**100.0**		100.0	99.0
Thr123	60	**100.0**		100.0	32.0
	180		**36.0**		
	-60		**64.0**		68.0
His45	60				26.0
	180			97.7	62.0
	-60	**100.0**		2.3	12.0
His114	60		**6.0**		
	180			53.5	88.0
	-60	**100.0**	**94.0**	46.5	12.0
His124	60		**2.0**		2.0
	180		**26.0**		
	-60	**100.0**	**71.0**	100.0	98.0
His141	60		**13.0**		
	180				1.0
	-60	**100.0**	**87.0**	100.0	99.0
His149	60		**1.0**		
	180				
	-60	**100.0**	**99.0**	100.0	100.0

Table S3: *χ_1 rotamer population data for DHFR complex E:THF as calculated by MC-SCE (continued)*

Residue type -TRP		*Experimental Population*		*MC-SCE Populations*	
	χ_1 Class	*CC χ_1*	*NMR χ_1*	*CC backbone*	*Thermal backbone*
Trp22	60	**100.0**	**85.0**	100.0	100.0
	180		**15.0**		
	-60				
Trp30	60				
	180	**100.0**	**87.0**	95.4	78.0
	-60		**13.0**	4.6	22.0
Trp47	60		**8.0**		
	180	**100.0**	**92.0**	100.0	100.0
	-60				
Trp74	60		**3.0**		7.0
	180				26.0
	-60	**100.0**	**97.0**	100.0	67.0
Trp133	60		**20.0**		
	180				
	-60	**100.0**	**80.0**	100.0	100.0
Phe31	60				
	180	**100.0**	**98.0**	100.0	100.0
	-60		**2.0**		
Phe103	60		**8.0**		
	180				
	-60	**100.0**	**92.0**	100.0	100.0
Phe125	60				
	180	**100.0**		69.8	34.0
	-60			30.2	66.0
Phe137	60		**30.0**		
	180	**100.0**	**70.0**	100.0	89.0
	-60				11.0
Phe140	60		**17.0**		
	180	**100.0**	**54.0**	100.0	90.0
	-60		**30.0**		10.0
Phe153	60		**3.0**		
	180		**2.0**	4.7	14.0
	-60	**100.0**	**95.0**	95.3	86.0

Table S4. *Thermodynamic rankings and Z-scores of the native X-ray crystal structure relative to low RMSD misfolded structures for 55 of the 57 Rosetta protein decoy sets.* Our MC-SCE algorithm was unable to find any non-clashing side chains for 2 proteins in the original Rosetta decoy set and thus have not been reported. Definitions for the thermodynamic quantities are given in the text. Proteins with asterisks are discussed in the text and some analyzed in more detail in Table 2.1 of the main text.

Protein	E_{single} ***Rank***	E_{single} ***Z-score***	E_{best} ***Rank***	E_{best} ***Z-score***	***F Rank***	***F Z-score***	$\%\ \chi_1$ ***changes***
1a19	1	-3.11	1	-3.33	1	-3.34	23.4
1a32	1	-3.29	1	-3.53	1	-2.74	27.6
1a68	1	-3.17	1	-2.72	1	-2.93	24.7
1acf	1	-2.94	1	-3.61	1	-3.63	21.1
1ail*	62	0.31	9	-1.22	3	-1.73	35.5
1aiu	1	-2.3	1	-2.95	1	-2.9	26.1
1b3a	1	-3.63	1	-3.3	1	-3.34	18.0
1bgf	1	-4.18	1	-4.33	1	-4.07	26.9
1bk2	1	-5.69	1	-5.15	1	-4.98	21.6
1bkr	1	-5.98	1	-5.7	1	-4.48	24.7
1bm8	1	-2.89	1	-2.78	1	-2.72	18.8
1bq9	1	-4.49	1	-2.83	1	-2.55	25.6
1c8c*	47	-0.21	3	-2.01	2	-2.34	23.5
1c9o	1	-3.47	1	-3.84	1	-3.9	20.7
1cc8	1	-4.02	1	-3.71	1	-3.4	15.1
1cei	1	-3.03	1	-3.19	1	-2.76	32.0
1cg5	1	-3.47	1	-5.8	1	-5.61	22.9
1ctf	1	-3.8	1	-3.71	1	-3.74	27.6
1dhn*	2	-2.1	2	-2.4	2	-2.01	26.7
1e6i	1	-2.84	1	-2.58	1	-2.38	19.8
1elw	1	-4.96	1	-5.11	1	-3.25	28.1
1enh*	81	1.12	14	-1.03	13	-1.02	44.9
1ew4	1	-3.46	1	-3.94	1	-3.34	30.4
1eyv	1	-4.38	1	-5.13	1	-4.02	27.7
1fkb	1	-4.93	1	-4.3	1	-4.32	22.4
1gvp*	8	-1.55	21	-0.81	18	-1.04	38.2
1hz6*	7	-1.77	3	-2.03	3	-2.22	17.0
1ig5	8	-1.42	1	-3.12	1	-2.61	27.9
1iib	1	-3.17	1	-3.77	1	-3.65	26.8
1kpe	1	-5.11	1	-5.23	1	-5.16	23.3
1lis	1	-2.01	1	-4.71	1	-3.86	21.1
1lou	1	-3.81	1	-3.87	1	-3.62	26.8
1nps	1	-4.29	1	-3.74	1	-3.44	13.6
1opd	1	-2.41	1	-3.17	1	-2.27	21.7
1pgx*	2	-1.84	3	-2.29	2	-2.23	20.0
1ptq	1	-2.36	1	-3.19	1	-2.72	22.2
1rnb*	93	1.49	90	1.18	89	1.17	19.6

1scj	1	-2.59	1	-3.00	1	-2.74	38.2
1shf	1	-3.3	1	-4.63	1	-4.52	27.4
1ten	1	-5.27	1	-6.23	1	-5.97	31.2
1tig	2	-2.02	1	-4.08	1	-3.23	26.3
1tul	1	-3.76	1	-5.08	1	-4.78	22.5
1ubi*	9	-1.33	10	-1.23	5	-1.38	18.4
1ugh	1	-3.18	1	-4.16	1	-3.98	25.0
1urn	1	-4.27	1	-3.71	1	-3.79	23.4
1utg*	94	1.45	81	0.94	75	0.73	18.4
1vcc*	4	-1.71	3	-1.99	2	-2.06	21.4
1vie	1	-5.63	1	-5.29	1	-5.01	14.3
1vls*	1	-2.23	75	0.45	98	2.11	31.1
1who	1	-4.67	1	-4.59	1	-4.2	22.5
2acy	1	-4.72	1	-4.62	1	-4.07	26.2
2chf	15	-1.03	1	-2.49	1	-2.32	24.3
2ci2	4	-1.46	2	-1.89	1	-2.17	33.9
5cro	1	-3.85	1	-3.32	1	-3.54	36.3
256b	1	-3.72	1	-2.07	3	-1.65	30.0

Chapter 3

The role of side chain entropy and mutual information for improving the *de novo* design of Kemp eliminases KE07 and KE70

Side chain entropy and mutual entropy information between residue pairs have been calculated for two *de novo* designed Kemp eliminase enzymes, KE07 and KE70, and for their most improved versions at the end of laboratory directed evolution (LDE). It was found that entropy, not just enthalpy, helped to destabilize the preference for the reactant state complex of the designed enzyme as well as favoring stabilization of the transition state complex for the best LDE enzymes. Furthermore, residues with the highest side chain couplings as measured by mutual information, when experimentally mutated, were found to diminish or annihilate catalytic activity, some of which were far from the active site. In summary, these findings demonstrate how side chain fluctuations and their coupling can be an important design feature for *de novo* enzymes, and furthermore could be utilized in the computational steps in lieu of or in addition to the LDE steps in future enzyme design projects as will be shown in Chapter 4. This chapter is based on the following publication

A. Bhowmick, S. Sharma, H. Honma, T. Head-Gordon (2016). The role of side chain entropy and mutual information for improving the de novo design of Kemp Eliminases KE07 and KE70. *Phys. Chem. Chem. Phys., 18, 19386*

3.1 INTRODUCTION

The ability to control for protein structure, energetics and dynamical motions is a fundamental problem that limits our ability to rationally design catalysts for new chemical reactions not known to have a natural biocatalyst. Current computational approaches for *de novo* enzyme design seek to engineer a small catalytic construct into an accommodating protein scaffold, as exemplified by the Rosetta strategy applied to the design of many different catalytic motifs[1, 2]. In this study we consider the Rosetta design of the Kemp elimination reaction[3] involving the deprotonation of a small ligand substrate (5-nitro benzisoxazole) by a base (Figure 3.1a), in which the designed catalytic construct was engineered into a TIM barrel scaffold[2]. Two well-studied *de novo* enzymes for this reaction are KE07 and KE70, in which some minimal activity was observed in the designed enzymes and proved an important validation of the Rosetta approach. Nonetheless the catalytic activity was very low, and a number of follow-on studies have provided some important insight into the active site energetic features that limited the catalytic activity of the original designs of KE07 and KE70[4-7].

What proved more beneficial to improving the catalytic performance of KE07 and KE70 was application of laboratory directed evolution (LDE)[8-10], an experimental strategy based on the principle of natural selection[11]. The goal of LDE is to alter the protein sequence through multiple rounds of mutagenesis and selection to isolate the few new sequences that exhibit enhanced catalytic performance. Given the limitations of our understanding of the structure-function relationship[12], LDE provides an attractive alternative to rational design approaches to

biocatalysis, is highly flexible in application to different biocatalysis reactions, and provides an effective way of improving upon *de novo* enzymes generated from computational designs[9, 13]. Although LDE can be an opaque process because it offers no direct rationale as to why mutations are successful, many hypotheses and useful heuristics have been proposed and developed for improving binding selectivity or protein stability using LDE[14-17]. For example, previous efforts to rationalize and ultimately decrease the sequence space for LDE focused on the interplay of sequence site entropy, i.e. the plasticity for evolutionary-driven substitutions, and the likelihood that these sites would thus be more prone to increased structural flexibility[18, 19], and which was borne out by mutations that reduced the entropy of these sites[20, 21]. For KE07 and KE70, LDE improved the Michaelis-Menten specificity constant k_{cat}/K_M by a factor of ~200 and ~400, respectively, in the best evolved enzymes.

The primary question we address in this work is what is missing in the original computational *de novo* design that is captured instead during the LDE process to improve the Michaelis-Menten specificity constant k_{cat}/K_M for KE07 and KE70? Using the framework of transition state theory[22], biocatalytic improvements as measured by k_{cat}/K_M should arise through reduction in the activation free energy, $\Delta G_T^\dagger = \Delta G^\dagger + \Delta G_{EL}$, where E and L represent the enzyme and ligand, respectively. The activation free energy is comprised of a positive $\Delta G^\dagger = G_{EL^\dagger} - G_{EL}$ that quantifies the catalytic barrier between the reactant EL state and transition $EL^\dagger$ complex, and therefore relates directly to k_{cat}; in addition $\Delta G_{EL} = G_{EL} - G_{E+L}$ measures the binding affinity of the ligand to the enzyme active site and thus relates to K_M. Therefore knowing $\Delta G_T^\dagger$ or the activation enthalpy, $\Delta H_T^\dagger$, and activation entropy, $\Delta S_T^\dagger$, components, we can connect directly to the k_{cat}/K_M ratio through

$$\frac{k_{cat}}{K_M} = \frac{kT}{h}\exp\left(\frac{-\Delta G_T^\dagger}{RT}\right) = \frac{kT}{h}\exp\left(\frac{-\Delta H_T^\dagger}{RT}\right)\exp\left(\frac{\Delta S_T^\dagger}{R}\right) \quad (1)$$

and therefore the success of the LDE process applied to KE07 and KE70 must have a rational thermodynamic basis via Eq. (1).

While it is broadly accepted that optimizing enthalpic interactions is paramount for good substrate binding and lowering of the transition state barrier to the chemical reaction, the role of dynamics for improving catalytic performance is more controversial. One aspect of the controversy pertains to the definition of dynamics, for example whether it refers to equilibrium statistical fluctuations[23-25], dynamical coupling[26] and/or maximizing the reactive flux through the transition state surface[27]. Probably the most commonly implied definition of important functional motions for biocatalysis is a thermodynamic one, i.e. statistical fluctuations that are embodied in an entropy change that along with enthalpy contributes to the changes in the free energy state function as per Eq. (1).

In order to support the design of good enthalpic interactions between the substrate and the enzyme, it would seem desirable to impose some limits on the conformational flexibility to aid the catalytic function[28, 29]. A survey of 178 enzymes led to the conclusion that active site residues of naturally occurring enzymes are the least flexible within a sequence, supported by their low B-factors in the crystalline environment[30]. At the same time, evidence also exists that increased conformational flexibility can also be a factor in improved biocatalytic performance. Room temperature X-ray crystallography[31], in good agreement with NMR[32, 33], has shown that protein interiors are very fluid, especially at the level of side chain motions, and that alternate side chain conformers in ligand binding and catalysis can be critical for function[34], and conformational

flexibility forms the basis of computational approaches to conformational selection in allostery[35-37]. Hence, even though configurational entropy may well be important for biocatalysis, it still remains poorly understood how statistical fluctuations can be utilized to improve the *de novo* design process.

In this study we consider the question of how LDE improvements in the catalytic activity of KE07 and KE70 changes the active site energetics as well as side chain entropy and side chain coupling captured through mutual information. We find that the best KE07 and KE70 enzymes at the end of LDE process exhibit enthalpies *and* entropies that both destabilize the reactive state and stabilize the transition state with respect to the designed enzymes, showing that the original enzymes were over-designed for the EL reactant state, whereas the LDE process created enzymes that preferred the EL† complex instead, especially for the KE70 enzyme. Furthermore, we find that residues with the highest mutual information proved to be critical for enzyme catalysis, which we tested on the best evolved enzyme for KE07. We show that new amino acid chemistries with high mutual information in the active site, some of which have not been reported in previous studies of the same enzyme, proved critical to function since experimental mutations at these sites destroyed enzyme activity. Of greater interest is that other residues identified as having high mutual information that are far from the active site were found to diminish or annihilate catalytic activity when mutated in the best evolved KE07 enzyme. In summary, our findings demonstrate how differences in not only energetics, but side chain fluctuations and their coupling, can be an important design feature for *de novo* enzymes, and furthermore could be utilized in future computational enzyme design projects.

Transition State Theory

We rely on the analysis of enzyme performance using transition state theory via Eq. (1)[22]. For the calculation of the enthalpy, we assume that the PV term is negligible such that it can be quantified using only potential energy calculations. We therefore calculate all protein-protein interactions for KE07 and KE70 using the generalized Amber force field, while the model for all protein and 5-nitro benzisoxazole interactions with aqueous solvent is based on our GB-HPMF implicit solvent model, which has been well-validated in previous work[38,39]. We use electrostatic models of the 5-nitro benzisoxazole ligand in the reactant state and transition state based on partial charges as reported by Frushicheva and co-workers[6], and long molecular dynamics calculations have confirmed that the ligand charges in the two states are compatible and thus stable within the protein modeled using a classical force field. The state enthalpy is evaluated as an average across an ensemble of backbone conformations, each of which has a large ensemble of side chain packings, such that we define $H = \langle H \rangle_{SC,BB}$ for a given state: the EL† complex, the EL complex, and apo state of the enzyme E.

The state entropy term defined in Eq. (1) can be further decomposed into sums over (i) contributions from the individual residues in the enzyme, as well as (ii) contributions from correlated motion between side chains of residues[40, 41], averaged over the backbone configurations

$$S \sim \sum_{i}^{N_{res}} \left\langle S^{(i)} \right\rangle_{SC,BB} - \sum_{i}^{N_{res}-1} \sum_{j=i+1}^{N_{res}} \left\langle I^{(i,j)} \right\rangle_{SC,BB} + ... \qquad (2)$$

and similarly Eq. (2) can be used to define the entropy of EL†, EL, and E states. Thus, we see that the catalytic power of an enzyme as measured from k_{cat}/K_M, can ultimately be related to

entropy contributions from individual residues, mutual information between residue pairs, or even higher order correlations, when defining the total entropy change.

3.2 MATERIALS AND METHODS

Computational Methods

Generating backbone ensembles for the apo, EL and EL† states of KE07 and KE70: Although we mostly focused on the two end state sequences, i.e. the two designed enzymes and the final LDE rounds for KE07 and KE70, some results in the SI material also consider the intermediate rounds of LDE for each of the enzymes. The initial backbone structures and initial definition of the side chain rotameric state of the KE07 apo enzyme for the initial design and LDE rounds 4 and 6 were taken from the PDB database[42]. Apo state structures for rounds without PDB structures were generated using Modeller with the KE07 design as the backbone/side chain template. For KE70, the apo structure of the initial design was taken from the computational model reported elsewhere[2]. For round 2, the apo state structure was taken from the PDB (ID: 3NPX) and rounds 4, 5 and 6 variants were generated by Modeller using the KE70 design as template.

Modeller was used to generate the EL state structure using the apo state as the template for the original designs and all LDE rounds for KE07 and KE70. For the EL state of the KE07 and KE70 designs, we used the docked structure definition of the ligand as reported elsewhere[2]. The ligand was then kept fixed in its modeled position for all subsequent backbone perturbations and MC-SCE calculations. The substrate geometry for the EL† state was kept the same as in the EL complex, and only TS charges were changed to reflect the transition state of the bound complex.

Using each of these PDB/modeled structures for the backbone in the apo and ligand bound states, we then used the backrub algorithm implemented in Rosetta[21] to run 50 independent simulations, each generating 10,000 trial moves using the C_α atoms as pivot residues, to generate uncorrelated backbone ensembles. From each simulation the lowest energy structure was saved and these 50 low energy backrub structures were selected, and divided into 5 backbone ensembles with 10 structures in each ensemble; this was done for all the rounds for both apo and ligand bound states. Since the backbone scaffolds for KE07 and KE70 are quite rigid, we believe the backbone variations we have generated are adequate.

Generating side chain ensembles for the apo, EL and EL† states of KE07 and KE70: We have recently developed a Monte Carlo Side Chain Ensemble method (MC-SCE)[43] to create large side chain ensembles to calculate the terms in Eq. (2). The MC-SCE method has been validated across a large number of proteins and protein complexes, in which it was found to be highly accurate when compared against high quality X-ray crystallography and NMR J-coupling data for side chain rotameric preferences[43]. The MC-SCE use a Rosenbluth chain growth algorithm to generate an ensemble of side chain packings for a given protein backbone. From the bare backbone conformation m, and for subsequent steps i, the side chain rotamer, r_k, for residue k is selected according to the following probability

$$p_i^{(m,r_k)} = \frac{P_{r_k}^{(PDB)} e^{-\beta E_i^{(m,r_k)}}}{\sum_{\{v_k\}} P_{v_k}^{(PDB)} e^{-\beta E_i^{(m,v_k)}}} \tag{3}$$

where $\{v_k\}$[44] are the possible side chain conformations for residue k, using the values reported in the recent backbone-dependent Dunbrack library[45], which we have augmented by allowing for

dihedral angle variations that are Gaussian distributed about a given rotamer value and weighted by its probability of occurrence in the PDB, $P_{r_k}^{(PDB)}$. $E_i^{(m,r_k)}$ is the energy of interaction of side chain conformation r_k of residue k with the backbone and all protein side chains grown so far (step i), using the energy function described above, and all residues are grown with ideal bond lengths and angles. Once the side chain of a residue is placed, the process is repeated until all the side chains are grown, thereby creating one complete protein structure. Each complete structure m is then assigned a weight $W(m)$ in order to adjust for sampling bias due to the chain growth as well as to account for energetic solvent effects

$$W(m) = e^{-\beta F_{solv}^{(m)}} \prod_{i=1}^{N} \frac{\sum_{\{v_k\}} P_{v_k}^{(PDB)} e^{-\beta E_i^{(m,v_k)}}}{P_{v_k}^{(PDB)}} \tag{4}$$

For unsuccessful chain growths, the partially grown structure is considered dead and its weight is set to zero. This process is repeated in order to create ~20,000 side chain ensemble on the given backbone.

Since we use a total of 5 independent backbone ensembles, each comprised of 10 backbones, our ensemble for each state are comprised of a total of 1,000,000 fully grown structures. For each of the independent backbone ensembles we calculate the probability $p_{v_k}^{(k)}$ of each rotameric state ν_k using equation (5)

$$p_{v_k}^{(k)} = \frac{\sum_{m=1}^{M} W(m) \delta_{r_k,v_k}^{(m)}}{\sum_{m=1}^{M} W(m)} \tag{5}$$

where M=200,000 and the Kronecker delta is 1 if the side chain conformation r_k that was picked for the residue k in the m-th structure is v_k and 0 otherwise. The probabilities in Eq. (5) are then used to calculate side chain entropy (SCE) of each residue k using the Gibbs probabilistic definition, with SCE values in units of k_BT.

$$S_{SC}^{(k)} = \sum_{\{v_k\}} p_{v_k}^{(k)} \log p_{v_k}^{(k)} \tag{6}$$

We estimated the mean and standard deviation for the SCE values from the 5 independent backbone ensembles for the apo, EL and EL for each protein for each round.

Given our MC-SCE method, we can also calculate mutual information, $I^{(i,j)}$. It is defined as the amount of information residue k has about another residue j based on the amount of coupled side chain dihedral angle fluctuations. In units of k_BT, this can be written as

$$I_{SC}^{(k,j)} = \sum_{\{v_k\}} \sum_{\{v_j\}} p_{v_k,v_j}^{(k,j)} \log \left(\frac{p_{v_k,v_j}^{(k,j)}}{p_{v_k}^{(k)} p_{v_j}^{(j)}} \right) \tag{7}$$

where in analogy to Eq. (5)

$$p_{v_k,v_j}^{(k,j)} = \frac{\sum_{m=1}^{M} W(m) \delta_{r_k,v_k}^{(m)} \delta_{r_j,v_j}^{(m)}}{\sum_{m=1}^{M} W(m)} \tag{8}$$

Thus Eq. (7) can be further simplified to

$$I_{SC}^{(k,j)} = (S_{SC}^{(k)} + S_{SC}^{(j)}) - S_{SC}^{(k,j)} \quad (9)$$

in which the individual entropy $S_{SC}^{(k)}$ and joint entropy, $I_{SC}^{(k,j)}$, is calculated using the probabilistic definition of entropy via Eq. (6), and thus Eq. (9) can be interpreted as the degree of coupling of torsional motions of residues *k* and *j*.

In practice, a background error persists in mutual information calculations since two completely uncorrelated variables will never be zero given a finite simulation time. In order to correct for this, we modified the strategy used by Dubay and Geissler[37] to subtract out the erroneous extra mutual information that persists due to finite time scales. We first carry out our MC-SCE chain growth with the full energy function over all backbones in an ensemble, and using Eq. (8) we calculate the mutual information for the *N* structures obtained using the complete energy model, $I_{SC}^{(k,j)}$.

We then use our MC-SCE method to create structures where side chains for each residue are grown independent of the environment, i.e. clashes are ignored and the energy (and hence probability of chain growth) of each side chain conformer ν_k of residue k is given by

$$-\beta E_{uncorr}^{(r_k)} = \log\left(p_{\nu_k}^{(k)}\right) \quad (10)$$

where the energy in Eq. (10) used in the Rosenbluth sampling is replaced by the log of the probabilities ($p_{\nu_k}^{(k)}$) determined from Eq. (8) from the full energy MC-SCE simulation to calculate $I_{SC,uncorr}^{(k,j)}$ for *n* structures that lie beyond the energy cutoff. This value reflects the background error due to the chain growth process and can be cancelled out to yield the true mutual information value as given in Eq. (12).

$$I_{SC}^{(k,j)}(N,n) = I_{SC}^{(k,j)}(N,n) - I_{SC,uncorr}^{(k,j)}(N,n) \quad (11)$$

In this paper, all mutual information (MI) values reported are background corrected.

Reproducibility and Error Analysis. The reproducibility of SCE and MI values was tested on a randomly selected backbone ensemble of R7 and carried out 5 independent times. The data is shown in SI Table S1. SCE values are consistent and the background corrected MI values are reproducible within a reasonable error. The MI values without background correction is also included to give an estimate of the amount of spurious error possible in these calculations. Error bars shown in this paper are standard error of the mean calculated from the backbone ensembles (5 ensembles each for both apo and ligand bound states). As an example, to determine the error in side chain entropy for a set of residues [4], variances resulting from backbone fluctuation (σ^2) as well as intrinsic error of MC-SCE method (σ'^2) were added up as given in Eq. (12).

$$\sigma_{SCE}^{\{k\}} = \left[\overbrace{\sum_{\{k\}} \left(\sigma_{apo,k}^2 + \sigma_{lig,k}^2\right)}^{\text{Backbone variability}} + \overbrace{\sum_{\{k\}} \left(\sigma'^2_{apo,k} + \sigma'^2_{lig,k}\right)}^{\text{Intrinsic error}} \right]^{1/2} \quad (12)$$

Intrinsic error data was taken from the backbone ensemble used to test MI/SCE reproducibility above.

Experimental Methods

The ligand 5-nitrobenzisoxazole was synthesized by following an earlier published method[46], and its improved version from the Hilvert laboratory[47]. The KE07 R7-2 plasmids were kindly provided by the David Baker laboratory at University of Washington, Seattle, WA, and variants studied in this work were generated by site-directed mutagenesis using a Quik Change II site-

directed mutagenesis kit (Stratagene; Agilent Technologies, Santa Clara) using appropriate PCR primers (Table S2). After the mutagenesis PCR reactions, the mutated plasmids were transformed into XL-10 gold cells and the plasmids encoding individual mutations were isolated. The identity of the mutated plasmids were confirmed by sequencing the plasmid from both forward and reverse directions using T7 forward and T7 reverse primers at UC Berkeley Sequencing facility. The individual mutated plasmids were transformed into expression cell line BL21 (DE3) gold.

A single colony from the transformed cells containing individual variant was used to inoculate a starter culture of 20 mL LB medium supplemented with 50 μg/mL kanamycin and the resulting culture incubated with shaking overnight at 37°C. This starter culture was used to inoculate 500 mL LB medium with 50 μg/mL kanamycin and incubated for ~3h at 37°C until OD600 reached ~1.2. The culture was then induced with 1mM IPTG for overproduction and the culture was further grown with shaking at 37°C for 4h. The cells from the liquid culture were harvested and stored at -80°C until used for the isolation. In general, roughly 2 g of the wet cells were routinely obtained from 0.5L culture.

The harvested cells were thawed, re-suspended in 35 mL lysis buffer (25 mM Hepes, pH 7.25 containing 100 mM NaCl, 5% glycerol), lysed by sonication, centrifuged to remove insoluble debris and the soluble fraction loaded into pre-washed NI-NTA column (5mL resin, His-Pur, Thermo-Fisher). The NI-NTA resin with the bound proteins were washed first with 10 column volume of lysis buffer followed by 15 column volume of 20 mM NaPi, pH 7.4, 500 mM NaCl, 30 mM Imidazole to remove nonspecific and weakly bound proteins. The bound His-tagged fusion protein was then eluted from the NI-NTA resin with 20-25 mL of 500mM Imidazole buffer solution (20 mM NaPi pH 8.0, 500 mM NaCl, 500 mM Imidazole). The eluted fusion protein were extensively dialysed in lysis buffer, concentrated through Amicon filters (30,000 MWCO, Millipore), its concentration estimated by measuring the absorbances at 280 nm and stored at -80°C in smaller aliquots. This purification protocol yielded over 90% pure protein (assessed through the visible bands in SDS-PAGE) and routinely produced 18-23 mg of His-tagged KE07 proteins.

The enzymatic characterization of the KE07 R7 variants was performed similar to previously published work[42] with some modification in the Cary 50 spectrophotometer (Varian) that used a quarz cuvette. In short, the kinetic analysis were performed in 25 mM Hepes, pH 7.25, 100 mM NaCl, 5% glycerol with 5-nitrobenzisoxazole concentration ranging from 5-1500 μM with the co-solvent acetonitrile concentration equalized to 1.5% (v/v) in a micro-cuvette capable of monitoring reaction at 200 μL. A known amount of dry 5-nitroxybenzisoxazole was dissolved in acetonitrile to have 100mM substrate stock. From this stock a series of dilutions of the substrate were made in acetonitrile to achieve the concentration ranges in the kinetic assay. The reaction was initiated by the addition of small amount of the enzyme aliquot (final concentration from 0.2-1.0 μM in the assay) and the product formation was monitored spectrophotometrically at 380 nm ($\Delta\varepsilon$ = 15,800 M-1, cm-1). Steady-state parameters were obtained after fitting the data to the Michelis-Menten equation.

3.3 RESULTS

For KE07 (Figure 3.1b), the key intended active site residues include Glu101 as the catalytic base, Lys222 for stabilizing the developing negative charge on oxygen in the transition state, and Trp50 as a π-stacking residue to orient the 5-nitro-benzisoxazole ligand (Figure 3.1c). In addition, 10 other positions in the original scaffold (1THF) were changed to accommodate the engineered active site, culminating in a total of 13 designed residues for KE07. The initial design

exhibited very poor activity (k_{cat}/K_m = 12 $M^{-1}s^{-1}$) but after 7 rounds of LDE, a two-order improvement in catalytic performance was obtained for KE07-R7[42]. Table S3 lists the KE07 designed residues and the sequence changes made during LDE, as well as the corresponding improvements in k_{cat} and K_M for each round.

Enzyme KE70 (Figure 3.1d) also utilized a TIM barrel scaffold but one that differed from KE07 (deoxyribose phosphate aldolase from E. coli, PDB 1JCL). KE70 was designed using a His17-Asp45 dyad as the catalytic base, Ser138 as the charge stabilizing residue and Tyr48 as the π-stacking residue (Figure 3.1e). In addition, 12 other positions were designed to support the incorporation of the new active site. In terms of catalytic performance, the original KE70 design was an order of magnitude better than KE07 (k_{cat}/K_m = 126 $M^{-1}s^{-1}$) and with LDE KE70 reached a peak performance in round 6 (KE70-R6) that led to a further 450 factor improvement over its starting sequence[48]. Table S4 summarizes the original design, the mutations from straight DE (i.e. random mutagenesis), and later rounds using "spiked" DE through recombination of new design features (R2, R4 and R6) and the corresponding improvements in k_{cat} and K_M for each round.

Nearly all of the LDE changes in KE07 were satellite residues in the undesigned regions of the scaffold, with only one designed residue being mutated in the first round of LDE (Asn224Asp). In stark contrast to the LDE results for KE07, the designed residues in KE70 were directly targeted for change such that the best R6 variant mutated 7 of the originally designed residues, some of which were in the active site. While this might imply that the KE07 design was robust, our MD and MC-SCE simulations found that the overall active site chemistry was quite different than that shown in Figure 3.1c. Although Lys222 was a designed residue whose role is to stabilize the charged ligand in the transition state, instead we found that the heavy atom distances for Lys222N_ζ to the ligand oxygen was greater than 5.0 Å in all KE07 enzyme constructs; this is consistent with previous studies[4, 5] that showed that Lys222 is never in spatial proximity to the ligand to fulfill this role. Instead we find that Lys222 often forms a hydrogen bond with Ser48, as well as with residues Glu46 and Ile7 or its replacement in LDE R4 with Asp7; we find that catalytic activity is annihilated when we perform site mutagenesis at positions Ser48 and Lys222 (Table 3.1), as was true for mutation of Asp7 reported elsewhere[42]. This supports the reasoning of Khersonsky et al. that Asp7 serves to tether Lys222 so that it does not have unproductive interactions with the catalytic base[42], although we find a more extended network of Lys222 interactions. Hence, although Lys222 never fulfilled its intended design role, it is involved in interactions that nonetheless support the catalytic purpose of KE07[42].

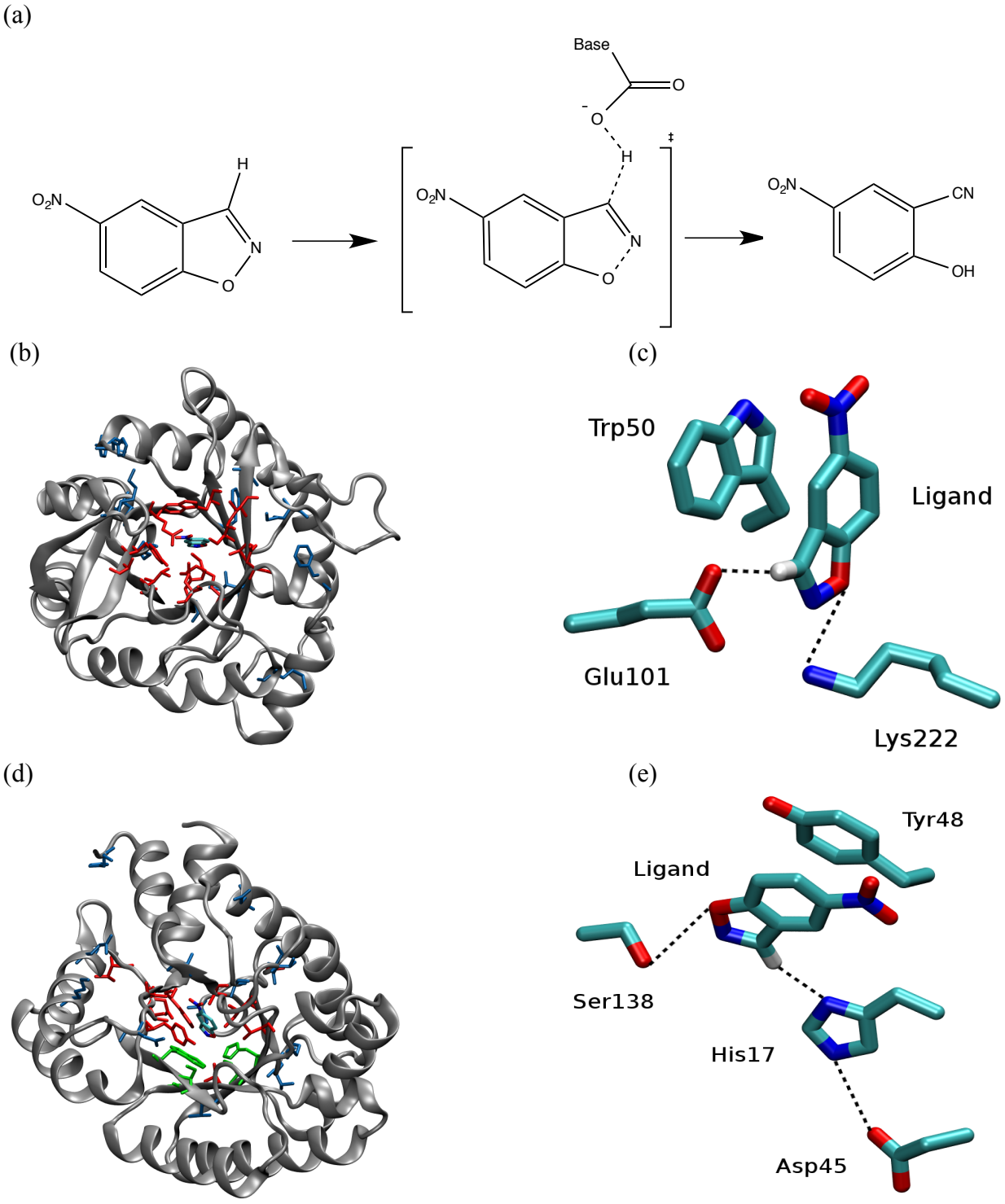


Figure 3.1. *The Kemp elimination KE07 and KE70 designs*. (a) The one-step reaction scheme involving the abstraction of hydrogen from 5-nitro benzisoxazole by a catalytic base. Shown is the transition state that has a partial negative charge on the substrate oxygen with cleavage of the O-N bond. (b) KE07 involved residues mutated from the original scaffold (red) as well as

mutations introduced by LDE shown in blue. (c) Relative orientation of the key catalytic residues with respect to the ligand in the ideal active site of KE07. (d) KE70 involved residues mutated from the original scaffold (red) as well as mutations made during laboratory DE shown in blue. Additional design mutations via a recombination DE strategy are shown in green (e) Relative orientation of the key catalytic residues with respect to the ligand in the ideal active site of KE70.

Table 3.1. *Experimental validation of effect of mutating network hubs in R7 on catalytic activity.* Steady-state measurements were recorded in 20 mM Hepes at pH 7.25 containing 100 mM NaCl, 5% glycerol at 20°C. The substrate (5-nitroxybenzisoxazole) was dissolved in acetonitrile and the enzymatic assay contained final concentration of acetonitrile at 1.5% (v/v).

KE07 Variant	**k_{cat}, s^{-1}**	**K_M, mM**	**K_{cat}/K_M, $M^{-1}s^{-1}$**	**% Activity relative to R7**
R7	0.81(0.01)	407(15)	1990(79)	100.0
Active Site Residues				
R7, S48N	0.1	689.7	145	7.3
R7, Y128F	-	-	-	~0
R7, H201A[a]	-	-	108(11)	5.4
R7, H201K[a]	0.562	5411.4	104	5.2
R7, K222A[a]	-	-	40 (6)	2.0
Distance Residues with high MI				
R7, R16Q	0.57(0.02)	589(46)	968(83)	49
R7, N25S	0.58(0.03)	479(57)	1221(157)	61
R7, L52A	0.51	514	992	50
R7, M62A	0.64	542	1181	59
R7, H84Y	0.77	497	1549	78
R7, K132N	0.75	560	1339	67
R7, I199S	0.33	771	428	22
R7, I199F	0.26	564	461	23
R7, I199A	0.23	1467	155	7.8
Controls				
R7, K132M	0.72	352	2045	116
R7, K162A	0.72	419	1718	86
R7, L170A	0.65(0.01)	338(19)	1929(115)	98
R7, E185A	0.89(0.02)	430(19)	2065(98)	104

[a] These variants did not exhibit substrate saturation and only sub-saturating substrate concentration data points were used to estimate k_{cat}/K_M.

Instead, we find that His201 is closest to the oxygen of the substrate heterocycle, with heavy atom distances between His201N_ε and the ligand oxygen found to be ~3.5-4.0 Å; Table 3.1 reports the experimental mutation at His201Ala and confirms that it destroys all enzyme activity. Furthermore the Gly202Arg mutation introduced in all rounds of LDE resulted in a very stable hydrogen bond between the Arg202-N_ζ and the nitro group of the ligand, and the designed Tyr128 forms a hydrogen bond with Arg202 that appears to further stabilize that interaction; in fact when Tyr128 is mutated to Phe, all enzyme activity is destroyed (Table 3.1). Similar "re-purposing" of other scaffold residues to aid in ligand positioning or charge stabilization has also been observed in crystal structures of another *de novo* designed Kemp eliminase, HG3.17 with a substrate analog[49]. Figure 3.2 shows the rotamer flexibility found in the greater network of the active site region of the best performing R7 variant for KE07, which stands in contrast to the static truncated active site assumed during the design process (Figure 3.1c). Further details pertaining to Figure 3.2 are given in Table S7.

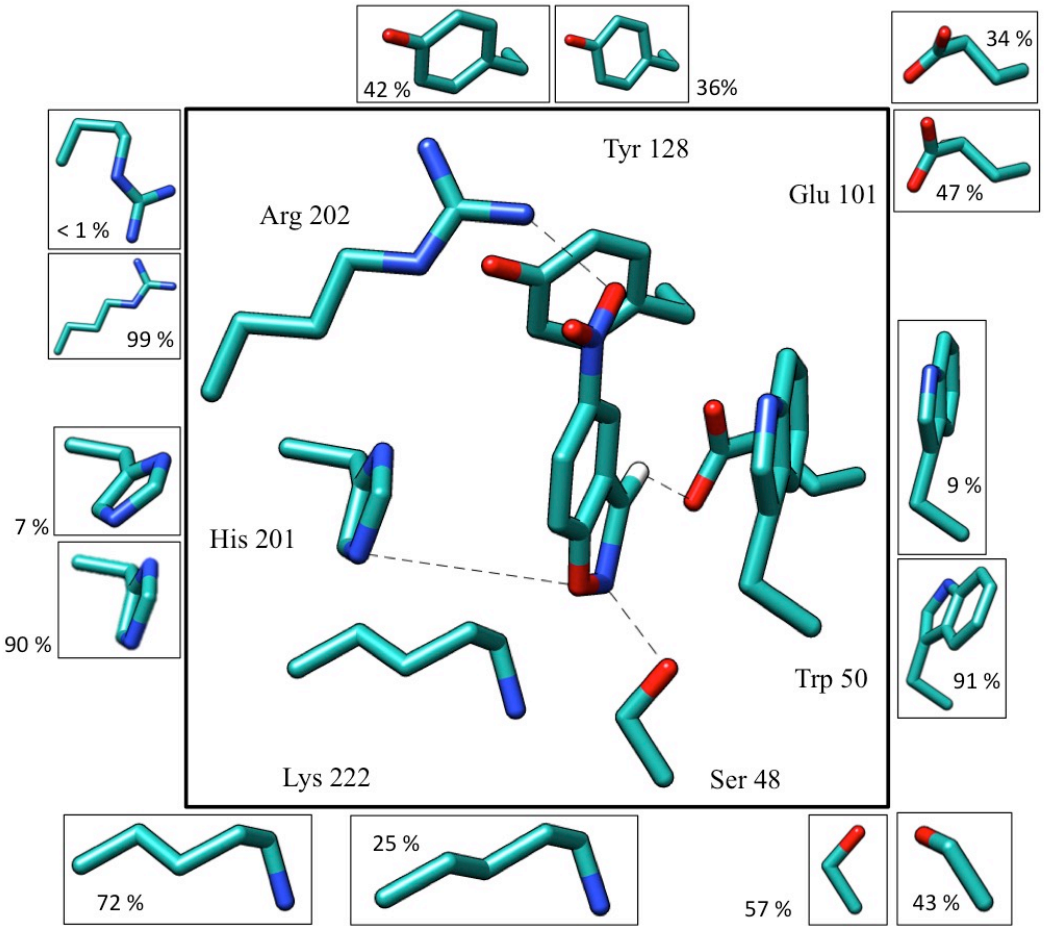


Figure 3.2. *The Kemp elimination KE07 active site in the best R7 variant*. The percentage represents the occupation of each rotamer as determined from the side chain ensemble of KE07-R7. We note that no one has reported on the importance of either His201 nor Tyr128 for the active site chemistry in KE07, which has been confirmed by experimental site mutagenesis in Table 3.1.

We next consider an overall thermodynamic analysis of the Michaelis-Menten scheme and the enthalpy and entropy breakdowns for the relative free energy of stabilization of the apo state, EL reactive complex and the EL† transition state complex (Table 3.2) for the designed enzymes and their best evolved variant KE07-R7 and KE70-R6. Note that for numerical calculations of free energy we ignore mutual information contributions due to the poor convergence of Eq. (2) where higher order correlations are clearly needed. Although we account for ligand solvation free energies by evaluating the ligand in our implicit solvent model, we are also missing explicit solvation or other types of solvent reorganization contributions that will stabilize each state differently. Furthermore, we model the transition state classically using altered partial charges that attempt to describe the electrostatics of bond-making and bond-breaking of the true quantum mechanical process. As such the absolute thermodynamic values for each state should be taken with caution, as we would require these additional contributions to connect to the experimental k_{cat} and K_M numbers. The idea behind the free energy analysis is instead to show how the individual contributions of side chain entropy and enthalpy reproduce the overall trends in these quantities, and yield a fairly suggestive picture as to why the KE07-R7 and KE70-R6 enzymes proved to be better biocatalysts than their designed counterparts.

Table 3.2. *Evaluation of the free energy under the Michaelis-Menten scheme for KE07 and KE70.* Calculated enthalpy and entropy differences between apo, EL and EL† states and their summed free energies, all in kcal/mole. We use a linear response approximation to evaluate the energy and entropy contributions for the transition state that involves the addition of an adiabatic step followed by enzyme reorganization (see text) in order to define the total free energy change. Note that we ignore mutual information contributions due to the poor convergence of the total entropy in Eq. (2), and we can't reliably account for explicit solvent free energy contributions, and hence we can't make direct or quantitative contact with k_{cat} and K_M values. We can only describe the qualitative trends in side chain entropy and enthalpy as shown.

State Function	KE07	KE07-R7		KE70	KE70-R6
EL Stabilization					
ΔH_{EL}	-9.9	-3.6		-13.5	5.7
$-T\Delta S_{EL}$	-6.3	4.5		-4.4	0.7
ΔG_{EL} [a]	-33.7	-16.4		-35.4	-11.0
EL† Barrier (Adiabatic)					
$\langle \Delta H^{\dagger} \rangle_{EL}$	11.6	10.5		15.7	8.7
$\langle -T\Delta S^{\dagger} \rangle_{EL}$	0	0		0	0
EL† Barrier (Reorganization)					
$\langle \Delta H^{\dagger} \rangle_{EL^{\dagger}}$	-2.0	-3.3		1.5	-3.0
$\langle -T\Delta S^{\dagger} \rangle_{EL^{\dagger}}$	-0.7	-3.8		3.6	1.7
EL† Barrier (Total)					
$\Delta G^{\dagger}$	8.9	3.4		20.8	7.4

[a] Includes the ligand solvation free energy, calculated from our model to be 17.5 kcal/mole.

We find that the enthalpy change between the EL complex and the apo state of the enzyme, $\Delta H_{EL} = \langle H_{EL} \rangle - \langle H_E \rangle$ is destabilized in the best evolved KE07-R7 and KE70-R6 enzymes compared to the original designs, consistent with what has been reported previously using EVB calculations[6]. However, we find the same destabilization trend is also observed for the entropy as well, since both designed enzymes exhibit $-T\Delta S_{EL} = -T\left(\langle S_{EL} \rangle - \langle S_E \rangle\right) < 0$; this means that there is greater conformational flexibility when the enzyme binds the ligand relative to the apo state, thereby stabilizing the enzyme-substrate complex. However, the introduction of new mutations in successive rounds of LDE leading to KE07-R7 and KE70-R6 contributes to reduction in the favorable entropy of the EL state ($-T\Delta S_{EL} > 0$), and hence the entropy also contributes to destabilization of the EL complex in the best LDE enzymes.

We also evaluate the enthalpy and entropy of the $EL^{\dagger}$ complex and how that changes with respect to the EL state based on a linear response approximation. We first assume an adiabatic step in which the $EL^{\dagger}$ complex is averaged over the EL ensemble to isolate the enthalpy, and then a subsequent step to account for enthalpic and entropic contributions due to enzyme reorganization in response to the change in ligand charges by averaging over the $EL^{\dagger}$ ensemble.

$$\Delta H^{\dagger} = \left\langle \Delta H^{\dagger} \right\rangle_{EL} + \left\langle \Delta H^{\dagger} \right\rangle_{EL^{\dagger}} \tag{13}$$

$$\Delta S^{\dagger} = \left\langle \Delta S^{\dagger} \right\rangle_{EL^{\dagger}} \tag{14}$$

Based on the linear response approximation using Eqs. (13) and (14), we find a very small stabilization of the adiabatic enthalpy for KE07-R7 relative to the original design, consistent with previous EVB calculations[6]. By contrast the large number of active site modifications made on the KE70 enzyme is consistent with the fact that the adiabatic enthalpy barrier is nearly halved in the KE70-R6 enzyme. However, by considering the reorganization terms as well, we find that there is transition state stabilization not only through the enthalpy, but that the entropy further lowers the catalytic barrier of the best enzymes relative to the original designs for both KE07 and KE70. Thus our thermodynamic calculations summarized in Table 3.2 supports the view that the active site of the original KE07 and KE70 enzymes were over-designed for the binding affinity of the EL state, whereas the LDE process created enzymes that unambiguously preferred the $EL^{\dagger}$ complex instead, especially for the KE70 enzyme.

Although the higher order terms in the entropy expansion in Eq. (2) may be directly related to k_{cat}/K_M, they can't currently be included for numerical calculations for free energies since higher order correlations are required for convergence of the total entropy. Nonetheless, we show that mutual information can yield even further insight as to why the evolved Kemp eliminases are better enzymes by focusing on residues with the largest mutual information with other residues; more specifically, such a position is defined as a "network hub" when it has a large ΔI as measured by |Z-scores| > 2 with at least 25 other residues throughout the scaffold. While this definition is somewhat arbitrary, it does quantify the residues with the strongest correlations with a large number of other residues, i.e. those with highest MI are always found using other definitions. One of the important features of the network hubs is that they are all mutually coupled, i.e. they each count as one of their connections all of the other hub residues. We shall see that network hubs are often identified as active site residues as well as mutational hot spots during the directed evolution process (Tables S5 and S6).

Figures 3.3 and 3.4 show how the network hubs are distributed over the scaffold in the apo state, EL state, as well as the $EL^{\dagger}$ state, for the original designs and the best KE70-R6 and KE07-R7 enzymes, respectively. As evident from Figure 3.3 and tabulated in Table 3.3, the

designed KE70 enzyme has high MI in the EL state and low MI in the apo and transition state. Furthermore we identify the His17-Asp45 dyad as 2 network hubs whose motions are strongly correlated with residues in KE70 that were subsequently mutated during the LDE process (23, 29, 48, 74, 166; see Tables S5). However, by the end of LDE the strongly correlated network in the EL state has been destroyed in favor of high MI in the apo state and transition state instead (Table 3.3). For KE07, there are no active site residues that are identified as network hubs in the designed enzyme for any of the states (Table 3.4). However, 7 out of the 13 LDE mutations were classified as a network hub at some point during the LDE process (Table S6), so that by the end of LDE the best evolved KE07-R7 enzyme exhibits network hubs involving active site residues 7, 50, 128, 201, 202, and 222 in the apo and/or EL† states. In turn the active site residues are highly correlated with *other* network hub residues, some of which are located far away from active site (> 10 Å) for the R7 variant of KE07.

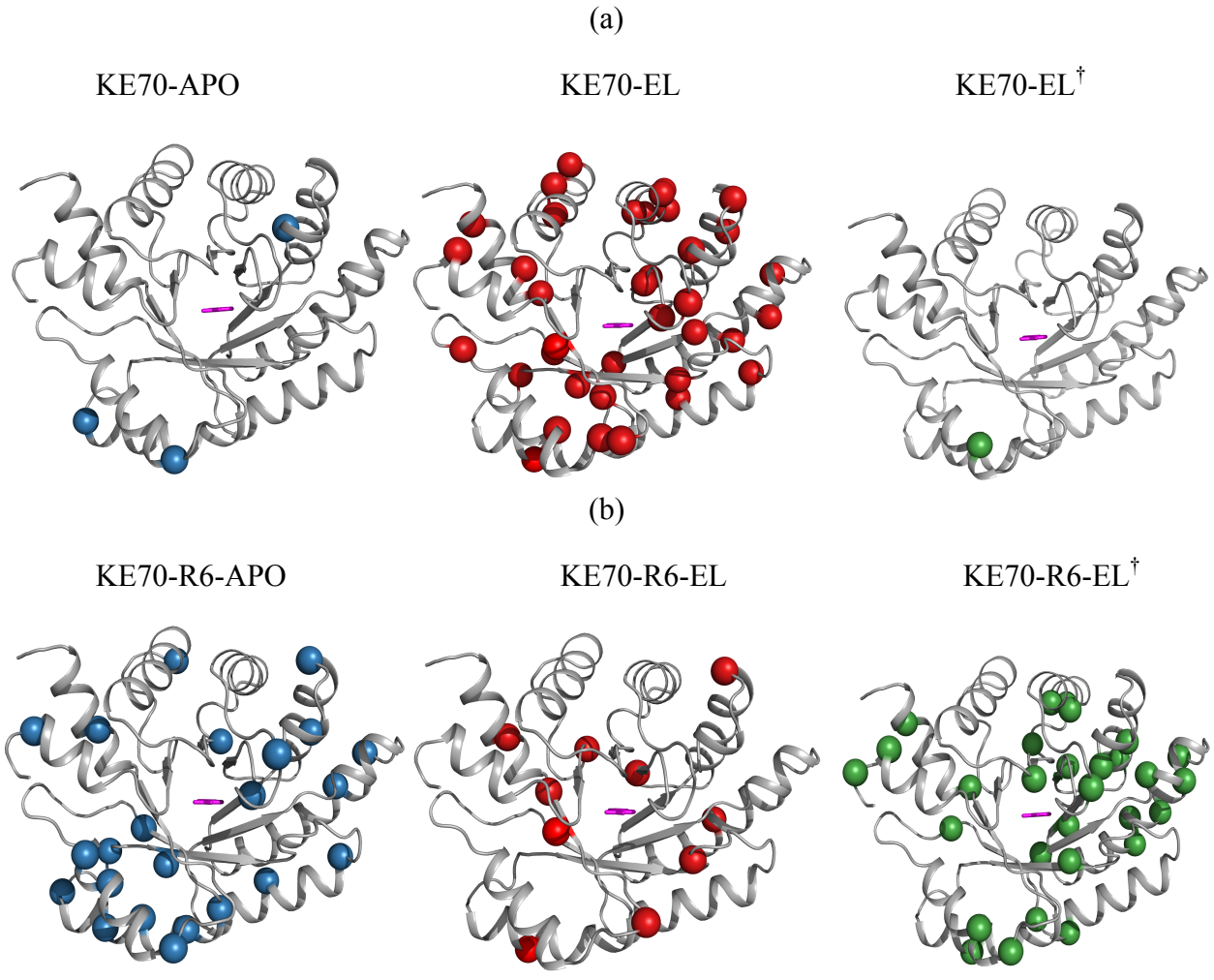


Figure 3.3. *Change in high mutual information hubs for the apo state, EL state, and EL† state for (a) designed KE70 and (b) the KE70-R6 variant.* The spheres represent residues that are high mutual information hubs (centered at the Cα position). We have uploaded an interactive visualizer of these hubs on our website at http://thglab.berkeley.edu

Table 3.3. *Residues that were determined to be network hubs with high mutual information for KE70*. Residues colored red were designed into the scaffold of 1JCL and residues colored blue were mutated during the course of LDE; the only exceptions are residues 43, 48, 74, and 166 that were both a designed and mutated residue.

Round	Highest MI in Apo state	Highest MI in EL complex	Highest MI in EL† complex
Design	27, 64, 143	6, 11, 14, 17, 23, 24, 29, 38, 45, 48, 58, 67, 70, 74, 83, 90, 100, 104, 115, 117, 121, 142, 147, 153, 154, 166, 167, 170, 173, 184, 186, 188, 191, 193, 216, 217, 221, 247	28
R6	11, 18, 25, 33, 35, 45, 50, 52, 56, 59, 64, 67, 70, 83, 90, 115, 118, 148, 154, 170, 174, 198, 223, 247	10, 15, 17, 22, 58, 76, 123, 148, 165, 232	6, 14, 25, 28, 43, 58, 73, 95, 97, 100, 107, 109, 115, 120, 123, 124, 136, 141, 154, 160, 166, 173, 185, 186, 193, 196, 204, 247, 249

To test the robustness of whether these other network hub residues are catalytically important due to their connection to the active site residues, we experimentally mutated network hubs for KE07-R7 (Table 3.4). The identified networks hubs included and were mutated as follows: Arg16Gln, Asn25Ser, Leu52Ala, Met62Ala, His84Tyr, Lys132Asn, and finally Ile199 to Ser, Phe, and Ala (Table 3.1). In all cases activity was diminished, with k_{cat}/K_M values anywhere between 10% to 78% of the KE07-R7 result, highlighting that residues located far away from the active site can also affect catalytic activity. We also performed two types of control experiments, in which a residue not identified to be a network hub is mutated (Lys162Ala, Leu170Ala and Glu185Ala) or in one case replacing a network hub residue with another residue that was also found to be a network hub and correlated to the active site (Lys132Met). We found in all four control experiments that catalytic activity was unaffected, even though one position Leu170 is within close proximity to the active site and substrate. This result clearly illustrates that residues with high mutual information are critical in the improved enzymatic activity of the KE07-R7 variant and by extension to the KE70 enzyme as well.

Table 3.4. *Residues that were determined to be network hubs with high mutual information for KE07*. Residues colored red were designed into the scaffold of 1THF and residues colored blue were mutated during the course of LDE. The bold faced residues identified as network hubs in R7 were subjected to mutagenesis to confirm that they reduced enzyme activity.

Round	Highest MI in Apo state	Highest MI in EL complex	Highest MI in EL† complex
Design	4, 10, 63, 66, 71, 87, 118, 212	16, 19, 58, 68, 85, 86, 139, 163, 174, 175, 185, 230, 232, 235	19, 51, 58, 64, 68, 91, 123, 161

R7	12, **16**, **25**, 42, **52**, 74, 94, 95, 128, **132**, 133, 149, 201, 202, 209, 222, 230	5, 19, **62**, 63, 71, 73, **84**, 87, 91, 92, 118, 148, 155, 159, **199**, 244, 247	7, 12, 37, 42, 50, 52, 58, 68, **84**, 92, 137, 148, 159, 182, 201, 208, 222, 230, 238

(a)

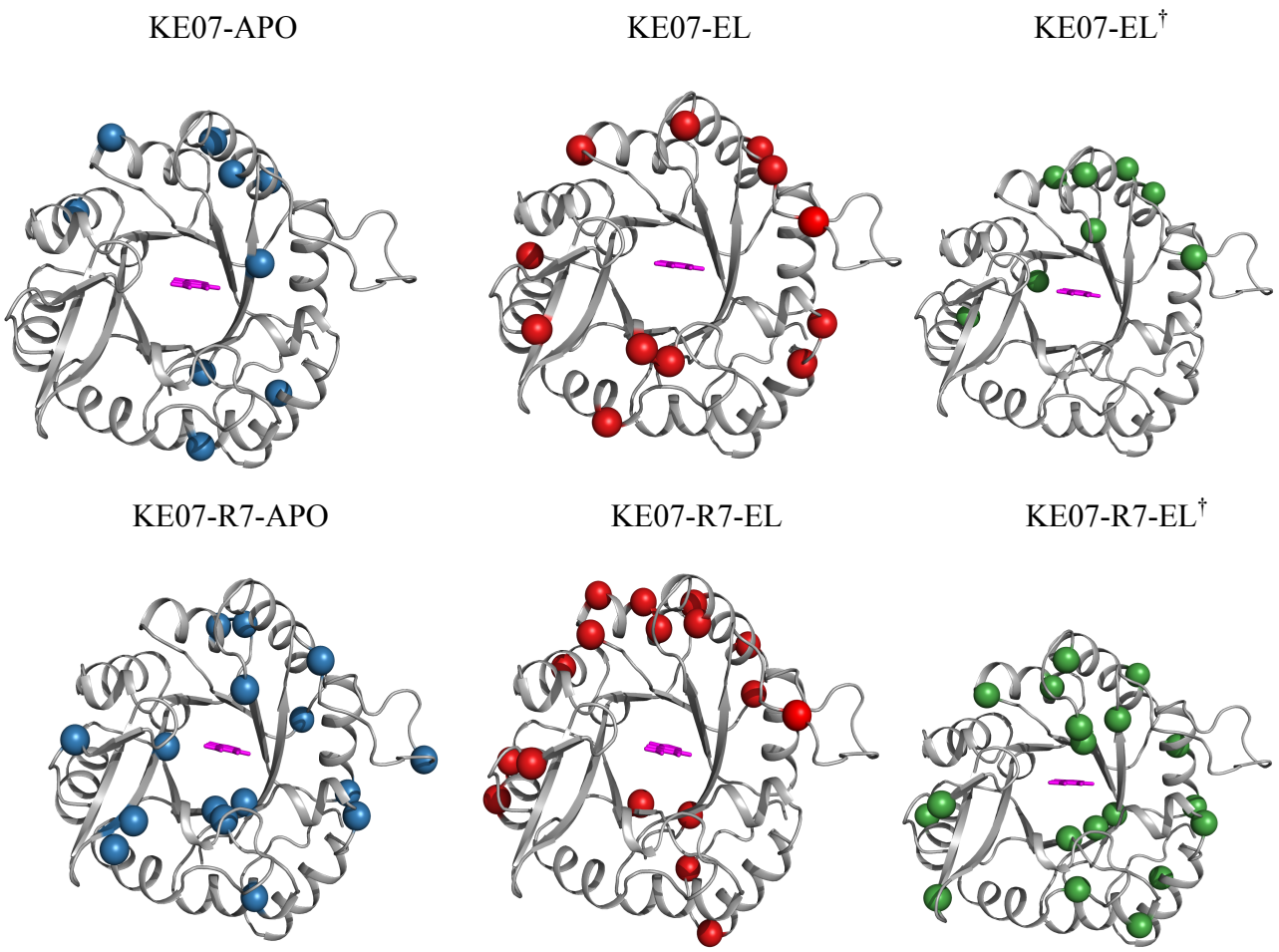


Figure 3.4. *Change in high mutual information hubs for the apo state, EL state, and EL† state for (a) designed KE07 and (b) the KE07-R7 variant.* The spheres represent residues that are high mutual information hubs (centered at the Cα position). We have uploaded an interactive visualizer of these hubs on our website at http://thglab.berkeley.edu

3.4 CONCLUSIONS

Given our current limitations in developing robust enzyme designs, laboratory directed evolution provides an attractive addition to rational computational design approaches since it is highly flexible in application to different biocatalysis reactions. Nonetheless although often highly successful, LDE is an opaque process because it offers no complete rationale as to why the mutations were successful, and therefore stands outside our ability to systematically reach novel

catalysis outcomes. To bridge this design gap, we have used side chain entropy and mutual information metrics applied to two different *de novo* enzymes and their LDE variants to better understand how conformational flexibility influences catalysis, which is central to many prominent proposals about the origin of enzyme activity[50-52]. Our analysis showed that by the end of the LDE process that changes in entropy, as well as enthalpy, helped to destabilize the EL complex in favor of stabilization of the EL† complex for both KE07-R7 and KE70-R6 when compared to the designed enzymes. Furthermore, we identified new active site players in KE07-R7, and using site mutagenesis showed that residues with large mutual information are catalytically important in KE07-R7 even though they may be remote from the active site.

There are two prominent but competing proposals as to what are the most important considerations in optimizing enzyme performance. Warshel and co-workers have emphasized that electrostatic pre-organization is the primary strategy by which enzymes achieve their remarkable catalytic activity compared to the uncatalyzed reaction[6, 53]. To recapitulate that argument, the electrostatic environment of the enzyme active site is structurally optimized in the apo state such that it is pre-organized to preferentially bind the transition state over reactants or products, thereby lowering $\Delta G_T^{\dagger}$ relative to that of the uncatalyzed reaction that must fold in the cost of reorganization factors (polarization) that raise the catalytic barrier. The other proposal is that conformational motion can also be key to catalytic performance by lowering $\Delta G_T^{\dagger}$, where equilibrium statistical fluctuations[23-25], dynamical coupling[26] and maximizing the reactive flux through the transition state surface[27] have emerged as potentially important dynamical aspects of the success of natural enzymes.

We believe that our results presented here on side chain entropy and mutual information are consistent with both the dynamical picture and the electrostatic pre-organization principle long advocated by Warshel and co-workers[6, 53]. For KE07-R7, network hubs in the apo state including Tyr128, His201 and Arg202 formed direct electrostatic interactions with the substrate, or the remote residues were charged or polar residues (Arg16, Asp25 and Lys132) whose long-ranged electrostatic effects clearly played some role in lowering $\Delta G_T^{\dagger}$ given that our experimental results showed reduced activity when these residues were mutated.

Furthermore we note a very interesting observation that there are high mutual information hubs in the EL state with much fewer MI hubs in the apo and EL† transition state complex in both the KE07 and KE70 designs. This we believe could be a signature of the problem of over-design of the EL state using the Rosetta strategy, and that LDE intervened to create new residue correlations in the apo and EL† transition state complex in the most improved enzyme variants. While highly speculative, it may be evidence for pre-organization signatures in the apo state, and a network of interactions that favor the EL† complex instead of the EL complex. At this point we can only quantify the importance of these changes in high MI sites through experimental mutagenesis.

Given that the active site residues of both KE07 and KE70 proved to have high mutual information and were strongly networked to other residues that also have extensive networks of strong side chain couplings, we believe that it should be possible to use computation to propose new sequence mutations that will improve the catalytic activity of other Kemp Eliminase enzymes for which LDE has not been performed. Furthermore, our method needs to be generalized to enzymes with substrates that are larger and more flexible than 5-nitroxybenzisoxazole, and estimating the enthalpy and entropy contribution from water would require explicit treatment of water, currently an area of research in our lab.

3.5 REFERENCES

1. 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*, 1387-91.
2. Röthlisberger, D.; Khersonsky, O.; Wollacott, A. M.; Jiang, L.; DeChancie, J.; Betker, J.; Gallaher, J. L.; Althoff, E. a.; Zanghellini, A.; Dym, O.; Albeck, S.; Houk, K. N.; Tawfik, D. S.; Baker, D., Kemp Elimination Catalysts by Computational Enzyme Design. *Nature* **2008,** *453*, 190-5.
3. Casey, M.; Kemp, D., Physical Organic Chemistry of Benzisoxazoles. I. Mechanism of the Base-Catalyzed Decomposition of Benzisoxazoles. *J. Org. Chem.* **1973,** *58*, 33-34.
4. Alexandrova, A. N.; Röthlisberger, D.; Baker, D.; Jorgensen, W. L., Catalytic Mechanism and Performance of Computationally Designed Enzymes for Kemp Elimination. *J. Amer. Chem. Soc.* **2008,** *130*, 15907-15.
5. Kiss, G.; Röthlisberger, D.; Baker, D.; Houk, K. N., Evaluation and Ranking of Enzyme Designs. *Protein science* **2010,** *19*, 1760-73.
6. Frushicheva, M. P.; Cao, J.; Chu, Z. T.; Warshel, A., Exploring Challenges in Rational Enzyme Design by Simulating the Catalysis in Artificial Kemp Eliminase. *Proc Natl Acad Sci U S A* **2010,** *107*, 16869-74.
7. Labas, a.; Szabo, E.; Mones, L.; Fuxreiter, M., Optimization of Reorganization Energy Drives Evolution of the Designed Kemp Eliminase Ke07. *Biochim. Biophys. Acta* **2013,** *1834*, 908-17.
8. Jackel, C.; Kast, P.; Hilvert, D., Protein Design by Directed Evolution. *Annu Rev Biophys* **2008,** *37*, 153-73.
9. Bolon, D. N.; Voigt, C. A.; Mayo, S. L., De Novo Design of Biocatalysts. *Curr Opin Chem Biol* **2002,** *6* (2), 125-9.
10. Taylor, S. V.; Kast, P.; Hilvert, D., Investigating and Engineering Enzymes by Genetic Selection. *Angew Chem Int Ed Engl* **2001,** *40* (18), 3310-3335.
11. Arnold, F. H., Design by Directed Evolution. *Acc Chem Res* **1998,** *31*, 125-131.
12. Romero, P. A.; Arnold, F. H., Exploring Protein Fitness Landscapes by Directed Evolution. *Nat Rev Mol Cell Biol* **2009,** *10* (12), 866-76.
13. Bolon, D. N.; Mayo, S. L., Enzyme-Like Proteins by Computational Design. *Proc Natl Acad Sci U S A* **2001,** *98* (25), 14274-9.
14. Cherry, J. R.; Lamsa, M. H.; Schneider, P.; Vind, J.; Svendsen, A.; Jones, A.; Pedersen, A. H., Directed Evolution of a Fungal Peroxidase. *Nature Biotechnol.* **1999,** *17*, 379-384.
15. Voigt, C. A.; Mayo, S. L.; Arnold, F. H.; Wang, Z.-G., Computational Method to Reduce the Search Space for Directed Protein Evolution. *Proc Natl Acad Sci USA* **2001,** *98*, 3778-3783.
16. Wu, S.; Acevedo, J. P.; Reetz, M. T., Induced Allostery in the Directed Evolution of an Enantioselective Baeyer-Villiger Monooxygenase. *Proceedings of the National Academy of Sciences of the United States of America* **2010,** *107*, 2775-80.

17. Fox, R. J.; Davis, S. C.; Mundorff, E. C.; Newman, L. M.; Gavrilovic, V.; Ma, S. K.; Chung, L. M.; Ching, C.; Tam, S.; Muley, S.; Grate, J.; Gruber, J.; Whitman, J. C.; Sheldon, R. A.; Huisman, G. W., Improving Catalytic Function by Prosar-Driven Enzyme Evolution. *Nature Biotechnol* **2007,** *25*, 338-344.
18. Voigt, C. A.; Mayo, S. L.; Arnold, F. H.; Wang, Z. G., Computational Method to Reduce the Search Space for Directed Protein Evolution. *Proc Natl Acad Sci U S A* **2001,** *98* (7), 3778-3783.
19. Saunders, C. T.; Baker, D., Evaluation of Structural and Evolutionary Contributions to Deleterious Mutation Prediction. *J. Mol. Bio.* **2002,** *322* (4), 891-901.
20. Fleishman, S. J.; Khare, S. D.; Koga, N.; Baker, D., Restricted Sidechain Plasticity in the Structures of Native Proteins and Complexes. *Protein Sci* **2011,** *20* (4), 753-7.
21. Smith, A. J.; Müller, R.; Toscano, M. D.; Kast, P.; Hellinga, H. W.; Hilvert, D.; Houk, K. N., Structural Reorganization and Preorganization in Enzyme Active Sites: Comparisons of Experimental and Theoretically Ideal Active Site Geometries in the Multistep Serine Esterase Reaction Cycle. *J. Amer. Chem. Soc.* **2008,** *130* (46), 15361-15373.
22. Fersht, A., *Enzyme Structure and Mechanism*. WH Freeman: New York, New York, 1985.
23. Klinman, J., *Dynamics in Enzyme Catalysis*. Springer Berlin Heidelberg: Berlin, Heidelberg, 2013; Vol. 337.
24. Wong, K. F.; Selzer, T.; Benkovic, S. J.; Hammes-Schiffer, S., Impact of Distal Mutations on the Network of Coupled Motions Correlated to Hydride Transfer in Dihydrofolate Reductase. *Proc Natl Acad Sci U S A* **2005,** *102*, 6807-6812.
25. Boehr, D. D.; Nussinov, R.; Wright, P. E., The Role of Dynamic Conformational Ensembles in Biomolecular Recognition. *Nature Chem. Bio.* **2009,** *5*, 789-96.
26. Bhabha, G.; Lee, J.; Ekiert, D. C.; Gam, J.; Wilson, I. a.; Dyson, H. J.; Benkovic, S. J.; Wright, P. E., A Dynamic Knockout Reveals That Conformational Fluctuations Influence the Chemical Step of Enzyme Catalysis. *Science* **2011,** *332*, 234-8.
27. Boekelheide, N.; Salomón-Ferrer, R.; III, T. F. M., Dynamics and Dissipation in Enzyme Catalysis. *Proc Natl Acad Sci USA* **2011,** *108*, 16159.
28. Henzler-Wildman, K.; Kern, D., Dynamic Personalities of Proteins. *Nature* **2007,** *450*, 964-72.
29. Henzler-Wildman, K. a.; Lei, M.; Thai, V.; Kerns, S. J.; Karplus, M.; Kern, D., A Hierarchy of Timescales in Protein Dynamics Is Linked to Enzyme Catalysis. *Nature* **2007,** *450*, 913-6.
30. Bartlett, G. J.; Porter, C. T.; Borkakoti, N.; Thornton, J. M., Analysis of Catalytic Residues in Enzyme Active Sites. *J. Mol. Bio.* **2002,** *324*, 105-121.
31. Fraser, J. S.; van den Bedem, H.; Samelson, A. J.; Lang, P. T.; Holton, J. M.; Echols, N.; Alber, T., Accessing Protein Conformational Ensembles Using Room-Temperature X-Ray Crystallography. *Proc Natl Acad Sci U S A* **2011,** *108*, 16247-52.
32. Frederick, K. K.; Marlow, M. S.; Valentine, K. G.; Wand, a. J., Conformational Entropy in Molecular Recognition by Proteins. *Nature* **2007,** *448*, 325-9.
33. Tuttle, L. M.; Dyson, H. J.; Wright, P. E., Side-Chain Conformational Heterogeneity of Intermediates in the Escherichia Coli Dihydrofolate Reductase Catalytic Cycle. *Biochemistry* **2013,** *52*, 3464-77.

34. Fraser, J. S.; Clarkson, M. W.; Degnan, S. C.; Erion, R.; Kern, D.; Alber, T., Hidden Alternative Structures of Proline Isomerase Essential for Catalysis. *Nature* **2009,** *462*, 669-73.
35. Gunasekaran, K.; Ma, B.; Nussinov, R., Is Allostery an Intrinsic Property of All Dynamic Proteins? *Proteins* **2004,** *57*, 433-443.
36. Monod, J.; Wyman, J.; Changeux, J. P., On the Nature of Allosteric Transitions: A Plausible Model. *J. Mol. Bio.* **1965,** *12*, 88-118.
37. Dubay, K. H.; Bothma, J. P.; Geissler, P. L., Long-Range Intra-Protein Communication Can Be Transmitted by Correlated Side-Chain Fluctuations Alone. *PLoS Comp. Bio.* **2011,** *7*, e1002168.
38. Lin, M. S.; Fawzi, N. L.; Head-Gordon, T., Hydrophobic Potential of Mean Force as a Solvation Function for Protein Structure Prediction. *Structure* **2007,** *15*, 727-40.
39. Lin, M. S.; Head-gordon, T., Reliable Protein Structure Refinement Using a Physical Energy Function. *J. Comp. Chem.* **2011,** *32*, 709-717.
40. McClendon, C. L.; Friedland, G.; Mobley, D. L.; Amirkhani, H.; Jacobson, M. P., Quantifying Correlations between Allosteric Sites in Thermodynamic Ensembles. *J Chem Theory Comput* **2009,** *5* (9), 2486-2502.
41. Killian, B. J.; Yundenfreund Kravitz, J.; Gilson, M. K., Extraction of Configurational Entropy from Molecular Simulations Via an Expansion Approximation. *J Chem Phys* **2007,** *127* (2), 024107.
42. Khersonsky, O.; Röthlisberger, D.; Dym, O.; Albeck, S.; Jackson, C. J.; Baker, D.; Tawfik, D. S., Evolutionary Optimization of Computationally Designed Enzymes: Kemp Eliminases of the Ke07 Series. *J. Mol. Bio.* **2010,** *396*, 1025-42.
43. Bhowmick, A.; Head-Gordon, T., A Monte Carlo Method for Generating Side Chain Structural Ensembles. *Structure* **2015,** *23* (1), 44-55.
44. Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C., Comparison of Multiple Amber Force Fields and Development of Improved Protein Backbone Parameters. *Proteins* **2006,** *65* (3), 712-725.
45. Shapovalov, M. V.; Dunbrack, R. L., A Smoothed Backbone-Dependent Rotamer Library for Proteins Derived from Adaptive Kernel Density Estimates and Regressions. *Structure* **2011,** *19*, 844-58.
46. Lindemann, H.; Thiele, H., Zur Chemie Des Benz-A,B-Isoxazols. *Justus Liebigs Ann. Chem.* **1926,** *449*, 63-81.
47. Hollfelder, F.; Kirby, A. J.; Tawfik, D. S.; Kikuchi, K.; Hilvert, D., Characterization of Proton-Transfer Catalysis by Serum Albumins. *J. Amer. Chem. Soc.* **2000,** *122* (6), 1022-1029.
48. Khersonsky, O.; Röthlisberger, D.; Wollacott, A. M.; Murphy, P.; Dym, O.; Albeck, S.; Kiss, G.; Houk, K. N.; Baker, D.; Tawfik, D. S., Optimization of the in-Silico-Designed Kemp Eliminase Ke70 by Computational Design and Directed Evolution. *J. Mol. Bio.* **2011,** *407*, 391-412.
49. Blomberg, R.; Kries, H.; Pinkas, D. M.; Mittl, P. R. E.; Grütter, M. G.; Privett, H. K.; Mayo, S. L.; Hilvert, D., Precision Is Essential for Efficient Catalysis in an Evolved Kemp Eliminase. *Nature* **2013,** *503*, 418-21.
50. Villà, J.; Štrajbl, M.; Glennon, T. M.; Sham, Y. Y.; Chu, Z. T.; Warshel, A., How Important Are Entropic Contributions to Enzyme Catalysis? *Proc Natl Acad Sci U S A* **2000,** *97* (22), 11899-11904.

51. Stone, M. J., Nmr Relaxation Studies of the Role of Conformational Entropy in Protein Stability and Ligand Binding. *Acc. Chem. Res.* **2001,** *34*, 379-88.
52. Wand, A. J., Dynamic Activation of Protein Function: A View Emerging from Nmr Spectroscopy. *Nature Struct. Bio.* **2001,** *8* (11), 926-931.
53. Frushicheva, M. P.; Cao, J.; Warshel, A., Challenges and Advances in Validating Enzyme Design Proposals : The Case of Kemp Eliminase Catalysis. *Biochemistry* **2011,** *50* (18), 3849-3858.

3.6 APPENDIX

Table S1. *Convergence of Side chain entropy (SCE) and Mutual Information(MI)*. 5 independent trial simulations were done on a random backbone ensemble from R7 variant. Reported are values for SCE, MI and MI without correction. Values reported are in units of k_BT.

	Trial-1	Trial-2	Trial-3	Trial-4	Trial-5	Avg	Stdev
SCE	355.6	349.2	354.0	353.0	352.0	352.7	2.39
MI	975.6	986.7	889.6	843.6	938.6	926.8	60.0
MI - uncorrected	3397.6	3805.2	2841.8	2437.8	3311.1	3158.7	528.7

Table S2. List of primers used for mutagenesis reaction to generate the specific mutations in R7-2 template[a]

Variants	Primers
H201A	Forward 5' CA TTG CCG ATC ATT GCA GCG AGG GGA GCT GGC AAG ATG 3' Reverse 5' CAT CTT GCC AGC TCC CCT CGC TGC AAT GAT CGG CAA TG 3'
K222A	Forward 5' GT GCA GAC GCG GCT GCG GCC GAT TCG GTT TTT C 3' Reverse 5' G AAA AAC CGA ATC GGC CGC AGC CGC GTC TGC AC 3'
R16Q	Forward 5' CAT TAA TAA TGA AGG ATG GCC AGG TTG TCA AAG GTA GC 3' Reverse 5' GCT ACC TTT GAC AAC CTG GCC ATC CTT CAT TAT TAA TG 3'
N25S	Forward 5' GTA GCA ATT TTG AAA GCC TGC GTG ACT CTG 3' Reverse 5' CAG AGT CAC GCA GGC TTT CAA AAT TGC TAC 3'
L170A	Forward 5' CC GGC GAA ATT GTG GCG GGT TCA ATT GAC CGC 3' Reverse 5' GCG GTC AAT TGA ACC CGC CAC AAT TTC GCC GG 3'
Q185A	Forward 5' CC GGC GAA ATT GTG GCG GGT TCA ATT GAC CGC 3' Reverse 5' GCG GTC AAT TGA ACC CGC CAC AAT TTC GCC GG 3'
[a] Changed nucleotides are shown in red text.	

Table S3. *Mutations made in various rounds of directed evolution of KE07*. The computationally designed residues (red) and mutated residues introduced by LDE of a given round (black) have been listed in the table below. The experimental k_{cat} and K_M values and representative variant names have been taken from (Khersonsky et al., 2010)

Sequence Position (Directed Evolution Round)	KE07 design	R2 11/10D	R3 I3/10A	R4 1E/11H	R5 10/3B	R6 3/7F	R7 10/11G
ILE 7			Gln	Asp	Asp	Asp	Asp
ALA 9							
ILE 11							
VAL 12					Met		Met
LYS 19		Glu				Glu	
SER 48							
TRP 50							
PHE 77							Ile
HIS 84							
PHE 86			Leu				
GLU 101 (catalytic base)							
ILE 102							Phe
GLN 123		Arg					
TYR 128							
ALA 130							
LYS 146		Thr	Thr	Glu		Thr	Thr
VAL 169							
GLY 171							
LEU 176							
HIS 201							
GLY 202		Arg	Arg	Arg	Arg	Arg	Arg
MET 207							
LYS 222							
ASN 224		Asp	Asp	Asp	Asp	Asp	Asp
PHE 229			Ser				Ser
k_{cat} (s^{-1})	**0.02**	**0.02**	**0.21**	**0.70**	**0.49**	**0.60**	**1.37**
K_M (mM)	**1.40**	**0.31**	**0.48**	**2.40**	**0.59**	**0.69**	**0.54**
k_{cat} / K_M ($M^{-1}s^{-1}$)	**12.2**	**66.0**	**425**	**291**	**836**	**872**	**2590**

Table S4. *Mutations made in various rounds of directed evolution of KE70*. The computationally designed residues(red) , other mutated residues introduced by LDE of a given round (black) and residues after which insertions took place (green) have been listed in the table below. The experimental k_{cat} and K_M values and representative variant names have been taken from [1]

Sequence Position (Directed Evolution Round)	KE70 Design	R2 7/12F	R4 4/1B	R5 7/4A	R6 4/8B
HIS 17 (catalyzing)					
ALA 19					
THR 20			Ser		Ser
ALA 21					
ASP 23		Gly			
LYS 29			Asn	Asn	Asn
THR 43			Asn	Asn	Asn
ASP 45 (catalyzing)					
TYR 48		Phe	Phe	Phe	Phe
TRP 72			Cys	Cys	Cys
SER 74					Gly
GLY 101				Ser	Ser
ALA 103					
SER 138			Ala	Ala	Ala
HIS 166				Asn	Asn
VAL 168					
THR 171			Pro	Pro	
GLY 177					
ALA 178					Ser
LYS 197					Asn
THR 198					Ile
ILE 202					
ALA 204			Val	Val	Val
ASP 212		Glu			
ALA 231				Ser	
ALA 235					
SER 239			Ser	Ser	Ala
HIS 251		Tyr			
k_{cat} (s^{-1})	**0.14**	**0.32**	**1.66**	**5.38**	**5.00**
K_M (mM)	**1.11**	**0.24**	**0.18**	**0.14**	**0.09**
k_{cat} / K_M ($M^{-1}s^{-1}$)	**126**	**1330**	**9240**	**37800**	**57300**

Table S5. *Residues that were determined to be network hubs with high mutual information for KE70 as a function of LDE round*. Residues colored red were designed and residues colored blue were mutated during the course of LDE; the only exception is that residue 48 was both a designed and mutated residue. The network hubs identified for KE70, 23 and 48 were mutated in R2, hub residue 29 in R2 was mutated in R4, and hub residue 197 was mutated in R6.

Round	Higher MI in EL complex state	Higher MI in Apo state
KE70	6, 11, 14, 17, 23, 24, 38, 45, 48, 58, 67, 70, 74, 83, 100, 104, 115, 117, 121, 147, 153, 166, 167, 173, 184, 186, 188, 191, 193, 216, 217, 221, 247	27, 64, 143
R2	11, 14, 17, 29, 30, 58, 67, 84, 122, 153, 188, 189	5, 64, 68, 70, 100, 109, 135, 141, 146, 147, 191, 193, 208, 215, 222
R4	18, 25, 33, 50, 64, 116, 121, 147, 170, 174, 189, 191, 197	5, 16, 22, 35, 58, 70, 90, 123, 209, 215, 233
R5	5, 24, 41, 64, 68, 79, 82, 83, 95, 153, 187, 196, 221, 232, 247	11, 49, 52, 77, 92, 115, 147, 165, 173
KE70-R6	10, 15, 17, 22, 58, 76, 123, 165, 232	11, 18, 25, 33, 35, 45, 50, 52, 56, 59, 64, 67, 70, 83, 90, 115, 118, 148, 154, 170, 174, 198, 223, 247

Table S6. *Residues that were determined to be network hubs with high mutual information for KE07 as a function of LDE round*. Residues colored red were designed into the scaffold of 1THF and residues colored blue were mutated during the course of LDE; the only exception is 224 that was both a designed and mutated residue in KE07. The network hub residues 19 and 86 identified in the designed enzyme were mutated in R2 and R3, respectively. In R2, we observe that residues 7 and 224 are network hubs that were subsequently mutated in R3. In R3, we observe that residues 7, 146, and 224 are hubs that were subsequently mutated in R4. In further improved variants like R5, we see that residues 19 and 102 appeared as network hubs, and were subsequently mutated in later rounds.

Round	Higher MI in EL complex state	Higher MI in Apo state
KE07	16, 19, 58, 68, 85, 86, 139, 163, 174, 175, 185, 230, 232, 235	4, 10, 63, 66, 71, 87, 118, 212
R2	58, 163, 222, 224, 242	6, 7, 10, 22, 34, 64, 87, 102, 113, 185, 187, 231, 236, 244
R3	42, 52, 60, 61, 62, 174	7, 22, 24, 46, 50, 59, 71, 72, 95, 146, 154, 179, 188, 193, 208, 212, 222, 224, 228, 238
R4	6, 42, 51, 59, 62, 63, 65, 85, 146, 159, 174, 175, 185, 201, 206, 230, 232, 243, 244	31, 34, 50, 72, 95, 156, 163, 191, 235, 238
R5	26, 52, 58, 62, 63, 64, 67, 102, 132, 133, 137, 148, 155, 167, 175, 208, 212, 239, 242	10, 19, 50
R6	41, 52, 91, 185, 212, 214, 236	10, 50, 62, 86, 99, 209, 222, 235, 238, 249
KE07-	5, 19, 62, 63, 71, 73, 84, 87, 91, 92, 118,	12, 16, 25, 42, 52, 74, 94, 95,

R7	148, 155, 159, 199, 244, 247	128, 132, 133, 149, 201, 202, 209, 222, 230

Table S7. Side chain dihedral angles for the residues highlighted in Figure 3.2 of the main text. Note that the rotamer classification has been done using the Dunbrack 2007 library.

Residue	Occupation %	χ_1	χ_2	χ_3	χ_4
Ser 48	57	176			
	43	-64			
Trp 50	91	180	-88		
	9	180	-123		
Glu 101	47	180	180	88	
	34	180	180	-60	
Tyr 128	42	180	80		
	36	180	102		
His 201	90	180	-103		
	7	180	-136		
Arg 202	99	-66	180	180	180
	< 1	-62	-75	75	78
Lys 222	72	180	180	-70	-70
	25	-65	180	180	-64

Chapter 4

Improving a Designed Kemp Eliminase without Directed Evolution

Laboratory directed evolution (LDE) is currently the standard procedure to improve performance of minimally competent design enzymes. Despite significant success, the process is highly labor intensive and belies the original goal of developing highly active biocatalysts rationally. Here, I report a systematic and rational improvement of a Kemp eliminase KE15 attained by selecting mutations through a newly developed computational screening method. Building of our understanding of other improved Kemp eliminases, the method uses side chain mutual information to predict sequence positions that can be targeted. To predict what the target positions should be mutated to, a combination of metrics like reactant state destabilization, transition state stabilization and optimization of positions with high mutual information was used. Starting from the design that has a k_{cat}/K_M of 27 $M^{-1}s^{-1}$, beneficial mutations were added in a stepwise manner. The most improved mutant had 3 mutations and k_{cat}/K_M of 304 $M^{-1}s^{-1}$. Unlike other rational improvement strategies, almost all the improvement came through k_{cat}, indicative of a direct impact on the chemical step. This work raises the prospect of designing new enzymes that achieve better efficiency with minimal experimental intervention.

4.1 INTRODUCTION

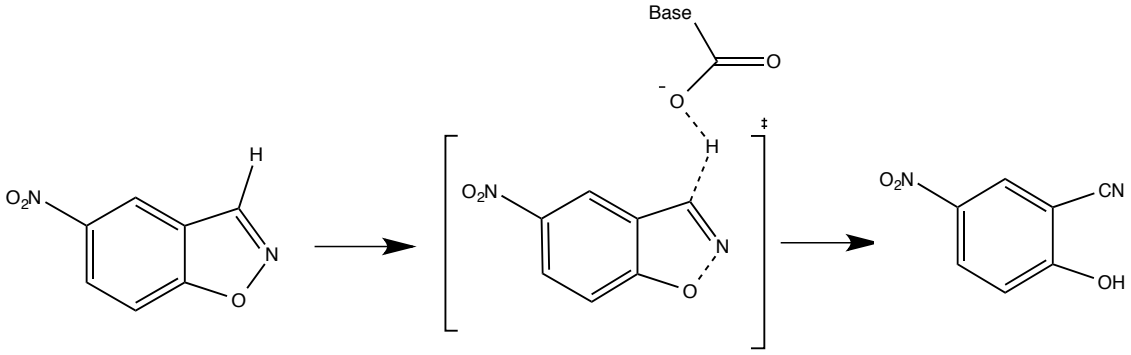
Advances in protein folding [1], superior scaffold matching algorithms [2] and improved active site modeling [3 4] have catalyzed the field of enzyme design in the last two decades. Representative examples include designed enzymes built for non-native reactions without any natural enzymes like the Kemp elimination [5 6 7], Retro-Aldol condensation [8] and Diels-Alders reaction [9]. Although current designed enzymes are only of academic interest, potential applicability of custom made enzymes in areas like biotechnology and sustainable energy is well appreciated and anticipated [10]. The first step in the design process is to construct a theozyme using quantum chemistry that includes 2-3 amino acids stabilizing the transition state. Next, the database of known protein scaffolds are searched using techniques like RosettaMatch, to look for motifs that can support this active site chemistry. Usually such grafting of the theozyme requires re-engineering some parts of the envisaged active site, leading to additional mutations. This leads to the final designed enzyme that typically differs from the starting scaffold by 10-15 mutations on average. When tested in the lab, some of these designs show experimental activity, validating the design protocol. Unfortunately, most active designs to date have exhibited very little catalytic competence compared to their native counterparts. A case in point is the Kemp elimination reaction, a one-step proton transfer reaction involving 5-nitrobenzisoxazole by a base leading to breaking of the 5-membered ring and forming the final product, alpha-cyanophenol (Figure 4.1). Kemp eliminases KE07 and KE70 [11] designed to catalyze this reaction had efficiencies (measured by k_{cat}/K_M) of 12 and 126 $M^{-1}s^{-1}$ respectively. Natural enzymes typically show efficiency in the range of 10^6-10^8 $M^{-1}s^{-1}$, typifying the scope for improvements in enzyme design. Despite such low efficiencies, they provide a starting point for further improvement by using laboratory directed evolution (LDE). Inspired by the rules of natural evolution, directed evolution generates multiple clones (on the order of 1000) of the starting sequence, each containing mutations introduced by techniques like random mutagenesis, shuffling etc. These

clones are then screened to pick out the most catalytically improved variants (selection pressure) and the process is repeated on the sequence of these refined variants. For the Kemp eliminases, LDE yielded a 200 [6] and 400-fold [7] improvement of efficiency in KE07 and KE70 respectively. Thus the success of LDE has led to much optimism of achieving efficiencies that rival natural enzymes. However, the very philosophy of LDE runs counter to the concept of rationally designing competent enzymes. Being an opaque process, there is very little explanation for most of the mutations [12 13]. Furthermore, LDE experiments are subject to chance, leading to a problem of reproducibility. Finally, LDE is very labor-intensive process with abysmal efficiency. On an average, only 1-10 mutants in 1000 clones show improvements in catalytic activity. Most of the improvements conferred by LDE highlight drawbacks of the design process. Thus in principle if understood quantitatively, these improvements could be introduced in a systematic way into the design.

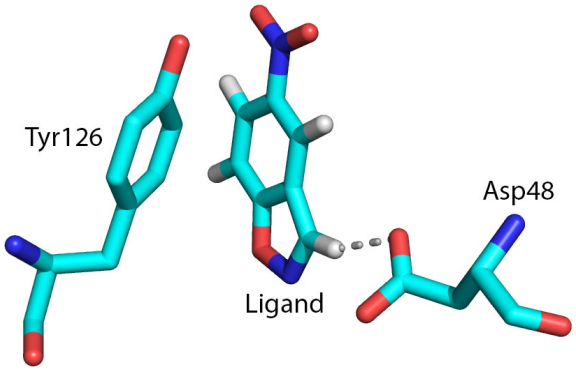
The effect of side chain conformational variability in improving designed KE07 and KE70 enzymes was reported in chapter 3. It was shown that improved variants used a combination of destabilization of reactant state (EL) and stabilization of transition state ($EL^{\dagger}$) to enhance activity. This was done jointly by enthalpy as well as side chain entropy. Furthermore, about half of the mutational hotspots during LDE had a high value of mutual information, a thermodynamic metric for quantifying level of correlation between residue side chains. Moreover, these high mutual information sites (hubs) were present mostly in the reactant state and significantly less in the apo and transition state of the enzyme. In the most improved variants, the apo and transition state had more hub residues (including active site residues) with less concentration in the EL state (almost none in the active site). Thus, reactant state (EL) destabilization, transition state ($EL^{\dagger}$) stabilization and hub residues can serve as convenient descriptors to guide future directed evolution experiments.

In this paper, we apply the understanding gained from the previous study to improve another de novo enzyme KE15 without resorting to laboratory directed evolution. KE15 was designed with the same Rosetta protocol, using Asp-48 as a base and a π-stacking residue in the form of Tyr-126 (Figure 4.1) with a TIM barrel scaffold (PDB ID: 1THF) as a backdrop. Thirteen additional design mutations were introduced into the scaffold to accommodate the active site. Experimentally, the enzyme has a k_{cat}/K_M of 27 $M^{-1}s^{-1}$ (k_{cat}=0.006 s^{-1}; K_M = 293 μM). Since no directed evolution was done on KE15, it serves as a good test system for improving catalytic efficiency in a rational and systematic way. Previous attempts at improving designed enzymes have had some success. The most notable one was the improvement of DA_20_10, an enzyme that catalyzes the Diers-Alders reaction, by crowdsourcing it to FoldIt players. Suggestions by the gamers led to a backbone remodeling and an 18-fold improvement in activity [14], almost all of which was due to K_M. Researchers have also tried smarter ways of constructing LDE libraries, using backrub motion [15], loop redesign [16], consensus mutations [17 18] etc. Even for natural enzymes, it is non-trivial to propose mutations that can modulate their functionality. Arnold, Mayo and coworkers have developed computational techniques like SCHEMA [19] that uses a genetic algorithm framework to propose mutations that don't lead to unstable folds. This is usually done by the evaluating the number of contacts that are broken/made before and after doing the proposed shuffling experiments (SCHEMA was developed for shuffling). Other attempts at predicting site-directed mutations have tried to minimize absolute entropy of the mutation site [20], again with an eye on stability. Very few (if any) of the methods that we are aware of attempt to propose mutations based on the actual reaction that is taking place.

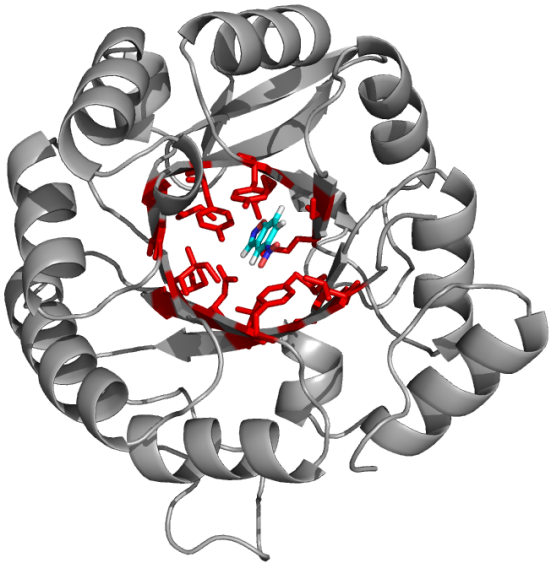
Frushicheva et. al proposed mutants for KE07 design based on electrostatic stabilization of the transition state but none of them seem to have been successful [21, 22].



(a)



(b)



(c)

Figure 4.1: *The Kemp Elimination reaction and KE15*. (a) The one-step reaction scheme involving the abstraction of hydrogen from the carbon of 5-nitrobenzisoxazole by a catalytic base. Shown is the transition state that has a partial negative charge on the substrate oxygen with cleavage of the O-N bond and nascent formation of a C≡N triple bond (b) Active site of KE15

design where the base is Asp-48 and Tyr-126 as π-stacking residue. (c) View of the overall KE15 enzyme. The active site and other designed residues are shown in red with substrate in cyan.

Here we report a method that is capable of screening mutants *in silico* and identifying mutants with better catalytic activities. We use a side chain ensemble method (MCSCE) combined with a physical energy function to determine residues with high mutual information in each of the apo, EL and EL* state of the enzyme. Residues with high mutual information as well as glycine/alanine that are not captured by our side chain entropy method are targeted for mutation. By concentrating on these sites, many of which are likely to be mutational hotspots, we can dramatically reduce the sequence space that needs to be explored for KE15. Next, by computationally screening for mutants that satisfy properties like EL state destabilization, $EL^{\dagger}$ state stabilization and optimized hub features (see methods) – all of which were seen in improved KE07/KE70 variants, we can propose variants that should be tested in lab, further reducing the experimental workload. This process was repeated three times, each step adding a mutation to the previous best variant. The best variant, a triple mutant, has a k_{cat}/K_M of 304 $M^{-1}s^{-1}$ (k_{cat} = 0.10; K_M = 329 μM) with mutations in the lower barrel (Asp130Lys), pointing away from the active site (Ile168Met) and in the active site (Gly199Ala). All in all, fifteen experimental mutants, all predicted by computational screening, were tried in the lab to achieve this order of magnitude enhancement. Similar improvements in LDE experiments can take 2-7 rounds, needing about ~ 2000-7000 clones, underlying the benefits of this computational screening method. With the rising popularity of cloud computing and advent of exascale computing, screening methods like ours can be automated to select for potentially beneficial mutants on a massively parallel scale, rationally.

4.2 COMPUTATIONAL METHODS

Generating backbone ensembles for the apo, EL and $EL^{\dagger}$ states of KE15: The starting structure for KE15 was the Rosetta model published previously. Mutant structures were generated by Modeller using the KE15 model as template. For the EL and $EL^{\dagger}$ state, we used the docked structure definition of the ligand. The ligand was then kept fixed in its modeled position for all subsequent backbone perturbations and MC-SCE calculations. The substrate geometry for the $EL^{\dagger}$ state was kept the same as in the EL complex, and only TS charges were changed to reflect the transition state of the bound complex.

Using each of these modeled structures for the backbone in the apo and ligand bound states, we then used the backrub algorithm implemented in Rosetta to run 25 independent simulations, each generating 10,000 trial moves using the C_α atoms as pivot residues, to generate uncorrelated backbone ensembles. From each simulation the lowest energy structure was saved and these 25 low energy backrub structures were selected, and divided into 5 backbone ensembles with 5 structures in each ensemble; this was done for all the rounds for both apo and ligand bound states. Since the backbone scaffold for another kemp eliminase KE07 that uses the same scaffold is quite rigid, we believe the backbone variations we have generated are adequate.

Generating side chain ensembles for the apo, EL and $EL^{\dagger}$ states of KE15: We have recently developed a Monte Carlo Side Chain Ensemble method (MC-SCE) to create large side chain ensembles. The MC-SCE method has been validated across a large number of proteins and protein complexes, in which it was found to be highly accurate when compared against high

quality X-ray crystallography and NMR J-coupling data for side chain rotameric preferences. The MC-SCE uses a Rosenbluth chain growth algorithm to generate an ensemble of side chain packings for a given protein backbone. From the bare backbone conformation *m*, and for subsequent steps *i*, the side chain rotamer, r_k, for residue *k* is selected according to the following probability

$$\rho_i^{(m,r_k)} = \frac{P_{r_k}^{(PDB)} e^{-\beta E_i^{(m,r_k)}}}{\sum_{\{\nu_k\}} P_{\nu_k}^{(PDB)} e^{-\beta E_i^{(m,\nu_k)}}} \quad (1)$$

where $\{\nu_k\}$ are the possible side chain conformations for residue *k*, using the values reported in the recent backbone-dependent Dunbrack library, which we have augmented by allowing for dihedral angle variations that are Gaussian distributed about a given rotamer value and weighted by its probability of occurrence in the PDB, $P_{r_k}^{(PDB)}$. $E_i^{(m,r_k)}$ is the energy of interaction of side chain conformation r_k of residue *k* with the backbone and all protein side chains grown so far (step i), using the energy function described above, and all residues are grown with ideal bond lengths and angles. Once the side chain of a residue is placed, the process is repeated until all the side chains are grown, thereby creating one complete protein structure. Each complete structure *m* is then assigned a weight *W(m)* in order to adjust for sampling bias due to the chain growth as well as to account for energetic solvent effects

$$W(m) = e^{-\beta F_{solv}^{(m)}} \prod_{i=1}^{N} \frac{\sum_{\{\nu_k\}} P_{\nu_k}^{(PDB)} e^{-\beta E_i^{(m,\nu_k)}}}{P_{r_k}^{(PDB)}} \quad (2)$$

For unsuccessful chain growths, the partially grown structure is considered dead and its weight is set to zero. This process is repeated in order to create ~20,000 side chain ensemble on the given backbone.

Determining mutational hotspots in KE15: Target sequence positions are picked by determining residues with high level of side chain correlation. This is quantified by a thermodynamic metric called mutual information. Given our MC-SCE method, we can calculate mutual information $I^{(i,j)}$. It is defined as the amount of information residue *k* has about another residue *j* based on the amount of coupled side chain dihedral angle fluctuations. In units of $k_B T$, this can be written as

$$I_{SC}^{(k,j)} = \sum_{\{\nu_k\}} \sum_{\{\nu_j\}} p_{\nu_k,\nu_j}^{(k,j)} \log\left(\frac{p_{\nu_k,\nu_j}^{(k,j)}}{p_{\nu_k}^{(k)} p_{\nu_j}^{(j)}}\right) \quad (3)$$

For each of the independent backbone ensembles we calculate the probability $p_{\nu_k}^{(k)}$ of each rotameric state ν_k using equation (4)

$$p_{\nu_k}^{(k)} = \frac{\sum_{m=1}^{M} W(m) \delta_{r_k,\nu_k}^{(m)}}{\sum_{m=1}^{M} W(m)} \quad (4)$$

where M=200,000 and the Kronecker delta is 1 if the side chain conformation r_k that was picked for the residue k in the m-th structure is v_k and 0 otherwise. Analogously, the joint probability distribution between residues k and j can be determined by Eq. (5)

$$p_{v_k,v_j}^{(k,j)} = \frac{\sum_{m=1}^{M} W(m)\delta_{r_k,v_k}^{(m)}\delta_{r_j,v_j}^{(m)}}{\sum_{m=1}^{M} W(m)} \tag{5}$$

The probabilities in Eq. (4) are then used to calculate side chain entropy (SCE) of each residue k using the Gibbs probabilistic definition, with SCE values in units of $k_B T$.

$$S_{SC}^{(k)} = -\sum_{\{v_k\}} p_{v_k}^{(k)} \log p_{v_k}^{(k)} \tag{6}$$

We estimated the mean for the SCE values from the 5 independent backbone ensembles for the apo, EL and EL for each protein for each round.

A natural extension of (6) gives us a joint entropy. We can thus rewrite Eq. (3) as

$$I_{SC}^{(k,j)} = \left(S_{SC}^{(k)} + S_{SC}^{(j)}\right) - S_{SC}^{(k,j)} \tag{7}$$

in which the individual entropy $S_{SC}^{(k)}$ and joint entropy, $I_{SC}^{(k,j)}$, is calculated using the probabilistic definition of entropy via Eq. (6), and thus Eq. (7) can be interpreted as the degree of coupling of torsional motions of residues *k* and *j*.

In practice, a background error persists in mutual information calculations since two completely uncorrelated variables will never be zero given a finite simulation time. In order to correct for this, we modified the strategy used by Dubay and Geissler to subtract out the erroneous extra mutual information that persists due to finite time scales. We first carry out our MC-SCE chain growth with the full energy function over all backbones in an ensemble, and using Eq. (7) we calculate the mutual information for the *N* structures obtained using the complete energy model, $I_{SC}^{(k,j)}$.

We then use our MC-SCE method to create structures where side chains for each residue are grown independent of the environment, i.e. clashes are ignored and the energy (and hence probability of chain growth) of each side chain conformer ν_k of residue k is given by

$$-\beta E_{uncorr}^{(v_k)} = \log\left(p_{v_k}^{(k)}\right) \tag{8}$$

where the energy in Eq. (8) used in the Rosenbluth sampling is replaced by the log of the probabilities ($p_{v_k}^{(k)}$) determined from Eq. (4) from the full energy MC-SCE simulation to calculate $I_{SC,uncorr}^{(k,j)}$ for *n* structures that lie beyond the energy cutoff. This value reflects the background error due to the chain growth process and can be cancelled out to yield the true mutual information value as given in Eq. (9).

$$I_{SC}^{(k,j)}(N,n) = I_{SC}^{(k,j)}(N,n) - I_{SC,uncorr}^{(k,j)}(N,n) \quad (9)$$

In this paper, all mutual information (MI) values reported are background corrected.

4.3 EXPERIMENTAL METHODS

The ligand 5-nitrobenzisoxazole was synthesized by following an earlier published method[23], and its improved version from the Hilvert laboratory[24]. The KE15 plasmids were kindly provided by the David Baker laboratory at University of Washington, Seattle, WA, and variants studied in this work were generated by site-directed mutagenesis using a Quik Change II site-directed mutagenesis kit (Stratagene; Agilent Technologies, Santa Clara) using appropriate PCR primers (Chapter 3, Table S2). After the mutagenesis PCR reactions, the mutated plasmids were transformed into XL-10 gold cells and the plasmids encoding individual mutations were isolated. The identity of the mutated plasmids were confirmed by sequencing the plasmid from both forward and reverse directions using T7 forward and T7 reverse primers at UC Berkeley Sequencing facility. The individual mutated plasmids were transformed into expression cell line BL21 (DE3) gold.

A single colony from the transformed cells containing individual variant was used to inoculate a starter culture of 20 mL LB medium supplemented with 50 μg/mL kanamycin and the resulting culture incubated with shaking overnight at 37°C. This starter culture was used to inoculate 500 mL LB medium with 50 μg/mL kanamycin and incubated for ~3h at 37°C until OD600 reached ~1.2. The culture was then induced with 1mM IPTG for overproduction and the culture was further grown with shaking at 37°C for 4h. The cells from the liquid culture were harvested and stored at -80°C until used for the isolation. In general, roughly 2 g of the wet cells were routinely obtained from 0.5L culture.

The harvested cells were thawed, re-suspended in 35 mL lysis buffer (25 mM Hepes, pH 7.25 containing 100 mM NaCl, 5% glycerol), lysed by sonication, centrifuged to remove insoluble debris and the soluble fraction loaded into pre-washed NI-NTA column (5mL resin, His-Pur, Thermo-Fisher). The NI-NTA resin with the bound proteins were washed first with 10 column volume of lysis buffer followed by 15 column volume of 20 mM NaPi, pH 7.4, 500 mM NaCl, 30 mM Imidazole to remove nonspecific and weakly bound proteins. The bound His-tagged fusion protein was then eluted from the NI-NTA resin with 20-25 mL of 500mM Imidazole buffer solution (20 mM NaPi pH 8.0, 500 mM NaCl, 500 mM Imidazole). The eluted fusion protein were extensively dialysed in lysis buffer, concentrated through Amicon filters (30,000 MWCO, Millipore), its concentration estimated by measuring the absorbances at 280 nm and stored at -80°C in smaller aliquots. This purification protocol yielded over 90% pure protein (assessed through the visible bands in SDS-PAGE) and routinely produced 18-23 mg of His-tagged KE07 proteins.

The enzymatic characterization of the KE15 variants was performed similar to previously published work[6] with some modification in the Cary 50 spectrophotometer (Varian) that used a quarz cuvette. In short, the kinetic analysis were performed in 25 mM Hepes, pH 7.25, 100 mM NaCl, 5% glycerol with 5-nitrobenzisoxazole concentration ranging from 5-1500 μM with the co-solvent acetonitrile concentration equalized to 1.5% (v/v) in a micro-cuvette capable of monitoring reaction at 200 μL. A known amount of dry 5-nitroxybenzisoxazole was dissolved in acetonitrile to have 100mM substrate stock. From this stock a series of dilutions of the substrate were made in acetonitrile to achieve the concentration ranges in the kinetic assay. The reaction was initiated by the addition of small amount of the enzyme aliquot (final concentration from

0.2-1.0 μM in the assay) and the product formation was monitored spectrophotometrically at 380 nm (Δε = 15,800 M-1, cm-1). Steady-state parameters were obtained after fitting the data to the Michelis-Menten equation.

4.4 RESULTS

The starting point for our screening process was KE15 design. As with KE07 and KE70 design, high reactant state stabilization and low transition state stabilization was seen from MC-SCE simulations. Additionally, high mutual information sites (hub residues) were concentrated in the EL state and very sparsely located in the apo/EL† states Thus, it would seem likely to improve KE15 with the same metrics learnt from KE07/KE70. Using hub residues as target positions, we attempted to pick out the most beneficial mutation through computational screening and propose mutations. Once we were able to identify the most improved variant in the laboratory, we kept that mutation and redid the screening, this time with hub residues of the improved variant. This iterative process was done 3 times, thus adding 3 mutations to KE15 design. Details of the starting variant in each round are provided in Table 4.1 and Hub residues (and *in silico* targets) for each variant are tabulated in Table 4.2.

Of the residues listed in Table 4.2, mostly 1st and 2nd shell of the enzyme were pursued for generating improved variants. The reason behind restricting to just these regions is to maximize the possibility of selecting beneficial mutants within the constraints of exploring limited sequence space with the computer time we had. On an average, 8-10 alternative amino acids were tried at the chosen sequence positions with an emphasis on diversifying charge, hydrophobicity and size. A combination of hub mutations was not tried, again due to the explosion of sequence space. In total, we tested about 350 mutants. To give a sense of the compute time this took, each mutant was simulated in the apo, EL and EL† state and each state requires about 3600 computer hours when run on our local computer cluster (AMD Opteron Processor 6274 (2.2 Ghz) cores). Thus, approximately 4 million compute hours were used up for the 3 rounds.

Table 4.1*: Details of the starting variant in each round of the screening process*. The most beneficial mutation from each iteration was added to move to the next round of screening. Only a handful of mutations were tested in the lab, dramatically reducing the labor that usually goes into improving designed enzymes. The best performing triple mutant is also listed.

Round	Starting Variant	K_{cat}/K_M ($M^{-1}s^{-1}$)	Number of mutants tested *in silico*	Number of experimental mutants tested
1	KE15	27	120	8
2	Asp130Lys-KE15	62	80	5
3	Ile168Met-Asp130Lys-KE15	150	110	2
-	Gly199Ala-Ile168Met-Asp130Lys-KE15	304	NA	NA

Table 4.2: *High mutual information sites (hub residues) in each starting variant and the best variant.* Hub residues were targeted for computational screening. Position in italics were tested in-silico and those in bold were designed and active site residues (within 5 Å of substrate)

Starting Variant	Highest MI in apo state	Highest MI in EL state	Highest MI in EL† state
KE15	240	*19*, 24, 34, *52*, 58, 59, 61, 62, *73, 74*, 84, 95, *102*, 116, *125, 130*, 133, 137, 148, *156*, 159, 161, 163, *173*, 182, 188, 190, 207, 212, *219*, 231, 235, 242, 243, 244	
Asp130Lys-KE15	16, 31, 55, 119, 175, *201*	*11*, 24, 28, ***50***, *52*, 64, 67, 68, 71, 84, 87, 93, 108, 112, *123*, 152, 154, 159, *168*, 185, 208, 215, 228, 231, 239, 244	*19*
Ile168Met-Asp130Lys-KE15	11, 27, 31, 42, 59, 68, 85, 86, 163, 176, 191, 228, 230	***5***, 14, 24, 41, ***50***, 64, 65, 71, 74, 83, 108, 118, 138, 168, *173*, 208, 212, 231	51, *103*, 163
Gly199Ala-Ile168Met-Asp130Lys-KE15	4, 118, 147, 175, 230	11, 17, 18, 19, 27, 42, 51, 52, 58, 59, 62, 79, 83, 85, 91, 103, 133, 149, 161, 182, 208, 209, 228, 233, 235	26, 68, 73, 87, 91, 239

The primary selection criterion for the mutations was EL state destabilization, EL† destabilization and optimization of hub features. Optimizing hub feature includes (a) Minimizing hub residues in EL state (b) Maximizing hub residues in apo and EL† states (c) Maximizing 1st shell hub residues in apo/ EL† states and removing 1st shell hubs in EL state. We preempted that not all selection criteria would likely be satisfied by a single mutation in such a rudimentary design. Thus we tried to choose mutants that exhibited atleast one of the criteria.

For Round 1, we screened for twelve mutational positions listed in Table 4.2. Of the approximately 120 mutants, eight partially satisfied metrics indicative of an improved enzyme. 3 of these mutants when tested in the lab showed improved performance either through an enhancement in k_{cat} or k_{cat}/K_M (Table 4.3). Position 130 that is originally an Aspartate gave 2 improved mutants, one being Asparagine and the other Lysine. The reasons for choosing these 2 mutants were (a) More destabilization of EL state (b) Further stabilization of EL† state (c) Apo/EL† hub residues showing up. It should be pointed out that the number of EL hubs did not

decrease substantially but in both cases, these mutants were the best we could screen for. The mutation Asp130Lys gave the best variant with a k_{cat}/K_M of 62 $M^{-1}s^{-1}$ with bulk of the improvement seen in k_{cat}. Structure wise, Asp130 is located at the bottom of the barrel (Figure 4.2) and thus this is a long-range effect on the chemical step. The other mutation that showed an improvement in k_{cat} was Ile173Gln. Unfortunately, the 5-fold k_{cat} improvement was accompanied by a 5-fold increase in K_M, leading to negligible change in efficiency.

In line with proposed strategy, we retained the most beneficial mutation Asp130Lys and moved to round 2. As evinced by Table 4.2, the EL state still has a high number of hubs. Another problematic aspect is that Ile50 shows up as a hub in the EL state. Instead of enriching 1st shell hubs in the EL state, improved variants tend to enrich them in the apo/EL† states. Round 2 mutations tried to address some of these concerns. Of the 5 mutations tested in lab, 2 showed enhanced activity. Asp11Tyr mutation improved slightly (k_{cat}/K_M = 80 $M^{-1}s^{-1}$) and also removed position 11 from the list of hubs in EL state. However, the best mutant was Ile168Met. This mutation further destabilized the reactant state and stabilized the transition state. The number of hubs in the unbound and transition states increased along with a decrease in the EL state hubs. Most of the performance improvement came from K_M, suggesting tuning of the binding of substrate. Although Ile168 is in the 1st shell of residues, it faces the opposite direction, indicating an indirect effect on catalysis.

As with the Asp130Lys-KE15 mutation, the double mutant also has the same issue of two active site hubs in EL state – Leu5 and Ile50. Given the lack of hub positions other than 5 and 50 in the 1st and 2nd shell, in round 3 we also expanded the positions considered for mutation by including proximal Ala and Gly residues – amino acids with no rotameric degrees of freedom. Of the 110 mutations tried, 4 were tried in the lab. None of the mutations at 5 or 50 improved catalytic activity. However, the mutation at Gly199Ala improved k_{cat}/K_M 2-fold (k_{cat} = 0.10 s^{-1}; K_M = 329 μM). This mutation was chosen because of the increase in hubs in the transition state as well as a reduction in the active site hubs. This was again not a perfect mutation prediction since 2 other metrics – free energy of stabilization of EL and EL† states showed inconsistent trends. We have not attempted any further mutation on the triple mutant till date. The location of the 3 mutations in KE15 is shown in Figure 4.2.

All in all, fifteen laboratory mutations were tried with five of them showing improvement from the starting variant. The non-trivial number of false positives highlights the limited resolution of the model as well as missing physics like that of solvent. In cases like Ile50Asp, a proposed mutation, the enzyme showed no activity, suggesting a catastrophic change in the enzyme machinery. Such effects most likely can be explained with a more sophisticated model and further work is underway to do just that. We save that autopsy of the failed mutations for later work.

In conclusion, I would like to emphasize a key accomplishment of this work that sets it apart from several other enzyme improvement ventures – increasing the k_{cat} instead of just K_M. This is a reflection of tuning the chemical step, considered to be a much harder problem than the binding problem (i.e improving K_M).

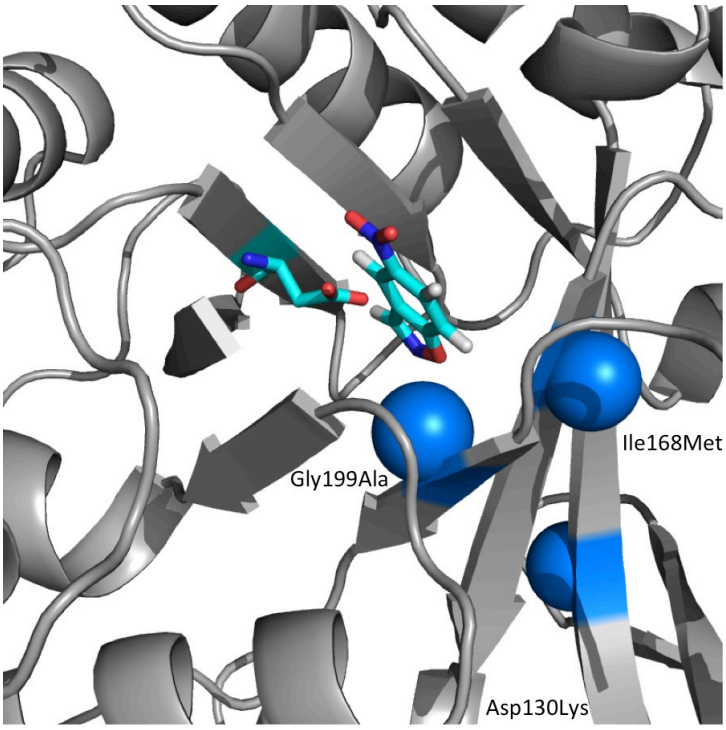


Figure 4.2: *Location of the 3 mutations in the best variant of KE15.* Asp130Lys lies in the lower barrel of the scaffold and Ile168Met is in the 1st shell but facing the opposite direction. Gly199Ala mutation is in direct contact with the substrate.

Table 4.3: *Details of the various successful mutants obtained for KE15*. The table highlights trends in the 3 criteria we used to improve KE15.

Mutant	ΔG_S (kcal/mol)	$\Delta G^{\dagger}$ (kcal/mol)	Apo hubs	EL Hubs	$EL^{\dagger}$ Hubs	k_{cat} (s^{-1})	K_M (μM)
Design KE15	-29.4	13.9	1	35	0	0.008	293
Round-1							
Asp130Lys	-25.8	3.6	6	26	1	0.031	474
Round-2							
Ile168Met-Asp130Lys	-14	0	13	18	3	0.036	227
Round-3							
Gly199Ala-Ile168Met-Asp130Lys	-23.2	10.4	5	25	6	0.101	329

4.5 DISCUSSIONS AND CONCLUSIONS

The greater than ten-fold performance enhancement reported here without using directed evolution is a positive step in the field of designed enzymes. By concentrating on a select few residues determined by physically reasonable metrics like mutual information, we are able to dramatically reduce the sequence space to be explored. Shifting the burden of trying out 100s of mutations in the lab to the computer further saved time, resources and manpower. At a minimum, the results give us confidence that improving designed enzymes need not be a dark art and by rationally putting back missing physics to optimize the reaction process, performance can be enhanced. I hope this screening method can find use in future enzyme design projects.

However, enzymes are complicated machines and one or two descriptors may not suffice to elevate designed enzymes to the scale of efficiency natural enzymes have. Even within the limited number of mutations we tried, only 30 % showed improved performance. In cases like Ile50 that is directly in the active site, predicted mutations to Asp completely annihilated any catalytic response, underlining shortcoming in the model. These problems are well acknowledged and point to deficiencies in the current treatment of electrostatics and solvent physics. Moreover, even with KE07 and KE70, about 50% of the mutations were picked up by this screening method. Picking up the remaining mutations would inevitably require a different approach. As was evident from Table 4.2, we were starting to see a lack of hubs that would likely influence catalytic activity, forcing us to also consider residues like Gly/Ala that don't have side chain dihedral degrees of freedom. Most of these hubs have high mutual information due to their presence on the surface, allowing them more freedom. Going forward, it will be important to understand other key aspects of the physics that would help us improve KE15 by another couple of decades at the very least. In the next chapter, I will describe how electrostatic fields can point to deficiencies in the design and how they can be ameliorated.

4.6 REFERENCES

1. Rohl, C. A.; Strauss, C. E.; Misura, K. M.; Baker, D., Protein Structure Prediction Using Rosetta. *Methods in enzymology* **2004,** *383*, 66-93.
2. Malisi, C.; Kohlbacher, O.; Höcker, B., Automated Scaffold Selection for Enzyme Design. *Proteins: Structure, Function, and Bioinformatics* **2009,** *77* (1), 74-83.
3. Tantillo, D. J.; Jiangang, C.; Houk, K. N., Theozymes and Compuzymes: Theoretical Models for Biological Catalysis. *Current opinion in chemical biology* **1998,** *2* (6), 743-750.
4. Kiss, G.; Çelebi‐Ölçüm, N.; Moretti, R.; Baker, D.; Houk, K., Computational Enzyme Design. *Angewandte Chemie International Edition* **2013,** *52* (22), 5700-5725.
5. Khersonsky, O.; Kiss, G.; Röthlisberger, D.; Dym, O.; Albeck, S.; Houk, K. N.; Baker, D.; Tawfik, D. S., Bridging the Gaps in Design Methodologies by Evolutionary Optimization of the Stability and Proficiency of Designed Kemp Eliminase Ke59. *Proc Natl Acad Sci U S A* **2012,** *109*, 10358-63.
6. Khersonsky, O.; Röthlisberger, D.; Dym, O.; Albeck, S.; Jackson, C. J.; Baker, D.; Tawfik, D. S., Evolutionary Optimization of Computationally Designed Enzymes: Kemp Eliminases of the Ke07 Series. *J. Mol. Bio.* **2010,** *396*, 1025-42.
7. Khersonsky, O.; Röthlisberger, D.; Wollacott, A. M.; Murphy, P.; Dym, O.; Albeck, S.; Kiss, G.; Houk, K. N.; Baker, D.; Tawfik, D. S., Optimization of the in-Silico-Designed Kemp Eliminase Ke70 by Computational Design and Directed Evolution. *J. Mol. Bio.* **2011,** *407*, 391-412.

8. 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*, 1387-91.
9. Siegel, J. B.; Zanghellini, A.; Lovick, H. M.; Kiss, G.; Lambert, A. R.; Clair, J. L. S.; Gallaher, J. L.; Hilvert, D.; Gelb, M. H.; Stoddard, B. L., Computational Design of an Enzyme Catalyst for a Stereoselective Bimolecular Diels-Alder Reaction. *Science* **2010,** *329* (5989), 309-313.
10. Strohmeier, G. A.; Pichler, H.; May, O.; Gruber-Khadjawi, M., Application of Designed Enzymes in Organic Synthesis. *Chemical Reviews* **2011,** *111* (7), 4141-4164.
11. Röthlisberger, D.; Khersonsky, O.; Wollacott, A. M.; Jiang, L.; DeChancie, J.; Betker, J.; Gallaher, J. L.; Althoff, E. a.; Zanghellini, A.; Dym, O.; Albeck, S.; Houk, K. N.; Tawfik, D. S.; Baker, D., Kemp Elimination Catalysts by Computational Enzyme Design. *Nature* **2008,** *453*, 190-5.
12. Brustad, E. M.; Arnold, F. H., Optimizing Non-Natural Protein Function with Directed Evolution. *Current opinion in chemical biology* **2011,** *15* (2), 201-210.
13. Tobin, M. B.; Gustafsson, C.; Huisman, G. W., Directed Evolution: The 'Rational'basis for 'Irrational'design. *Current opinion in structural biology* **2000,** *10* (4), 421-427.
14. Eiben, C. B.; Siegel, J. B.; Bale, J. B.; Cooper, S.; Khatib, F.; Shen, B. W.; Stoddard, B. L.; Popovic, Z.; Baker, D., Increased Diels-Alderase Activity through Backbone Remodeling Guided by Foldit Players. *Nature biotechnology* **2012,** *30* (2), 190-192.
15. Tokuriki, N.; Tawfik, D. S., Stability Effects of Mutations and Protein Evolvability. *Current opinion in structural biology* **2009,** *19* (5), 596-604.
16. Bershtein, S.; Tawfik, D. S., Advances in Laboratory Evolution of Enzymes. *Current opinion in chemical biology* **2008,** *12* (2), 151-158.
17. Romero, P. A.; Arnold, F. H., Exploring Protein Fitness Landscapes by Directed Evolution. *Nature Reviews Molecular Cell Biology* **2009,** *10* (12), 866-876.
18. Goldsmith, M.; Tawfik, D. S., Directed Enzyme Evolution: Beyond the Low-Hanging Fruit. *Current opinion in structural biology* **2012,** *22* (4), 406-412.
19. Voigt, C. A.; Martinez, C.; Wang, Z.-G.; Mayo, S. L.; Arnold, F. H., Protein Building Blocks Preserved by Recombination. *Nature Structural & Molecular Biology* **2002,** *9* (7), 553-558.
20. Voigt, C. A.; Mayo, S. L.; Arnold, F. H.; Wang, Z. G., Computational Method to Reduce the Search Space for Directed Protein Evolution. *Proc Natl Acad Sci U S A* **2001,** *98* (7), 3778-3783.
21. Frushicheva, M. P.; Cao, J.; Warshel, A., Challenges and Advances in Validating Enzyme Design Proposals : The Case of Kemp Eliminase Catalysis. *Biochemistry* **2011,** *50* (18), 3849-3858.
22. Frushicheva, M. P.; Cao, J.; Chu, Z. T.; Warshel, A., Exploring Challenges in Rational Enzyme Design by Simulating the Catalysis in Artificial Kemp Eliminase. *Proc Natl Acad Sci U S A* **2010,** *107*, 16869-74.
23. Lindemann, H.; Thiele, H., Zur Chemie Des Benz-A,B-Isoxazols. *Justus Liebigs Ann. Chem.* **1926,** *449*, 63-81.
24. Hollfelder, F.; Kirby, A. J.; Tawfik, D. S.; Kikuchi, K.; Hilvert, D., Characterization of Proton-Transfer Catalysis by Serum Albumins. *J. Amer. Chem. Soc.* **2000,** *122* (6), 1022-1029.

Chapter 5

The Importance of the Scaffold for *de Novo* Enzymes

In this chapter, I report electric field values relevant to the reactant and transition states of designed Kemp eliminases KE07 and KE70, and their improved variants from laboratory directed evolution (LDE), using atomistic simulations with the AMOEBA polarizable force field. The catalytic base residue contributes the most to the electric field stabilization of the transition state of the LDE variants of the KE07 and KE70 enzymes, whereas the electric fields of the remainder of the enzyme and solvent *disfavo*r the catalytic reaction. These results suggest that LDE is ultimately a limited strategy for improving *de novo* enzymes since it is largely restricted to optimization of chemical positioning in the active site, thus yielding up to a ~3 order magnitude improvement that is an upper bound estimate based on LDE applied to comparable *de novo* Kemp Eliminases, whereas the electrostatic environment is thought to play a large role in stabilization of the transition state for naturally occurring enzymes. Instead *de novo* enzymatic reactions would most productively benefit from optimization of the electrostatics of the protein scaffold in early stages of the computational design, utilizing electric field optimization as guidance.

5.1 INTRODUCTION

Although the design of new biocatalysts has not yet reached the level of proficiency of naturally occurring enzymes, there is optimism that further progress toward that goal is realistic and within reach as our understanding deepens on why current efforts have fallen short[1] and what makes natural enzymes so exceptional[2-3]. In this work we consider *de novo* enzyme design whereby a small catalytic "theozyme" is placed into an accommodating native protein scaffold, i.e. one that remains stable. While minimal activity was observed for these *de novo* designed enzymes, it is still orders of magnitude below the activity typically seen in natural enzymes. While computation has provided insight[4-7] and useful improvements[8-10], the majority of the improvement comes from laboratory directed evolution (LDE)[11] by altering the protein sequence through multiple rounds of mutagenesis and selection to isolate the few new sequences that exhibit enhanced catalytic performance.[12-16]

This process is well-illustrated by the popular *de novo* design of the Kemp elimination (KE) reaction[17], involving the deprotonation of a small ligand substrate 5-nitrobenzisoxazole by a catalytic base (Figure 5.1)[17], with corresponding electronic rearrangements that break the C-H and N-O bonds while forming a C≡N triple bond, engineered into related TIM barrel scaffolds and usually further optimized with LDE to create different KE catalytic motifs such as KE07[12], KE70[13], KE59[16], HG3[8], HG3.17.[15] For KE07 and KE70, the focus of our study here, the majority of catalytic performance was obtained after 6-7 rounds of LDE, which improved the k_{cat}/K_M by a factor of ~200 (KE07.R7) and ~400 (KE70.R6), respectively, in the best evolved enzymes.[12-13] It is noteworthy that while most of the catalytic improvement for KE07 resulted from increases in k_{cat}, the improvements in KE70 were derived equally from k_{cat} and K_M, and would suggest that LDE took very different strategies in the optimization of the two enzymes.[12-13]

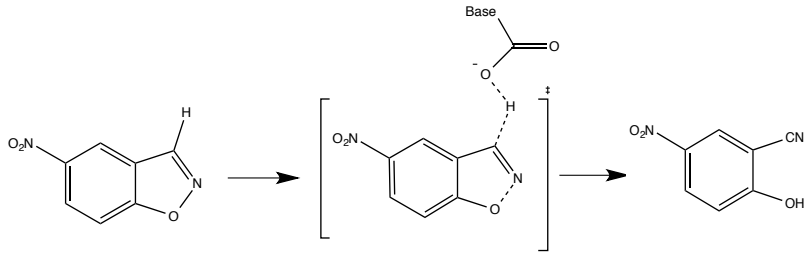


Figure 5.1: *The Kemp elimination reaction*. The one-step reaction scheme involving the abstraction of hydrogen from the carbon of 5-nitrobenzisoxazole by a catalytic base. Shown is the transition state that has a partial negative charge on the substrate oxygen with cleavage of the O-N bond and nascent formation of a C≡N triple bond.

Almost all design protocols for Kemp Eliminases[18-21] have taken a minimalist strategy of placing a base in a hydrophobic pocket, thus increasing the pK_a. For example, catalytic antibody 34E4 can catalyze the Kemp elimination reaction with efficiencies comparable to the KE07.R7 variant using a simplified active site motif of a functional base surrounded by hydrophobic residues[22]. Such rudimentary Kemp eliminases have also been designed not only in TIM barrels[8, 12-13, 15-17], but into scaffolds of calmodulin[23] and T4-lysozyme[24]. Thus, regardless of the fold involved, a basic level of activity can be obtained for this reaction.[1] However, to reach the level of natural enzymes, there needs to be synergism between multiple functional groups[22] that includes not only hydrophobicity but beneficial electrostatic contributions[2-3]. Furthermore, electrostatic stabilization comes not only from the proximity of a few residues in the active site[25-26], but also the rest of the protein scaffold[27-29] as well as the surrounding solvent[30-31].

Advances in vibrational Stark effect (VSE) spectroscopy[29] have enabled researchers to probe the electric field in the active site of enzymes in order to quantify their contribution to the observed acceleration of reaction rates over the uncatalyzed reaction in aqueous solvent. An electric field can have a catalytic effect if it adopts a sustained direction that specifically stabilizes the transition state in preference to the reactant state—an effect that in principle is better optimized in the pre-organized state of an enzyme relative to bulk aqueous environment[2-3, 28, 32]. Using VSE for the ketosteroid isomerase (KSI) enzyme and its inhibitor 19-nortestosterone (19NT), which has a C=O group located in the same position as the carbonyl group of it's natural substrate in the active site, Boxer and co-workers have shown that the large electric fields (~100-140 Mv/cm) exerted on this bond were linearly correlated with the activation free energies of the wild type and mutated variants.[28, 33] Although precise chemical positioning of the Asp40 base in the active site for proton abstraction from the substrate is important for KSI[26], leading to transition state stabilization that contributes 2-3 orders of magnitude to the observed accelerated rate, the analysis of the VSE data suggests that ~5 orders of magnitude improvement in k_{cat} comes about due to the reduction in the catalytic barrier that arises from the electrostatic *environment* of the KSI protein[28, 34]. Although there is disagreement on the relative orders of magnitude that electrostatics contributes to the chemical base positioning vs. the scaffold and solvent contributions[26, 35], there is no question that each are highly important for the catalytic performance of naturally occurring enzymes.

Presumably natural enzymes like KSI have developed highly optimized structural folds, including surfaces that invoke additional favorable orientations of solvent dipoles, which together contribute to a long-ranged and organized electrostatic environment for biocatalysis.[2-3]

However, for *de novo* designed enzymes it is reasonable to assume that they suffer from both non-optimal chemical positioning as well as a poorly concocted electrostatic environment, since the scaffold merely serves as a "backdrop" for containing the designed active site. In this work, we decipher the role of electrostatic pre-organization and transition state stabilization in the designed Kemp enzymes KE07 and KE70, and to demonstrate how the electrostatics are further tuned by LDE to improve the catalytic activity for both. Using atomistic computer simulations with an advanced polarizable force field, we measure the electric field at the 3 bonds that are made or broken in the ligand bound enzyme (EL) and transition state ($EL^{\dagger}$) as shown in Figure 5.2. Furthermore, we have formulated the calculations such that they allow us to decompose electric fields into contributions from each residue[32] as well as solvent to better distinguish between "chemical positioning" of the catalytic base in the active site, and the contributions that arise from the longer-ranged electrostatic environment from the protein and solvent.

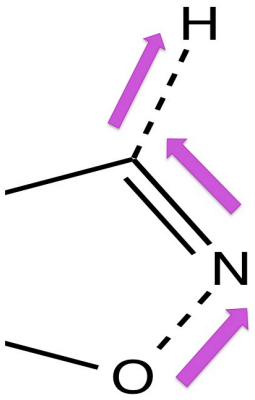


Figure 5.2: *Electric field projection onto the C-H, C-N, and O-N bond dipoles of 5-nitrobenzisoxazole and sign convention used.* Electric fields are calculated at the C, H, N and O of the ligand, in which the critical chemical step of the reaction is the breaking of the C-H and O-N bonds and the making of the C-N triple bond. The positive field direction shown by arrows is chosen to conform to the opposite direction of movement of electrons in the Kemp elimination reaction, a favorable field direction that supports the transition state.

We find that electrostatic fields are far greater in the active site of the enzymes relative to bulk solution when projected onto the relevant bonds, and a significant change in electrostatic pre-organization was found when going from the designed enzyme to the most improved variant for KE07 but not KE70. We find that chemical positioning, i.e. the optimization of the active site base that interacts directly with the substrate, contributes the most to the electric field environments of the best enzymes for KE07, whereas the pre-organization effect of the electrostatic field is still present but smaller in the designed KE70 enzyme, and diminishes in the LDE variant due to many mutations to hydrophobic amino acids that promote substrate affinity for the active site instead.

But in all cases, whether designed or LDE optimized, the electrostatic fields of the remainder of the enzyme and solvent largely disfavor the catalytic reaction. The underlying premise of the design approach – construction of a new catalytic "theozyme" that is placed into an *arbitrary* protein fold – suggests that the limitations of the *de novo* strategy is the restriction of the LDE search to optimization of chemical positioning in the active site, with an upper bound

of ~2-3 orders of magnitude estimated from natural enzymes like KSI. Instead *de novo* enzymatic reactions would most productively benefit from optimization of the protein scaffold[36] in earlier stages of the computational design, utilizing electric field optimization as guidance, to recover the many missing orders magnitude improvements from electric field environments.

5.2 METHODS

Generating backbone and side chain ensembles for EL and EL† states of KE07 and KE70. For both enzymes, the initial design was modeled using the structures reported in reference [12-13] with the ligand docked in the appropriate position. Starting structures for improved variants for both cases (R7 for KE07 and R6 for KE70) were generated using Modeller. Using each of these PDB/modeled structures for the backbone in the ligand bound state, we then used the backrub algorithm implemented in Rosetta to run 25 independent simulations, each generating 10,000 trial moves using the C_α atoms as pivot residues, to generate uncorrelated backbone ensembles. From each simulation the lowest energy structure was saved. Since the backbone scaffolds for KE07 and KE70 are quite rigid, we believe the backbone variations we have generated are adequate.

With these 25 backrub structures, we then used a recently developed Monte Carlo Side Chain Ensemble (MC-SCE) method[37] to create large side chain ensembles for each structure. MC-SCE has been validated across a large number of proteins and protein complexes in which it performed extremely well in predicting observables reported in high quality X-ray crystallography and NMR J-coupling experiments.[37] We note that for the MC-SCE part, the substrate was kept fixed in the docked position in both the EL and EL† state. The substrate geometry for the EL† state was the same as in the EL state with only the charges changed to reflect the transition state nature. The resulting structural ensembles for KE07 and KE70 represent sampling on the microsecond to millisecond timescale as estimated from repacking of the amino acid sidechains on different backbones.

Molecular dynamics simulations with AMOEBA. From the MC-SCE simulations on each backbone for the EL and EL† states of KE07 and KE70, we save the lowest energy structure which is then used as the starting point for molecular dynamics simulations with the AMOEBA polarizable force field[38-40]. The AMOEBA model is described using a permanent multipole expansion up to quadrupoles, and polarization effects are explicitly accounted for by calculating induced dipoles in a self-consistent manner. Due to the sophistication of electrostatics and short-ranged anisotropic interactions, AMOEBA should provide an excellent model for the electric fields in enzymes.

All the MD simulations in this study were performed using TINKER software. The tleap module in AMBER was used to solvate the system with a 10 Å spacing between the solute and the nearest box edge. Minimization was then performed using an LBFGS scheme with gradient RMS cutoff of 0.01. After minimization, an NPT simulation was performed with a timestep of 1fs integrated by the Beeman scheme. The temperature was maintained at 298 K with a Nose-Hoover thermostat. The PME real space cutoff and Van der Waals cutoff was set to 8 Å. Induced dipoles were iterated until the root-mean-square change was less than 10^{-5} Debye/atom. Given the ensemble of structures from the molecular dynamics and MC-SCE calculations described above, which provides effective sampling over much longer timescales than an individual and standard tens of nanosecond trajectory, we run 25 independent 100 ps trajectories of which we discard the first 50 ps and then collect statistics for the remaining 50 ps at intervals of 1 ps. Electric field values were calculated at the 4 atoms in the ligand involved in the breaking and

making of chemical bonds in the substrate, namely C, H, N and O as shown in Figure 5.2. This was done for EL and EL† states in both designed enzymes and best LDE variants.

Electric field calculations. In the AMOEBA framework, the permanent and induced electric fields at atom *i* due to another atom *j* can be written as follows

$$E_{perm,\alpha}^{(i,j)} = -T_\alpha q^{(j)} + T_{\alpha\beta}\mu^{(j,\beta)} - \frac{1}{3}T_{\alpha\beta\gamma}\Theta^{(j,\beta,\gamma)} \quad (1a)$$

$$E_{ind,\alpha}^{(i,j)} = T_{\alpha\beta}\mu_{ind}^{(j,\beta)} \quad (1b)$$

where q, μ, Θ correspond to point charge, point dipole, and point quadrupole permanent electrostatics, μ_{ind} is the polarizable dipole, and the tensor T is expressed in a compact format as

$$T_{\alpha\beta\ldots\nu} = \frac{1}{4\pi\epsilon_0}\nabla_\alpha\nabla_\beta \ldots \nabla_\nu\left(\frac{1}{R}\right) \quad (1c)$$

Although during the dynamical simulation the long-ranged electrostatics of the many-body polarization are evaluated under Ewald, in order to break down the electric field contributions from specific residues, we do an extra calculation where the induced dipoles are again calculated to convergence but using the real-space interactions only, with no cutoff's, and then Eq. (1a-1c) is calculated. When we add up all real-space contributions from all residues j to define the total electric field at the *i*=C, N, O, and H atoms of the substrate,

$$E_\alpha^{(i)} = \sum_{[j]} E_\alpha^{(i,j)} \quad (2)$$

we determine errors of ~1.0% when we compare to the full Ewald calculation.

Once we know the electric field values at atomic site i due to site j, the electric field values at a bond are then evaluated as the arithmetic mean of the field values at the 2 atoms forming the bond. For example, along coordinate axis α, the average field at the bond b_{ik} comprised of atoms i and k due to residue j is

$$E_\alpha^{(b_{ik},j)} = \left(E_\alpha^{(i,j)} + E_\alpha^{(k,j)}\right)/2 \qquad \alpha = x, y, z \quad (3)$$

Field values along a bond are then calculated by taking the dot product between the electric field vector at the bond (Eq 3) and the unit vector of the bond with positive direction illustrated in Fig 2. These values have been reported in all Tables and Figures. In all the 3 bonds studied here, we chose the positive direction of the field to be opposite to the direction of movement of electron in the bond breaking or bond making process. This is shown in Figure 5.2 with the arrows illustrating the positive field direction for each bond.

5.3 RESULTS

We first calculate the activation free energy stabilization of the transition state EL† relative to the reactant state EL due to electrostatics, $\Delta G_{elec}^{\neq}$

$$\Delta G_{elec}^{\neq} = \Delta_{EL\rightarrow EL^\dagger}\left(\vec{\mu}\cdot\vec{E}\right) \quad (4)$$

where $\vec{\mu}$ is the bond dipole and $\vec{E}$ is the electric field, in order to determine its contribution to the observed rate enhancements, i.e. on k_{cat}

$$k_{cat} = \frac{kT}{h}e^{-\beta\Delta G_{elec}^{\neq}}e^{-\beta\Delta G_{other}^{\neq}} \quad (5)$$

By convention, field directions that are aligned with the breakage of the C-H and N-O single bonds, and fields aligned in the opposite direction for the formation of a C≡N triple bond, would contribute to free energy stabilization of the transition state (Eq. (4)). Using the transition state

structure reported in [41] for an acetate base for the same ligand, and using the AMOEBA electrostatic parameters for charges and fixed dipoles, we can assign a $G^{\neq}_{elec}$ contribution to the C-H, N-O, and C≡N bond dipoles in the EL and EL† states (see the SI material for details). It is important to note that we are missing other contributions to the total free energy barrier, $\Delta G^{\neq}_{other}$, such as the entropic effects arising from desolvation (although the enthalpic interactions are likely accounted for in part by the solvent electrostatic field contributions in $G^{\neq}_{elec}$). In addition, we have shown that side chain entropy played a significant role in the observed k_{cat} trends for which the active site of the original KE07 and KE70 enzymes were over-designed for the binding affinity of the EL state, whereas the LDE optimized enzymes stabilized the EL† complex instead.[7] Therefore, while experimentally the $\Delta\Delta G^{\neq}_{total}$= –2.6 kcal/mol accounts for the ~70X improvement in k_{cat} for the best KE07 variant compared to the design, and the corresponding free energy barrier reduction $\Delta\Delta G^{\neq}_{total}$= –2.1 kcal/mol accounts for the ~35X improvement in k_{cat} for the best KE70 variant, we are only analyzing electric field contributions $\Delta G^{\neq}_{elec}$ and thus do not expect to reproduce these total activation free energy values.

Table 5.1: Free energy stabilization of the transition state. Reduction in activated free energies are calculated using $\Delta G^{\ddagger} = -0.048(F_{TS}\bullet\mu_{TS} - F_S\bullet\mu_S)$. Electric field values along the 3 bonds of the substrate 5-nitrobenzisoxazole in the EL and EL† states of KE07 and KE70 designed enzymes and best LDE variants, as well as in aqueous solvent. Positive field indicates favorable contribution and fields are reported in units of Mv/cm. Standard error of the means are in parentheses. Bond dipole moments are estimated from AMOEBA charges and fixed dipoles (see SI material) in the EL and EL† complexes; for C-H μ_{TS} = 1.0D, μ_S = –0.7D; for C≡N μ_{TS} = 0.4D, μ_S =2.0D; for O-N μ_{TS} = 2.3D, μ_S =–1.7D;

Enzyme Construct and $\Delta G^{\ddagger}$ transition state stabilization		**Fields generated for each bond**		
		C-H	**C≡N**	**O-N**
Designed KE07	EL	47.6 (3.9)	43.9 (2.0)	3.7 (2.7)
	EL†	68.7 (7.3)	58.8 (4.1)	22.7 (2.2)
$\Delta G^{\ddagger}$ = –4.6 kcal/mole		–4.9 kcal/mole	3.1 kcal/mole	–2.8 kcal/mole
LDE R7 Variant KE07	EL	81.5 (11.0)	49.1 (4.3)	7.2 (3.7)
	EL†	108.2 (12.9)	77.8 (9.6)	30.3 (3.7)
$\Delta G^{\ddagger}$ = –8.7 kcal/mole		–7.9 kcal/mole	3.2 kcal/mole	–4.0 kcal/mole
Designed KE70	EL	53.3 (3.6)	48.7 (2.4)	8.8 (1.7)
	EL†	77.6 (3.7)	62.2 (1.8)	28.1 (1.2)
$\Delta G^{\ddagger}$ = –5.8 kcal/mole		–5.5 kcal/mole	3.5 kcal/mole	–3.8 kcal/mole
LDE R6 Variant KE70	EL	54.0 (3.8)	29.7 (1.7)	6.9 (1.1)
	EL†	76.7 (2.6)	37.0 (1.6)	16.8 (0.9)
$\Delta G^{\ddagger}$ = –5.8 kcal/mole		-5.5 kcal/mole	2.1 kcal/mole	–2.4 kcal/mole
Substrate in water	EL	27.3	36.8	-10.5
	EL†	48.8	66.7	15.8
$\Delta G^{\ddagger}$ = –1.8 kcal/mole		-3.2 kcal/mole	2.3 kcal/mole	–0.9 kcal/mole

Table 5.1 reports the total electrostatic field values along the 3 relevant bonds of the substrate 5-nitrobenzisoxazole in the EL and EL† states of the KE07 and KE70 designed

enzymes, and their corresponding best LDE variants, as well as the fields acting on the reactant and transition state in aqueous solvent. If we assume that the electrostatic contribution to $\Delta G^{\neq}_{elec}$ arises from the additive contributions from the 3 bonds, then we can draw several immediate conclusions. The first is that the designed enzymes and their LDE variants help focus and enhance the electric fields along these bonds relative to the electric fields in bulk solvent, and overall the transition state is stabilized in preference to the reactant state regardless of enzyme variants (and which is true even in bulk solvent). In addition, the electric field stabilization is better for the designed KE70 relative to the designed KE07 enzyme, consistent with the fact that the k_{cat} of the former is an order of magnitude better than the latter. Furthermore, it is apparent that the electric fields are different in the transition state compared to the reactant state – indicating that activation free energies changes are attributable to electric field reorganization, and not just changes in the bond dipoles, as is often assumed[28-29]. Finally, while for KE07 there is a very clear trend of increasing electric field strength going from the designed enzyme to the best R7 variant in both EL and EL† states for all relevant substrate chemical bonds, the KE70 enzyme exhibits no net activated free energy decrease in going from the designed to the best R6 LDE variant.

In the case of KE07, when we break down the contributions to the total electric field for the C-H, C ≡ N, and N-O bonds from individual residues (Table S1), we see that the overwhelming contribution comes from the catalytic base Glu-101 (Figure 5.3a), and the field strength contributed by the Glu-101 in the R7 variant increases by an additional ~25-40 Mv/cm over the designed KE07 enzyme, a huge improvement for transition state stabilization. There are also significant stabilizing electric field contributions (> k_bT ~10 Mv/cm) arising from His201, and in the best LDE R7 variant from the GlyArg202 substitution, which we have shown in previous work interacts directly with the substrate to aid in chemical positioning of the base[7]. However, in both the KE07 design and R7 variant the designed residues Lys-222 and Ser-48, originally intended to stabilize the charge of the substrate in the transition state, have electrostatic fields that negatively impact the activation free energy. We and others have shown that Lys222 often forms a hydrogen bond with Ser48, as well as with residues Glu46 and Ile7 or its replacement in LDE R4 with Asp7, that help support the catalytic purpose of KE07 by removing unproductive interference with the base positioning[5, 7]. The Asn224Asp mutation is also unproductive in regards electric field stabilization[42], although a possible purpose for the Asp224 mutation is to better complex with water, as seen in the crystal structure of the R7 variant. Even so, these alternate roles for Lys222, Ser48, and Asp224 come with sacrifices to activation free energy stabilization afforded by constructive electric field effects on the substrate.

For KE70, there is virtually no overall optimization of the electrostatic fields going from the designed enzyme to the best R6 variant in both EL and EL† states (Table 5.1), a result that is largely orthogonal to the LDE optimization path taken for KE07. When we break down the largest contribution to the activation free energy by residue, there is ~25 Mv/cm enhancement from the His-Asp dyad for proton abstraction from carbon in the best R6 enzyme (Figure 5.3b), although the majority of the net ~80 Mv/cm field strength for the EL† state primarily comes from histidine (Table S2). This is not surprising since the main negative electric field contribution at this bond comes from Arg70 that is known to form unfavorable interactions with Asp45, thus reducing the pKa of the His-Asp dyad. Otherwise, the electrostatic field due to the His-Asp catalytic base contributes negligibly to the stabilization of the other bonds of the substrate, suggesting that the electric field in KE70 is not as highly optimized as it is in KE07. What modest gains are made in electric field stabilization of the EL† state for the primary reactive step

of C-H bond breaking in the designed KE70 enzyme are diminished by active site mutations to more hydrophobic groups (Trp72Cys and Ser138Ala) in the best LDE R6 variant. For KE70, it appears that other factors like productive binding of the substrate played a more significant role than electrostatics in the LDE improvement, captured experimentally through an order of magnitude reduction in K_M. In fact the pKa(k_{cat}) ~ 6.2 in both the designed KE70 and R6 LDE variant, whereas for KE07, where the majority of the improvement came through electrostatic stabilization, the pKa(k_{cat}) changed from < 4.5 in the design to 5.9 in the best R7 LDE variant.

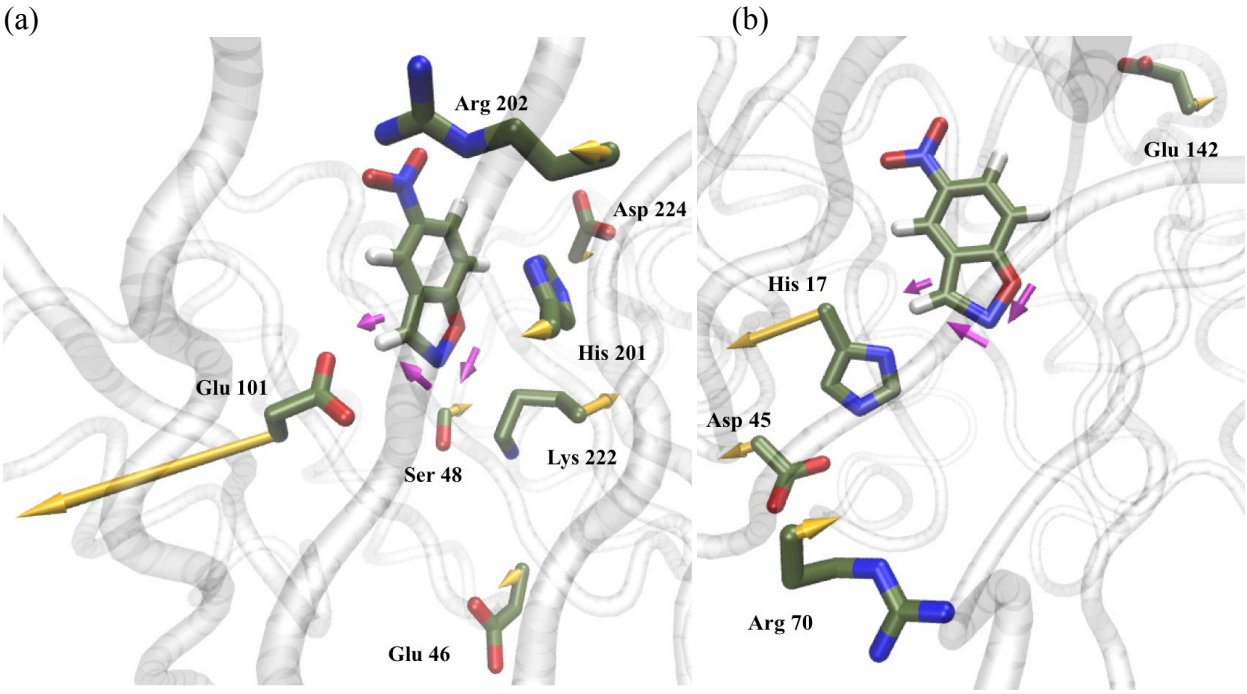


Figure 5.3: *The electric field projection onto the C-H bond dipole of 5-nitrobenzisoxazole from key residues in the active of (a) KE07 and (b) KE70.* The yellow arrows indicate the field direction/magnitude and the ones in magenta indicate dipole directions for each bond studied. All residues shown have a field > 10Mv/cm (~ k_bT) in the transition state of the best variant.

Table 5.2: *Chemical Positioning vs. Electric Field Environment at the C-H Bond*. The magnitude of the electric field in either the EL and EL† states for the designed KE07 and KE70 enzymes and the best LDE variants. The active site is defined by residues within 5 Å from the center of the substrate, while the protein environment is summed over all residues outside this region. Solvent includes waters in the neck of the TIM barrel as well as the surrounding hydration and bulk water. Positive sign indicates field supporting bond breaking. Fields are reported in units of Mv/cm

Region	**KE07 Design**		**KE07 R7 Variant**			**KE70 Design**		**KE70 R6 Variant**	
	EL	**EL†**	**EL**	**EL†**		**EL**	**EL†**	**EL**	**EL†**
Base	86.3	103.6	142.2	144.3		46.1	65.1	61.4	80.1
Active	1.0	11.2	2.0	8.3		16.7	23.1	2.3	2.9
Solvent	-15.6	-19.2	-22.6	-20.2		2.9	0.7	1.9	2.8
Protein	-24.1	-26.8	-40.1	-24.1		-12.2	-11.3	-11.6	-9.1

A lack of stabilization of the oxy-anion is thought to be a bottleneck for the catalytic reaction executed in catalytic antibody 34E4, and appears to be a problem for both Kemp Eliminases studied here as can be seen from the electric fields projected onto the N-O bond dipole. Although LDE improved the C-H fields considerably in KE07, the improvements in N-O field were considerably less, ~1.2 kcal/mol of additional stabilization for the KE07.R7 variant. For KE70, this bond breaking was destabilized by LDE, quite possibly due to the complete removal of Ser-138 whose primary intent was stabilizing the oxy-anion. As already stated elsewhere,[22] oxy-anion stabilization may be as critical as the chemical positioning involving the proton abstraction step.

However the important optimization of chemical positioning and active site improvements may also require further electrostatic stabilization of the transition state by the scaffold. For natural enzymes such as KSI, Boxer has shown that the major contribution to lowering the activation barrier comes from the electrostatic environment of the protein scaffold, as opposed to the contributions of residues that interact directly with the substrate or residues that aid in better chemical positioning of the catalytic base. For KSI it was estimated that $10^{2.5}$ fold improvement in k_{cat} was due to chemical positioning whereas an additional ~10^5 fold improvement was attributable to the electrostatic "environment" of the protein scaffold and surrounding solvent. While the relative percent contributions due to chemical positioning vs. electrostatic environment may be questioned[26, 35], there is no argument that enzyme folds have optimized an electrostatic environment that aids the catalytic reaction. This is clearly not the case for both the designed and LDE optimized KE07 or KE70 enzymes. Table 5.2 shows that the electric fields from the protein scaffold and solvent are mostly counterproductively aligned with the C-H bond for KE07, or effectively negligible in the case of KE70, a result that generalizes to the other bonds as well (Table S3 and S4).

These observations on KE07 and KE70 go a long way to explain why *de novo* enzymes are so poor to begin with, and why LDE is such a limited strategy for improving them. By using an "arbitrary" protein scaffold as a container for the active site theozyme, that also orients water solvent in such a way that are optimized for the scaffold and not the reactive chemistry, it should not be surprising the electric field environments are *highly* non-optimized for stabilizing the transition state. Hence the only tractable LDE strategy is to optimize the electrostatic fields locally at the active site, as was done for KE07, or utilize other chemical positioning strategies or ways to increase the basicity of the catalytic base through creation of a more non-polar active site, as found for KE70.

5.4 DISCUSSION

At present computational approaches have yielded *de novo* enzyme designs that are minimally competent, and therefore there is a necessary reliance on laboratory directed evolution to bridge the performance gap to compete at the level of catalytic antibodies, but even then they are certainly nowhere near the catalytic efficiencies of natural enzymes. An important aspect that helps explain the incredible performance of natural enzymes is that they have optimized folded structures that create favorable electric fields from the entire protein and surrounding solvent, not just the active site, to stabilize the transition state. In order to understand a natural enzyme's high catalytic proficiency, Warshel has suggested that an enzyme structural fold creates a pre-organized electrostatic environment, not found in bulk aqueous solution, that preferentially stabilizes the transition state charge distribution compared to the substrate reactant.

While we draw more specific conclusions pertaining to the Kemp Eliminases KE07 and KE70 below, in this discussion section we place these results in a greater context that provide

more general considerations for advancing *de novo* enzyme design. We conjecture that LDE is ultimately a limited strategy for improving *de novo* enzymes since it would require wholesale reengineering of most of the sequence of the scaffold; if such sequences prove to be unstable for maintaining the fold, it would extend the need to the creation of a new protein fold, that is beyond the capacity of any realistically sized LDE libraries, not to mention human time and patience. This vast reduction in the optimizable sequence space then is largely now restricted to chemical positioning in the active site.

If we were to take KSI as a reference point for the free energy stabilization attributable to local active site organization, we would expect at most a 3 order of magnitude improvement using LDE. At present all known attempts to further optimize the artificial Kemp Eliminases biocatalysts using LDE have yielded as little as one order of magnitude (the result for KE07 and KE70 after 6-7 LDE rounds beyond which no improvement was realized), to the best result obtained after 17 LDE rounds applied to the *in silico* design Kemp Eliminase HG3[8], yielding a k_{cat} for HG3.17 that is ~1000 times better than the design[15]. We do not mean to diminish what is clearly a success story in these recent successes, but we believe that it is unlikely for any designed enzyme to further improve through greater active site precision using LDE, and one must now venture further into the greater protein scaffold to find the next orders of magnitude improvements.

While *de novo* enzymatic reactions would most productively benefit from optimization of the protein scaffold utilizing electric field optimization as guidance, it should happen in earlier stages of the computational design. For KE07 the cluster of interactions involving Lys222, Ser48, Ile7 (Asp7) and Asp224-water have allowed for better positioning of the Glu101 base to act on the substrate, but with counterproductive electric field effects on the substrate that raises the activation free energy. The primary problem in their removal is that these residues are "baked in" to perform other benefits to support the catalytic purpose of KE07, but the evolved enzyme has to develop even more optimized catalytic base electric fields to compensate. Both KE07 and KE70 may have reached a cul de sac in regards further improvement in the active site after 6-7 rounds of LDE due to such electric field compensations.

5.5 CONCLUSIONS

In this study we have used a robust model for electrostatics, the AMOEBA polarizable force field, to calculate electric fields for the designed KE07 and KE70 enzymes and for the best variants that were improved under laboratory directed evolution. By calculating the field directions that are productively aligned with the breakage of the C-H and N-O single bonds and the formation of a C≡N triple bond for the small ligand substrate 5-nitrobenzisoxazole, we can assess the electrostatic free energy stabilization of the transition state relative to the reactant state. For KE07 it was found that the enhanced catalytic activity of the best R7 LDE variant stemmed from mutations that improved the electric fields locally in the active site, mostly attributed to the catalytic base, for stabilizing the transition state, while in KE70 the electric field enhancements to the transition state for its best LDE variant were more modest and completely isolated to the catalytic His17-Asp45 dyad. Finally, regardless of the Kemp Eliminase construct (i.e. designed or LDE optimized), we showed that the electrostatic environment of the protein and solvent are counterproductive in their contribution to stabilizing the transition state.

We suggest that LDE is ultimately a limited strategy for improving *de novo* enzymes since it is largely restricted to optimization of chemical positioning in the active site, thus yielding up to a ~3 order magnitude improvement that we offer is an upper bound estimate based

on the best known *de novo* Kemp Eliminase HG3.17[15], as well as based on estimates made on naturally occurring enzymes such as KSI[28]. Therefore *de novo* enzymatic reactions could take a different tack by focusing on optimization of the protein scaffold in early stages of the computational design, utilizing electric field optimization as guidance. One simple optimization strategy would scan a range of known protein scaffolds with the *theozyme* present, and ranking them according to their electric field contributions, followed by selecting the best scoring protein scaffold after each round of LDE to capture the maximum positive electric field contributions. Widening the repertoire of folds considered for the design of Kemp eliminases beyond TIM barrel is also likely to be beneficial. For example in the design process of the Kemp Eliminases, TIM barrels show up disproportionately (71% of low-energy structures) compared to their occurrence in natural enzymes (10%). Considering motifs of enzymes like KSI that catalyze a proton transfer reaction involving a labile hydrogen from an aromatic motif with high efficiency might be considered as an alternative scaffold. Even with the current TIM barrels used in the Kemp Eliminases one can imagine a better enzyme scaffold optimization by focusing on polar or charged residue mutations on the protein surface to better pre-organize solvent dipoles, whose integrated electric field could be quite large; every ~30 Mv/cm improvement in electric field alignment of the solvent on the active site would result in an order of magnitude of improvement in the catalytic rate.

ACKNOWLEDGEMENTS. This work was supported by the Director, Office of Science, Office of Basic Energy Sciences, Chemical Sciences Division of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. This research used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

5.6 REFERENCES

1. Korendovych, I. V.; DeGrado, W. F. Catalytic efficiency of designed catalytic proteins. *Current Opinion in Structural Biology* **2014,** *27*, 113-121.
2. Warshel, A. Electrostatic origin of the catalytic power of enzymes and the role of preorganized active sites. *J. Biol. Chem.* **1998,** *273* (42), 27035-27038.
3. Warshel, A.; Sharma, P. K.; Kato, M.; Xiang, Y.; Liu, H. B.; Olsson, M. H. M. Electrostatic basis for enzyme catalysis. *Chemical Reviews* **2006,** *106* (8), 3210-3235.
4. Ruscio, J. Z.; Kohn, J. E.; Ball, K. A.; Head-Gordon, T. The influence of protein dynamics on the success of computational enzyme design. *Journal of the American Chemical Society* **2009,** *131*, 14111-5.
5. Alexandrova, A. N.; Röthlisberger, D.; Baker, D.; Jorgensen, W. L. Catalytic mechanism and performance of computationally designed enzymes for Kemp elimination. *J. Amer. Chem. Soc.* **2008,** *130*, 15907-15.
6. Boekelheide, N.; Salomón-Ferrer, R.; III, T. F. M. Dynamics and dissipation in enzyme catalysis. *Proc Natl Acad Sci USA* **2011,** *108*, 16159.
7. Bhowmick, A.; Sharma, S. C.; Honma, H.; Head-Gordon, T. The role of side chain entropy and mutual information for improving the de novo design of Kemp eliminases KE07 and KE70. *Physical Chemistry Chemical Physics* **2016,** *18* (28), 19386-19396.

8. Privett, H. K.; Kiss, G.; Lee, T. M.; Blomberg, R.; Chica, R. A.; Thomas, L. M.; Hilvert, D.; Houk, K. N.; Mayo, S. L. Iterative approach to computational enzyme design. *Proceedings of the National Academy of Sciences* **2012,** *109* (10), 3790-3795.
9. Bolon, D. N.; Mayo, S. L. Enzyme-like proteins by computational design. *Proc Natl Acad Sci U S A* **2001,** *98* (25), 14274-9.
10. Bolon, D. N.; Voigt, C. A.; Mayo, S. L. De novo design of biocatalysts. *Curr Opin Chem Biol* **2002,** *6* (2), 125-9.
11. Arnold, F. H. Design by directed evolution. *Acc Chem Res* **1998,** *31*, 125-131.
12. Khersonsky, O.; Röthlisberger, D.; Dym, O.; Albeck, S.; Jackson, C. J.; Baker, D.; Tawfik, D. S. Evolutionary optimization of computationally designed enzymes: Kemp eliminases of the KE07 series. *J. Mol. Bio.* **2010,** *396*, 1025-42.
13. Khersonsky, O.; Röthlisberger, D.; Wollacott, A. M.; Murphy, P.; Dym, O.; Albeck, S.; Kiss, G.; Houk, K. N.; Baker, D.; Tawfik, D. S. Optimization of the in-silico-designed kemp eliminase KE70 by computational design and directed evolution. *J. Mol. Bio.* **2011,** *407*, 391-412.
14. Giger, L.; Caner, S.; Obexer, R.; Kast, P.; Baker, D.; Ban, N.; Hilvert, D. Evolution of a designed retro-aldolase leads to complete active site remodeling. *Nat Chem Biol* **2013,** *9* (8), 494-498.
15. Blomberg, R.; Kries, H.; Pinkas, D. M.; Mittl, P. R. E.; Grutter, M. G.; Privett, H. K.; Mayo, S. L.; Hilvert, D. Precision is essential for efficient catalysis in an evolved Kemp eliminase. *Nature* **2013,** *503* (7476), 418-421.
16. Khersonsky, O.; Kiss, G.; Röthlisberger, D.; Dym, O.; Albeck, S.; Houk, K. N.; Baker, D.; Tawfik, D. S. Bridging the gaps in design methodologies by evolutionary optimization of the stability and proficiency of designed Kemp eliminase KE59. *Proc Natl Acad Sci U S A* **2012,** *109*, 10358-63.
17. Röthlisberger, D.; Khersonsky, O.; Wollacott, A. M.; Jiang, L.; DeChancie, J.; Betker, J.; Gallaher, J. L.; Althoff, E. a.; Zanghellini, A.; Dym, O.; Albeck, S.; Houk, K. N.; Tawfik, D. S.; Baker, D. Kemp elimination catalysts by computational enzyme design. *Nature* **2008,** *453*, 190-5.
18. Malisi, C.; Kohlbacher, O.; Höcker, B. Automated scaffold selection for enzyme design. *Proteins: Structure, Function, and Bioinformatics* **2009,** *77* (1), 74-83.
19. Nosrati, G. R.; Houk, K. SABER: A computational method for identifying active sites for new reactions. *Protein Science* **2012,** *21* (5), 697-706.
20. Zanghellini, A.; Jiang, L.; Wollacott, A. M.; Cheng, G.; Meiler, J.; Althoff, E. A.; Röthlisberger, D.; Baker, D. New algorithms and an in silico benchmark for computational enzyme design. *Protein Science* **2006,** *15* (12), 2785-2794.
21. Zhu, X.; Lai, L. A novel method for enzyme design. *Journal of computational chemistry* **2009,** *30* (2), 256-267.
22. Seebeck, F. P.; Hilvert, D. Positional Ordering of Reacting Groups Contributes Significantly to the Efficiency of Proton Transfer at an Antibody Active Site. *Journal of the American Chemical Society* **2005,** *127* (4), 1307-1312.
23. Korendovych, I. V.; Kulp, D. W.; Wu, Y.; Cheng, H.; Roder, H.; DeGrado, W. F. Design of a switchable eliminase. *Proceedings of the National Academy of Sciences of the United States of America* **2011,** *108* (17), 6823-6827.

24. Merski, M.; Shoichet, B. K. Engineering a model protein cavity to catalyze the Kemp elimination. *Proceedings of the National Academy of Sciences of the United States of America* **2012,** *109*, 16179-83.
25. Lamba, V.; Yabukarski, F.; Pinney, M.; Herschlag, D. Evaluation of the Catalytic Contribution from a Positioned General Base in Ketosteroid Isomerase. *Journal of the American Chemical Society* **2016,** *138* (31), 9902-9909.
26. Natarajan, A.; Yabukarski, F.; Lamba, V.; Schwans, J. P.; Sunden, F.; Herschlag, D. Comment on "Extreme electric fields power catalysis in the active site of ketosteroid isomerase". *Science* **2015,** *349* (6251), 936.
27. Huang, X.; Xue, J.; Lin, M.; Zhu, Y. Use of an Improved Matching Algorithm to Select Scaffolds for Enzyme Design Based on a Complex Active Site Model. *PloS one* **2016,** *11* (5), e0156559.
28. Fried, S. D.; Bagchi, S.; Boxer, S. G. Extreme electric fields power catalysis in the active site of ketosteroid isomerase. *Science* **2014,** *346* (6216), 1510-1514.
29. Fried, S. D.; Boxer, S. G. Measuring Electric Fields and Noncovalent Interactions Using the Vibrational Stark Effect. *Accounts of Chemical Research* **2015,** *48* (4), 998-1006.
30. Dielmann-Gessner, J.; Grossman, M.; Nibali, V. C.; Born, B.; Solomonov, I.; Fields, G. B.; Havenith, M.; Sagi, I. Enzymatic turnover of macromolecules generates long-lasting protein–water-coupled motions beyond reaction steady state. *Proceedings of the National Academy of Sciences* **2014,** *111* (50), 17857-17862.
31. Grossman, M.; Born, B.; Heyden, M.; Tworowski, D.; Fields, G. B.; Sagi, I.; Havenith, M. Correlated structural kinetics and retarded solvent dynamics at the metalloprotease active site. *Nature structural & molecular biology* **2011,** *18* (10), 1102-1108.
32. Liu, C. T.; Layfield, J. P.; Stewart, R. J.; French, J. B.; Hanoian, P.; Asbury, J. B.; Hammes-Schiffer, S.; Benkovic, S. J. Probing the Electrostatics of Active Site Microenvironments along the Catalytic Cycle for Escherichia coli Dihydrofolate Reductase. *Journal of the American Chemical Society* **2014,** *136* (29), 10349-10360.
33. Wu, Y.; Boxer, S. G. A Critical Test of the Electrostatic Contribution to Catalysis with Noncanonical Amino Acids in Ketosteroid Isomerase. *Journal of the American Chemical Society* **2016,** *138* (36), 11890-11895.
34. Fried, S. D.; Boxer, S. G. Response to Comments on "Extreme electric fields power catalysis in the active site of ketosteroid isomerase". *Science* **2015,** *349* (6251), 936.
35. Chen, D.; Savidge, T. Comment on "Extreme electric fields power catalysis in the active site of ketosteroid isomerase". *Science* **2015,** *349* (6251), 936.
36. Altamirano, M. M.; Blackburn, J. M.; Aguayo, C.; Fersht, A. R. Directed evolution of new catalytic activity using the α/β-barrel scaffold. *Nature* **2000,** *403* (6770), 617-622.
37. Bhowmick, A.; Head-Gordon, T. A Monte Carlo Method for Generating Side Chain Structural Ensembles. *Structure* **2015,** *23* (1), 44-55.
38. Albaugh, A.; Boateng, H. A.; Bradshaw, R. T.; Demerdash, O.; Dziedzic, J.; Mao, Y.; Margul, D. T.; Swails, J.; Zeng, Q.; Case, D. A.; Eastman, P.; Essex, J. W.; Head-Gordon, M.; Pande, V. S.; Ponder, J. W.; Shao, Y.; Skylaris, C.-K.; Todorov, I. T.; Tuckerman, M. E.; Head-Gordon, T. Advanced Potential Energy Surfaces for Molecular Simulation. *J Phys Chem B (Feature article)* **2016,** *in press*.
39. Ponder, J. W.; Wu, C.; Ren, P.; Pande, V. S.; Chodera, J. D.; Schnieders, M. J.; Haque, I.; Mobley, D. L.; Lambrecht, D. S.; DiStasio, R. A., Jr.; Head-Gordon, M.; Clark, G. N.; Johnson, M.

E.; Head-Gordon, T. Current status of the AMOEBA polarizable force field. *J Phys Chem B* **2010,** *114* (8), 2549-64.
40. Ren, P.; Wu, C.; Ponder, J. W. Polarizable Atomic Multipole-based Molecular Mechanics for Organic Molecules. *J Chem Theory Comput* **2011,** *7* (10), 3143-3161.
41. 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. *Journal of the American Chemical Society* **2004,** *126* (26), 8197-8205.
42. Fuxreiter, M.; Mones, L. The role of reorganization energy in rational enzyme design. *Current Opinion in Chemical Biology* **2014,** *21*, 34-41.

5.7 APPENDIX

Parameterization of substrate in reactant and transition state. In order to perform simulations with 5-nitrobenzisoxazole using the AMOEBA force field, we need to obtain parameters for all atoms in the substrate molecule in the reactant and transition states. For parameterizing the transition state of this molecule, the structure of the substrate used for parameterization was reported in Ref [41], in which the C-H and N-O bonds are partially broken and the C-N bond is somewhere between a double and triple bond as shown in Figure 5.2 of the main text. Since the transition state structure is not at it's energy minimum, we do not minimize the structure as done in the original protocol (the reactant state structure is minimized). As can also be seen in Figure 5.1 of the main text, the system used for the parameterization includes not only the ligand but also a base (acetate) to better model the transition state. The overall system has a net charge of -1e.

We then use the protocol described by Ponder and Ren [40] which has 2 main components – first finding the electrostatic parameters and second, finding 'valence' parameters (bond lengths, bond angles, dihedrals). The electrostatic component is described briefly in 6 steps below.

1. Run a single point quantum mechanics-based calculation on the transition state structure using Gaussian g09 at the MP2/6-311G(1D, 1P) level of theory. This calculation returns the electron density as obtained at this relatively low level of theory.
2. Find approximate charges, dipoles and quadrupoles by running the distributed multipole analysis using GDMA on the electron density.
3. Once we have the approximate multipoles, use Tinker's POLEDIT program to break the dipole moments into permanent contributions that act between polarization groups and mutual contributions that act within and between polarization groups.
4. A second Gaussian g09 calculation is run at the MP2/6-311G(2D, 2P) level of theory to obtain a electrostatic potential.
5. In order to obtain a electrostatic potential, we create a spatial grid on which to calculate the potential using Tinker's POTENTIAL program and then compute the potential using the Gaussian CUBEGEN program.
6. Finally, using Tinker's POTENTIAL program, we fit the atomic multipoles to the MP2/6-311G(2D, 2P) electrostatic potential.

After finishing the first step, the 'valence' parameters are assigned from similar, previously parameterized organic compounds. Thus, we model the transition state of the substrate with transition state electrostatics and energy minimized state valence parameters

Calculation of dipole moment of the 3 bonds in EL and EL† states for 5-nitrobenzisoxazole.

We used the monopoles and dipole moments of the parameterized 5-nitrobenzisoxazole in AMOEBA to calculate the dipole moment of the 3 bonds in each state. Table S1 lists the parameters used to calculate the dipole moment of each bond. The positive direction is as shown in Fig 5.2 of the main text. Since the net charge is not zero, we used Δq instead of q to calculate the dipole contribution from the monopoles.

C-H bond (+ve axis from C to H)	**EL**	**EL†**
Permanent dipole (Debye)	-0.93	0.1
Charge on C (e^- units)	0.05	0.05
Charge on H (e^- units)	0.04	0.20
Charge difference (e^- units)	-0.01	0.15
Distance (Å)	1.09	1.31
Net dipole (Debye)	-1.0	1.0

C-N bond (+ve axis from N to C)	**EL**	**EL†**
Permanent dipole (Debye)	-0.1	-0.2
Charge on C (e^- units)	0.05	0.05
Charge on N (e^- units)	-0.27	-0.05
Charge difference (e^- units)	-0.32	-0.1
Distance (Å)	1.36	1.25
Net dipole (Debye)	2.0	0.4

N-O bond (+ve axis from O to N)	**EL**	**EL†**
Permanent dipole (Debye)	0.7	-1.1
Charge on O (e^- units)	0.09	-0.44
Charge on N (e^- units)	-0.27	-0.05
Charge difference (e^- units)	-0.36	0.39
Distance (Å)	1.4	1.8
Net dipole (Debye)	-1.7	2.3

Supplementary Figures

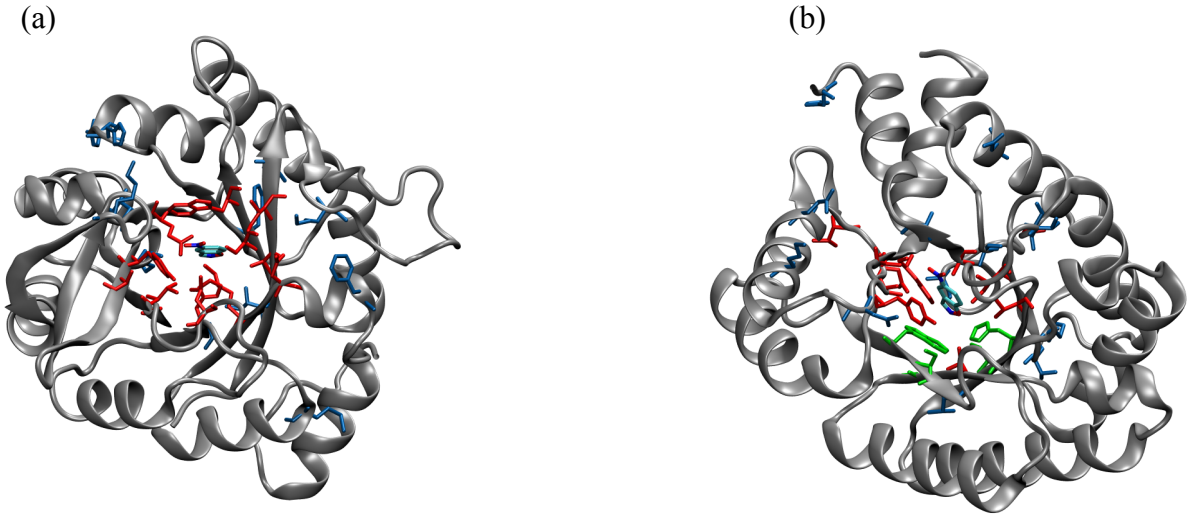


Figure S1. *The Kemp elimination KE07 and KE70 designs*. (a) KE07 involved residues mutated from the original scaffold (red) as well as mutations introduced by LDE shown in blue. (b) KE70 involved residues mutated from the original scaffold (red) as well as mutations made during laboratory DE shown in blue. Additional design mutations via a recombination DE strategy are shown in green.

Table S1: List of top residues that contribute >10 Mv/cm electric field by magnitude at the C-H, C-N, and NO bond in either the EL and EL† states for the designed enzyme KE07 enzyme and the best LDE R7 variant. Positive sign indicates field supporting bond breaking (C-H and N-O) and bond-making (C-N).

KE07 Design			C-H Bond	KE07 R7 Variant		
	Electric Field				Electric Field	
Residue	EL	EL†		Residue	EL	EL†
Glu-101	86.3 (11)	103.6 (11)		Glu-101	142.2 (12)	144.3 (16)
His-201	8.7 (1)	11.2 (1)		His-201	4.1 (1)	7.3 (1)
Gly-202	1.1 (0.1)	1.5 (0.1)		Arg-202	18.7 (1)	18.6 (2)
Glu-46	6.6 (1)	5.3 (0.5)		Glu-46	11.2 (1)	12.1 (2)
Ser-48	-7.6 (2)	-4.3 (2)		Ser-48	-17.1 (2)	-16.1 (3)
Lys-222	-43.5 (6)	-46.3 (6)		Lys-222	-51.8 (6)	-41.5 (6)
Asn-224	1.2 (0.5)	1.7 (0.2)		Asp-224	-16.1 (2)	-10.4 (2)

KE07 Design			C-N Bond	KE07 R7 Variant		
	Electric Field				Electric Field	
Residue	EL	EL†		Residue	EL	EL†
Glu-101	15.5 (7)	18.7 (7)		Glu-101	47.4 (7)	54.4 (12)
His-201	3.9 (2)	5.5 (2)		His-201	16.3 (2)	21.2 (3)
Lys-222	10.6 (7)	16.8 (6)		Lys-222	-3.8 (3)	4.0 (7)
Asn-224	2.2 (0.6)	2.9 (0.5)		Asp-224	-19.8 (2)	-13.6 (2)

KE07 Design			N-O Bond	KE07 R7 Variant		
	Electric Field				Electric Field	
Residue	EL	EL†		Residue	EL	EL†
Glu-101	53.0 (6.0)	65.1 (2.4)		Glu-101	58.5 (2.5)	60.0 (1.3)
His-201	12.7 (2.5)	15.4 (0.7)		His-201	7.3 (2)	7.5 (1)
Gly-202	1.1 (0.1)	1.5 (0.1)		Arg-202	21.4 (2)	21.8 (2.9)
Glu-46	7.3 (1)	6.8 (0.3)		Glu-46	10.2 (1)	10.6 (0.3)
Ser-48	-5.8 (1)	-3.5 (1)		Ser-48	-10.0 (2)	-7.7 (3)
Lys-222	-54.5 (6.1)	-59.5 (3.2)		Lys-222	-45.6 (3.9)	-40.6 (2.4)
Asn-224	1.8 (0.7)	2.6 (0.5)		Asp-224	-19.2 (2.9)	-11.2 (2.3)

Table S2: List of top residues that contribute >10 Mv/cm electric field by magnitude at the C-H, C-N, and NO bond in either the EL and EL† states for the designed enzyme KE70 enzyme and the best LDE R6 variant. Positive sign indicates field supporting bond breaking (C-H and N-O) and bond-making (C-N).

KE70 Design			C-H Bond	KE70 R6 Variant		
	Electric Field				Electric Field	
Residue	EL	EL†		Residue	EL	EL†
His-17	30.0 (4)	47.4 (5)		His-17	48.0 (4)	65.0 (4)
Asp-45	16.1 (2)	17.8 (1)		Asp-45	13.5 (1)	15.1 (1)
Arg-70	-12.1 (2)	-13.4 (1)		Arg-70	-10.9 (1)	-11.5 (1)
Trp-72	9.0 (2)	10.8 (1)		Cys-72	0.8 (1)	0.2 (1)
Ser-138	5.3 (0.5)	8.5 (0.3)		Ala-138	0.7 (0.1)	0.7 (0.1)
Glu-142	-7.6 (0.5)	-6.9 (0.2)		Glu-142	-6.8 (~0)	-7.4 (~0)

KE70 Design			C-N Bond	KE70 R6 Variant		
	Electric Field				Electric Field	
Residue	EL	EL†		Residue	EL	EL†
His-17	22.7 (2)	28.9 (0.5)		His-17	20.5 (2)	25.7 (1)
Asp-45	6.7 (1)	6.7 (0.5)		Asp-45	3.0 (1)	4.4 (0.5)
Arg-70	-1.9 (0.8)	-1.3 (0.6)		Arg-70	2.3 (1)	0.1 (0.4)
Trp-72	9.0 (0.9)	10.8 (0.5)		Cys-72	3.8 (1)	2.2 (0.4)
Ser-138	20.1 (2)	29.3 (1)		Ala-138	2.3 (0.2)	2.8 (0.3)
Glu-142	-0.9 (1)	-1.0 (0.3)		Glu-142	-0.1 (0.4)	-1.3 (0.2)

KE70 Design			N-O Bond	KE70 R6 Variant		
	Electric Field				Electric Field	
Residue	EL	EL†		Residue	EL	EL†
His-17	4.9 (1)	7.7 (1)		His-17	10.5 (1)	12.9 (1)
Asp-45	10.7 (1)	12.5 (0.5)		Asp-45	10.8 (0.6)	11.6 (0.3)
Arg-70	-10.1 (1)	-12.0 (0.6)		Arg-70	-11.7 (1)	-11.2 (0.5)
Trp-72	7.8 (1)	11.9 (0.5)		Cys-72	-3.6 (0.8)	-1.9 (1)
Ser-138	-8.4 (1)	-2.2 (0.7)		Ala-138	-0.5 (0.3)	-0.9 (0.3)
Glu-142	-9.9 (1)	-8.7 (0.3)		Glu-142	-8.4 (0.6)	-9.5 (0.4)

Table S3: *Chemical Positioning vs. Electric Field Environment at the C-H, C-N and O-N Bonds*. The magnitude of the electric field in either the EL and EL† states for the designed KE07 enzyme and the best R7 variant. The active site is defined by residues within 5 Å from the center of the substrate, while the protein environment is summed over all residues outside this region. Solvent includes waters in the neck of the TIM barrel as well as the surrounding hydration and bulk water. Positive sign indicates field supporting bond breaking. Fields are reported in units of Mv/cm

KE07 Design			C-H Bond	KE07 R7 Variant		
	Electric Field				Electric Field	
Region	EL	EL†		Residue	EL	EL†
Base	86.3 (11)	103.6 (12)		Base	142.2 (12)	144.3 (17)
Active Site	1.0 (1)	11.2 (2)		Active Site	2.0 (1)	8.3 (1)
Solvent	-15.6 (4)	-19.2 (1)		Solvent	-22.6 (1)	-20.2 (2)
Protein	-24.1 (7)	-26.8 (7)		Protein	-40.1 (7)	-24.1 (7)

KE07 Design			C-N Bond	KE07 R7 Variant		
	Electric Field				Electric Field	
Region	EL	EL†		Residue	EL	EL†
Base	15.4 (7)	18.7 (7)		Base	47.4 (8)	54.4 (14)
Active Site	18.3 (3)	26.3 (3)		Active Site	26.3 (3)	35.3 (4)
Solvent	6.8 (3)	7.7 (2)		Solvent	2.1 (2)	0.3 (2)
Protein	3.3 (7)	6.2 (6)		Protein	-26.5 (5)	-12.3 (8)

KE07 Design			N-O Bond	KE07 R7 Variant		
	Electric Field				Electric Field	
Region	EL	EL†		Residue	EL	EL†
Base	53.0 (7)	64.1 (3)		Base	58.5 (5)	60.0 (4)
Active Site	4.1 (1)	14.5 (1)		Active Site	11.5 (2)	17.2 (2)
Solvent	-21.4 (3.8)	-19.9 (1.5)		Solvent	-27.4 (2.4)	-23.7 (3)
Protein	-32.0 (7)	-37.0 (4)		Protein	-35.4 (5)	-23.2 (4)

Table S4: *Chemical Positioning vs. Electric Field Environment at the C-H, C-N and O-N Bonds*. The magnitude of the electric field in either the EL and EL† states for the designed KE70 enzyme and the best R6 variant. The active site is defined by residues within 5 Å from the center of the substrate, while the protein environment is summed over all residues outside this region. Solvent includes waters in the neck of the TIM barrel as well as the surrounding hydration and bulk water. Positive sign indicates field supporting bond breaking. Fields are reported in units of Mv/cm

KE70 Design			C-H Bond	KE70 R6 Variant		
	Electric Field				Electric Field	
Region	EL	EL†		Residue	EL	EL†
Base	46.1 (5)	65.1 (5)		Base	61.4 (4)	80.1 (4)
Active Site	16.7 (2)	23.1 (2)		Active Site	2.3 (1)	2.9 (1)
Solvent	2.9 (1)	0.7 (1)		Solvent	1.9 (1)	2.8 (1)
Protein	-12.2 (4)	-11.3 (2)		Protein	-11.6 (2)	-9.1 (1)

KE70 Design			C-N Bond	KE70 R6 Variant		
	Electric Field				Electric Field	
Region	EL	EL†		Residue	EL	EL†
Base	29.4 (3)	35.6 (2)		Active Site	23.5 (2)	30.1 (1)
Active Site	23.2 (4)	32.0 (4)		Active Site	6.1 (2)	8.9 (1)
Solvent	4.6 (1)	6.0 (1)		Solvent	5.9 (1)	4.7 (1)
Protein	-8.6 (4)	-11.4 (2)		Protein	-5.8 (2)	-6.7 (2)

KE70 Design			N-O Bond	KE70 R6 Variant		
	Electric Field				Electric Field	
Region	EL	EL†		Residue	EL	EL†
Base	15.5 (2)	20.2 (1)		Base	21.3 (2)	24.4 (2)
Active Site	2.0 (~0)	16.6 (1)		Active Site	-2.7 (1)	1.9 (1)
Solvent	1.1 (1)	-0.6 (1)		Solvent	0.9 (1)	1.2 (1)
Protein	-9.8 (3)	-8.1 (2)		Protein	-12.6 (2)	-10.7 (1)

Table S5. *Design and laboratory directed evolution mutations for KE07 and KE70.* The computationally designed residues (red) and mutated residues introduced by LDE of a given round (black) have been listed in the table below.

Sequence	KE07 Design	KE07 Best LDE Variant	KE70 Design	KE70 Best LDE Variant
	ILE 7	Asp	HIS 17	
	ALA 9		ALA 19	
	ILE 11		THR 20	Ser
	VAL 12	Met	ALA 21	
	LYS 19		ASP 23	
	SER 48		LYS 29	Asn
	TRP 50		THR 43	Asn
	PHE 77	Ile	ASP 45	
	HIS 84		TYR 48	Phe
	PHE 86		TRP 72	Cys
	GLU 101		SER 74	Gly
	ILE 102	Phe	GLY 101	Ser
	GLN 123		ALA 103	
	TYR 128		SER 138	Ala
	ALA 130		HIS 166	Asn
	LYS 146	Thr	VAL 168	
	VAL 169		THR 171	
	GLY 171		GLY 177	
	LEU 176		ALA 178	Ser
	HIS 201		LYS 197	Asn
	GLY 202	Arg	THR 198	Ile
	MET 207		ILE 202	
	LYS 222		ALA 204	Val
	ASN 224	Asp	ASP 212	
	PHE 229	Ser	ALA 231	
			ALA 235	
			SER 239	Ala
	0.02 1.40 12.2	1.37 0.54 2590	HIS 251 0.14 1.11 126	5.00 0.09 57300

Chapter 6

Conclusion

In this work, I have developed 2 different approaches of studying designed enzymes. The 1st approach, premised on side chain conformational variability showed that both entropy and enthalpy played a coherent role in improving performance of designed KE07 and KE70 enzymes. In addition to the usual transition state stabilization, the calculations showed reactant state destabilization, a less common strategy used by enzymes. Further, it was found that high mutual information sites could serve as a descriptor for picking mutational hotspots. About 50% of the mutations in both enzymes were high information sites. Putting these ideas to test, in chapter 4 we were able to improve another *de novo* enzyme KE15 by an order of magnitude. This is a significant advancement given similar improvements through laboratory directed evolution take 2-7 rounds and a big investment in time and resources.

Enzymes are complicated machines and thus it would require more sophisticated treatment to figure out other metrics of note. I chose electrostatic field stabilization going off recent work in the field of vibrational stark spectroscopy. While the base was found to contribute substantially to the catalytic process, the lack of participation (and even detrimental contribution) by scaffold residues and solvent was a big revelation. Unlike natural enzymes that are known to have optimized scaffolds that help promote the reaction using favorable electrostatic fields, the arbitrary scaffolds used for accommodating the theozyme do not provide good electrostatic relief. The takeaway is that future design attempts should try to incorporate the scaffold electrostatics as well instead of just assuming it to be an inert motif.

The field of enzyme design has come a long way. Given the effect I have shown of side chain fluctuations and electrostatics, it is important that they be considered part of the design protocol. My hope is in the next few years designing efficient enzymes will become a routine procedure and computers will lead the way.