Title
Computational Methods for Analysis and Redesign of Enzyme Active Sites

Author
Nosrati, Geoffrey Ryan

Publication Date
2012

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

COMPUTATIONAL METHODS FOR ANALYSIS AND REDESIGN OF ENZYME ACTIVE SITES

A dissertation submitted in partial satisfaction

of the requirements for the degree Doctor of Philosophy

in Biochemistry & Molecular Biology

by

Geoffrey Ryan Nosrati

2012

ABSTRACT OF THE DISSERTATION

COMPUTATIONAL METHODS FOR ANALYSIS AND REDESIGN OF ENZYME ACTIVE SITES

by

Geoffrey Ryan Nosrati

Doctor of Philosophy in Biochemistry & Molecular Biology

University of California, Los Angeles, 2012

Professor Kendall N. Houk, Chair

The rational design of enzymes and understanding of enzyme mechanisms both present a tremendous challenge in terms of both computational and experimental methodologies. In this work, we review the current state-of-the-art techniques in enzyme design and present a new computational method for active site redesign, called SABER (Selection of Active/Binding sites for Enzyme Redesign). This program was used to analyze both an existing enzyme redesign from the literature, o-succinyl benzoate synthase, and to predict new scaffolds that might be redesigned function similarly to a designed Kemp elimination enzyme. The logic behind the program is discussed in detail with examples taken from the code. Next, SABER was applied to the problem of predicting catalytic functionality in enzyme active sites. This methodology was used to analyze the active sites from two well-studied enzyme families, the serine proteases

and the tyrosine phosphatases, and was found to predict catalytic function with high accuracy. It was then used to analyze the active site of orotidine 5'-monophosphate decarboxylase, an enzyme with an unknown mechanism, and used to support of an imminium-based mechanism of catalysis. Finally, a very challenging rational enzyme design process to catalyze an aromatic Claisen rearrangement is discussed. This led to the design and synthesis of six enzymes, none of which showed any rate acceleration versus the background reaction. During this work, an aromatic Claisen substrate that rearranges rapidly in water was discovered and characterized. This was used to highlight the difficulties in both the enzyme design process and in prediction of hydrogen-bond catalyzed reaction rates in water.

The dissertation of Geoffrey Ryan Nosrati is approved.

Steven G. Clarke

David S. Eisenberg

Jon Fukuto

Jing Huang

Kendall N. Houk, Committee Chair

University of California, Los Angeles

2012

DEDICATION

This dissertation is dedicated to my mother and father.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Chapter 1

This chapter includes material written by G. Nosrati from:

Kiss, G., Johnson, S. A., Nosrati, G., Çelebi-Ölçüm, N., Kim, S., Paton, R. and Houk, K. N. (2011) Computational Design of New Protein Catalysts, in Modeling of Molecular Properties (ed P. Comba), Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany.

Permission was given for use of this work in this dissertation.

K.N. Houk was the PI for this work.

Chapter 2

This chapter is a version of:

Nosrati, G. R., and Houk, K. N. (2012) SABER: A computational method for identifying active sites for new reactions, *Protein Sci 21*, 697-706.

Permission was given for use of this work in this dissertation.

K.N. Houk was the PI for this work.

Chapter 4

This work has been accepted for publication in the journal *Biochemistry*.

K.N. Houk was the PI for this work.

Funding

This work was supported by the National Institute of General Medical Science, National Institutes of Health (GM 75962 and T32 GM008496-16), and the Department of Defense, Defense Advanced Research Projects Agency.

VITA

GEOFFREY R. NOSRATI

EDUCATION:

Institution:	Degree:	Field of Study:
Duke University	B.S.	Chemistry

RESEARCH TRAINING:

10/00 - 10/02 Dr. Vernon Jiang, Genesoft, Inc., South San Francisco, CA.
Design and development of computational models to complement traditional drug-discovery experiments.

3/03 - 6/05 Dr. Daniel Santi, Kosan Biosciences, Hayward, CA.
Design and characterization of novel polyketide synthase complexes.

9/05 – 6/12 Prof. K.N. Houk, Department of Chemistry and Biochemistry, UCLA.
Rational design of enzymes using quantum mechanics and structural bioinformatics. Computational analysis of enzymatic reaction mechanisms.

FELLOWSHIPS AND AWARDS:

2005 Summer Research Fellowship, UCLA
2005 University Research Fellowship, UCLA
2006 First Year Academic and Research Award, UCLA
2007-2010 Chemistry-Biology Interface Program, Trainee
2008-2012 Majeti Research Fellow, UCLA

TEACHING EXPERIENCE:

2005-2011 Biochemical Methods Laboratory course 154L
Biochemistry 153A
Chemistry of Biology 105A/205A
General Chemistry 11A/12A

PUBLICATIONS:

Kiss, G., Johnson, S. A., Nosrati, G., Çelebi-Ölçüm, N., Kim, S., Paton, R. and Houk, K. N. (2011) Computational Design of New Protein Catalysts, in Modeling of Molecular Properties (ed P. Comba), Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany.

Nosrati, G. R., and Houk, K. N. (2012) SABER: A computational method for identifying active sites for new reactions, *Protein Sci 21*, 697-706.

Nosrati, G. R., and Houk, K. N. (2012) Using Catalytic Atom Maps to Predict the Catalytic Functions Present in Enzyme Active Sites, *Biochemistry,* in press.

CONFERENCES:

G. Nosrati and K.N. Houk. Catalytic Atom Mapping: A Scaffold Selection Method for the Rational Redesign of Active Sites. Poster presented at: ACS 2010. 239th National Meeting of the American Chemical Society, 2010 Mar 23, San Francisco, CA.

G. Nosrati and K.N. Houk. Rational Redesign of Active Sites Using Structural Bioinformatics. Poster presented at: UCLA Molecular Biology Institute Conference, 2009 Oct 17, Lake Arrowhead, CA.

G. Nosrati and K.N. Houk. Gess: A Computational Method for the Analysis of Catalytic Geometries in Active Sites. Poster presented at: David S. Sigman Symposium, 2009 Jan 22, Los Angeles, CA.

G. Nosrati and K.N. Houk. Investigating the Reaction Mechanism of ODCase using Structural Bioinformatics. Poster presented at: UCLA Molecular Biology Institute Conference, 2008 Oct 18, Lake Arrowhead, CA.

CHAPTER 1. ENZYME DESIGN USING ROSETTA

1.1 Introduction

This chapter is an introduction to the basic concepts and techniques for rational enzyme redesign using the Rosetta software, developed by David Baker's group at the University of Washington. For a more thorough discussion of this topic, please see the chapter written by me and members of our research group in the book *Modeling of Molecular Properties*. (*1*) This chapter was written by Gert Kiss, Scott Johnson, me, Nihan Çelebi-Ölçüm, Seonah Kim, Robert Paton, and Prof. K.N. Houk.

It is important to note that enzyme design using Rosetta is significantly different from the protein design process using Rosetta software, which has been used to generate novel protein folds, such as TOP7. (*2*) Rosetta-based enzyme design takes existing protein structures with known active and/or binding sites and then grafts a new set of catalytic groups into these sites. Thus, these enzyme designs are best described as active site designs rather than as full-scale enzyme designs, as the vast majority of the protein structure is unchanged during the enzyme design process. The most up-to-date Rosetta documentation is available online at the RosettaCommons website, and in reference (*3*).

The remainder of this chapter will provide an overview of the current state-of-the-art for Rosetta-based enzyme design. The discussion will begin with the process of designing a new active site, followed by methods for placing the designed active site and transition state into an existing protein structure, and methods for refining an initial

enzyme design into a final candidate for experimental testing. The experimental evaluation process will also be described, as will the three successful published designs.

1.2 Active Site Design

The active site of a designed enzyme can come from a variety of sources. Optimal placement of the catalytic residues can be calculated using quantum mechanical methods to generate a theozyme, as has been the case for the three published designs. (*4*) It is also possible to design active sites based on existing catalysts, such as catalytic antibodies or other enzymes. The latter is useful in cases where a catalytic function needs to be moved into a different protein scaffold.

In general, the use of quantum mechanics has been the most successful approach to designing active sites for new enzymes. These calculations are performed using the transition state of the target reaction, along with mimics of the amino acid sidechains and/or main chain functionalities that are predicted to stabilize the transition state. An example of a Kemp elimination theozyme can be seen in Chapter 2. The theozyme provides the ideal geometric placement of the catalytic groups relative to the transition state.

Theozymes are generally calculated at the B3LYP/6-31g(d) level of theory and basis set. An implicit solvation model, CPCM, is usually employed with a dielectric

constant of 4, to simulate the protein environment the transition state and catalytic residues are likely to experience. The rate of the background reaction in water is also calculated using the same methodology, but without any stabilizing groups and with the dielectric constant of the implicit solvation model set to 80, to predict the activation barrier for the uncatalyzed reaction in pure water. Comparison of the calculated activation barrier using the theozyme to the activation barrier in waters allows for prediction of the maximum possible rate acceleration for a given theozyme.

1.3 The Match Stage: Incorporation of Designed Active Sites into Existing Protein Structures

Once the target active site geometry has been determined, either through quantum mechanical calculations or based on an existing catalytic site, the next step in the design of a new enzyme is to incorporate the designed active site into an existing protein. This is generally referred to as the matching stage. During this stage, the goal is simply to place the new transition state and catalytic groups in an existing binding or active site. The incorporation of binding groups and the remaining residues in the active site will be placed during the design stage.

The placement of the theozyme and catalytic groups can be accomplished using the RosettaMatch program. (*2*) This program uses a scaffold library with defined active sites for redesign. At the start of the matching stage, all of the residues in the site to be redesigned are converted into glycine residues, to allow the active site to accommodate

the new trasition state. Next, RosettaMatch is used to seek a match of a low-energy rotamer of each catalytic group to the backbone of the protein in such a way so as to achieve a theozyme-like geometry in the designed active site. The catalytic residue placement is done sequentially, which limits the number of catalytic residues that can be placed accurately to 3 at present.

Catalytic residue placement can also be accomplished using the SABER program for enzyme design, which was developer here and is discussed extensively in Chapter 2. Briefly, instead of trying to graft the theozyme to an existing library of protein scaffolds, SABER searches the entire Protein Data Bank for proteins that already have theozyme-like geometries in their active and/or binding sites. This approach allows for a very large number of potential scaffolds to be explored. It is dependent on there being an existing structure in the Protein Data Bank with a theozyme-like arrangement of catalytic groups, however.

Whatever method is used for placement of the transition state and catalytic residues, the key criterion for evaluation at the match stage is geometric fidelity to the theozyme. A good match, then, is one that can both accommodate the new transition state and can place the catalytic groups in their ideal positions for catalysis.

1.4 The Match Stage: Considerations for Maximizing Diversity

Given that enzyme design is far from trivial, it should be expected that a large and diverse set of designs will have to be generated in order to locate even a single design that will be active during experimental characterization. Diversity can be maximized by exploring three aspects of design: 1) catalytic groups, 2) catalytic group placement in the theozyme, and 3) scaffold selection.

1.4.1 Catalytic Group Selection

Even a simple theozyme like the base/H-bond donor used for catalysis of the Kemp elimination described in Chapter 2 provides some diversity in terms of the catalytic groups that can be chosen. For example, the base used in the theozyme is a carboxylate sidechain, which could be either from an aspartate or glutamate residue. Likewise, the H-bond donor could be from a protonated lysine residue, or from any of the residues with an alcohol group on their sidechains (serine, threonine, tyrosine). All combinations of potential base and H-bond donor should be explored, as long as they are predicted to provide significant rate acceleration.

1.4.2 Catalytic Group Placement

As is the case with catalytic group selection, the placement of these groups relative to the transition state can also be explored to maximize match diversity. If the transition state has rotatable bonds that can be sampled without increasing the activation free energy, this allows for different placements of the catalytic groups in the theozyme. Since it is impossible to predict which catalytic group placements will be compatible with a given scaffold, diversifying the locations of the catalytic groups can lead to greater match diversity.

1.4.3 Scaffold Selection

Finally, the diversity of scaffolds that can accommodate the theozyme should also be explored. It is impossible to know if one particular scaffold (for example, β-propeller) will be problematic for a given theozyme during the protein expression and purification phases. As such, creating high-quality matches in as many different types of scaffolds as possible is advantageous. To date, there have been successful enzyme designs in the TIM barrel, jelly roll, and β-propeller scaffolds. Until a process for selecting the ideal scaffold for a designed enzyme is developed, the odds of success are maximized by generating matches in a number of different scaffolds. It is also essential to select scaffolds that express well in an organism that makes protein expression relatively easy (usually *E. coli*).

1.5 The Design Stage: Filling the Remainder of the Active Site (Non-catalytic residues)

After the transition state and the catalytic residues are placed during the match process, there will still be non-catalytic residues that need to be changed. The RosettaDesign program is used to rebuild the remainder of the active site, removing any steric clashes, optimizing packing around the catalytic residues and transition state, and adding binding interactions. After a single round of RosettaDesign, the potential designs are evaluated according to a number of criteria:

a) Geometric fidelity to the theozyme (can change during the design process)
b) RosettaEnergy score
c) Rosetta Packing score
d) Rosetta Ligand energy score

Measuring geometric fidelity to the theozyme is the same as during the match stage. The RosettaEnergy score is a measure of the overall energy of the designed enzyme, and is compared to the energy score of the wild-type scaffold. A significant change in the energy score may indicate a protein that has been destabilized and will be less likely to fold correctly in solution. Similarly, the packing score is compared to that of the wild-type protein. As with the energy score, a significant change in the packing score is an indication that the protein might express and/or fold poorly. The

ligand energy score should be ≤ 0, indicating that there are no repulsive interactions between the designed enzyme and the transition state.

1.6 The Design Stage: Mutation Reversion and Additional Rounds of Design

The design process continues with the best candidates identified during the first round of design. At this point, it is necessary to manually inspect each potential design and examine the mutations made by RosettaDesign. These designs are further improved in three stages: 1) reversion of unnecessary mutations, 2) additional rounds of design for better sequence sampling, and 3) unconstrained evaluation of final enzyme designs.

1.6.1 Reversion of Unnecessary Mutations

During the inspection of each design, it is likely that mutations will be found that are distant from the transition state. These mutations are generally reverted back to their original (wild-type) residues. In addition, any mutations to proline, arginine, or tryptophan should be scrutinized carefully. Prolines are problematic because they affect the protein backbone flexibility directly and can add kinks to α-helices. Arginine and tryptophan have bulky sidechains that are often difficult to pack the protein structure around. These residues are also often reverted to wild-type, or converted to residues

with less bulky sidechains. Finally, any residue converted to a glycine should be mutated into an alanine. Once these changes are made, a second round of design should be performed, and the new design should be evaluated according to the geometric, energetic, and packing criteria described above.

1.6.2 Additional Rounds of Design

Any designs that make it past the second round should be subjected to an additional round using RosettaDesign, to allow for additional sampling of the possible mutations in the active site. The residues sampled can be controlled using a resfile (see Rosetta documentation for details). Since each round of design starts with a random seed, this allows for the exploration of more potential structures and fine-tuning of the energy, packing, and ligand energy scores. The third round of design should be set to output many (20-100) structures, to maximize this sampling. Generally, these designs will converge on a few sequences, the best of which can be selected according to the evaluation criteria.

1.6.3 Unconstrained Evaluation of Designs

This is the final stage of design evaluation using RosettaDesign. Each design is originally created using a constraint (.cst) file that forces the design to maintain a theozyme-like geometry whenever possible. The third round designs should be evaluated by setting the constraints for the theozyme to 1, and then performing a single

round of repacking using RosettaDesign with no mutations. If the transition state stays in place, the design passes this test. If it moves out of the protein structure, this indicates significant repulsion and the design will need additional work.

1.7 Post-design Evaluation

The enzyme designs that satisfy all of the third-round criteria are generally subjected to additional analysis via molecular dynamics. This technique has been shown to discriminate active designs from inactive designs, and has been used to convert an inactive enzyme design into an active one. (*5*)This methodology is discussed extensively in the book listed at the beginning of this chapter, so it will not be discussed in detail here.

1.8 Experiments

A final concern is the experimental stage of the enzyme design process. Once designs are complete, it is necessary to convert them into DNA sequences that can be synthesized. At this stage, it often makes sense to make additional modifications to the protein sequence for convenience in purification, such as adding a His tag to one end of the protein. After the genes have been received, they must be cloned into expression vectors and transformed into a suitable host, usually *E. coli*. Following transformation,

the protein must be expressed and purified. If soluble, the purified protein is analyzed via circular dichroism to verify the presence of secondary structure, which indicates a protein that is at least partially folded.

Once the designed enzyme has been purified, activity assays can be performed. These experiments vary, depending on the substrate and product. Ideally, the product will have an optical property that can be easily and rapidly measured by a technique such as UV/vis spectrophotometry or fluorimetry, as was the case with two of the successful enzyme designs. Alternatively, it is possible to use techniques such as LC/MS (liquid chromatography/mass spectrometry) if the products are optically inactive. This method has much lower throughput, but can be used when necessary.

1.9 Published Designs

1.9.1 Kemp Elimination

This series of designed enzymes accelerates the reaction shown in Figure 1.1. A catalytic base is used to abstract a hydrogen from the benzisoxazole ring, leading to ring opening and the formation of the product. A hydrogen bond donor is also used to stabilize the transition state. Of the hundreds of designs that were generated, 59 were expressed, and of the expressed designs, 8 were found to be active. The top design had a k_{cat}/k_{uncat} of 2.5 x 10^5, which was further improved via directed evolution. (*6-8*)

One of the moderately active Kemp designs, KE59, has been extensively optimized via directed evolution to obtain kcat/kuncat values of ~10^6, showing that this technique can be effective in improving the activity of designed enzymes. (*9*)

A molecular dynamics protocol has been developed to separate active from inactive designed enzymes, and has been applied extensively to the Kemp elimination. (*5*) The designed enzymes for this reaction have also been studied using quantum mechanics/molecular mechanics (QM/MM) and free energy perturbation (FEP) methods. (*10*) Finally, the Mayo and Houk groups have used an iterative computational method to convert an inactive designed Kemp eliminase into an active enzyme for this reaction. (*11*)

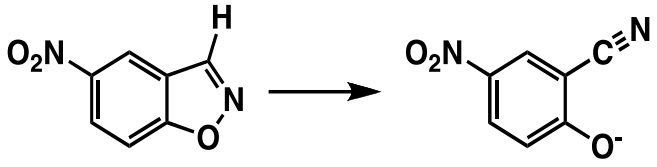


Figure 1.1 The Kemp elimination reaction.

1.9.2 Retro-Aldol

These enzymes catalyze the retro-aldol reaction shown in Figure 1.2. These enzymes are noteworthy in that they form a covalent intermediate with the substrate, using a lysine nucleophile. (*12*) As with the Kemp elimination, hundreds of designs were generated computationally. 72 of these designs were expressed, and 32 were

found have detectable rate accelerations over the background reaction. The best design had a k_{cat}/k_{uncat} of 2.4 x 10^4. This work has continued since the publication of the original design, adding 33 additional active designs in a number of new scaffolds. (*13*) The active sites of the designed retro-aldolase enzymes have been studied extensively using mutational analysis. (*14, 15*)

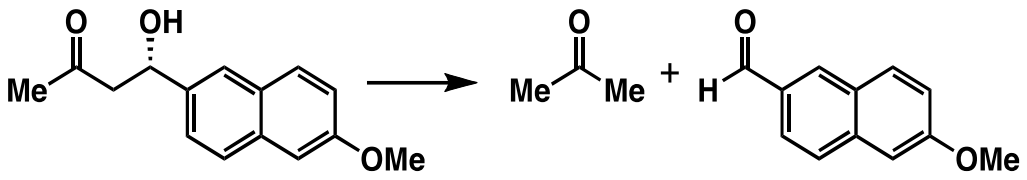


Figure 1.2. The retro-aldol reaction catalyzed by designed enzymes.

1.9.3 Stereoselective Diels-Alder

A third successful enzyme design catalyzed a stereoselective Diels-Alder reaction, in which only one of the 4 possible products was observed. (*16*) This reaction is shown in Figure 1.3. Like the previous designed enzymes, hundreds of designs were generated computationally, and two were found to be active via experiment. The rate acceleration over the background reaction was ~10^2-fold for the best design. However, the stereoselectivity of this enzyme is extremely good, as only one of the products was detected via LC/MS. The activity of the best design has since been enhanced 18-fold as a result of the crowdsourced design effort of the FoldIt game, from the Baker group. (*17*)

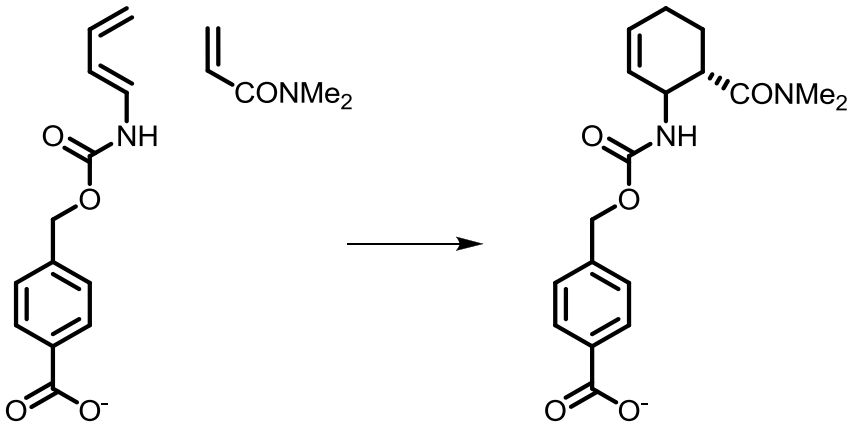


Figure 1.3. The stereoselective Diels-Alder reaction.

1.10 Conclusions

The rational design of new enzymes continues to present an enormous challenge for computational chemists and biochemists. As can be seen from the three examples in Section 1.9, the current rate accelerations over the background reaction are modest, in the range of 10^4 - 10^5. These accelerations are in the same range as catalytic antibodies and other catalysts that generally use non-covalent interactions to stabilize transition states. For comparison, an average natural enzyme has an acceleration over background of 10^{12}. In addition, the chance of a designed enzyme being active experimentally is low, and for this and other reasons, going from a theozyme to a successful design is a time-consuming process.

In the future, we hope to expand upon this work, making more active designs with a higher success rate. We have developed the SABER program, described in Chapter 2, to maximize the diversity and theozyme-likeness of designs. We also plan to incorporate more advanced active sites, using covalent and/or metal based mechanisms, in order to achieve greater rate accelerations.

1.11 References

1. Kiss, G., Johnson, S. A., Nosrati, G., Çelebi-Ölçüm, N., Kim, S., Paton, R., and Houk, K. N. (2011) Computational Design of New Protein Catalysts, In *Modeling of Molecular Properties* (Comba, P., Ed.), Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany.

2. Rohl, C. A., Strauss, C. E. M., Misura, K. M. S., and Baker, D. (2004) Protein structure prediction using rosetta, *Methods Enzymol. 383*, 66-93.

3. Richter, F., Leaver-Fay, A., Khare, S. D., Bjelic, S., and Baker, D. (2011) De Novo Enzyme Design Using Rosetta3, *Plos One 6*, 1-12.

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

5. Kiss, G., Rothlisberger, D., Baker, D., and Houk, K. N. (2010) Evaluation and ranking of enzyme designs, *Protein Sci. 19*, 1760-1773.

6. Rothlisberger, 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., and Baker, D. (2008) Kemp elimination catalysts by computational enzyme design, *Nature 453*, 190-195.

7. Khersonsky, O., Rothlisberger, D., Dym, O., Albeck, S., Jackson, C. J., Baker, D., and Tawfik, D. S. (2010) Evolutionary Optimization of Computationally Designed Enzymes: Kemp Eliminases of the KE07 Series, *J. Mol. Biol. 396*, 1025-1042.

8. Khersonsky, O., Rothlisberger, D., Wollacott, A. M., Murphy, P., Dym, O., Albeck, S., Kiss, G., Houk, K. N., Baker, D., and Tawfik, D. S. (2011) Optimization of the In-Silico-Designed Kemp Eliminase KE70 by Computational Design and Directed Evolution, *J. Mol. Biol. 407*, 391-412.

9. Khersonsky, O., Kiss, G., Rothlisberger, D., Dym, O., Albeck, S., Houk, K. N., Baker, D., and Tawfik, D. S. (2012) 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. 109*, 10358-10363.

10. Alexandrova, A. N., Rothlisberger, D., Baker, D., and Jorgensen, W. L. (2008) Catalytic Mechanism and Performance of Computationally Designed Enzymes for Kemp Elimination, *J. Am. Chem. Soc. 130*, 15907-15915.

11. Privett, H. K., Kiss, G., Lee, T. M., Blomberg, R., Chica, R. A., Thomas, L. M., Hilvert, D., Houk, K. N., and Mayo, S. L. (2012) Iterative approach to computational enzyme design, *Proc. Natl. Acad. Sci. U.S.A. 109*, 3790-3795.

12. Jiang, L., Althoff, E. A., Clemente, F. R., Doyle, L., Rothlisberger, D., Zanghellini, A., Gallaher, J. L., Betker, J. L., Tanaka, F., Barbas, C. F., 3rd, Hilvert, D., Houk, K. N., Stoddard, B. L., and Baker, D. (2008) De novo computational design of retro-aldol enzymes, *Science 319*, 1387-1391.

13. Althoff, E. A., Wang, L., Jiang, L., Giger, L., Lassila, J. K., Wang, Z. Z., Smith, M., Hari, S., Kast, P., Herschlag, D., Hilvert, D., and Baker, D. (2012) Robust design and optimization of retroaldol enzymes, *Protein Sci. 21*, 717-726.

14. Lassila, J. K., Baker, D., and Herschlag, D. (2010) Origins of catalysis by computationally designed retroaldolase enzymes, *Proceedings of the National Academy of Sciences of the United States of America 107*, 4937-4942.

15. Wang, L., Althoff, E. A., Bolduc, J., Jiang, L., Moody, J., Lassila, J. K., Giger, L., Hilvert, D., Stoddard, B., and Baker, D. (2012) Structural Analyses of Covalent Enzyme-Substrate Analog Complexes Reveal Strengths and Limitations of De Novo Enzyme Design, *J. Mol. Biol. 415*, 615-625.

16. Siegel, J. B., Zanghellini, A., Lovick, H. M., Kiss, G., Lambert, A. R., St Clair, J. L., Gallaher, J. L., Hilvert, D., Gelb, M. H., Stoddard, B. L., Houk, K. N., Michael, F. E., and Baker, D. (2010) Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction, *Science 329*, 309-313.

17. Eiben, C. B., Siegel, J. B., Bale, J. B., Cooper, S., Khatib, F., Shen, B. W., Players, F., Stoddard, B. L., Popovic, Z., and Baker, D. (2012) Increased Diels-Alderase activity through backbone remodeling guided by Foldit players, *Nat. Biotechnol. 30*, 190-192.

CHAPTER 2. SABER: A COMPUTATIONAL METHOD FOR IDENTIFYING ACTIVE SITES FOR NEW REACTIONS

2.1 Abstract

A software suite, SABER (Selection of Active/Binding sites for Enzyme Redesign), has been developed for the analysis of atomic geometries in protein structures, using a geometric hashing algorithm (Barker et al. (2003), *Bioinformatics 19*, 1644-1649). SABER is used to explore the Protein Data Bank (PDB) to locate proteins with a specific 3D arrangement of catalytic groups to identify active sites that might be redesigned to catalyze new reactions. As a proof-of-principle test, SABER was used to identify enzymes that have the same catalytic group arrangement present in o-succinylbenzoate synthase (OSBS). Among the highest-scoring scaffolds identified by the SABER search for enzymes with the same catalytic group arrangement as OSBS were L-Ala D/L-Glu epimerase (AEE) and muconate lactonizing enzyme II (MLE), both of which have been redesigned to become effective OSBS catalysts, demonstrated by experiments. Next, we used SABER to search for naturally existing active sites in the PDB with catalytic groups similar to those present in the designed Kemp elimination enzyme KE07. From over 2000 geometric matches to the KE07 active site, SABER identified 23 matches that corresponded to residues from known active sites. The best of these matches, with a 0.28 Å catalytic atom RMSD to KE07, was then redesigned to be compatible with the Kemp elimination using RosettaDesign. We also used SABER to search for potential Kemp eliminases using a theozyme predicted to provide a greater rate acceleration than the active site of KE07, and used Rosetta to create a design based on the proteins identified.

2.2 Introduction

The rational design of enzymes for new reactions is one of the greatest challenges in computational biology. There have been four major recent successes for the design of enzymes that are capable of accelerating reactions not found in nature. Nevertheless, many hurdles must be overcome in order to make the rational design process robust and general. (*1*-*6*)

The most successful computational enzyme designs thus far involve catalysis of a Kemp elimination, a retro-aldol reaction, a stereoselective Diels-Alder reaction, and nitric oxide reduction. (*3*-*6*) The first three enzymes were designed using the "inside-out" approach, in which active sites are designed by quantum mechanical methods; transition states with catalytic functionality from appropriate side chains of amino acids are optimized to find a suitable geometrical arrangement that is predicted to accelerate the reaction. This is referred to as a theoretical enzyme, or theozyme. (*7*) Subsequently, a protein structure is designed that is predicted to fold to form this active site. Because the *de novo* design of folds is not routinely feasible, the Rosetta programs developed by the Baker laboratory are used to find a suitable fold into which the theozyme can be incorporated. (*8, 9*) RosettaMatch is used to determine whether the theozyme can be grafted into one or more of the scaffolds in a scaffold library. This must be achieved with low energy conformations of the side chains involved in the theozyme. Following this step, RosettaDesign is then used to fill in the remaining side

chains in the active site around the theozyme, optimizing protein packing and transition state binding. (*9*)

A critical step in enzyme design is the proper placement of the catalytic residues. The importance of the positioning of the residues in the active site has been discussed extensively and is clearly a feature of proficient enzymes. (*10*) Many examples have been described in the literature: Warshel has proposed that active site preorganization and electrostatic stabilization of the transition state are the principal factors controlling enzyme catalysis. (*11-13*) Preorganization involves the correct spatial positioning of catalytic groups. Hilvert has demonstrated that mutating a catalytic Glu residue to an Asp in the 34E4 Kemp eliminase catalytic antibody has a significant (> 2 kcal/mol transition state destabilization) effect on catalysis, indicating the need for precise placement of catalytic groups. (*14*) A recent investigation of serine esterases has shown that their active sites are preorganized into geometries that allow the reaction to be carried out with a minimal rearrangement of catalytic residues in the many steps of the catalytic cycle. These geometries are very close to the optimum geometries computed using quantum mechanics. (*15*) Significant deviations from the optimum catalytic arrangement of residues are generally not found in nature; computational tests have been carried out on a wide variety of enzymes to show that evolution leads to active sites with optimum catalytic distances, according to comparisons with quantum mechanical calculations. (*16*)

For the reasons discussed above, the selection of a scaffold that can support correctly positioned and oriented catalytic groups is an essential feature of enzyme design. In addition, an ideal scaffold should provide an environment such that the pK_a values of catalytic acids or bases and the reactivties of nucleophilic or electrophilic groups are optimal. Finally, the designed active site must be sufficiently isolated from bulk aqueous solvent to allow for efficient catalysis to take place in an optimum environment.

Mutations distant from the active site can also have a significant impact on enzyme catalysis. (*17*) In addition, dynamic effects, such as loop motions, may also play a key role in determining the catalytic power of a designed enzyme, as suggested by molecular dynamics simulations of designed retro-aldolases. (*18*) However, whether dynamical motions coupled to the reaction coordinate impact enzyme catalysis is still a subject of considerable debate. (*19-22*) All of these factors contribute to the enormous challenge inherent in rational enzyme design, as even minor changes can have significant effects on catalysis. As these effects are difficult to predict, one way to limit their impact is to keep the number of mutations required for active site redesign to a minimum. The use of active sites that require minimal engineering to catalyze new reactions has been the subject of two recent reviews by the Hilvert, Gerlt, and Babbitt laboratories. (*23, 24*)

We have explored an alternative to grafting catalytic groups into a protein scaffold. Instead, we search for a natural protein that has the necessary catalytic groups

appropriately positioned for catalysis of the desired reaction. Essentially, we are searching for active sites that might be promiscuous for a target reaction, or might be engineered to be promiscuous for the target reaction with mutations to the non-catalytic residues in the active site.

This approach involves searching the public repositories of protein structures, such as the Protein Data Bank (PDB), to determine if there are enzymes that have active site residues in a geometry capable of catalyzing the reaction of interest. If a protein with the necessary preorganized catalytic groups is found, mutations necessary to accommodate the new transition state, and to provide additional stabilizing groups, can then be incorporated through protein design programs such as RosettaDesign, Dezymer, or Phoenix, or by experimental approaches such as directed evolution. (*8, 25-27*) This approach can also have the benefit of providing an active site with catalytic groups in an environment that provides suitable pK_a values for catalysis without extreme redesign.

We have developed a method for the identification of active sites with geometries suitable for a desired reaction. This procedure uses the program we have developed, SABER, and a Catalytic Atom Map (CAM), a truncated model of the ideal geometry of the key catalytic residues in the active site. SABER is used to search the Protein Data Bank for known enzymes with CAM-like geometric arrangements of functional groups.

The program combines geometric searches of the PDB to locate CAM-like geometries in known protein structures and informatics tools to evaluate the suitability of these geometries to function as catalytic sites. We have validated this methodology using a known example from the literature in which the activity of one enzyme, o-succinyl benzoate synthase, has been transplanted into two other known enzymes. We have also used SABER in conjunction with RosettaDesign to computationally design a new enzyme that is closely related to a previously designed and active Kemp eliminase, KE07.

2.3 Results

SABER as a tool for enzyme design

SABER is described in detail in the Materials and Methods section. Briefly, this program differs from other enzyme design approaches in terms of how the active site residues are located and identified. SABER is used to locate protein structures that have their catalytic groups arranged in an optimal geometry for the catalysis of the target reaction. The program rapidly searches and scores the entire Protein Data Bank, providing a very large pool of potential active or binding sites for redesign. The scoring functions of SABER rapidly identify the most promising sites for redesign, based both on the geometric placement of the proposed catalytic residues and whether or not they are part of a known active or binding site and have predicted pK_a values of catalytic residues appropriate for the target reaction. We describe how SABER can identify the best set of active site redesign candidates, which we term “pre-designs”, from a pool of thousands of geometric matches to the Catalytic Atom Map.

Using Catalytic Atom Maps to select active sites capable of catalyzing new reactions

The Gerlt and Babbitt laboratories have altered the active sites of two enzymes with different catalytic functions to perform the reaction catalyzed by o-succinyl benzoate synthase (OSBS). The two enzymes redesigned to perform the OSBS reaction were L-Ala D/L-Glu epimerase (AEE) and muconate lactonizing enzyme II

(MLE). The key catalytic residues were unchanged by the mutations, and catalysis was verified by experiment. (*28*) We used this example to test the ability of SABER to locate catalytic functionality in cases where experimentalists have already proven the effectiveness of a modified enzyme to catalyze a reaction not found in the native enzyme. A Catalytic Atom Map based on the catalytic residues of OSBS was used to search the Protein Data Bank for scaffolds that have their catalytic groups arranged appropriately to catalyze this reaction.

OSBS, AEE, and MLE all share geometrically similar active sites, with 3 conserved Asp residues, used to bind Mg^{2+}, and 2 conserved Lys residues. (*28*) The essential Mg^{2+} stabilizes an enediolate intermediate in each enzyme; however, the reaction catalyzed by each wild-type enzyme is quite different. AEE catalyzes the epimerization at the alpha carbon of its target amino acids. MLE catalyzes the lactonization of muconate, and OSBS catalyzes a dehydration to form o-succinyl benzoate. These reactions are depicted in Figure 2.1. It has been demonstrated experimentally that OSBS activity can be introduced into both AEE and MLE, through mutations to their active sites, even though neither of these enzymes shows any native OSBS activity. (*28-30*) These mutations leave the key catalytic residues intact, but optimize the rest of the active site for the new substrate and transition state.

L-ala/D-glu epimerase (AEE)

It has been demonstrated experimentally that wild-type AEE does not catalyze the OSBS reaction. (*28*) However, this enzyme contains a similar active site, as indicated by the Catalytic Atom Map search. OSBS activity was installed by Gerlt et al. in the E. coli AEE via a single mutation, changing Asp297 to a glycine. This single residue change improves k_{cat} for the OSBS reaction from undetectable to 2.5 x 10^{-3} sec^{-1}. (*28*) This is a modest acceleration compared to a wild-type OSBS enzyme, which has a k_{cat} of 24 sec^{-1} in *E. coli*. (*28*) However, it still represents an acceleration of approximately seven orders of magnitude over the uncatalyzed reaction. Further improvements in k_{cat} were achieved with two additional mutations (I19F and R24W), which improved it to 1.6 x 10^{-1} sec^{-1}, which is a 10^{9}-fold rate enhancement versus the uncatalyzed reaction. (*29, 30*)

Muconate lactonizing enzyme (MLE)

MLE has two catalytic lysines and three catalytic carboyxlates arranged very similarly to the active site of OSBS. This enzyme has been re-engineered by Gerlt and coworkers to perform the OSBS reaction via a single amino acid change. In this case, MLE II from *Pseudomonas sp.* P51 was able to catalyze the OSBS reaction after changing Glu323 to a glycine. (*28*) This single residue change increased k_{cat} from undetectable to 1.5 sec^{-1}, compared to the activity of wild-type *E. coli* OSBS, at 24 sec^{-1}. This represents an approximately 10^{10}-fold enhancement of the rate versus the

background reaction. Unlike the AEE case, additional mutation experiments to improve the rate have not been published. However, this single amino acid change produces an enzyme that is within an order of magnitude of wild-type activity.

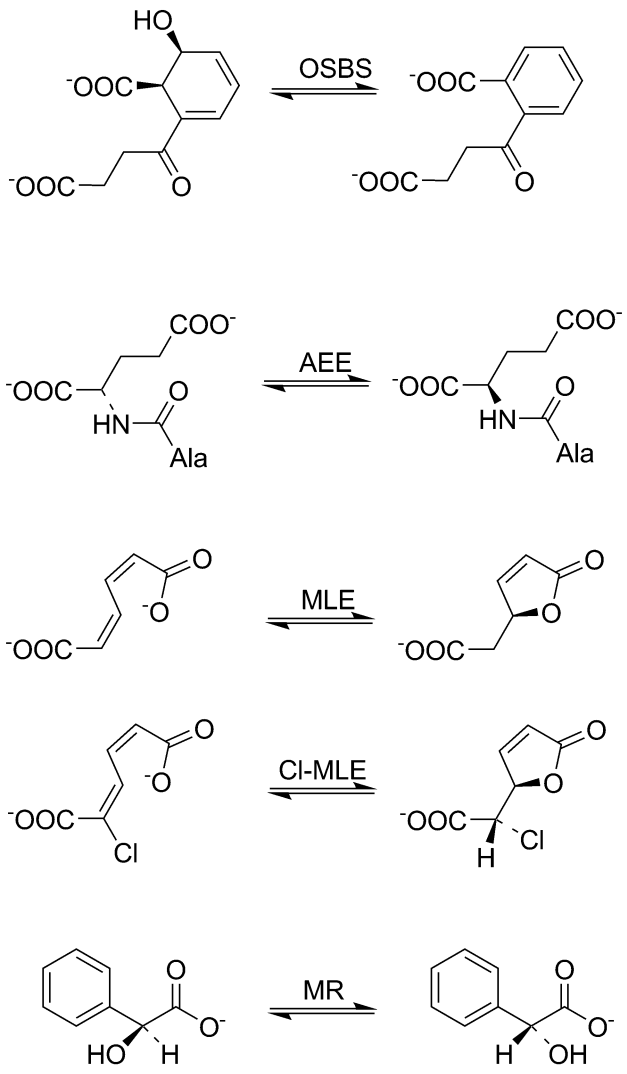


Figure 2.1. The reactions catalyzed by o-succinyl benzoate synthase (OSBS), L-ala/D-glu epimerase (AEE), muconate lactonizing enzyme (MLE), chloromuconate lactonizing enzyme (Cl-MLE), and mandelate racemase (MR).

We used these examples to test the effectiveness of SABER at identifying suitable candidates for active site redesign. We searched all structures in the PDB90 data set with a resolution ≤ 2.0 Å to locate proteins with arrangements of atoms matching the CAM of the OSBS active site. A five atom map was constructed for the target active site to represent the 3 carboxylic acids and 2 lysines in the OSBS active site. This is shown in Figure 2.2. These atoms were constrained by atom type, residue type, and interatomic distances.

The three carboxylate ligands for Mg^{2+} are defined by the three oxygen atoms in the Catalytic Atom Map. These oxygens must be from an aspartate or glutamate (PDB atom codes OD1, OD2, OE1, OE2), and the nitrogen atoms must be from lysine residues. The Catalytic Atom map specifies that the only nitrogen matches must be lysine ε-amino nitrogens, as naturally occurring OSBS enzymes use lysine exclusively in this role. The match radius for each atom was set at 2.0 Å.

All of the SABER pre-designs located using the OSBS Catalytic Atom Map where the RMSD was ≤ 0.6 Å were analyzed. The search generated five pre-designs within this RMSD range. As there were no available high resolution structures for an L-Ala/D-Glu epimerase in the PDB90 data set, one structure of this enzyme (PDB code: 1JPM) was manually analyzed, for completeness.

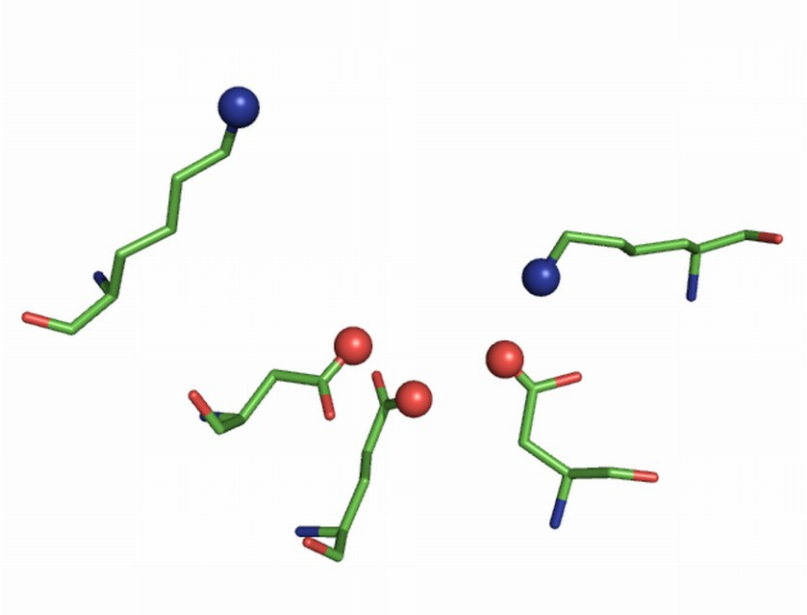


Figure 2.2. The OSBS active site from crystal structure 1FHV, as full Lys/Asp/Glu/Asp/Lys residues. Catalytic atoms are shown as spheres.

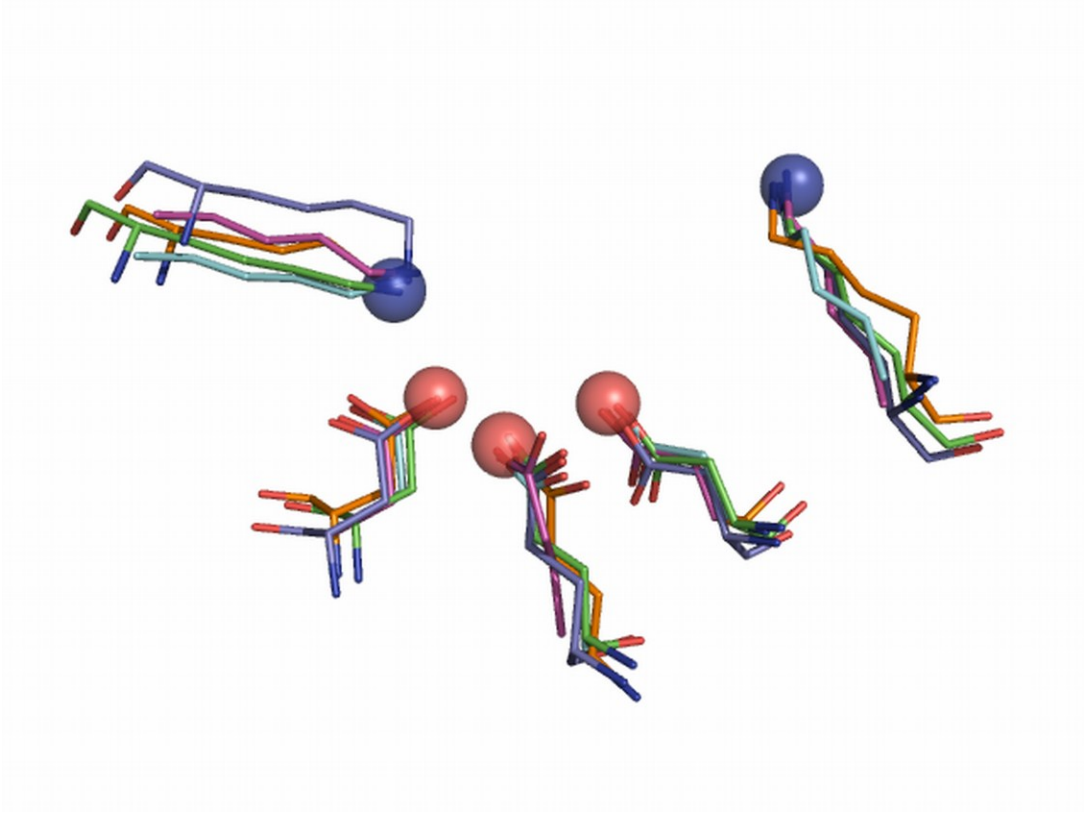


Figure 2.3. The top pre-designs identified by the OSBS catalytic atom map (rendered as spheres). Pre-designs by color: MLE (cyan, PDB code: 1MUC), AEE (magenta, 1JPM), Cl-MLE (green, 1NU5), MR (purple, 2QDE), MR (orange, 2RDX).

Table 2.1. Active sites with appropriate catalytic group positioning for the OSBS reaction, identified by the 1FHV catalytic atom map for o-succinyl benzoate synthase.

PDB code	Protein	RMSD	Residues identified (Mg^{2+})[1]	Residues identified (Lys)
1FHV	o-succinyl benzoate synthase	0.00	ASP161, GLU190, ASP213	LYS133, LYS235
1MUC	Muconate lactonizing enzyme	0.26	ASP198, GLU224, ASP249	LYS169, LYS273
1NU5	Chloromuconate lactonizing enzyme	0.33	ASP194, GLU220, ASP245	LYS165, LYS269
2QDE	Mandelate racemase/MLE	0.38	ASP195, GLU221, ASP246	LYS167, LYS270
1JPM[2]	L-ala/D-glu epimerase	0.43	ASP191, GLU219, ASP244	LYS162, LYS268
2RDX	Mandelate racemase/MLE	0.52	ASP193, GLU218, ASP241	LYS165, LYS265

[1]These residues compose the carboxylate triad predicted to coordinate with the Mg^{2+} atom.

[2]There was no structure of AEE available with a resolution $\leq$ 2.0 Å, so this pre-design was included manually to demonstrate the geometric match with the active site of this enzyme.

All of the pre-designs identified within the specified RMSD range were known active site homologs of OSBS. Each of the pre-designs catalyzes one of the following 3 reactions: the o-succinyl benzoate synthase reaction, L-ala/D-glu epimerization, or muconate lactonization. The 1NU5 pre-design is very similar to MLE, but this enzyme prefers chloromuconate as a substrate, instead of muconate. Two of the entries, 2RDX and 2QDE, were deposited in the Protein Data Bank without full papers to describe

them. Both of these proteins are described as mandelate racemase/muconate lactonizing enzyme family proteins in their PDB entries, so it is reasonable to expect that their active sites would be compatible with the OSBS reaction. The geometry of the active site for each of these proteins is in good agreement with the catalytic atom map for OSBS, lending further credence to this proposal.

This analysis indicates that the OSBS Catalytic Atom map is able to identify other enzymes that could be modified to be catalysts for the OSBS reaction. Two of the enzymes identified, MLE and AEE, have already been demonstrated experimentally to be capable of catalyzing the OSBS reaction. A key point is that the redesigned forms of both MLE and AEE leave the catalytic residues that match the OSBS CAM intact. The changes made in these enzymes to allow for catalysis of the OSBS reaction only create a binding pocket that can accommodate the new substrate.

Using Catalytic Atom Maps to search for active sites geometrically similar to KE07

In order to evaluate the use of SABER in selecting active sites for redesign, we also performed a search for proteins that could potentially catalyze a Kemp elimination of 5-nitrobenzisoxazole, shown in Figure 2.4. We based the Catalytic Atom Map on the active site of KE07, an enzyme designed specifically for this reaction by the Baker and Houk laboratories. (*3*) KE07 used a Glu residue as a general base and a Lys H-bond donor to catalyze the Kemp elimination, as shown in Figure 2.5. We sought to identify

enzymes with similar active site geometries in the Protein Data Bank, and then use Rosetta to redesign their active sites to be compatible with the Kemp transition state.

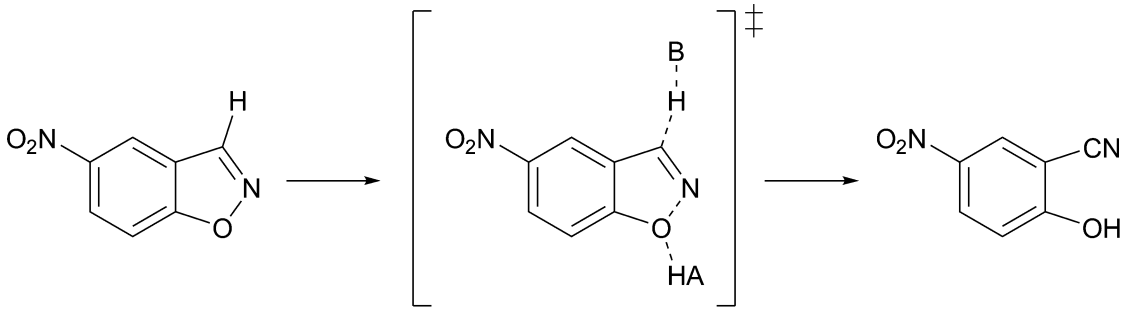


Figure 2.4. The Kemp elimination catalyzed by the KE07 designed enzyme. In the enzyme, B indicates a residue functioning as a general base, while HA indicates a residue functioning as a hydrogen bond donor.

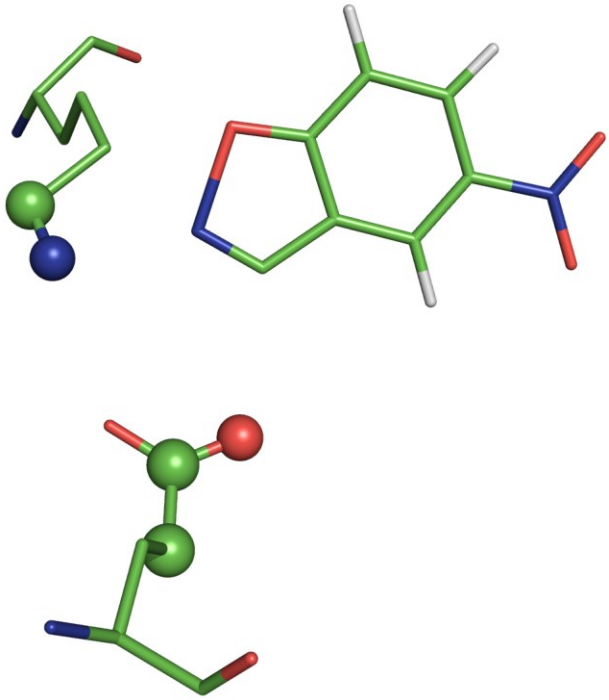


Figure 2.5. The designed active site for KE07, featuring the catalytic base (Asp/Glu) and a Lys hydrogen bond donor. Catalytic atoms are rendered as spheres.

The Catalytic Atom Map in Figure 2-5 was used for the SABER search. The atoms from the general base were allowed to match on the corresponding atoms from either Asp or Glu residues. Otherwise, the search tolerances were identical to the previous SABER search, with the match radius for each catalytic atom set at 2.0 Å.

The search generated 2259 pre-designs with RMSD values ≤ 0.4 Å of the KE07 Catalytic Atom Map. Two of the pre-designs had an ActiveSiteScore of 2, while another 21 had an ActiveSiteScore of 1. These should be the most promising choices for redesign, because the active sites have already been annotated in the Catalytic Site Atlas. There were an additional 286 pre-designs that were identified by BindingSiteFinder as having heteroatoms near the residues identified by the Catalytic Atom Map, which also potential candidates for redesign. The breakdown of the SABER pre-designs is summarized in Figure 2-6. The top pre-design with a predicted catalytic base pK_a in the range of 5-9 from each category (ActiveSiteScore = 2, ActiveSiteScore = 1, and ActiveSiteScore = 0 with a BindingSiteFinder flag) is shown in Table 2.2.

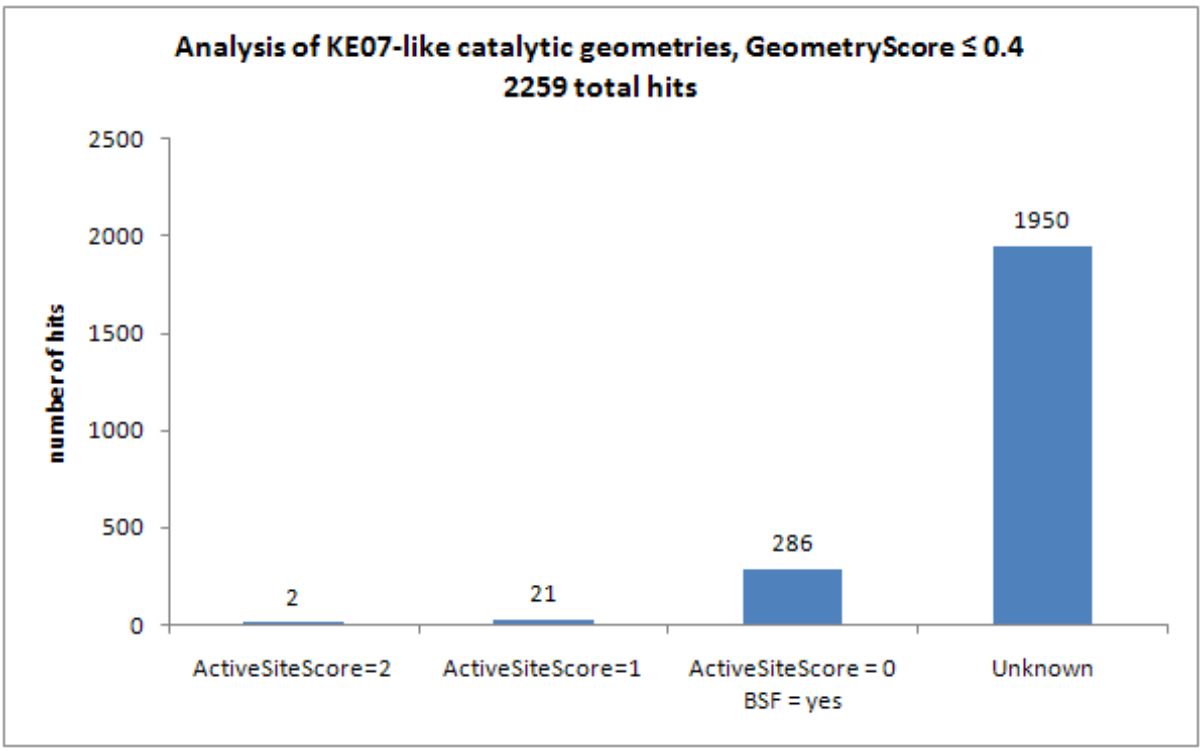


Figure 2.6. Analysis of the results from the PDB search using the KE07-based Catalytic Atom Map in Figure 5. Data is shown for pre-designs with an RMSD ≤ 0.4.

Table 2.2. Active sites identified by the CAM search as having catalytic groups arranged in a geometry similar to that of the KE07 design. The top pre-design from each category (ActiveSiteScore = 2, ActiveSiteScore = 1, ActiveSiteScore = 0 & BindingSiteFinder = yes) is shown.

PDB	GeometryScore	ActiveSiteScore	BindingSiteFinder	Base pKa	Residues
1X0L	0.28	2	None	5.1	Asp204, Lys171
1F0L	0.30	1	Yes	5.1	Glu189, Lys162
2C14	0.22	0	Yes	7.2	Asp131, Lys229

As a proof-of-principle test of SABER's ability to identify scaffolds for redesign, the best pre-design from the ActiveSiteScore = 2 category, PDB code 1X0L, was redesigned to accommodate the Kemp transition state using the RosettaDesign program. 1X0L is an isocitrate dehydrogenase enzyme from *T. thermophilus*. As shown in Figure 2.7, the catalytic residues from the KE07 crystal structure and those residues identified by SABER in the 1X0L crystal structure are in very similar orientations.

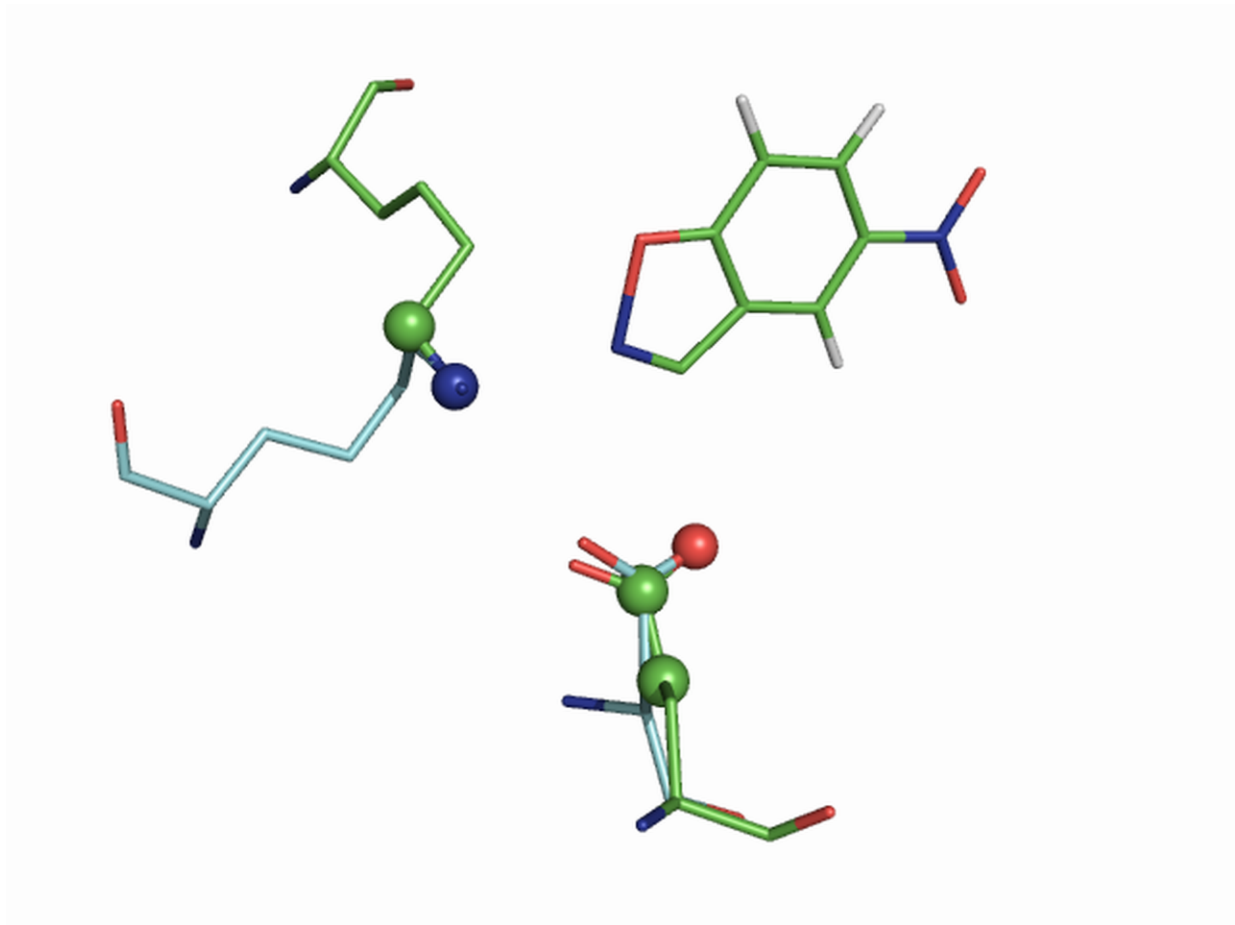


Figure 2.7. Superposition of the 1X0L active site residues identified by SABER and the KE07 designed active site. The KE07 design in shown in green, while the 1X0L residues are shown in blue.

The starting structure for RosettaDesign was generated by superimposing the 5-nitrobenzisoxazole Kemp elimination transition state from the KE07 design on the 1X0L active site. The transition state was placed in the orientation identified during the SABER analysis. This resulted in a large number of steric clashes in the active site, as shown in Figure 2.8A. After a single round of constrained optimization, redesign, and repacking, RosettaDesign was able to incorporate the Kemp elimination transition state into the 1X0L active site with 9 mutations of non-catalytic residues. This compares favorably with the original KE07 design in the 1THF scaffold, which required 13 mutations. Most importantly, the catalytic residues are from the original scaffold and thus did not require any changes to the protein structure to be in the optimal position for catalysis of the new reaction. The redesigned active site of 1X0L is shown in Figure 2.8B, with the steric clashes in the active site removed during the course of resdesign.

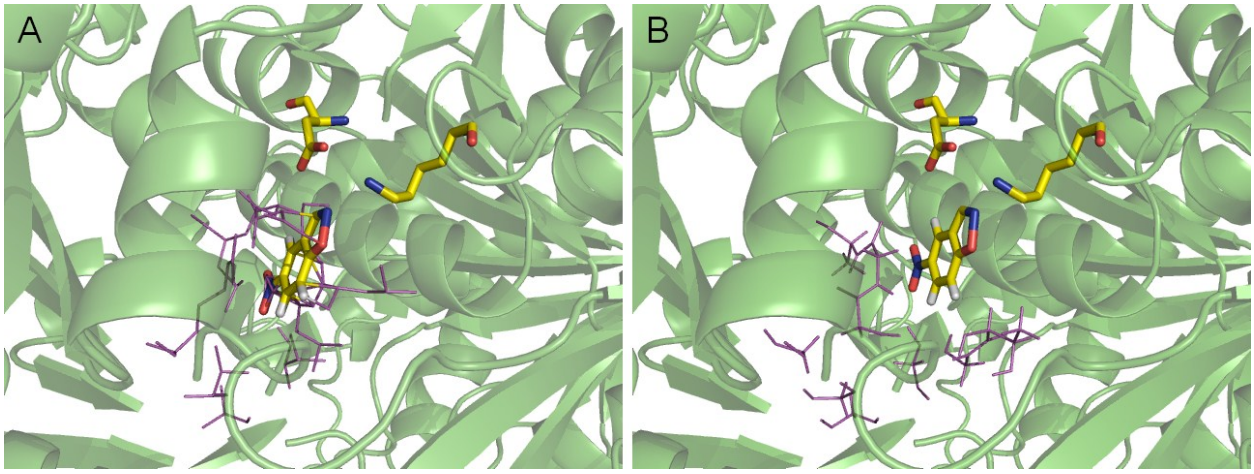


Figure 2.8. The active site of 1X0L, redesigned to catalyze the Kemp elimination. The catalytic residues and transition state are shown in yellow. The catalytic base (Asp204) and the hydrogen bond donor (Lys171) are unchanged from the original isocitrate dehydrogenase protein scaffold. (A) The original 1X0L active site, with the residues that clash with the Kemp elimination transition state shown in magenta. (B) The 1X0L active site after redesign, showing the 9 mutations from RosettaDesign in magenta. Mutations: R118G, A169G, V203G, T222G, L225T, L226G, D228G, I229G, L230A (1X0L numbering).

As a final validation of SABER's usefulness in the enzyme design process, one more Kemp elimination design was produced. Examination of Figure 2.4 shows that the ideal placement for the hydrogen bond donor should have it oriented toward the oxygen atom of the nitrobenzisoxazole transition state. However, as can be seen in Figure 2.5, the KE07 design has the hydrogen bond donor oriented towards the nitrogen atom. We used SABER and a CAM based on a theozyme calculated in our laboratory with a more ideal orientation of the H-bond donor to search the PDB for active sites that might be readily redesigned to function as more proficient Kemp eliminases. This theozyme, like the KE07 active site, employs a carboxylate (Asp/Glu) as the catalytic base, but uses an alcohol (Ser/Thr/Tyr) as the hydrogen bond donor, a motif found in other Kemp designs from reference 3. This theozyme is shown in Figure 2.9. It should be noted that unlike the KE07 active site, this theozyme places the carboxylate base such that both oxygen atoms can interact with the transition state. This catalytic arrangement for the Kemp elimination has been calculated to be the most efficient using quantum mechanics. (*31*)

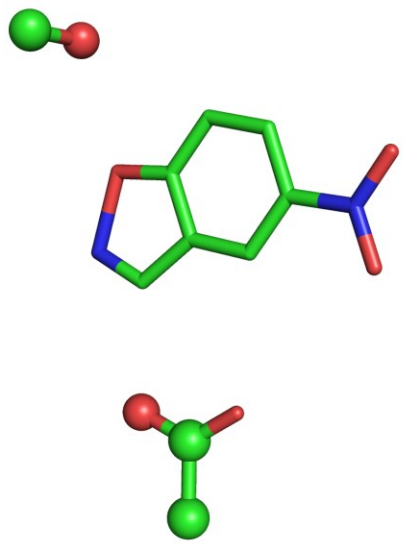


Figure 2.9. The theozyme used in the SABER search to locate pre-designs with a more ideal geometry for catalysis of the Kemp elimination. Note the alternative arrangement of the catalytic base vs. that shown in Figure 5. The atoms used in the Catalytic Atom Map are shown as spheres.

Other than the use of a new Catalytic Atom Map, the SABER search was performed in the same manner as with the KE07-based CAM. The SABER analysis resulted in 2603 pre-designs with an RMSD ≤ 0.4 Å. Of these 2603 designs, 94 had ActiveSiteScore values > 0. The best pre-design candidate was 1TRB, a thioredoxin reductase protein from *E. coli*. This predesign had a GeometryScore of 0.13, an ActiveSiteScore of 1, and a predicted catalytic base pK_a of 5.8, with Glu160 serving as the proposed catalytic base and Ser138 as the H-bond donor. As with the KE07-based design, the RosettaDesign software was used to redesign the FAD-binding site of 1TRB to accommodate the Kemp elimination transition state. The results of this enzyme redesign are shown in Figure 2.10. The 1TRB pre-design required 6 mutations in the FAD-binding site region of the protein in order to accommodate the new transition state.

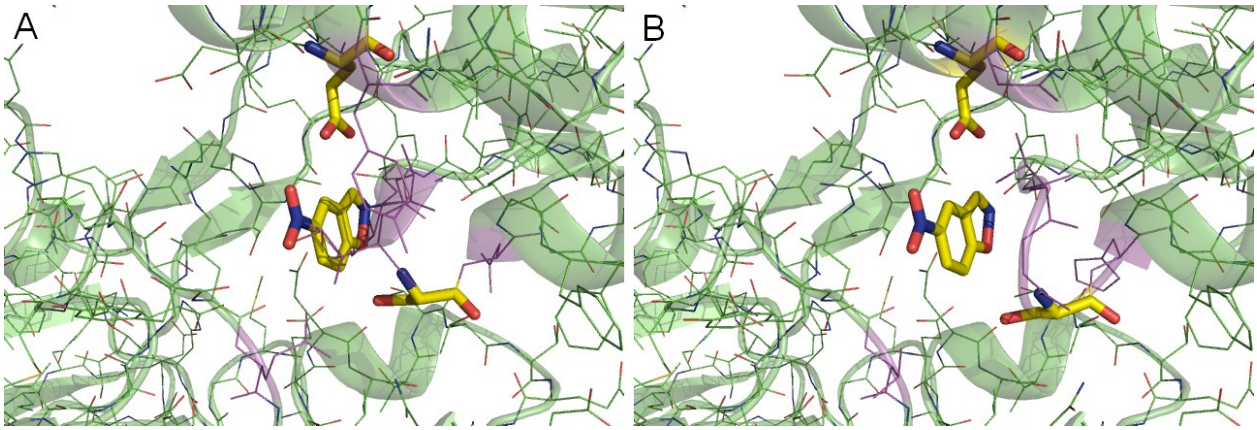


Figure 2.10. The active site of 1TRB, redesigned to catalyze the Kemp elimination. The catalytic residues and transition state are shown in yellow. The catalytic base (Glu160) and the hydrogen bond donor (Ser138) are unchanged from the original thioredoxin reductase protein scaffold. (A) The original 1TRB active site, with the residues that clash with the Kemp elimination transition state shown in magenta. (B) The 1TRB active site after redesign, showing the mutations from RosettaDesign in magenta. Mutations: C135G, A136T, T137G, D139H, Y163G, Q294T (1TRB numbering).

2.4 Discussion

The Protein Data Bank now contains over 70,000 structures. This represents an enormous wealth of data to draw upon when pursuing new enzyme designs. With SABER, it is possible to screen the entire PDB for geometric matches to a catalytic atom map, and also to identify the pre-designs that are part of known active sites. This program provides a method for the rapid identification of scaffolds that have catalytic residues placed in a theozyme-like geometry within an active site. Since there will be no need to position the catalytic residues and make mutations to hold them in place, this approach has the potential to significantly reduce the number of mutations required to modify a scaffold to carry out a new reaction.

In this paper, we have demonstrated a new technique, catalytic atom mapping, for selecting active sites for rational, minimalistic redesign. We have validated this technique using a known active site redesign example in the literature, the conversion of MLE and AEE into o-succinyl benzoate synthases. SABER was able to correctly identify the ideal scaffolds for redesign from a very large dataset. In addition, we have used the catalytic atom mapping methodology to identify active sites that could be remodeled to match that of a computationally design enzyme for a Kemp elimination, KE07. We have demonstrated that designs based on the active sites identified by SABER are complementary to the designs based on the standard Rosetta method. Studies are currently in progress to synthesize and evaluate enzymes designed using this methodology.

2.5 Materials and Methods: Computational Details

SABER Overview

Each section describes an element of the SABER program and how it integrates into the overall methodology for searching the PDB for active sites compatible with new reactions. Each protein identified during the search process is called a “pre-design”, to indicate a protein structure with appropriately placed catalytic groups, but requiring additional modification to the active site to accommodate a new transition state. An example of the output of a typical SABER pre-design is listed in Table 2.3 and described later.

Search protocol

This method first involves the creation of a Catalytic Atom Map from a 3D geometrical arrangement of atoms determined from a quantum mechanical model or from an enzyme crystal structure. The Catalytic Atom Map is then used to scan protein structures in the PDB for similar arrangements of atoms.

A Catalytic Atom Map is set of coordinates of atoms that are involved in the chemistry of the reaction, termed “catalytic atoms”; this can also involve other atoms that orient the position of the catalytic atoms. For example, in a serine protease, the minimal catalytic atom definition of the serine nucleophile is the γ-oxygen atom of the

alcohol in the sidechain. To fix the orientation of the γ-oxygen, the β-carbon can also be included. The remainder of the triad could defined using the ring nitrogens of the histidine residue and one or both of the carboxylate oxygens from the acid residues. A more extended CAM would include the oxyanion hole nitrogens of the NH hydrogen bond donors. The Catalytic Atom Map can be based on a known active site geometry from experiment, or on a theoretical model of an active site, such as a theozyme. (*7*)

Catalytic Atom Map Parameters

The Catalytic Atom Map has a number of parameters that can be set for each atom. In addition to the geometric constraints, each atom in the Catalytic Atom Map can be specified to match with only certain residue and/or atom types during the search. For example, a catalytic atom defined by the trypsin γ-oxygen from serine can be set to match with only other serine γ-oxygens, or can be set to match with either a serine γ-oxygen, or another potentially nucleophilic oxygen, or even a cysteine γ-sulfur. The latter case would be useful if the Catalytic Atom Map were intended to identify both serine and cysteine proteases. The radius of each atom in the map is set individually, making it possible to set a different tolerance for geometric deviation from the Catalytic Atom Map for individual atoms.

Once the Catalytic Atom Map is created, the Jess algorithm is used to search the Protein Data Bank via geometric superposition. (*32*) The Jess algorithm searches through the protein structures one by one, finding 3D matches to the arrangement of

atoms, defined by a Catalytic Atom Map. Jess uses a geometric hashing function to maximize the superposition of the Catalytic Atom Map and the atoms in the proteins being searched. (*32*) It is possible to have multiple matches to the Catalytic Atom Map in a single protein structure. The Jess output consists of the atoms involved in the match, as well as the RMSD of the match to the Catalytic Atom Map and the maximum deviation from the Catalytic Atom Map.

A Jess search of the entire Protein Data Bank will typically yield thousands to hundreds of thousands of geometric matches to the Catalytic Atom Map. This is especially true when loose tolerances are set for the catalytic atoms or when a small number of catalytic atoms is used. The additional steps in SABER identify the geometric matches that are most likely to be of interest, that is, the matches that are in known binding sites or active sites.

ActiveSiteFinder and Score

ActiveSiteFinder determines if the identified catalytic residues are part of an active site already contained in the Catalytic Site Atlas (CSA), which lists the known or putative active site residues for a given enzyme. (*33*) The Catalytic Site Atlas has been compiled by the Thornton group, and provides a publicly available database of known active site residues for proteins in the Protein Data Bank. The database contains active site residue information from sequence alignments as well as from experimental data in the literature. If the identified catalytic residues are part of an active site in the Catalytic

Site Atlas, the active site residues from the CSA are compared to the residues that matched the Catalytic Atom Map. The ActiveSiteScore is the number of catalytic atoms that are part of a known active site in the CSA.

For example, if the residues identified by the Catalytic Atom Map are: Asp100 Lys120 His140, and the known active site residues in the Catalytic Site Atlas are: Asp100 Lys120 Glu160 His180, the ActiveSiteScore is 2. Two of the residues in the pre-design, Asp100 and Lys120, have been annotated as part of an active site in the Catalytic Site Atlas. In the case of proteins for which no active site data are available, the score is 0.

BindingSiteFinder

BindingSiteFinder is an alternative method for locating active/binding sites in proteins that have not been identified in the Catalytic Site Atlas. This program analyzes the PDB file for every pre-design and searches for non-water PDB heteroatoms within 5 Å of the residues matching the Catalytic Atom Map. If any such heteroatoms are present, BindingSiteFinder will display the name of the ligand, as well as the identity of the heteroatom closest to the residues identified by the Catalytic Atom Map. The idea is that these heteroatoms will be from inhibitors, substrate analogs, or cofactors present in the crystal structure, and this will show that the identified site is capable of binding a substrate.

Prediction of pK_a values

PROPKA, from the Jensen group, is used to predict pK_a values for all of the residues identified as matching the Catalytic Atom Map. (*34, 35*) When acid-base catalysis is used, this gives information about the protonation states of the proposed catalytic groups.

Table 2.3. An example of a pre-design generated by SABER, along with the scoring functions used to identify optimal active sites for redesign.

Feature	Description	Example
PDB	PDB code for protein that contains matching residues	1EUA
RMSD	RMSD vs. Catalytic Atom Map (Å)	0.44
Maximum displacement	Largest individual deviation from Catalytic Atom Map (Å)	0.69
Protein name	Name listed in the Protein Data Bank	KDPG aldolase
Match residues	The residues identified in the protein as matching the Catalytic Atom Map	GluA45, LysA133
ActiveSiteScore	Number of residues matching the Catalytic Atom Map that are listed as active site residues in the Catalytic Site Atlas entry for the protein	2
BindingSiteFinder	The closest PDB-style heteroatom within 5 Å of the residues matching the CAM, along with the distance between the two atoms	Lys 133 NZ Pyr 2103 C2 1.23 Å
pK_a	PROPKA pKa values for the match residues. Also gives the PROPKA classification for each residue as buried or solvent-exposed.	GluA45, buried. $pK_a = 6.0$ LysA133, buried. $pK_a = 9.0$

2.6 Acknowledgements

This work was supported by the National Institute of General Medical Science, National Institutes of Health (GM 75962 and T32 GM008496-16), and the Department of Defense, Defense Advanced Research Projects Agency.

2.7 References

1. Kaplan, J., and DeGrado, W. F. (2004) De novo design of catalytic proteins, *Proc. Natl. Acad. Sci. U.S.A. 101*, 11566-11570.

2. Nanda, V., Rosenblatt, M. M., Osyczka, A., Kono, H., Getahun, Z., Dutton, P. L., Saven, J. G., and Degrado, W. F. (2005) De novo design of a redox-active minimal rubredoxin mimic, *J. Am. Chem. Soc. 127*, 5804-5805.

3. Rothlisberger, 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., and Baker, D. (2008) Kemp elimination catalysts by computational enzyme design, *Nature 453*, 190-195.

4. Jiang, L., Althoff, E. A., Clemente, F. R., Doyle, L., Rothlisberger, D., Zanghellini, A., Gallaher, J. L., Betker, J. L., Tanaka, F., Barbas, C. F., 3rd, Hilvert, D., Houk, K. N., Stoddard, B. L., and Baker, D. (2008) De novo computational design of retro-aldol enzymes, *Science 319*, 1387-1391.

5. Yeung, N., Lin, Y. W., Gao, Y. G., Zhao, X., Russell, B. S., Lei, L., Miner, K. D., Robinson, H., and Lu, Y. (2009) Rational design of a structural and functional nitric oxide reductase, *Nature 462*, 1079-1082.

6. Siegel, J. B., Zanghellini, A., Lovick, H. M., Kiss, G., Lambert, A. R., St Clair, J. L., Gallaher, J. L., Hilvert, D., Gelb, M. H., Stoddard, B. L., Houk, K. N., Michael, F. E., and Baker, D. (2010) Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction, *Science 329*, 309-313.

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

8. Rohl, C. A., Strauss, C. E. M., Misura, K. M. S., and Baker, D. (2004) Protein structure prediction using rosetta, *Methods Enzymol. 383*, 66-93.

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

10. Fersht, A. (1999) *Structure and mechanism in protein science : a guide to enzyme catalysis and protein folding*, W.H. Freeman, New York.

11. Warshel, A. (1978) Energetics of enzyme catalysis, *Proc. Natl. Acad. Sci. U.S.A. 75*, 5250-5254.

12. Warshel, A. (2003) Computer simulations of enzyme catalysis: methods, progress, and insights, *Annu. Rev. Biophys. Biomol. Struct. 32*, 425-443.

13. Warshel, A., Sharma, P. K., Kato, M., Xiang, Y., Liu, H., and Olsson, M. H. (2006) Electrostatic basis for enzyme catalysis, *Chem. Rev. 106*, 3210-3235.

14. Seebeck, F. P., and Hilvert, D. (2005) Positional ordering of reacting groups contributes significantly to the efficiency of proton transfer at an antibody active site, *Journal of the American Chemical Society 127*, 1307-1312.

15. Smith, A. J., Muller, R., Toscano, M. D., Kast, P., Hellinga, H. W., Hilvert, D., and Houk, K. N. (2008) 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. Am. Chem. Soc. 130*, 15361-15373.

16. Dechancie, J., Clemente, F. R., Smith, A. J., Gunaydin, H., Zhao, Y. L., Zhang, X., and Houk, K. N. (2007) How similar are enzyme active site geometries derived from quantum mechanical theozymes to crystal structures of enzyme-inhibitor complexes? Implications for enzyme design, *Protein Sci. 16*, 1851-1866.

17. Lassila, J. K., Keeffe, J. R., Kast, P., and Mayo, S. L. (2007) Exhaustive mutagenesis of six secondary active-site residues in Escherichia coli chorismate mutase shows the importance of hydrophobic side chains and a helix N-capping position for stability and catalysis, *Biochemistry 46*, 6883-6891.

18. Ruscio, J. Z., Kohn, J. E., Ball, K. A., and Head-Gordon, T. (2009) The influence of protein dynamics on the success of computational enzyme design, *J. Am. Chem. Soc. 131*, 14111-14115.

19. Olsson, M. H., Parson, W. W., and Warshel, A. (2006) Dynamical contributions to enzyme catalysis: critical tests of a popular hypothesis, *Chem. Rev. 106*, 1737-1756.

20. Nashine, V. C., Hammes-Schiffer, S., and Benkovic, S. J. (2010) Coupled motions in enzyme catalysis, *Curr. Opin. Chem. Biol. 14*, 644-651.

21. Kamerlin, S. C., and Warshel, A. (2010) At the dawn of the 21st century: Is dynamics the missing link for understanding enzyme catalysis?, *Proteins 78*, 1339-1375.

22. Boekelheide, N., Salomon-Ferrer, R., and Miller, T. F., 3rd. (2011) Dynamics and dissipation in enzyme catalysis, *Proc. Natl. Acad. Sci. U.S.A. 108*, 16159-16163.

23. Toscano, M. D., Woycechowsky, K. J., and Hilvert, D. (2007) Minimalist active-site redesign: teaching old enzymes new tricks, *Angew. Chem. Int. Ed. Engl. 46*, 3212-3236.

24. Gerlt, J. A., and Babbitt, P. C. (2009) Enzyme (re)design: lessons from natural evolution and computation, *Curr. Opin. Chem. Biol. 13*, 10-18.

25. Vizcarra, C. L., Zhang, N., Marshall, S. A., Wingreen, N. S., Zeng, C., and Mayo, S. L. (2008) An improved pairwise decomposable finite-difference Poisson-Boltzmann method for computational protein design, *J. Comput. Chem. 29*, 1153-1162.

26. Brustad, E. M., and Arnold, F. H. (2011) Optimizing non-natural protein function with directed evolution, *Curr. Opin. Chem. Biol. 15*, 201-210.

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

28. Schmidt, D. M., Mundorff, E. C., Dojka, M., Bermudez, E., Ness, J. E., Govindarajan, S., Babbitt, P. C., Minshull, J., and Gerlt, J. A. (2003) Evolutionary potential of (beta/alpha)8-barrels: functional promiscuity produced by single substitutions in the enolase superfamily, *Biochemistry 42*, 8387-8393.

29. Vick, J. E., Schmidt, D. M., and Gerlt, J. A. (2005) Evolutionary potential of (beta/alpha)8-barrels: in vitro enhancement of a "new" reaction in the enolase superfamily, *Biochemistry 44*, 11722-11729.

30. Vick, J. E., and Gerlt, J. A. (2007) Evolutionary potential of (beta/alpha)8-barrels: stepwise evolution of a "new" reaction in the enolase superfamily, *Biochemistry 46*, 14589-14597.

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

32. Barker, J. A., and Thornton, J. M. (2003) An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis, *Bioinformatics 19*, 1644-1649.

33. Porter, C. T., Bartlett, G. J., and Thornton, J. M. (2004) The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data, *Nucleic Acids Res. 32*, D129-D133.

34. Bas, D. C., Rogers, D. M., and Jensen, J. H. (2008) Very fast prediction and rationalization of pK(a) values for protein-ligand complexes, *Proteins-Structure Function and Bioinformatics 73*, 765-783.

35. Olsson, M. H. M., Sondergaard, C. R., Rostkowski, M., and Jensen, J. H. (2011) PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pK(a) Predictions, *J. Chem. Theory. Comput. 7*, 525-537.

CHAPTER 3: SABER PROGRAM STRUCTURE

3.1 Introduction

This chapter is both a description of core elements of the SABER program code and a practical guide for use of the software. The entire process for using the software, from generation of a Catalytic Atom Map (CAM) to final spreadsheet generation, is described. In addition, selections of code from certain SABER modules are presented. This is intended to provide insight into how SABER works from a programmer's point of view, and to allow changes to be made to the program as needed for future research. The code presented covers the core analytical functions of the module it is taken from. Basic functionality, like file input and output, is not discussed as this is common among all of the modules.

For the purposes of the examples in this chapter, a single theozyme designed to catalyze an aromatic Claisen rearrangement will be discussed. The reaction this theozyme is designed to catalyze is shown in Figure 3.1, and the theozyme and the transition state are shown in Figure 3.2. In this theozyme, the developing partial negative charge on the oxygen atom is being stabilized by two hydrogen bonds. One of the hydrogen bond donors is the sidechain of a protonated lysine residue, while the other is the sidechain of a tyrosine residue.

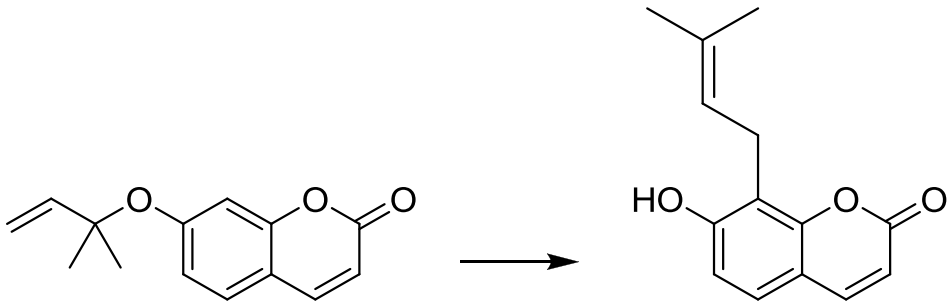


Figure 3.1 The aromatic Claisen rearrangement reaction used as an example in this chapter.

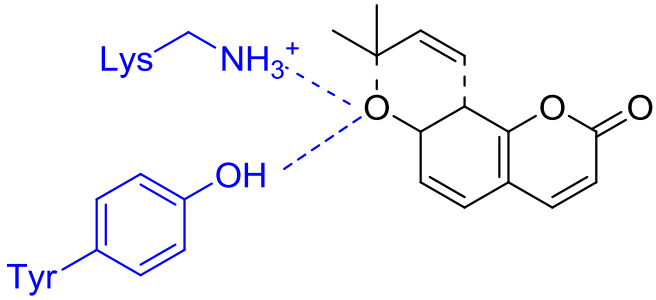


Figure 3.2 A theozyme for the stabilization of the transistion state of the aromatic Claisen rearrangement shown in Figure 3.1.

3.2 Catalytic Atom Map Construction

As described in Chapter 1, a theozyme is a truncated model of an enzyme active site paired with a transition state, calculated using quantum mechanics. (*1*) Theozymes are often used to construct Catalytic Atom Maps, as are the coordinates of enzyme active sites from X-ray crystal structures. This procedure, along with the selection of atoms to include in the Catalytic Atom Map, is described in Chapter 2. In order to

construct a Catalytic Atom Map that can be used with SABER, the following information is needed:

1) The xyz coordinates of each atom in the Catalytic Atom Map

2) The PDB atom code for each atom

3) The PDB residue name for each residue

This information needs to be entered into a text file, known as a Tess file. The Tess file must follow the format in Figure 3.3. The Tess file in Figure 3.3 is for a Catalytic Atom Map for the theozyme shown in Figure 3.2.

```
ATOM      0  O   LG1 A   1      0.139   1.008   0.073
ATOM      0  OH  TYR A   2      2.672   1.227  -0.759
ATOM      0  NZ  LYS A   3      1.683   0.171   2.093
END
```

Figure 3.3. The Tess file for a Catalytic Atom Map derived from the theozyme shown in Figure 3.2.

There are three atoms defined in the Tess file. The first row describes an oxygen (O) atom from the transition state, given the residue name LG1 in the PDB file. The second atom is a tyrosine OH atom, and the third is a lysine NZ atom. The 0 values in the second column dictate the logic of atom/residue matching used by Jess, which is will be described in the following section. The third column is the atom type, the fourth is residue name, the fifth is the chain ID, and the sixth is the residue number. The remaining columns are the xyz coordinates of the atoms.

The Tess file is used as input for the *tess2jess* script. This script will convert a Tess file into a human-readable XML input file suitable for use with Jess. Using the command:

[user@system~]$ tess2jess *Tess_filename*

will generate the XML output, as shown in Figure 3.4.

```
<?xml version="1.0"?>
<jess xmlns="http://www.ebi.ac.uk/~jbarker/Jess">
    <template name="yk_theo_ca.tess">
        <atom name="0">
            <select>
                and
                (
                    isAtomNamed("_O__"),
                    inResidueNamed("LG1")
                )
            </select>
            <xyz>0.139 1.008 0.073</xyz>
        </atom>
        <atom name="1">
            <select>
                and
                (
                    isAtomNamed("_OH_"),
                    inResidueNamed("TYR"),
                    inAnnulus(atom(#0),sub(2.68,$delta),add(2.68,$delta)),
                    inChainOf(atom(#0))
                )
            </select>
            <xyz>2.672 1.227 -0.759</xyz>
        </atom>
        <atom name="2">
            <select>
                and
                (
                    isAtomNamed("_NZ_"),
                    inResidueNamed("LYS"),
                    inAnnulus(atom(#0),sub(2.68,$delta),add(2.68,$delta)),
                    inChainOf(atom(#0)),
                    inAnnulus(atom(#1),sub(3.20,$delta),add(3.20,$delta))
                )
            </select>
```

```
            <xyz>1.683 0.171 2.093</xyz>
        </atom>
    </template>
</jess>
```

Figure 3.4. The Catalytic Atom Map from Figure 3.3, in Jess XML format.

The Jess XML file can be manually edited to allow for more diversity in the Catalytic Atom Map. For example, atom 1 (the tyrosine OH, since atom numbering starts at 0), could be allowed to match with a serine OH as well. In that case, the "inResidueNamed" line for atom 1 would be changed to:

```
inResidueNamed("TYR",”SER”),
```

and this would allow for matches on both tyrosine sidechain oxygen and serine sidechain oxygens. Note that it is possible to match PDB heteroatoms and other atoms that are not part of the standard twenty amino acids, as is the case with atom 0. This is especially useful when looking for heteroatoms near catalytic residues or for locating water molecules positioned in the active site. The $delta variable will be input during the next stage of SABER execution, and is a measure of the size of the sphere used to represent the match tolerance of each atom. Full documentation of the *tess2jess* script and matching logic can be found in the Jess manual.

Once the XML file describing the Catalytic Atom Map is complete, the next step is to run the Jess program to find matches to the CAM in a set a protein structures.

3.3 Jess

The Jess program uses geometric hashing to explore protein structure(s) and find matches to patterns of atoms, such as Catalytic Atom Maps. (*2*) Jess allows for rapid analysis of large data sets. It is possible to search the entire Protein Data Bank with a single Catalytic Atom Map overnight on a single-processor machine. Potentially, the search could be much faster, depending on the tolerances set during Catalytic Atom Map creation and during the execution of the program.

Some care must be taken with regard to data generation. A Catalytic Atom Map with very large geometric tolerances (e.g., > 2.0 Å) has the potential to generate an enormous number of matches, especially if it is used to search a large set of protein structures. With loose tolerances, it is easy to overload the filesystem and completely fill one's storage quota. It is advisable to periodically monitor the program while it is running, so as to control the amount of data being written to disk.

To execute the Jess program, a typical command line is:

[user@system~]$ jess –t *CAM.xml* –M *filelist.txt* –f –p –d delta=2.0

This command line uses the –t switch to indicate the Catalytic Atom Map to be searched. The –M switch indicates a file containing a list of PDB structures to search, with full path information. The –f switch turns on full output, which is necessary for

SABER. The –p switch limits Jess to writing the best permutation of a given match, and the –d switch is used to set the match radius of the sphere for each atom. It is possible to use different radii for each atom, so the –d switch might be followed by delta=1.5, gamma=2.0 if that were the case. The Catalytic Atom Map XML file would also have to be modified with these changes if multiple match radii were to be used. Additional information on these command line switches can be found in the Jess paper and documentation.

Once the Jess parameters have been tested, the output of the search should be written to a file for further analysis.

3.4 SummaryBuilder

This program is a bash script that converts the Jess output into a condensed text file for further analysis. The syntax for execution is:

[user@system~]$ bash summary_builder_v4.s <*Jess_file.txt* >*output_file.txt*

The output should appear as in Figure 3.5. This file will be immediately reprocessed by the next step in the SABER program, so there is no need to inspect it directly unless errors are encountered.

```
1.26    2.03   22GS   None       TYRA79  ARGA74  ASPB90
1.38    2.00   830C   None       SERB233  ASPB257  ASPB235
0.96    1.83   43C9   None       SERD40  LYSD43  ARGD38
1.00    2.45   43CA   None       SERF40  LYSF43  ARGF38
0.93    1.50   13GS   None       SERB65  ARGB13  ARGB100
1.67    3.71   13PK   None       SERA254  LYSC33  GLUA310
1.68    2.73   14GS   None       TYRB79  ASPB90  ASPB94
```

Figure 3.5. Typical output from successful execution of the jess_summary script.

Once the script is complete and the summary file is generated, the next stage of the SABER process can begin. The summary file will be used as the input for the next round of analysis.

3.5 ActiveSiteFinder

This program, written in Perl, takes the summary file as input and uses it to locate information regarding the residues identified during the Jess search. It performs a lookup of the protein's entry in the Catalytic Site Atlas (CSA) and compares the residues in the match to the known active site residues annotated in the CSA. (*3*) This information, along with the protein name and EzCat identifier (if present), is then appended to the output file for scoring by the next step in the SABER analysis. To execute this program, the command line syntax is:

```
[user@system~]$ perl as_finder_v6.pl
```

The program will request an input filename and an output filename. The input filename should be the file that was output by SummaryBuilder.

The Catalytic Site Atlas is searched using the get function, as shown in Figure 3.6. Note that the program writes the HTML to a file for future use, so that each protein entry needs to be looked up only once.

```
if ( $csa_html_exists == 0 ) {

      $csa_output = get $full_csa_url;

      print "Getting CSA data for $jess_prot, since csa_html_exists =
      $csa_html_exists\n";

#Write CSA html for hit into a file, protein.html

      open(HTMLFILE,">/u/home/campus/geoff/CSA_HTML/$jess_prot.html")
      or die "Can't open $jess_prot.html: $!";
      print HTMLFILE $csa_output;
      close HTMLFILE;

}
```

Figure 3.6. Using the get function to retrieve data from the Catalytic Site Atlas using the ActiveSiteFinder program.

Prior to running the function that retrieve data from the Catalytic Site Atlas, the local library is checked to see if the protein has had its entry analyzed previously, as shown in Figure 3.7. If this is the case, the local information is used. This makes the lookup process significantly faster, and reduces the data load on the CSA servers.

```
if ( -e "/u/home/campus/geoff/CSA_HTML/$jess_prot.html" ) {

      $csa_html_exists = 1;
#     print "For $jess_prot, csa_html_exists = $csa_html_exists\n";

      } else {

      $csa_html_exists = 0;
#     print "For $jess_prot, csa_html_exists = $csa_html_exists\n";

      }
```

Figure 3.7. Checking for locally stored Catalytic Site Atlas Data using the ActiveSiteFinder program.

This program has been able to successfully parse the vast majority of Catalytic Site Atlas entries correctly, but will occasionally put an extra symbol into a protein name, but the output is always clear. The output file written by ActiveSiteFinder has its fields delimited using the | symbol. This convention will be continued through the rest of the SABER analysis, and makes eventual import into Excel or other spreadsheet/database programs trivial.

3.6 ActiveSiteScore

ActiveSiteScore is a Perl program that is used for the next step in the SABER analysis sequence. This program takes the residues identified in the Jess search and compares them to the known active site residues in the protein, identified from data obtained from the Catalytic Site Atlas by ActiveSiteFinder. The command line to execute this program is:

[user@system~]$ perl saber as_score_v3.pl

Like ActiveSiteFinder, ActiveSiteScore will request both input and output filenames. The input file should be the one generated by ActiveSiteFinder.

The matches to the Catalytic Atom Map are scored according to how many residues match known active site residues from the Catalytic Site Atlas. Each time there is a match between a CAM-identified residue and a CSA residue (by both residue type and residue number), the score is incremented by 1. The default score for any match to the Catalytic Atom Map is 0, so any proteins without any annotations in the Catalytic Site Atlas will have a score of 0. The scoring element of the ActiveSiteScore program is shown in Figure 3.8.

```
#Loop to examine each of the CSA hits - CSA hits start at position 7

    for ( $n = 7 ; $n < scalar @line_fields ; ++$n ) {

      #Initialize site_score value at 0 for this CSA active site
      comparison

      $site_score = 0;

      @csa_residues = split ( ' ' , $line_fields[$n] );


      #Converts the CSA hit to numbers only

      for ( $o = 0 ; $o < scalar @csa_residues ; ++$o ) {
          @csa_residues[$o] =~ s/[A-Z]//g;

            }

      for ( $p = 0 ; $p < scalar @jess_residues ; ++$p ) {

          for ( $q = 0 ; $q < scalar @csa_residues ; ++$q ) {
```

```
            $jess_test_res = $jess_residues[$p];
            $csa_test_res = $csa_residues[$q];

            if ( $jess_test_res == $csa_test_res ) {
                $site_score = $site_score + 1;

            }

        }

    }
```

Figure 3.8. Identifying known active site residues using the ActiveSiteScore program.

The output of ActiveSiteScore will contain all of the information contained in the ActiveSiteFinder file, plus the scoring information for each match to the Catalytic Atom Map.

3.7 BindingSiteFinder

The next step in the SABER analysis uses the Perl program BindingSiteFinder. This program provides an alternative method for the identification of active and/or binding sites in a protein, as there are many proteins in the PDB for which no data is available beyond the structure itself. BindingSiteFinder looks for non-water PDB heteroatoms within 5 Å of the residues of the Catalytic Atom Map. If any non-water heteroatom is located in this radius, the match is flagged as such. Additional information reported includes the name of the heteroatom and the ligand, and the closest distance

between the heteroatom and the catalytic residues. To execute this program, the syntax is:

[user@system~]$ perl bsf_v6.pl

The program will request input and output filenames, as with the previous two modules. The input file must be a file generated by the ActiveSiteScore program.

Depending on the speed of the machine being used, this program can take a few hours to run on large data sets. This is due to the large number of distance measurements that need to be calculated. The for loop in Figure 3.9 shows the distance measurement process as the program iterates through the non-water heteroatoms in the protein:

```
for ( $het_atom_count = 0; $het_atom_count < scalar @het_test_array ;
++$het_atom_count ) {

           $het_atom_line = $het_test_array[$het_atom_count];

            $het_atom_no = substr ($het_atom_line,22,4);
            $het_atom_no =~ s/^\s+|\s+$//g;

            $het_atom_name = substr ($het_atom_line,12,4);
            $het_atom_name =~ s/^\s+|\s+$//g;

            $het_atom_res = substr ($het_atom_line,17,3);
            $het_atom_res =~ s/^\s+|\s+$//g;

            $ha_x = substr ($het_atom_line,30,8);
            $ha_x =~ s/^\s+|\s+$//g;
            $ha_y = substr ($het_atom_line,38,8);
            $ha_y =~ s/^\s+|\s+$//g;
            $ha_z = substr ($het_atom_line,46,8);
            $ha_z =~ s/^\s+|\s+$//g;
```

```
            $current_dist = (($ha_x - $ca_x)**2 + ($ha_y - $ca_y)**2 +
            ($ha_z - $ca_z)**2)**0.5;


      if ( $current_dist <= 5 && $current_dist < $best_dist ) {

                        $best_dist = $current_dist;
                        $best_pdb_res_type = $pdb_res_type;
                        $best_pdb_res_no = $pdb_res_no;
                        $best_pdb_atom_type = $pdb_atom_type;
                        $best_het_atom_no = $het_atom_no;
                        $best_het_atom_res = $het_atom_res;
                        $best_het_atom_name = $het_atom_name;
                        $bsid_flag = "Yes";

                        }

      }


}
```

Figure 3.9. The for loop for heteroatom distance analysis in the BindingSiteFinder program.

This for loop also highlights one of the difficulties of working with text files directly from the Protein Data Bank. It is critical to strip all whitespace and special characters from the various fields extracted from the PDB file. If this is not done, none of the calculations or string comparisons will return proper values and the program will give very confusing output, if it runs at all.

The output file from BindingSiteFinder serves as the input for the next module, SABER_wcn.

3.8 SABER_wcn

The SABER module is a Perl program written to perform weighted contact number (WCN) analysis. In order for this module to work with a given PDB identifier or set of identifiers, a WCN summary file for the protein(s) of interest must be generated. This is done using a separate program, wcn_filegen, discussed below. Using this file, saber_wcn reports the highest catalytic atom WCN, the average catalytic atom WCN, and the Cα WCN for a given residue. (*4, 5*) (A thorough discussion of WCN analysis is given in Chapter 4.) To execute this program, the following command line is used:

[user@system~]$ perl saber_wcn_v10.pl

In Figure 3.10, one of the WCN analysis loops in saber_wcn is shown, in this case the Cα loop. There are similar analysis loops for the catalytic atoms.

```
if ( $cnf_res_no eq $current_residue_no && $cnf_chain eq
$current_residue_chain && $cnf_atom_name eq "CA" ) {

      $ca_wcn_value = $cnf_atom_wcn;

      $ca_wcn_z = ($cnf_atom_wcn - $cnf_wcn_mean)/$cnf_wcn_stdev;

      $ca_ca_z = ($cnf_atom_ca_cn - $cnf_ca_cn_mean)/$cnf_ca_cn_stdev;

      $ca_ha_z = ($cnf_atom_ha_cn - $cnf_ha_cn_mean)/$cnf_ha_cn_stdev;

      $ca_wcn_z = sprintf("%.2f", $ca_wcn_z);

      $ca_ca_z = sprintf("%.2f", $ca_ca_z);

      $ca_ha_z = sprintf("%.2f", $ca_ha_z);
```

```
print "CAlph Z values wcn, ca_con, ha_con:  $ca_wcn_z    $ca_ca_z
$ca_ha_z","\n";

}
```

Figure 3.10. The Cα weighted contact number analysis loop in the saber_wcn module.

This module will report the three WCN values for each residue identified by the Catalytic Atom Map. These will be added, three per residue, in the order that the residues are listed in the input file. The first WCN value is the highest catalytic atom WCN, the second is the average catalytic atom WCN, and the final one is the Cα WCN. This ordering is conserved no matter how many catalytic residues are present, so determining which values belong to each residue is straightforward.

The output of this module can be run through the saber_pka module, or it can be imported into Excel and analyzed directly.

3.9 WCN_filegen

While this Perl program is not strictly a part of SABER, it used to support the saber_wcn module by generating pre-calculated files with WCN values of a protein of interest. This can be a slow process, depending on the number of heavy atoms in the protein. This program can be run in an embarrassingly parallel manner – in theory, it would be possible to run every protein in the PDB simultaneously if you had sufficient computational resources.

Once the program is used to analyze a given PDB file, the results are written to a .cnf file and stored, so that each PDB file need be analyzed only once. Use of the program is straightforward, as the only input is a list of PDB files to be analyzed. As the .cnf file is written at the beginning of the analysis process, it is possible to run multiple instances of this program in the same directory without overwriting files. There is a built-in check to see if a given .cnf file exists, and if it does, the program moves on to the next one down the list.

The syntax for execution of wcn_filegen is:

```
user@system~]$ perl wcn_filegen_v7.pl
```

We are currently in the process of building a library of these .cnf files to speed up future WCN analysis.

3.10 SABER_pKa

The final SABER module, saber_pka, is a Perl program used to predict the pK_a values for the residues identified by the Catalytic Atom Map. Like saber_wcn, this program relies on a library of files, one per PDB structure, with the predicted pK_a values for the residues in the protein. These files are generated using the PROPKA software.

(*6, 7*) The predicted pK_a value for each residue is reported. The syntax for execution of this program is:

```
user@system~]$ perl saber_pka_v5.pl
```

The program is currently configured to work best with PROPKA2 files, though it should be relatively straightforward to adapt it to parsing the output of future versions of the program. At this point, the SABER analysis process is complete.

3.11 Interpreting SABER Output

The output file from saber_wcn or saber_pka can be imported into Excel or another spreadsheet program very easily. Every field is delimited by the | character, so splitting on | will align the data into the cells properly. The cells will always be placed in the following order, up through 10. The remaining numbering will depend on the number of catalytic residues, since 3 WCN values will be provided for each one. The following table assumes there is only one catalytic residue in the CAM for simplicity.

Table 3.1. Data organization in SABER output.

Column	Contents
1	PDB identifier
2	RMSD vs. CAM (Å)

3	Maximum displacement from CAM (Å)
4	Protein name
5	ActiveSiteScore
6	EzCat identifier
7	BindingSiteFinder flag: yes/no
8	BindingSiteFinder data: Closest residue name, number, and atom type
9	BindingSiteFinder data: Heteroatom residue name, number, and atom type
10	BindingSiteFinder data: Heteroatom distance
11	WCN value of residue, heavy atom maximum
12	WCN value of residue, heavy atom average
13	WCN value of residue, Cα
14	Residues identified by the CAM
15	Predicted pKa of residue
16+	Known active site residues for this PDB ID, taken from the Catalytic Site Atlas

3.12 SABER Source File Locations

The SABER files are available to all UCLA users with Hoffman2 cluster accounts. The programs are located in the directories:

SABER: /u/home/campus/geoff/saber

Jess: /u/home/campus/geoff/Jess-1.1gamma

Additional libraries needed to run the program are listed below:

ActiveSiteFinder CSA data: /u/home/campus/geoff/CSA_HTML

SABER_wcn library: /u/home/campus/geoff/PDB_cnf

3.13 References

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

2. Barker, J. A., and Thornton, J. M. (2003) An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis, *Bioinformatics 19*, 1644-1649.

3. Porter, C. T., Bartlett, G. J., and Thornton, J. M. (2004) The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data, *Nucleic Acids Res. 32*, D129-D133.

4. Huang, S. W., Yu, S. H., Shih, C. H., Guan, H. W., Huang, T. T., and Hwang, J. K. (2011) On the Relationship Between Catalytic Residues and their Protein Contact Number, *Curr. Protein Pept. Sc. 12*, 574-579.

5. Lin, C. P., Huang, S. W., Lai, Y. L., Yen, S. C., Shih, C. H., Lu, C. H., Huang, C. C., and Hwang, J. K. (2008) Deriving protein dynamical properties from weighted protein contact number, *Proteins 72*, 929-935.

6. Olsson, M. H. M., Sondergaard, C. R., Rostkowski, M., and Jensen, J. H. (2011) PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pK(a) Predictions, *J. Chem. Theory. Comput. 7*, 525-537.

7. Bas, D. C., Rogers, D. M., and Jensen, J. H. (2008) Very fast prediction and rationalization of pK(a) values for protein-ligand complexes, *Proteins-Structure Function and Bioinformatics 73*, 765-783.

CHAPTER 4: USING CATALYTIC ATOM MAPS TO PREDICT THE CATALYTIC FUNCTIONS PRESENT IN ENZYME ACTIVE SITES

4.1 Abstract

Catalytic Atom Maps (CAMs) are minimal models of enzyme active sites. The structures in the Protein Data Bank (PDB) were examined to determine if proteins with CAM-like geometries in their active sites all share the same catalytic function. We combined the CAM-based search protocol with a filter based on the weighted contact number (WCN) of the catalytic residues, a measure of the "crowdedness" of the microenvironment around a protein residue. Using this technique, a CAM based on the Ser-His-Asp catalytic triad of trypsin was able to correctly identify catalytic triads in other enzymes within 0.5 Å RMSD of the Catalytic Atom Map with 96% accuracy. A CAM based on the Cys-Arg-(Asp/Glu) active site residues from the tyrosine phosphatase active site achieved 89% accuracy in identifying this type of catalytic functionality. Both of these Catalytic Atom Maps were able to identify active sites across different fold types. Finally, the PDB was searched to locate proteins with catalytic functionality similar to that present in the active site of orotidine 5'-monophosphate decarboxylase (ODCase), whose mechanism is not known with certainty. A CAM, based on the conserved Lys-Asp-Lys-Asp tetrad in the ODCase active site, was used to search the PDB for enzymes with similar active sites. The ODCase active site has a geometry similar to that of Schiff base-forming Class I aldolases, with lowest aldolase RMSD to the ODCase CAM at 0.48 Å. The similarity between this CAM and the aldolase active site suggests that ODCase has the correct catalytic functionality present in its active site for the generation of a nucleophilic lysine.

4.2 Introduction

The development of general methods for analysis of the chemical functionality present in active sites is an ongoing challenge for biochemists, given the rapid growth in protein structural data. In the last 3 years alone (2009-2011), more than 23,000 structures have been deposited in the Protein Data Bank. (*1*) In order to keep pace with the wealth of structural data being generated, there exists a need for high-throughput methods to analyze and identify the types of chemical reactions that an enzyme is likely to catalyze. We previously reported SABER (Selection of Active/Binding sites for Enzyme Redesign), a program used to search large sets of protein structures for specific arrangements of atoms that match the geometry described by a minimal model of an enzyme active site, known as a Catalytic Atom Map (CAM). (*2*) Here we introduce a method, using SABER and a Catalytic Atom Map, to identify the mechanisms available to an enzyme. We show how the precise geometric arrangement of catalytic groups such as nucleophiles, electrophiles, and general acids/bases can be used to predict the type of reaction catalyzed by a specific, preorganized active site. This is consistent with the recognition that much of the enormous proficiency of enzymes results from the precise organization of catalytic groups. (*3*-*5*)

To support our claim, we have explored two central questions. The first is whether a Catalytic Atom Map based on the conserved residues in an active site is sufficient to correctly identify the chemistry that occurs in that active site. We have created Catalytic Atom Maps for two enzyme families, one based on a Ser-His-Asp

catalytic triad, and the other from a tyrosine phosphatase active site. These CAMs were then used to search the Protein Data Bank for similar active sites. Proteins with close geometric matches to the Catalytic Atom Map were analyzed to determine if they catalyzed reactions with similar mechanisms. We also assessed the effectiveness of using a filter based on a residue's weighted contact number (WCN) value, a measure of the microenvironment occupied by a residue in the protein. This filter removes matches to the CAM that are unlikely to be part of an active site, because they are surface exposed. (*6, 7*)

In the second part of this study, we explored whether this technique could be used to identify the type of catalysis that is likely to operate in the active site of an enzyme that is known to catalyze a specific reaction, but for which the mechanism is unknown. We chose to study the enzyme orotidine 5'-monophosphate decarboxylase (ODCase), and created a Catalytic Atom Map based on the conserved residues in its active site. We then used this CAM to search the Protein Data Bank for other enzymes that have the catalytic atoms in their active sites placed in a geometry similar to that of ODCase. Based on this analysis, we gained evidence that ODCase has the correct catalytic functionality to generate a nucleophilic lysine in its active site.

4.3 Computational Details

SABER procedure and parameters

Catalytic Atom Map Generation: Coordinates for each of the Catalytic Atom Maps used in the SABER searches are provided in the Supporting Information. These CAMs were generated based on high-resolution X-ray structures from the Protein Data Bank (PDB codes: 1A0J, 1ZC0, and 3LTP).

SABER data collection: The radius of the sphere representing each atom in a CAM was set to 2.0 Å. A cutoff of 0.7 Å RMSD to each CAM was applied, and matches with RMSD values higher than this were discarded. Matches to each CAM were binned according to their geometric similarity to the Catalytic Atom Map in 0.1 Å increments, resulting in seven bins for each dataset: 0 – 0.10 Å, 0.11 – 0.2 Å, 0.21 – 0.30 Å, 0.31 – 0.40 Å, 0.41 – 0.50 Å, 0.51 – 0.60 Å, and 0.61 – 0.70 Å.

SABER scoring: After the data were binned, each bin was analyzed using the program's scoring functions. (*2*) First, each match was assigned an RMSD value to the Catalytic Atom Map. Second, using the ActiveSiteFinder module, matches were scored for overlap with known active site residues according to annotation in the Catalytic Site Atlas. (*8*) Matches with positive scores had at least one residue that was part of a known or putative active site. Third, each match was analyzed using the BindingSiteFinder module, an alternative method for active/binding site identification

based on detection of non-water PDB heteroatoms within 5 Å of the residues in the match. Finally, the SABER-WCN module was used to determine the weighted contact number (WCN) and WCN z-score for each residue. The WCN calculations are described below.

Weighted contact number (WCN) analysis in SABER

A new module was developed for SABER, called WCNcalc, to determine the weighted contact number values and z-scores for the residues identified in the SABER search. The WCN value of a residue is a measure of the crowdedness of the local protein environment around a residue.

The weighted contact number values for the individual residues were generated based on the WCN of the C_α atom for that residue. The WCN for atom *i* in a protein with N non-hydrogen atoms is defined as

$$WCN_i = \sum_{j \neq i}^{N} \frac{1}{r_{ij}^2}$$

as described in (*9*).

In order to normalize the WCN values used in this research, the z-score for each WCN was calculated using the formula

$$z_{wcn,i} = \frac{(\, wcn_i - \overline{wcn}\,)}{\sigma_{wcn}}$$

where wcn_i is the WCN of the C_α atom of the residue of interest, $\overline{wcn}$ is the average WCN for C_α atoms in the protein being analyzed, and σ_{wcn} is the standard deviation of the WCN of C_α atoms in the protein. The z_{wcn} value will often be referred to as the z-score for a residue in the text.

Analysis of matches to Catalytic Atom Maps

The SABER results, including WCNcalc results, were analyzed for all matches with an RMSD to the CAM of ≤ 0.7 Å. For the Ser-His-Asp triad CAM and the tyrosine phosphatase CAM, matches were identified as either true positives or false positives based on either their ActiveSiteScore or on literature analysis. Any match with an ActiveSiteScore equal to the number of residues in the Catalytic Atom Map was deemed a true positive. If a match had an ActiveSiteScore less than this number, the literature reference associated with the PDB code of the match was used to determine whether or not the residues identified by the Catalytic Atom Map were part of its active site and had the proposed chemical function in the protein. After this, the WCN z-score values of the residues contained in the true positive and false positive data sets were calculated individually.

For the analysis of ODCase-like active sites, it was not possible to bin these matches into true and false positives, as the mechanism of ODCase has not been conclusively proven. As such, we analyzed only the matches that had at least partial overlap with a known active site. Except where otherwise noted, these were defined as any hit within the 0.7 Å RMSD of the CAM and having an ActiveSiteScore > 0.

4.4 Results

Analysis of Ser-His-Asp catalytic triads using Catalytic Atom Maps

In order to determine whether or not a Catalytic Atom Map could be used to predict the presence of an active site that uses a catalytic triad to generate a nucleophilic serine hydroxyl group, we constructed at CAM based on a high-resolution structure of the serine protease trypsin (PDB code: 1A0J). We chose six atoms to represent the catalytic triad present in this enzyme, as shown in Figure 4.1. The Catalytic Atom Map parameters allowed the following atom types for each residue: serine (CB/OG), histidine (ND1/NE2), and aspartic acid (CG/OD1, OD2). The 1A0J CAM and the SABER program were used to search a subset of the Protein Data Bank for other catalytic triads. The PDB data set we used contained all X-ray structures of proteins with resolution ≤ 2.0 Å, with sequences that had ≥ 90% homology removed. We abbreviate this as the PDB90(2Å) dataset; this contained ~11,700 structures.

The SABER program was used to search the PDB90(2Å) dataset using the CAM described above. SABER uses the Jess geometric hashing algorithm to identify matches to the Catalytic Atom Map. (*10*) The program then uses a series of scoring functions to analyze each match to the CAM. For each match, it (a) provides the RMSD to the CAM, (b) scores for overlap with known active site residues annotated in the Catalytic Site Atlas, and (c) determines if any non-water heteroatoms are within 5 Å of any of the match residues. SABER is described more fully in the Computational Details section. A new module was added to the SABER that performs weighted contact number (WCN) analysis on the match residues, described below.

The weighted contact number (WCN) analysis was used to identify residues that were more likely to be part of an active site, and thus less likely to be either deeply buried in the protein structure or completely solvent exposed. The WCN is a measure of the average “crowdedness” of the environment occupied by a given residue or atom in a protein’s structure. This type of analysis has been used previously to characterize WCN values of catalytic residues vs. WCN values of non-catalytic residues, and the two classes have been show to have distinctly different distributions. (*7*) The z-scores of the WCN values were used to filter the geometric matches identified during the SABER search, so as to remove matches that were unlikely to be part of an active site.

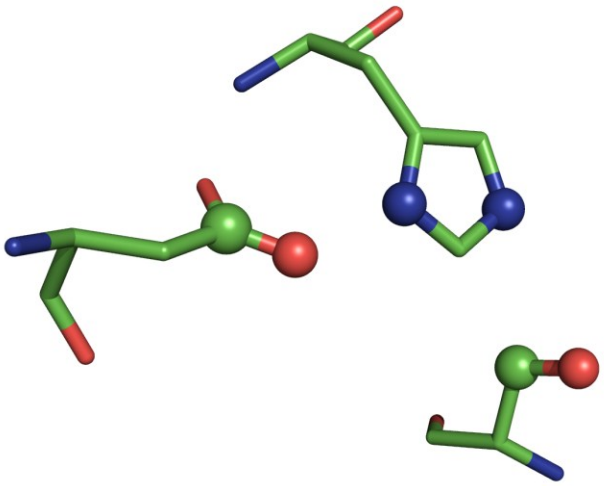


Figure 4.1. The Catalytic Atom Map for a Ser-His-Asp catalytic triad, based on the structure of trypsin (PDB code: 1A0J). The atoms used in the CAM are rendered as spheres.

After performing the SABER search, we analyzed the 127 structures identified during the search that had ≤ 0.7 Å RMSD to the Catalytic Atom Map. Each potential catalytic triad was categorized as either a true positive (part of a known triad) or a false positive (a match with a similar geometry to the CAM, but not a functioning catalytic triad). The results were sorted into bins by RMSD to the CAM in 0.1 Å increments, as shown in Figure 2. Within the 0.0 – 0.6 Å range, 79% of the geometric matches to the CAM were part of known catalytic triads. Beyond this range, there are still true positives, but the number of false positives rises dramatically. In the 0.51 – 60 Å bin, there is a 50% chance of being a false positive, and this probability rises to 73% in the 0.61 – 0.70

Å bin. We then used the SABER WCNcalc module to further analyze these matches, as shown in Table 4.1.

Table 4.1. Weighted contact number z-score analysis for matches to the 1A0J trypsin-based catalytic triad CAM.

Category	Ser z-score	His z-score	Asp z-score	Avg. z-score for triad
True positive	1.47 ± 0.34	0.65 ± 0.42	1.11 ± 0.37	1.08 ± 0.32
False positive	0.50 ± 1.16	0.58 ± 0.89	0.55 ± 1.01	0.55 ± 0.91

Based on the WCN analysis, it is clear that the catalytic residues in the true positive hits have significantly higher WCN values than the false positive hits, on average. This reflects the fact that catalytic triads, like other active site residues, tend to be in regions of a protein that are more crowded, have restricted motion, and are pre-organized for catalysis. (*6, 7*) We used the WCN z-score data to remove the false positive matches to the Catalytic Atom Map that were unlikely to be in a catalytic site. All of the matches that had an average z-score lower than the average true positive z-score minus one standard deviation (z-score < 0.76) were removed. The unfiltered false positives are shown in Figure 2 as red bars. We then eliminated the false positives with WCN z-scores < 0.76. These data are shown as green bars in Figure 4.2.

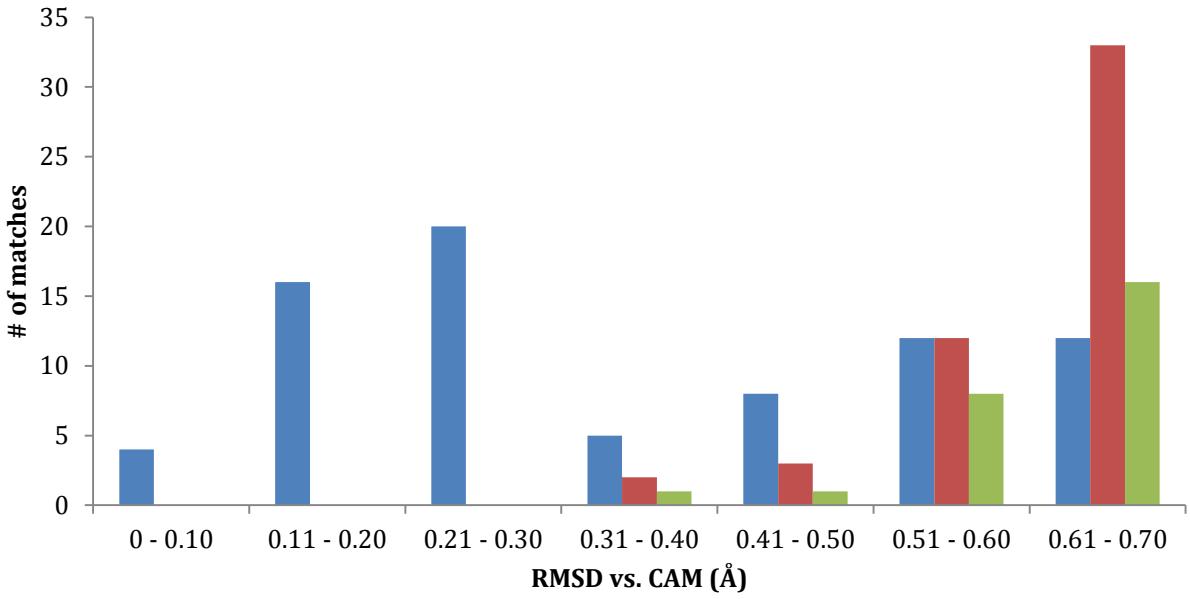


Figure 4.2. Results of the analysis of using the 1A0J Catalytic Atom Map to search the PDB90(2Å) data set. The blue bars indicate the number of correctly identified catalytic triads, while the red bars indicate groups of atoms that matched the CAM geometry but were not part of a catalytic triad. The green bars are incorrectly identified triads that remained after the WCN filter was applied.

After application of the WCN filter, 48% of the false positives were eliminated, as these matches were in regions of the protein unlikely to support a functional catalytic triad. Within 0.5 Å of the CAM, the triads were identified with 96% accuracy, and within 0.6 Å, 87% accuracy. Of the remaining false positive matches, the best was in a leukotriene A4 hydrolase enzyme (PDB code: 3B7S), with a 0.37 Å RMSD to the CAM. The Asp residue identified by the CAM, D375, is part of the substrate binding site of this enzyme. (*11*) The other two residues identified, Ser113 and His139, are not part of the active site. The WCN z-scores for these residues are 0.84 and 1.52, respectively, indicating a serine that is in a less crowded environment and a histidine that is in a much more crowded environment than those in the correctly identified triads. These

residues are well-positioned to function as a catalytic triad, but as the Ser and His are not part of the enzyme active site, the triad is inactive. The false positive with the next lowest RMSD, 0.44 Å, was an epoxide hydrolase (PDB code: 3KDA). (*12*) The potential triad again has reasonable geometric agreement with the CAM, but the residues are in a very crowded region of the protein, with an average WCN z-score of 1.52. The individual residues in the 3KDA match had WCN z-scores of 1.66, 1.62, and 1.27 for Ser, His, and Asp, respectively. The His and Asp residues are in a significantly more crowded protein environment than is typical for a catalytic triad. This makes catalytic function unlikely, as substrate access would be extremely limited, as would water access to the histidine imidazole ring for the release of the acyl-enzyme intermediate.

The correctly identified catalytic triads were not exclusively from trypsin-like enzymes. Catalytic triads were identified in the subtilisin fold and the α/β hydrolase fold as well, indicating that the CAM used is not fold-specific. In addition, the triads identified also included matches in ester hydrolases. The lack of reaction specificity shows that this CAM can be used to identify the catalytic machinery present in the active site, but not the specific reaction that takes place. This is not surprising, since catalytic triads are used to catalyze a number of different reactions in addition to peptide hydrolysis. In the α/β hydrolase fold alone, Ser-His-Asp triads are used for catalysis of peptide hydrolysis, ester hydrolysis, lipid hydrolysis, and haloperoxidase reactions. In every case, the catalytic triad activates the serine nucleophile for attack on an electrophilic carbon atom. (*13*)

Although the CAM used was able to identify Ser-His-Asp catalytic triads with high accuracy within the 0 – 0.60 Å RMSD range, there were true catalytic triads that fell outside this range. Some enzyme-catalyzed reactions associated with these triads, such as lipid hydrolysis and haloperoxidation, were not identified using the 1A0J Catalytic Atom map within the RMSD limits specified. The enzymes do not appear among the matches in Figure 2, as lipases first appear in the 0.71 – 0.80 Å RMSD bin, and haloperoxidases in the 0.81 – 0.90 Å RMSD bin. This result is not surprising, as we used a Catalytic Atom Map based on a single enzyme structure. While this CAM was able to correctly identify Ser-His-Asp catalytic triads with high accuracy, it does not locate every possible true catalytic triad in the PDB90(2Å) dataset.

Our search identified the enzymes that contain similar catalytic functionality in their active sites prearranged in a specific geometry. The combination of the geometry of the catalytic atoms and WCN scoring was sufficient to locate other active sites that are known to contain this serine nucleophile-generating functionality. We have found that is the RMSD to the CAM is ≤ 0.5 Å, the catalytic mechanism is very likely (> 90%) to operate in the same way. For enzymes with unknown mechanisms, this method has the potential to give insight into the catalytic mechanisms available to a given active site.

Analysis of tyrosine phosphatase active sites using Catalytic Atom Maps

Next, we sought to verify that a Catalytic Atom Map and the SABER program could be used to identify the chemical functionality of other types of active sites. The tyrosine phosphatases (TyrPs) are also known to have a highly conserved active site that exists in more than one fold type. (*14*) These enzymes catalyze the hydrolysis of of the phosphate group from phosporylated tyrosine residues, using a nucleophilic cysteine residue in the active site. This type of active site contains a highly conserved Cys-Arg-(Asp/Glu) triad, which we used the basis for the Catalytic Atom Map. The atoms in the CAM were constrained to match only the following atom types for each residue: Cys (CB/SG), Arg (CZ/NH1, NH2), and Asp/Glu (CG, CD/OD1, OD2, OE1, OE2).

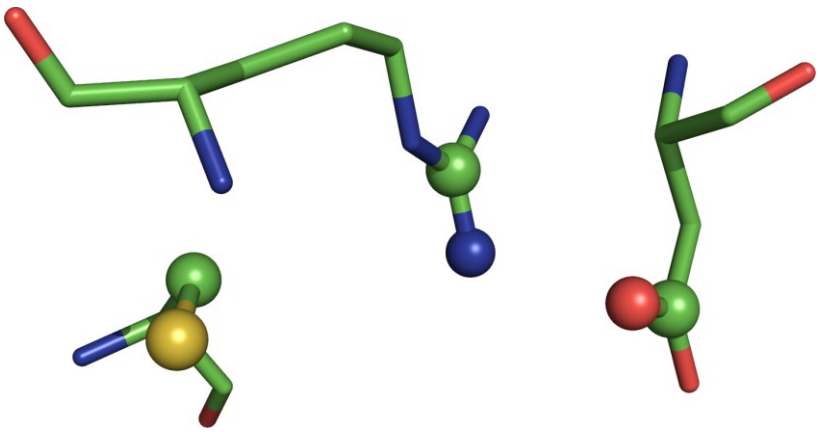


Figure 4.3. The Catalytic Atom Map for the Cys-Arg-(Asp/Glu) triad of tyrosine phosphatase, based on the structure of the human hematopoietic tyrosine phosphatase (PDB code: 1ZC0). Catalytic atoms are rendered as spheres.

Using SABER, we performed the same procedure as employed previously for the catalytic triad-based CAM search of the PDB90(2Å) dataset with the tyrosine

phosphatase CAM in Figure 4.3. This search identified a total of 27 proteins with tyrosine phosphatase-like triads with RMSD values ≤ 0.7 Å vs. the CAM. There were 11 hits in the 0.0 – 0.5 Å range, 73% of which were true TyrP catalytic triads. The WCN analysis of the catalytic residues was also performed, shown in Table 4.2.

Table 4.2. Weighted contact number z-score analysis for matches to the 1ZC0 TyrP-based catalytic triad CAM.

Category	Cys z-score	Arg z-score	Asp/Glu z-score	Avg. z-score for Cys/Arg
True positive	1.51 ± 0.39	1.51 ± 0.31	0.21 ± 0.43	1.51 ± 0.34
False positive	0.75 ± 0.68	0.40 ± 0.87	0.26 ± 1.14	0.58 ± 0.74

Unlike the Ser-His-Asp catalytic triads, the TyrP active site has the ability to tolerate significant variation in the packing around the Asp/Glu residue, and this residue is in a much less crowded environment than is typical for catalytic residues. Without an analysis of WCN z-scores for many different types of enzyme active sites, it is impossible to know if it is unusual for an active site residue to have such wide variation in its WCN z-score. The WCN z-score values for the Asp/Glu residue do not distinguish at all between catalytic and non-catalytic arrangements, but the WCN z-score from the Cys and Arg residues are excellent indicators. Only these z-scores from Cys and Arg were used in our analysis, shown in Figure 4.4.

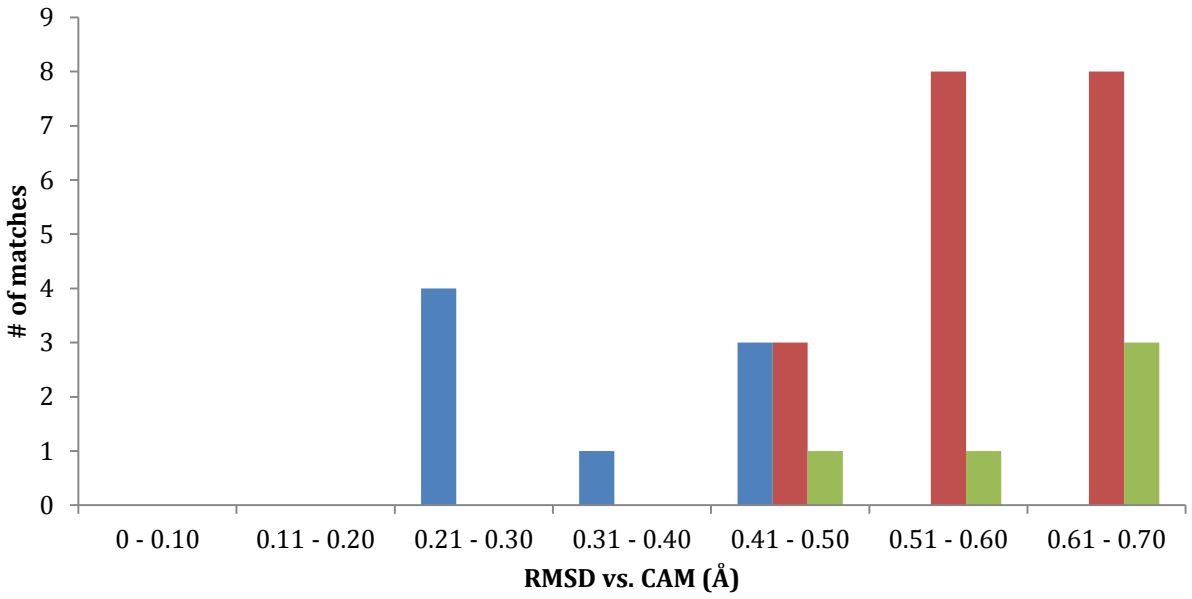


Figure 4.4. Results of the analysis of using the 1ZC0 Catalytic Atom Map to search the PDB90(2Å) data set. The blue bars indicate the number of correctly identified TyrP-like catalytic triads, while the red bars indicate groups of atoms that matched the CAM geometry but were not part of a TyrP catalytic triad. The green bars are incorrectly identified triads that remained after the WCN filter was applied.

As was the case with the Ser-His-Asp catalytic triads, the WCN z-score filter removed a significant number of false positives, reducing them by 74%. Within 0 – 0.5 Å RMSD to the CAM, true Cys-Arg-(Asp/Glu) triads were identified with 89% accuracy, and within 0 – 0.6 Å they were identified with 80% accuracy. The only false positive within the 0.0 – 0.5 Å range was the YdcF protein from *E. coli*, at 0.47 Å RMSD to the TyrP CAM. The function of this protein is not known, but it does bind S-adenosyl methionine. (*15*) All the residues identified by the TyrP CAM in this structure are in a very crowded region of the protein. The WCN z-scores for Cys, Arg, and Glu are 1.76, 1.52, and 0.93, respectively. This indicated a region of the protein that is likely to be completely inaccessible. The acid residue in particular is deeply buried compared to a

true tyrosine phosphatase. Beyond 0.5 Å, all of the matches were false positives. Also like the Ser-His-Asp catalytic triad CAM, the TyrP CAM was also to identify active sites across different fold types. TyrP enzymes with the Class I fold and the low molecular weight fold were both correctly identified.

Analysis of the ODCase active site using Catalytic Atom Maps

Based on the analysis of Ser-His-Asp catalytic triads (96% accuracy within 0.5 Å of the CAM) and the Cys-Arg-(Asp/Glu) residues of the tyrosine phosphatase active site (89% accuracy within 0.5 Å of the CAM), we concluded that this methodology could be used to identify enzymes with a common active site chemistry, in this case the deprotonation of serine or cysteine to activate them as nucleophiles. Next, we turned to investigating orotidine 5'-monophosphate decarboxylase (ODCase) in order to identify other enzymes with active sites having similar atomic geometries, and to use this information to predict the possible cataltyic function of the conserved groups present in the ODCase active site.

Despite more than two decades of research, the mechanism of ODCase has not been definitively established. The overall reaction catalyzed by this enzyme is shown in Figure 4.5. As described by Wolfenden in 1995, ODCase has an unusually high catalytic proficiency ($k_{cat}/K_M/k_{uncat}$) of 4.8 x 10^{22} M^{-1} and a k_{cat}/k_{uncat} of 7.1 x 10^{16}. At the time, ODCase was the most proficient enzyme yet discovered. (*16*) While there have been discoveries of enzymes with higher catalytic proficiencies, such as

uroporphyrinogen decarboxylase (*17*) and the S-O cleaving sulfatases (*18*), ODCase still represents a fascinating challenge to scientists because the source of its extreme catalytic proficiency has yet to be definitively established.

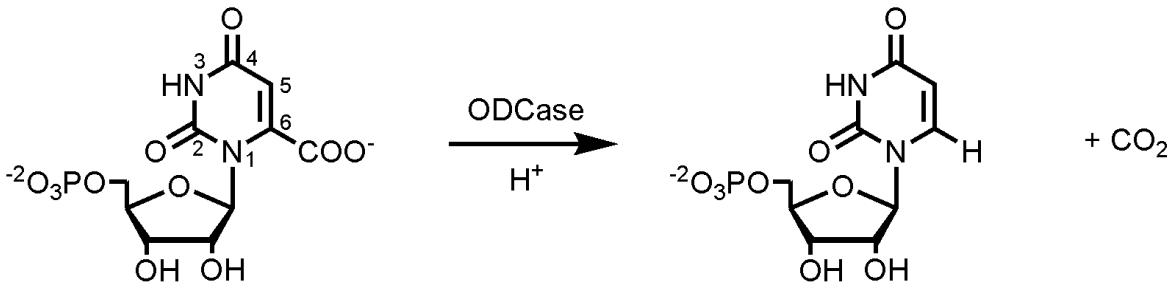


Figure 4.5. The reaction catalyzed by orotidine 5'-monophosphate decarboxylase (ODCase).

ODCase provides its enormous rate acceleration despite the fact that its active site does not utilize any cofactors, such as pyridoxal, or metal ions. (*19-21*) Several cofactor and ion-free mechanisms have been proposed, such as protonation at O2, O4, C5, or C6. (*22-27*) Nucleophilic mechanisms, such as iminium ion formation at C4 or a Michael addition at C5, have also been suggested. These mechanisms are generally discounted due to lack of oxygen exchange with ^{18}O-labeled water and kinetic isotope effects, respectively. (*28-30*) In addition, both electrostatic stabilization of the transition state and electrostatic destabilization of the substrate have been proposed as the source of the catalytic proficiency. (*31, 32*) These mechanistic proposals have all been reviewed in detail in reference (*33*).

Additional information about the ODCase mechanism has been obtained from structural data. The active sites of ODCases from many species has been extensively

studied using X-ray crystallography and subjected to a great deal of mutational analysis. (*34-37*) The catalytic functionality of the enzyme is controlled by a Lys-Asp-Lys-Asp tetrad in the active site. Mutation of any of these residues reduces the activity of ODCase by more than five orders of magnitude. (*38*) This active site tetrad is highly conserved, and is found across the ODCase enzymes from many species, including *E. coli*, *S. cerevisiae*, *M. thermoautotrophicum*, *B. subtilis*, and others. Beyond the Lys-Asp-Lys-Asp tetrad, other interactions with the substrate have also been implicated in ODCase catalysis. Binding of the substrate's phosphate group contributes significantly to catalysis in this enzyme. (*39-41*) A trio of hydrophobic residues in the active site have been shown to reduce k_{cat} by as much as 400-fold upon mutation to neutral hydrophilic residues, and mutation of a serine residue near the O4 atom of the substrate to a proline reduces k_{cat}/K_M by more than a factor of 10^6, consistent with proton transfer occurring within the vicinity of O4 in the transition state. (*42, 43*) Even with all of this information, there has been no definitive consensus on the mechanism of this enzyme.

As with the previous two active sites, we constructed a Catalytic Atom Map based on a high-resolution crystal structure of ODCase (PDB code: 3LTP). For this CAM, the nitrogen atoms were allowed to match only the sidechain nitrogen (NZ) atoms of Lys residues, and the oxygen atoms were allowed to match only the carboxylate oxygens of Asp or Glu (OD1, OD2, OE1, OE2).

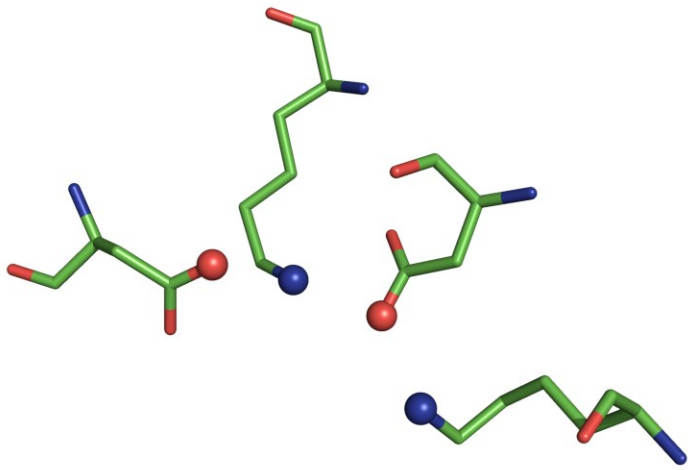


Figure 4.6. The Catalytic Atom Map for orotidine 5'-monophosphate decarboxylase, based on the structure of the *M. thermoautotrophicum* ODCase (PDB code: 3LTP). Catalytic atoms are rendered as spheres.

We set out to look for all functions of the ODCase Lys-Asp-Lys-Asp tetrad, as we had no way to distinguish true positives from false positives. We examined all of the geometric matches to the ODCase CAM that were within 0.7 Å RMSD and had an ActiveSiteScore > 0, indicating that the geometric match had at least partial overlap with a known active site. In addition, we used the BindingSiteFinder module in SABER to identify any potential active site matches with a metal ion within 2.5 Å of the atoms identified by the Catalytic Atom Map. As the ODCase active site is known to be free of metals, any match with a metal ion in such close proximity to the catalytic residues is unlikely to share a similar mechanism. (*19*) The results of this analysis are shown in Table 4.3.

Table 4.3. Results of the analysis of using the 3LTP Catalytic Atom Map to search the PDB90 data set. Data is shown for matches to the CAM with a GeometryScore ≤ 0.7 Å and an ActiveSiteScore > 0, except where noted.

PDB	Protein	RMSD (Å)	Avg. WCN z-score[a]	Metal within 2.5 Å?	Catalytic Element Identified
2FDS	Orotidine 5'-monophosphate decarboxylase	0.42	1.58	No	?
3HJZ[b]	Transaldolase	0.48	0.53	No	One lysine functions as a nucleophile; reaction proceed through a Schiff base intermediate
2PS2	Mandelate racemase	0.48	1.09	Yes	Carboxylate residues bind Mg^{2+}, lysines function as general acid/base
2CZD	Orotidine 5'-monophosphate decarboxylase	0.54	1.70	No	?
2HXT	L-fuconate dehydratase	0.61	1.39	Yes	Carboxylate residues bind Mg^{2+}, lysines function as general acid/base
3H12	Mandelate racemase	0.64	0.94	Yes	Carboxylate residues bind Mg^{2+}, lysines function as general acid/base
2A4A	Deoxyribose phosphate aldolase	0.65	1.42	No	One lysine functions as a nucleophile; reaction proceed through a Schiff base intermediate
2ZAD	Muconate cycloisomerase	0.65	1.63	Yes	Carboxylate residues bind Mg^{2+}, lysines function as general acid/base
1N7K	Deoxyribose phosphate aldolase	0.69	1.13	No	One lysine functions as a nucleophile; reaction proceed through a Schiff base intermediate

[a]WCN z-score is the average of the WCN z-scores for each of the residues identified by the Catalytic Atom Map.

[b]3HJZ was the aldolase structure with the lowest RMSD identified, but its ActiveSiteScore was 0, as it had not been annotated in the Catalytic Site Atlas at the time of this analysis.

It is immediately apparent that all of the active sites identified conform to the observation that active site residues are likely to be in more crowded regions of an enzyme's structure, as evidenced by the high WCN z-score values. Even the lowest average WCN z-score of 0.53 is more than one-half of one standard deviation higher in terms of WCN versus an average residue in that protein. Beyond that, the matches can be classified into three categories. The first of these categories was other ODCase enzymes, which we expected would be present, given the highly conserved geometry of

this enzyme's active site. The Lys-Asp-Lys-Asp catalytic tetrads from the 2FDS and 2CZD ODCase matches are shown in Figure 4.7, superimposed with the Catalytic Atom Map used in the SABER search of the PDB90(2Å) data set.

The second category consists of members of the enolase superfamily, known to include mandelate racemase, muconate cycloisomerase, and L-fuconate dehydratase. All of these enzymes use an Mg^{2+} ion bound by three carboxylate-containing sidechains to stabilize an enediolate intermediate, accompanied by other residues, often lysines, that act as general acids/bases. (*44-48*) Previous work has shown that a search of the PDB based on Catalytic Atom Map of the o-succinyl benzoate synthase (OSBS), another member of the enolase superfamily, does not locate the active site of ODCase, but does locate mandelate racemase and muconate cycloisomerase. (*2*) This is due to the ODCase active site's lack of a third carboxylate sidechain, in a geometry appropriate to bind the Mg^{2+} ion. Each of the enolase superfamily enzymes located using the ODCase CAM had a metal ion within 2.4 Å of one of the carboxylate sidechain oxygens. Despite the structural similarities of these active sites in terms of residue placement, the existence of the metal ion makes any sort of mechanistic overlap with ODCase extremely unlikely.

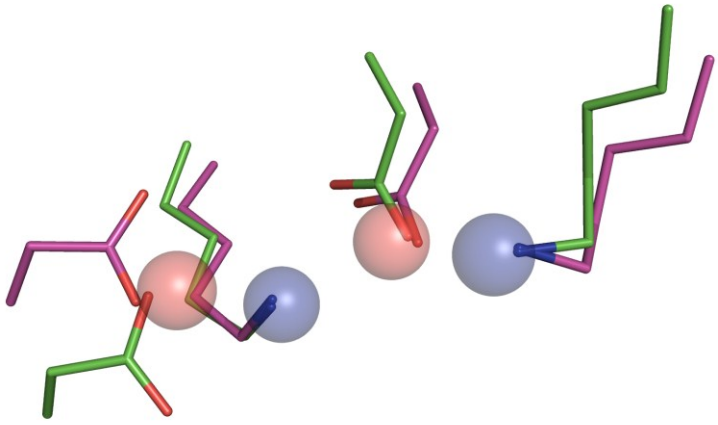


Figure 4.7. Superposition of the ODCase matches from the SABER search on the 3LTP-based ODCase Catalytic Atom Map. The ODCase active site from 2FDS is shown in green, while the active site from 2CZD is shown in magenta. The Catalytic Atom Map is shown as spheres.

The third category of matches to the ODCase Catalytic Atom Map were the Class I aldolases: transaldolase (PDB code: 3HJZ) and deoxyribose-phosphate aldolase (PDB codes: 2A4A and 1N7K). The Class I aldolases use a metal-free Schiff base mechanism to catalyze aldol condensation reactions and have a highly conserved Lys-Asp-Lys triad in the active site. (*49*) For example, the enzyme 2-dexoyribose-5-phosphate aldolase (DERA), which catalyzes an aldol condensation between acetaldehyde and glyceraldehyde-3-phosphate, uses the nucleophilic lysine in the Lys-Asp-Lys triad to attack the carbonyl carbon of acetaldehyde and form a Schiff base. This covalent intermediate has been observed via X-ray crystallography. (*50*) The SABER search revealed that the active sites of these Class I aldolase enzymes have a very similar geometry to the ODCase active site. The superposition of the three Class I aldolase matches with the ODCase Catalytic Atom Map is shown in Figure 4.8.

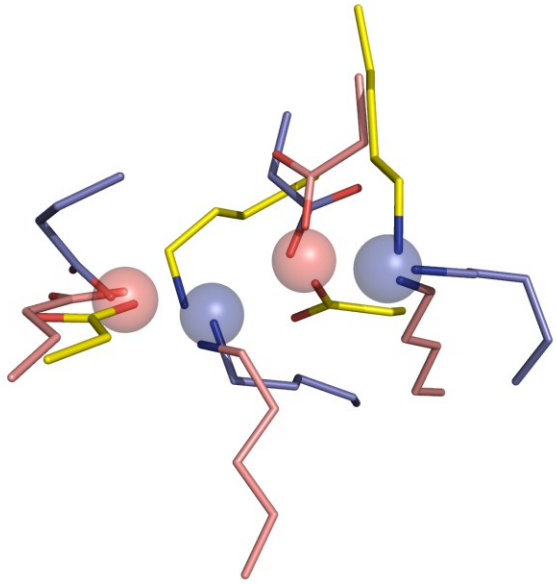


Figure 4.8. Superposition of the aldolase matches from the SABER search on the 3LTP-based ODCase Catalytic Atom Map. The aldolase active site from 2A4A is shown in yellow, the active site from 1N7K is shown in pink, and the active site from 3HJZ is shown in purple. The Catalytic Atom Map is shown as spheres.

This research does not prove that the ODCase active site functions through a similar mechanism, although the formation of an iminium ion at C4 has been suggested previously. (*29*) What the geometric similarity of these active sites does suggest, however, is that the active site of ODCase contains catalytic residues placed in a geometry appropriate for the generation of a nucleophilic lysine, in the same manner of the Class I aldolase active sites. This has been demonstrated experimentally, as covalent inhibitors have been developed for ODCase. 6'-iodouridine-5'-monophosphate (6-iodo-UMP) has been shown using X-ray crystallography to form a covalent bond with a Lys residue (*M. thermoautotrophicum* Lys72) in the ODCase active site. (*51*) 6'-azidouridine-5'-monophosphate (6-azido-UMP) has also been shown to be a covalent inhibitor of the ODCase enzymes from both *M. thermoautotrophicum* and *P. falciparum.*

(*52*) Unlike the iminium ion mechanism that was proposed that requires attack at C4, or the Silverman mechanism where nucleophilic attack would occur at C5, the nucleophilic lysine that has been observed in the ODCase active site reacts with the C6 atom, where the iodide leaving group was attached.

Although it is not possible to establish definitively the reaction mechanism of ODCase using this approach, its similarity to the Class I aldolase active site allows us to narrow the choice of potential mechanisms to investigate. Given the nucleophilic character of one of the lysines on the ODCase active site, this means it must spend a significant amount of time in the unprotonated and uncharged state. This has implications for the charge state, and thus the catalytic function, of the other residues in the ODCase catalytic tetrad. In addition to narrowing the choice of likely reaction mechanisms, this methodology is also useful for identification of additional functionality in an enzyme's active site that can be utilized for both inhibitor design and for active site redesign. The identification of a functional nucleophile in an enzyme's active site, whether or not it participates in the reaction mechanism, provides a target for inhibition. Likewise, a nucleophile identified in this manner could be utilized as part of a rational active site redesign process, in which the redesigned active site utilizes the nucleophile identified as part of its catalytic mechanism. This type of enzyme design strategy is discussed in reference (*2*).

As was the case with Ser-His-Asp catalytic triads, a geometric match to the Catalytic Atom Map does not determine the overall reaction catalyzed by the enzyme.

The CAM predicts the mode of catalysis that the functionality in an enzyme's active site can effect. The geometric similarity of the atoms in the ODCase CAM to the catalytic atoms in the aldolase active site shows that the ODCase active site has the correct chemical functionality to generate a lysine nucleophile. While this prediction is correct, as demonstrated by the covalent inhibitors described above, it does not establish that a nucleophilic lysine participates in the ODCase reaction mechanism. Given the placement of the lysine nuceophile in the active site of this enzyme, it is difficult to argue in support of the iminium ion mechanism based on these results alone. However, as we have no information regarding the nucleophilicity of the other lysine residue in the ODCase active site, this remains an open question.

4.5 Conclusion

We have presented a computational method for analysis of the chemical functionality present in the active site of an enzyme based on comparison to the active site of other enzymes with known mechanisms. As high-throughput structure determination projects such as the Protein Structure Initiative (*53*) expand our knowledge of the protein structures that exist in nature, this analysis too will aid in the prediction of enzyme functions. This is the second application for which we have used SABER. Previously, we have also demonstrated the utility of this program for enzyme design, and now it has been used successfully for the prediction of catalytic functionality in active sites. (*2*)

4.6 Acknowledgements

This work was supported by the National Institute of General Medical Science, National Institutes of Health (GM 75962 and T32 GM008496-16), and the Department of Defense, Defense Advanced Research Projects Agency.

4.7 References

1. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000) The Protein Data Bank, *Nucleic Acids Res. 28*, 235-242.

2. Nosrati, G. R., and Houk, K. N. (2012) SABER: A computational method for identifying active sites for new reactions, *Protein Sci. 21*, 697-706.

3. Warshel, A. (1998) Electrostatic origin of the catalytic power of enzymes and the role of preorganized active sites, *J. Biol. Chem. 273*, 27035-27038.

4. Warshel, A., Sharma, P. K., Kato, M., Xiang, Y., Liu, H., and Olsson, M. H. (2006) Electrostatic basis for enzyme catalysis, *Chem. Rev. 106*, 3210-3235.

5. Smith, A. J., Muller, R., Toscano, M. D., Kast, P., Hellinga, H. W., Hilvert, D., and Houk, K. N. (2008) 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. Am. Chem. Soc. 130*, 15361-15373.

6. Lin, C. P., Huang, S. W., Lai, Y. L., Yen, S. C., Shih, C. H., Lu, C. H., Huang, C. C., and Hwang, J. K. (2008) Deriving protein dynamical properties from weighted protein contact number, *Proteins 72*, 929-935.

7. Huang, S. W., Yu, S. H., Shih, C. H., Guan, H. W., Huang, T. T., and Hwang, J. K. (2011) On the Relationship Between Catalytic Residues and their Protein Contact Number, *Curr. Protein Pept. Sc. 12*, 574-579.

8. Porter, C. T., Bartlett, G. J., and Thornton, J. M. (2004) The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data, *Nucleic Acids Res. 32*, D129-D133.

9. Rothlisberger, 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., and Baker, D. (2008) Kemp elimination catalysts by computational enzyme design, *Nature 453*, 190-195.

10. Barker, J. A., and Thornton, J. M. (2003) An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis, *Bioinformatics 19*, 1644-1649.

11. Tholander, F., Muroya, A., Roques, B. P., Fournie-Zaluski, M. C., Thunnissen, M. M. G. M., and Haeggstrom, J. Z. (2008) Structure-Based Dissection of the Active

Site Chemistry of Leukotriene A4 Hydrolase: Implications for M1 Aminopeptidases and Inhibitor Design, *Chemistry & Biology 15*, 920-929.

12. Bahl, C. D., Morisseau, C., Bomberger, J. M., Stanton, B. A., Hammock, B. D., O'Toole, G. A., and Madden, D. R. (2010) Crystal Structure of the Cystic Fibrosis Transmembrane Conductance Regulator Inhibitory Factor Cif Reveals Novel Active-Site Features of an Epoxide Hydrolase Virulence Factor, *J. Bacteriol. 192*, 1785-1795.

13. Holmquist, M. (2000) Alpha/Beta-Hydrolase Fold Enzymes: Structures, Functions and Mechanisms, *Curr. Protein Pept. Sc. 1*, 209-235.

14. Moorhead, G. B. G., De Wever, V., Templeton, G., and Kerk, D. (2009) Evolution of protein phosphatases in plants and animals, *Biochem. J. 417*, 401-409.

15. Chao, K. L., Lim, K., Lehmann, C., Doseeva, V., Howard, A. J., Schwarz, F. P., and Herzberg, O. (2008) The Escherichia coli YdcF binds S-adenosyl-L-methionine and adopts an alpha/beta-fold characteristic of nucleotide-utilizing enzymes, *Proteins-Structure Function and Bioinformatics 72*, 506-509.

16. Radzicka, A., and Wolfenden, R. (1995) A proficient enzyme, *Science 267*, 90-93.

17. Lewis, C. A., and Wolfenden, R. (2008) Uroporphyrinogen decarboxylation as a benchmark for the catalytic proficiency of enzymes, *Proc. Natl. Acad. Sci. U.S.A. 105*, 17328-17333.

18. Edwards, D. R., Lohman, D. C., and Wolfenden, R. V. (2011) Catalytic Proficiency: The Extreme Case of S-O Cleaving Sulfatases, *J. Am. Chem. Soc. 134*, 525-531.

19. Cui, W., DeWitt, J. G., Miller, S. M., and Wu, W. (1999) No metal cofactor in orotidine 5'-monophosphate decarboxylase, *Biochem. Biophys. Res. Commun. 259*, 133-135.

20. Appleby, T. C., Kinsland, C., Begley, T. P., and Ealick, S. E. (2000) The crystal structure and mechanism of orotidine 5'-monophosphate decarboxylase, *Proc. Natl. Acad. Sci. U.S.A. 97*, 2005-2010.

21. Miller, B. G., Hassell, A. M., Wolfenden, R., Milburn, M. V., and Short, S. A. (2000) Anatomy of a proficient enzyme: the structure of orotidine 5'-monophosphate decarboxylase in the presence and absence of a potential transition state analog, *Proc. Natl. Acad. Sci. U.S.A. 97*, 2011-2016.

22. Beak, P., and Siegel, B. (1976) Mechanism of decarboxylation of 1,3-dimethylorotic acid. A model for orotidine 5'-phosphate decarboxylase, *J. Am. Chem. Soc. 98*, 3601-3606.

23. Lee, J. K., and Houk, K. N. (1997) A proficient enzyme revisited: the predicted mechanism for orotidine monophosphate decarboxylase, *Science 276*, 942-945.

24. Begley, T. P., Appleby, T. C., and Ealick, S. E. (2000) The structural basis for the remarkable catalytic proficiency of orotidine 5'-monophosphate decarboxylase, *Curr. Opin. Struct. Biol. 10*, 711-718.

25. Houk, K. N., Lee, J. K., Tantillo, D. J., Bahmanyar, S., and Hietbrink, B. N. (2001) Crystal structures of orotidine monophosphate decarboxylase: Does the structure reveal the mechanism of nature's most proficient enzyme?, *Chembiochem 2*, 113-118.

26. Lee, T. S., Chong, L. T., Chodera, J. D., and Kollman, P. A. (2001) An alternative explanation for the catalytic proficiency of orotidine 5'-phosphate decarboxylase, *J. Am. Chem. Soc. 123*, 12837-12848.

27. Lundberg, M., Blomberg, M. R. A., and Siegbahn, P. E. M. (2002) Density functional models of the mechanism for decarboxylation in orotidine decarboxylase, *J. Mol. Model. 8*, 119-130.

28. Silverman, R. B., and Groziak, M. P. (1982) Model Chemistry for a Covalent Mechanism of Action of Orotidine 5'-Phosphate Decarboxylase, *J. Am. Chem. Soc. 104*, 6434-6439.

29. Shostak, K., and Jones, M. E. (1992) Orotidylate Decarboxylase - Insights into the Catalytic Mechanism from Substrate-Specificity Studies, *Biochemistry 31*, 12155-12161.

30. Acheson, S. A., Bell, J. B., Jones, M. E., and Wolfenden, R. (1990) Orotidine-5'-Monophosphate Decarboxylase Catalysis - Kinetic Isotope Effects and the State of Hybridization of a Bound Transition-State Analog, *Biochemistry 29*, 3198-3202.

31. Warshel, A., Florian, J., Strajbl, M., and Villa, J. (2001) Circe effect versus enzyme preorganization: what can be learned from the structure of the most proficient enzyme?, *Chembiochem 2*, 109-111.

32. Wu, N., Mo, Y., Gao, J., and Pai, E. F. (2000) Electrostatic stress in catalysis: structure and mechanism of the enzyme orotidine monophosphate decarboxylase, *Proc. Natl. Acad. Sci. U.S.A. 97*, 2017-2022.

33. Houk, K. N., Tantillo, D. J., Stanton, C., and Hu, Y. F. (2004) What have theory and crystallography revealed about the mechanism of catalysis by orotidine monophosphate decarboxylase?, *Top. Curr. Chem. 238*, 1-22.

34. Miller, B. G., Hassell, A. M., Wolfenden, R., Milburn, M. V., and Short, S. A. (2000) Anatomy of a proficient enzyme: The structure of orotidine 5 '-

monophosphate decarboxylase in the presence and absence of a potential transition state analog, *Proceedings of the National Academy of Sciences of the United States of America 97*, 2011-2016.

35. Appleby, T. C., Kinsland, C., Begley, T. P., and Ealick, S. E. (2000) The crystal structure and mechanism of orotidine 5 '-monophosphate decarboxylase, *Proc. Natl. Acad. Sci. U.S.A. 97*, 2005-2010.

36. Harris, P., Poulsen, J. C. N., Jensen, K. F., and Larsen, S. (2000) Structural basis for the catalytic mechanism of a proficient enzyme: Orotidine 5 '-monophosphate decarboxylase, *Biochemistry 39*, 4217-4224.

37. Wu, N., Mo, Y. R., Gao, J. L., and Pai, E. F. (2000) Electrostatic stress in catalysis: Structure and mechanism of the enzyme orotidine monophosphate decarboxylase, *Proc. Natl. Acad. Sci. U.S.A. 97*, 2017-2022.

38. Miller, B. G., Snider, M. J., Wolfenden, R., and Short, S. A. (2001) Dissecting a charged network at the active site of orotidine-5'-phosphate decarboxylase, *J. Biol. Chem. 276*, 15174-15176.

39. Goryanova, B., Amyes, T. L., Gerlt, J. A., and Richard, J. P. (2011) OMP Decarboxylase: Phosphodianion Binding Energy Is Used To Stabilize a Vinyl Carbanion Intermediate, *J. Am. Chem. Soc. 133*, 6545-6548.

40. Barnett, S. A., Amyes, T. L., Wood, B. M., Gerlt, J. A., and Richard, J. P. (2008) Dissecting the total transition state stabilization provided by amino acid side chains at orotidine 5 '-monophosphate decarboxylase: A two-part substrate approach, *Biochemistry 47*, 7785-7787.

41. Amyes, T. L., Richard, J. P., and Tait, J. J. (2005) Activation of orotidine 5 '-monophosphate decarboxylase by phosphite dianion: The whole substrate is the sum of two parts, *J. Am. Chem. Soc. 127*, 15708-15709.

42. Iams, V., Desai, B. J., Fedorov, A. A., Fedorov, E. V., Almo, S. C., and Gerlt, J. A. (2011) Mechanism of the Orotidine 5 '-Monophosphate Decarboxylase-Catalyzed Reaction: Importance of Residues in the Orotate Binding Site, *Biochemistry 50*, 8497-8507.

43. Wood, B. M., Amyes, T. L., Fedorov, A. A., Fedorov, E. V., Shabila, A., Almo, S. C., Richard, J. P., and Gerlt, J. A. (2010) Conformational Changes in Orotidine 5 '-Monophosphate Decarboxylase: "Remote" Residues That Stabilize the Active Conformation, *Biochemistry 49*, 3514-3516.

44. Yew, W. S., Fedorov, A. A., Fedorov, E. V., Rakus, J. F., Pierce, R. W., Almo, S. C., and Gerlt, J. A. (2006) Evolution of enzymatic activities in the enolase superfamily: L-fuconate dehydratase from Xanthomonas campestris, *Biochemistry 45*, 14582-14597.

45. Gerlt, J. A., Babbitt, P. C., and Rayment, I. (2005) Divergent evolution in the enolase superfamily: the interplay of mechanism and specificity, *Arch. Biochem. Biophys. 433*, 59-70.

46. Hasson, M. S., Schlichting, I., Moulai, J., Taylor, K., Barrett, W., Kenyon, G. L., Babbitt, P. C., Gerlt, J. A., Petsko, G. A., and Ringe, D. (1998) Evolution of an enzyme active site: The structure of a new crystal form of muconate lactonizing enzyme compared with mandelate racemase and enolase, *Proc. Natl. Acad. Sci. U.S.A. 95*, 10396-10401.

47. Helin, S., Kahn, P. C., Guha, B. L., Mallows, D. G., and Goldman, A. (1995) The Refined X-Ray Structure of Muconate Lactonizing Enzyme from Pseudomonas-Putida Prs2000 at 1.85 Angstrom Resolution, *J. Mol. Biol. 254*, 918-941.

48. Neidhart, D. J., Howell, P. L., Petsko, G. A., Powers, V. M., Li, R. S., Kenyon, G. L., and Gerlt, J. A. (1991) Mechanism of the Reaction Catalyzed by Mandelate Racemase .2. Crystal-Structure of Mandelate Racemase at 2.5-a Resolution - Identification of the Active-Site and Possible Catalytic Residues, *Biochemistry 30*, 9264-9273.

49. Heine, A., Luz, J. G., Wong, C. H., and Wilson, I. A. (2004) Analysis of the class I aldolase binding site architecture based on the crystal structure of 2-deoxyribose-5-phosphate aldolase at 0.99 angstrom resolution, *J. Mol. Biol. 343*, 1019-1034.

50. Heine, A., DeSantis, G., Luz, J. G., Mitchell, M., Wong, C. H., and Wilson, I. A. (2001) Observation of covalent intermediates in an enzyme mechanism at atomic resolution, *Science 294*, 369-374.

51. Bello, A. M., Poduch, E., Fujihashi, M., Amani, M., Li, Y., Crandall, I., Hui, R., Lee, P. I., Kain, K. C., Pai, E. F., and Kotra, L. P. (2007) A potent, covalent inhibitor of orotidine 5 '-monophosphate decarboxylase with antimalarial activity (vol 50, pg 915, 2007), *J. Med. Chem. 50*, 2541-2541.

52. Bello, A. M., Poduch, E., Liu, Y., Wei, L. H., Crandall, I., Wang, X. Y., Dyanand, C., Kain, K. C., Pai, E. F., and Kotra, L. P. (2008) Structure-activity relationships of C6-uridine derivatives targeting Plasmodia orotidine monophosphate decarboxylase, *J. Med. Chem. 51*, 439-448.

53. Montelione, G. T. (2012) The Protein Structure Initiative: achievements and visions for the future, *F1000 Biol. Rep. 4*, 7.

CHAPTER 5: COMPUTATIONAL DESIGN OF ENZYMES TO CATALYZE AN AROMATIC CLAISEN REARRANGEMENT

5.1 Introduction

This chapter describes the computational design process for a set of enzymes to catalyze the aromatic Claisen rearrangement, shown in Figure 5.1. These designs were generated using the protocol described in Chapter 1, using the SABER-based methodology (see Chapter 2 for details). The goal of this research was to design an enzyme for a reaction without any natural or catalytic antibody precedent, using three different catalytic motifs.

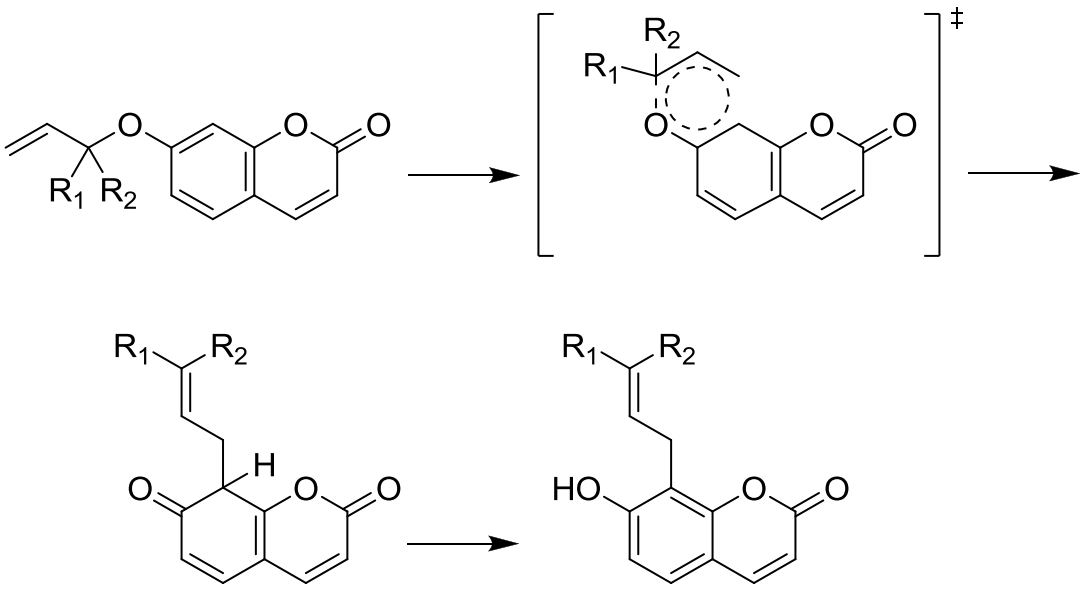


Figure 5.1. The aromatic Claisen rearrangement reaction studied in this research.

Two variants of the substrate were studied. The gem-dimethyl substrate, with methyl groups at positions R1 and R2, was the original choice. This substrate was designed to provide maximum charge separation in the transition state. For reasons that

will be discussed later in this chapter and in Chapter 6, another compound was also studied. This was the proteo substrate, with hydrogens instead of methyl groups at positions R1 and R2. Due to the smaller size of the proteo compound, it was assumed that any enzyme designed to catalyze the rearrangement of the gem-dimethyl substrate would also be able to accommodate the proteo substrate.

5.2 Theozyme Motifs

A number of theozyme motifs were explored to catalyze the aromatic Claisen rearrangement shown in Figure 5.1. The four motifs predicted to provide the largest rate accelerations all used two hydrogen bond donors to stabilize the partial negative charge generated at the ether oxygen in the transition state. These motifs were: Lys-(Asn/Gln), (Asp/Glu)-Tyr, Lys-Tyr, and (Asn/Gln)-Tyr. All theozyme energies were calculated using the PBE1PBE hybrid-GGA functional and the 6-31g(d) basis set. For theozymes, the CPCM implicit solvent model was used, with a dielectric constant of 3.8, to simulate the active site environment. The background (uncatalyzed) reaction rate was calculated in the same manner, except the CPCM dielectric constant was set to 80, to simulate a reaction occurring in water. (*1*) Of these four theozymes, designs meeting the criteria described in Chapter 1 were obtained for all except the (Asn/Gln)-Tyr motif. The theozymes are shown in Figures 5.2-5.4, and the predicted rate accelerations are shown in Table 5.1

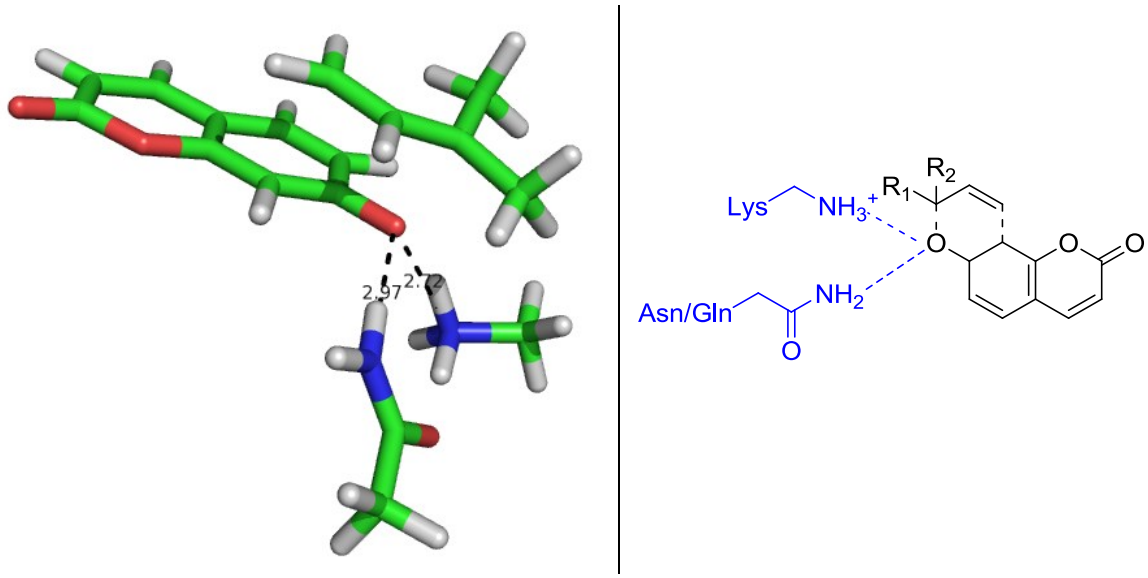


Figure 5.2 The Lys-(Asn/Gln) theozyme for acceleration of an aromatic Claisen rearrangement. (Left) The transition state, with key catalytic distances shown. (Right) Chemdraw representation of the theozyme.

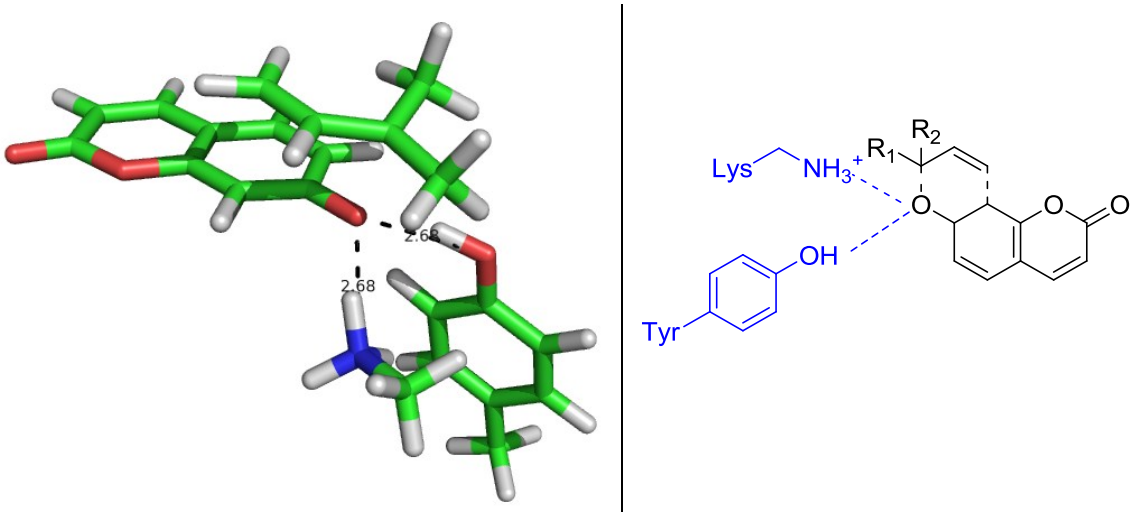


Figure 5.3 The Lys-Tyr theozyme for acceleration of an aromatic Claisen rearrangement. (Left) The transition state, with key catalytic distances shown. (Right) Chemdraw representation of the theozyme.

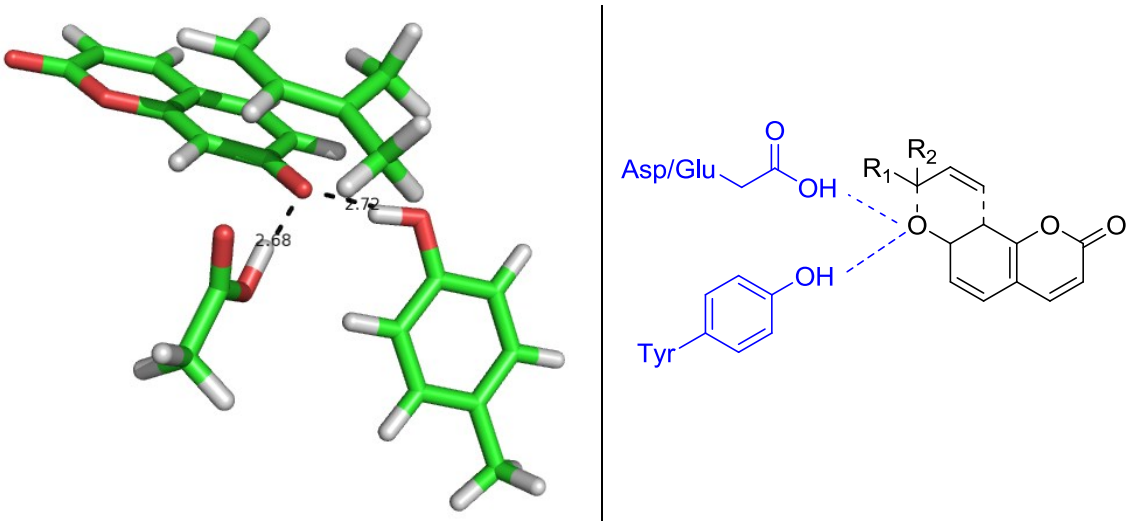


Figure 5.4 The (Asp/Glu)-Tyr theozyme for acceleration of an aromatic Claisen rearrangement. (Left) The transition state, with key catalytic distances shown. (Right) Chemdraw representation of the theozyme.

Table 5.1. Calculated activation free energies and rate enhancements for the top four theozymes designed to accelerate the aromatic Claisen rearrangement being studied. Calculations were performed using PBE1PBE/6-31g(d) and the CPCM implicit solvent model. The dielectric constant was set to 3.8 for theozymes, and 80 for the background reaction.

Theozyme	Predicted $\Delta G^{\ddagger}$ (kcal/mol)[1]	Predicted $\Delta\Delta G^{\ddagger}$ (kcal/mol)[1]	Rate enhancement[1] (vs. background)
Background reaction	27.0	-	-
Lys-Tyr	14.6	12.4	1.2×10^{9}
(Asn/Gln)-Tyr	16.6	10.4	4.0×10^{7}
Lys-(Asn/Gln)	17.1	9.9	1.8×10^{7}
(Asp/Glu)-Tyr	17.3	9.7	1.3×10^{7}

[1]As will be discussed extensively in Chapter 6, there are significant errors in these predictions when compared to experimental values. The $\Delta G^{\ddagger}$ for the background reaction is overestimated, and thus the experimental $\Delta\Delta G^{\ddagger}$ values and rate enhancements were much smaller than predicted values.

5.3 Enzyme Design Methodology and Evaluation Criteria

The enzyme design evaluation procedure described in Chapter 1 was used to rank the designs using each theozyme. Both the SABER design methodology described in Chapter 2 and the standard RosettaMatch-based inside-out methodology were used to produce these designs. The RosettaMatch designs are not discussed in this chapter, as that research was performed by a different member of the group.

In addition to the criteria described in Chapter 1, the EDGE (Enzyme Design Geometry Evaluation) score was used to measure the overall geometric fidelity of each designed active site to the theozyme. (*2*) The EDGE score was calculated using the equation shown in Figure 5.5. The lower the EDGE score, the smaller the deviation from the ideal theozyme. An EDGE score of 0 would indicate an ideal active site in which the theozyme was perfectly replicated. Given the hydrophobic nature of the substrate, hydrophobic residues were often favored during the design process.

$$\text{EDGE Score} = 0.5 \sum_{distances} (\Delta d)^2 + 0.3 \sum_{angles} (\Delta\theta)^2 + 0.3 \sum_{dihedrals} (\Delta\phi)^2$$

Figure 5.5. The formula used to generate EDGE scores for ranking the geometric fidelity of a designed active site to its theozyme.

All of the designs that made it past the first round of screening, which involved both numerical evaluation and empirical evaluation with the Baker group, were subjected to additional rounds of design as described in Chapter 1. Each step of the design process is shown, in order to make the logic of each iteration clear. The reports describing each design are grouped according to the theozyme used. Each report lists the key numeric criteria used to score each design, along with the mutations put in place during the RosettaDesign stage, and the rationale for exploration of other mutation inserting during each design stage. Each first-round design report shows a picture of the designed enzyme's active site, along with the key catalytic distances used in the calculation of the EDGE score.

Of the seven designs that made it past the initial evaluation, three were based on the Lys-Tyr theozyme, two were based on the Lys-(Asn/Gln) theozyme, and two were based on the ketosteriod isomerase inspired (Asp/Glu)-Tyr theozyme. Three of these designs, one from each theozyme, were expressed and tested for activity: 1GD9, 1HYO, and 1E3V. The results of these experiments are described in Chapter 6. The remaining designs were not expressed.

5.4 Design Reports: Lys-Tyr

5.4.1 Scaffold 1GD9, round 1

Reaction: aromatic Claisen rearrangement
Theozyme: Lys-Tyr
Scaffold: 1GD9
Catalytic Residues: Lys232, Tyr201

EDGE score: 0.08
Rosetta energy (design/native): -374.5 / -374.9
Rosetta packing (design/native): 0.569 / 0.577
Ligand score: -1.4

Protein sequence changes:

Residue	*Wild type*	*After design*	*Allowed changes*	*Rationale*
96	N	E	N	Close to dimer interface
121	F	G	AVL	Space for TS
122	V	R	V	Maintain hydrophobicity, remove Arg
123	S	G	AS	Sample possible substitutions
166	N	G	AVLN	Packs against TS – sampling
200	V	E	VL	Maintain hydrophobicity
317	A	G	A	Replace Gly

Active site:

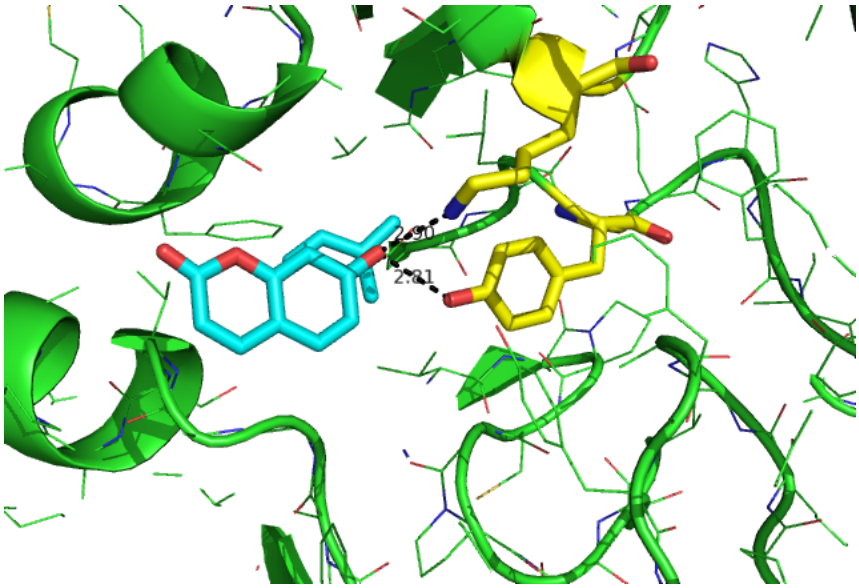


Figure 5.6. The first-round report on the 1GD9-based aromatic Claisen enzyme design.

5.4.2 Scaffold 1GD9, rounds 2-3

Reaction: aromatic Claisen rearrangement
Theozyme: Lys-Tyr
Scaffold: 1GD9
Catalytic Residues: Lys232, Tyr201

EDGE score: 0.08
Rosetta energy (design/native): -318.4 / -374.9
Rosetta packing (design/native): 0.572 / 0.577
Ligand score: -2.9

Protein sequence changes:

Residue	*Wild type*	*After design*	*Allowed changes*	*After design (2)*	*Allowed changes*	*After design (3)*	*Rationale*
33	G	N	AVG	A	-	A	Packs near TS – want to improve packing
96	N	E	N	N	-	N	Close to dimer interface
120	F	G	AVL	A	-	A	Space for TS
121	V	R	V	V	-	V	Maintain hydrophobicity, remove Arg
122	S	G	AS	S	-	S	Sample possible substitutions
166	N	G	N	N	-	N	Distant from TS
170	N	G	AVN	A	-	A	Space for TS
200	V	E	VL	L	V	V	Maintain hydrophobicity
317	A	G	A	A	-	A	Replace Gly

Figure 5.7. The final-round report on the 1GD9-based aromatic Claisen enzyme design.

5.4.3 Scaffold 1QM5, round 1

Reaction: aromatic Claisen rearrangement
Theozyme: Lys-Tyr
Scaffold: 1QM5
Catalytic Residues: Lys539, Tyr538

EDGE score: 0.13
Rosetta energy (design/native): -279.8 / -292.1
Rosetta packing (design/native): 0.485 / 0.494
Ligand score: -0.7

Protein sequence changes:

Residue	*Wild type*	*After design*	*Allowed changes*	*Rationale*
112	N	G	AVL	Space for TS
346	T	S	T	Distant from TS
533	K	G	AVLF	Space for TS
534	R	G	A	Replace Gly
536	H	A	AV	Room for TS
638	A	S	AVS	Contacts TS, doing more sampling

Active site:

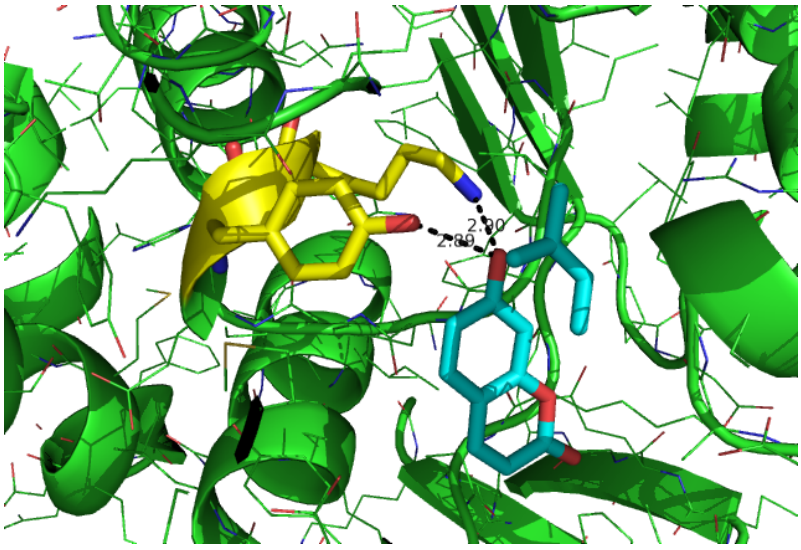


Figure 5.8. The first-round report on the 1QM5-based aromatic Claisen enzyme design.

5.4.4 Scaffold 1YIY, round 1

Reaction: aromatic Claisen rearrangement
Theozyme: Lys-Tyr
Scaffold: 1YIY
Catalytic Residues: Lys244, Tyr213

EDGE score: 0.05
Rosetta energy (design/native): 527.2 / 562.3 (seems bugged, still a native-like score)
Rosetta packing (design/native): 0.483 / 0.487
Ligand score: -2.5

Protein sequence changes:

Residue	*Wild type*	*After design*	*Allowed changes*	*Rationale*
99	A	S	AVL	Space for TS
100	Y	L	Y	Close to dimer interface
182	N	D	N	Distant from TS
212	V	S	AVL	Maintain hydrophobicity
241	S	D	S	Close to dimer interface

Active site:

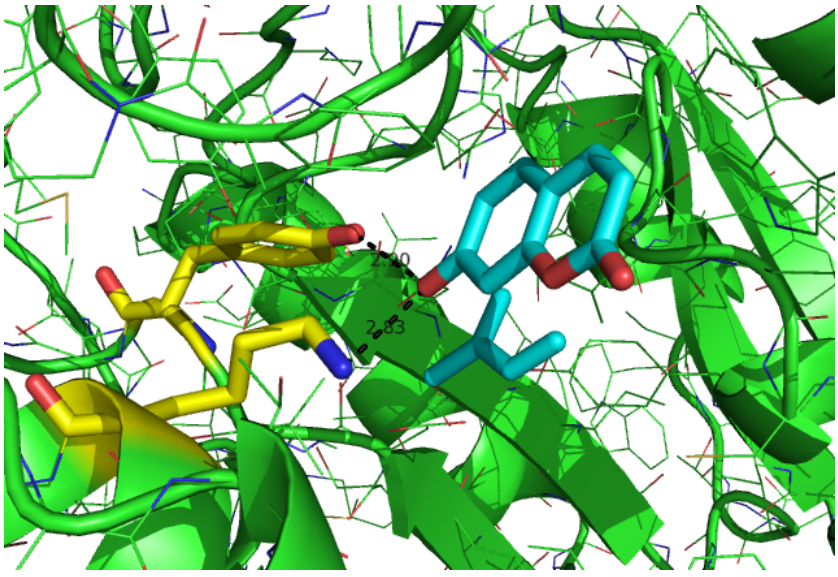


Figure 5.9. The first-round report on the 1YIY-based aromatic Claisen enzyme design.

5.5 Design Reports: Lys-(Asn/Gln)

5.5.1 Scaffold 1HYO, round 1

Reaction: aromatic Claisen rearrangement
Theozyme: Lys-(Asn/Gln)
Scaffold: 1HYO
Catalytic Residues: Lys255, Asn242

EDGE score: 0.04
Rosetta energy (design/native): -499.8 / -634.0
Rosetta packing (design/native): 0.499 / 0.507
Ligand score: -3.3

Protein sequence changes:

Residue	*Wild type*	*After design*	*Allowed changes*	*Rationale*
130	Y	G	Y	Distant from TS
139	V	A	A	Room for TS
142	M	G	VILM	Space for TS, some designs kept Met
143	F	R	AVL	Space for TS
201	E	A	ADE	Space for TS, some designs kept D/E
352	T	G	A	Space for TS
365	L	G	A	Space for TS

Active site:

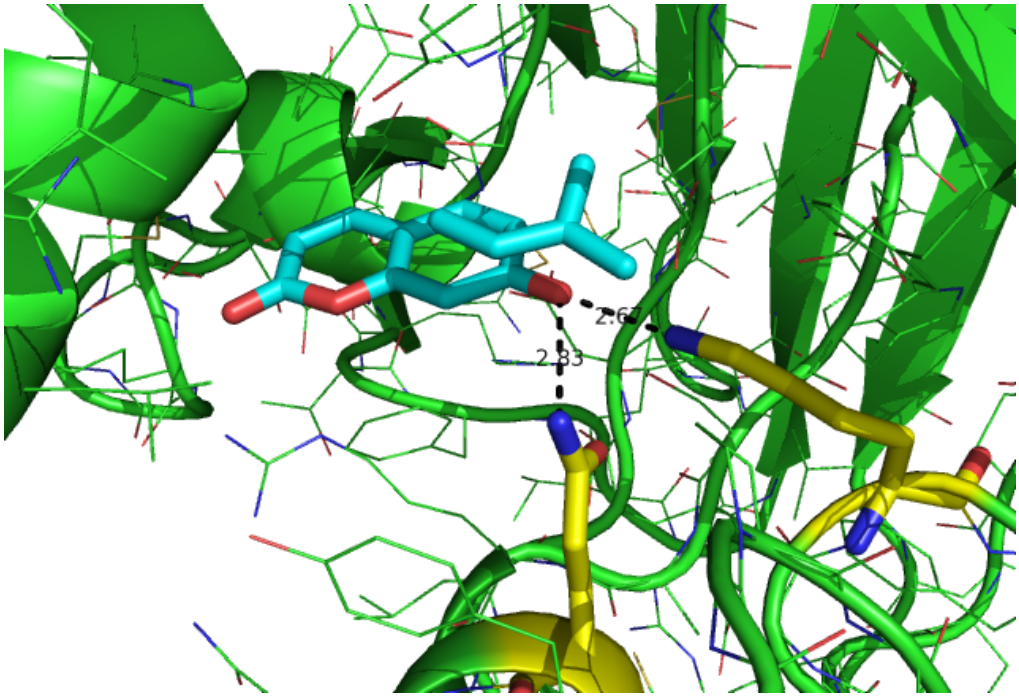


Figure 5.10. The first-round report on the 1HYO-based aromatic Claisen enzyme design.

5.5.2 Scaffold 1HYO, rounds 2-3

Reaction: aromatic Claisen rearrangement
Theozyme: Lys-(Asn/Gln)
Scaffold: 1HYO
Catalytic Residues: Lys255, Asn242

EDGE score: 0.04
Rosetta energy (design/native): -499.8 / -634.0
Rosetta packing (design/native): 0.499 / 0.507
Ligand score: -3.3

Protein sequence changes:

Residue	*Wild type*	*After design*	*Allowed changes*	*After design (2)*	*Allowed changes*	*After design (3)*	*Rationale*
130	Y	G	Y	Y	-	Y	Distant from TS
139	V	A	A	A	-	A	Room for TS
142	M	G	VILM	M	-	M	Space for TS, some designs kept Met
143	F	R	AVL	L	-	L	Space for TS
201	E	A	ADE	E	-	E	Space for TS, some designs kept D/E
352	T	G	A	A	-	A	Space for TS
365	L	G	A	A	-	A	Space for TS

Figure 5.11. The final-round report on the 1HYO-based aromatic Claisen enzyme design.

5.5.3 Scaffold 1MNS, round 1

Reaction: aromatic Claisen rearrangement
Theozyme: Lys-(Asn/Gln)
Scaffold: 1MNS
Catalytic Residues: Lys164, Asn195

EDGE score: 0.03
Rosetta energy (design/native): -579.9 / -584.5
Rosetta packing (design/native): 0.491 / 0.501
Ligand score: -1.7

Protein sequence changes:

Residue	*Wild type*	*After design*	*Allowed changes*	*Rationale*
50	F	V	V	Space for TS
52	Y	Q	YF	Maintain hydrophobicity
193	D	A	A	Space for TS
219	E	A	AV	Space for TS
220	E	H	H	Good packing and H-bonds
246	N	H	NH	More sampling of this position
295	H	D	D	Space for TS
296	L	S	AV	Replace with small hydrophobic

Active site:

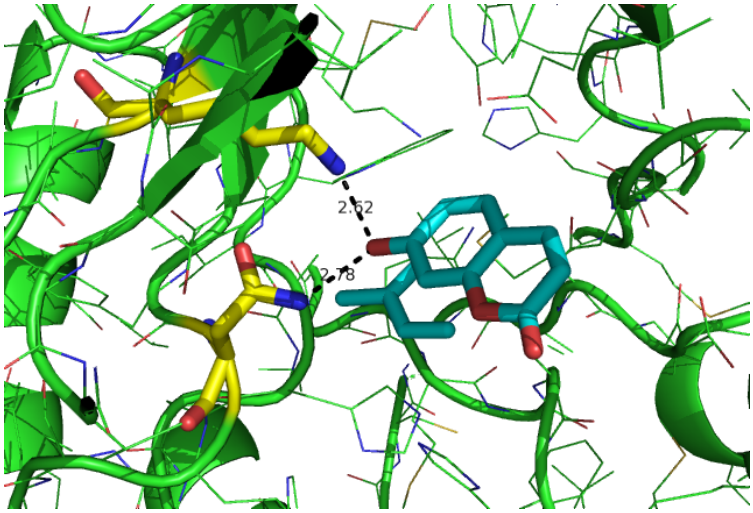


Figure 5.12. The first-round report on the 1MNS-based aromatic Claisen enzyme design.

5.6 Design Reports: (Asp/Glu)-Tyr

5.6.1 Scaffold 1DED, round 1

Reaction: aromatic Claisen rearrangement
Theozyme: (Asp/Glu)-Tyr
Scaffold: 1DED
Catalytic Residues: Asp229, Tyr100

EDGE score: 0.08
Rosetta energy (design/native): -405.3 / -498.0
Rosetta packing (design/native): 0.491 / 0.496
Ligand score: -0.6

Protein sequence changes:

Residue	*Wild type*	*After design*	*Allowed changes*	*Rationale*
98	H	G	A	Space for TS
194	L	D	AVL	Maintain hydrophobicity of active site
195	Y	G	AVL	Space for TS
230	A	S	A	Maintain hydrophobicity of active site
257	E	G	A	Space for TS
326	N	D	ND	Either residue can provide interaction
327	H	G	AV	Space for TS
328	D	S	AVS	Space for TS

Active site:

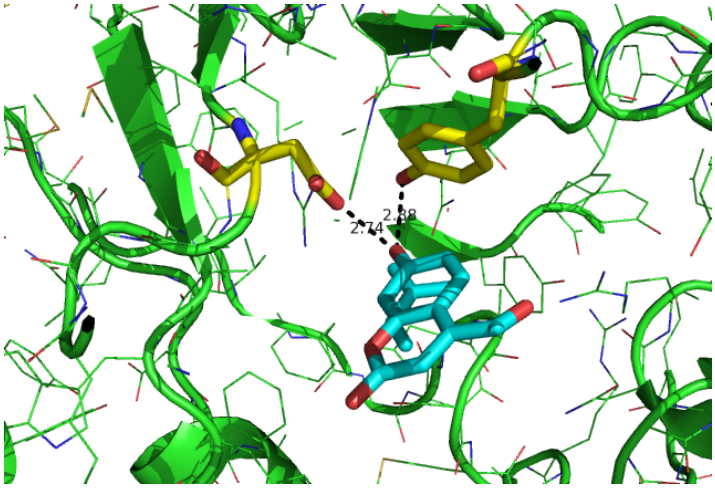


Figure 5.13. The first-round report on the 1DED-based aromatic Claisen enzyme design.

5.6.2 Scaffold 1E3V, round 1

Reaction: aromatic Claisen rearrangement
Theozyme: (Asp/Glu)-Tyr
Scaffold: 1E3V
Catalytic Residues: Asp102, Tyr15

EDGE score: 0.03
Rosetta energy (design/native): -146.4 / -165.8
Rosetta packing (design/native): 0.436 / 0.491
Ligand score: -2.8

Protein sequence changes:

Residue	*Wild type*	*After design*	*Allowed changes*	*Rationale*
55	F	G	F	Distant from TS
85	F	A	A	Space for TS
87	V	R	VLI	Space filling, replacing Arg
98	L	G	L	Distant from TS
100	V	G	A	Replacing Gly
117	A	S	AV	Replace with small hydrophobic
119	W	G	A	Space for TS

Active site:

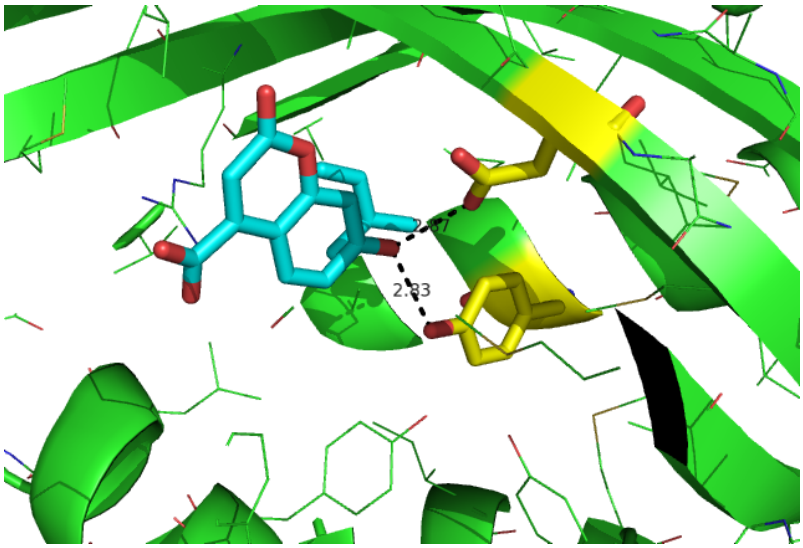


Figure 5.14. The first-round report on the 1E3V-based aromatic Claisen enzyme design.

5.6.3 Scaffold 1E3V, rounds 2-3

Reaction: aromatic Claisen rearrangement
Theozyme: (Asp/Glu)-Tyr
Scaffold: 1E3V
Catalytic Residues: Asp102, Tyr15

EDGE score: 0.03
Rosetta energy (design/native): -146.4 / -165.8
Rosetta packing (design/native): 0.436 / 0.491
Ligand score: -2.8

Protein sequence changes:

Residue	*Wild type*	*After design*	*Allowed changes*	*After design (2)*	*Allowed changes*	*After design (3)*	*Rationale*
55	F	G	F	F	-	F	Distant from TS
85	F	A	A	A	-	A	Space for TS
87	V	R	VLI	V	-	V	Space filling, replacing Arg
98	L	G	L	L	-	L	Distant from TS
100	V	G	A	A	-	A	Replacing Gly
117	A	S	AV	V	-	V	Replace with small hydrophobic
119	W	G	A	A	-	A	Space for TS

Figure 5.15. The final-round report on the 1E3V-based aromatic Claisen enzyme design.

5.7 References

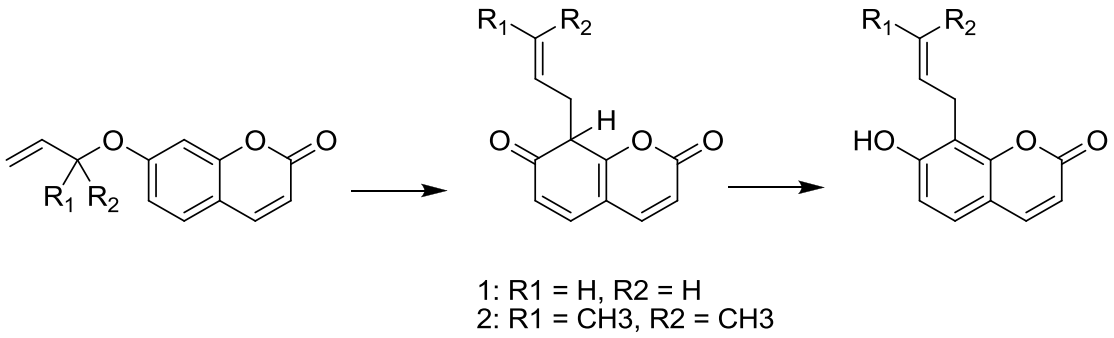
1. Bollot, G., and Houk, K. N. Unpublished results.

2. Johnson, S. A., and Houk, K. N. Unpublished results.

CHAPTER 6: EXPERIMENTAL CHARACTERIZATION OF AN UNCATALYZED AND ENZYME-CATALYZED AROMATIC CLAISEN REARRANGEMENT

6.1 Introduction

This chapter is a companion to Chapter 5, which detailed a series of enzymes designed to catalyze an aromatic Claisen rearrangement, shown in Figure 6.1. These enzymes were designed using the SABER protocol described in Chapter 2 (*1*), and scored according the criteria described in Chapters 1 and 5. Here, the experimental characterization of the background and enzyme-catalyzed reactions will be discussed in detail, with a focus on the lessons learned regarding the design of substrates for use with hydrogen bonding catalysts in aqueous solution.



1: R1 = H, R2 = H
2: R1 = CH3, R2 = CH3

Figure 6.1. The aromatic Claisen rearrangement studied during this research. Compound 1 is also referred to as the proteo compound, and compound 2 is referred to as the gem-dimethyl compound.

As stated in Chapter 5, of the seven designs produced using the SABER protocol, three were produced as synthetic genes, cloned into plasmids, expressed in *E. coli*, purified, and tested for activity. The background (uncatalyzed) reaction was also studied experimentally. Based on the quantum mechanical calculations shown in Chapter 5, the predicted free energy of activation for the background reaction in water

was 27.0 kcal/mol. (*2*) As will be discussed in Section 6.3, this prediction significantly overestimates the free energy of activation for the background reaction.

6.2 Materials and Methods

Both the gem-dimethyl and proteo substrates and products were analyzed using liquid chromatography/mass spectrometry (LC/MS). Solutions were prepared under the desired experimental conditions, and aliquots were taken for analysis directly from the stock solution using the HPLC's autoinjector. The HPLC used in these experiments was an Agilent 1100 with a thermally-controlled well-plate autosampler, which allowed for strict control of the temperature. The MS used in these experiments was a Finnigan LCQ with an ESI probe for ionization. These experiments were performed in the Baker laboratory at the University of Washington, in collaboration with Dr. Lucas Nivon, a research fellow in the Baker group.

HPLC conditions: Samples were analyzed using a Hypersil Gold C18 reverse-phase column, 2.1 x 100 mm. Column oven temperature was set to 40° C. The chromatography was performed under isocratic conditions, using a 50:50 mixture of acetonitrile and buffer (water + 0.1% formic acid). The flow rate was 0.2 mL/min. for 10 minutes, at which point the column was allowed to equilibrate for 1 minute, and then the next injection was performed. Injection volume was 20 μL, with a syringe draw speed of

83 μL/min. The UV photodiode array detector on the HPLC was set to measure absorbance at 274 nm and 330 nm.

MS conditions: TIC data was acquired for the full 10-minute experimental runtime. Data was acquired in a m/z range of 200-500 for standard assays and 100-500 for determination of mass spectra of standards.

Synthesis of aromatic Claisen substrates and products: These materials were synthesized by Roger Helgeson, a scientist in the Houk research group at the University of California, Los Angeles.

6.3 Results: Uncatalyzed Reaction

During the experimental analysis of the enzyme-catalyzed reactions, it became apparent that the background reaction of the gem-dimethyl compound, used as a control, was occurring much more rapidly than predicted. The substrate appeared to have a half-life of approximately 30-45 minutes in aqueous solution. Due to this observation, a number of experiments were performed to determine the kinetics of the background reaction. A number of hypotheses were explored to investigate the unexpectedly fast rate of reaction for the gem-dimethyl compound. These included the effect of substrate concentration on the background reaction rate, the effect of the choice of buffer on reaction rate, and the effect of temperature on the reaction rate.

These studies were performed using LC/MS, as described in the Materials and Methods section.

New batches of the proteo compound and the gem-dimethyl compound were prepared for this study, and their structures were confirmed by NMR, along with authentic product standards. We went so far as to investigate the effect of carrying out the reactions in glass vs. plastic tubes, but this was found to have no effect on the rate of reaction. The mass spectra of the gem-dimethyl substrate and product are shown in Figures 6.2 and 6.3, respectively. A typical HPLC trace showing the gem-dimethyl substrate and product is shown in Figure 6.4.

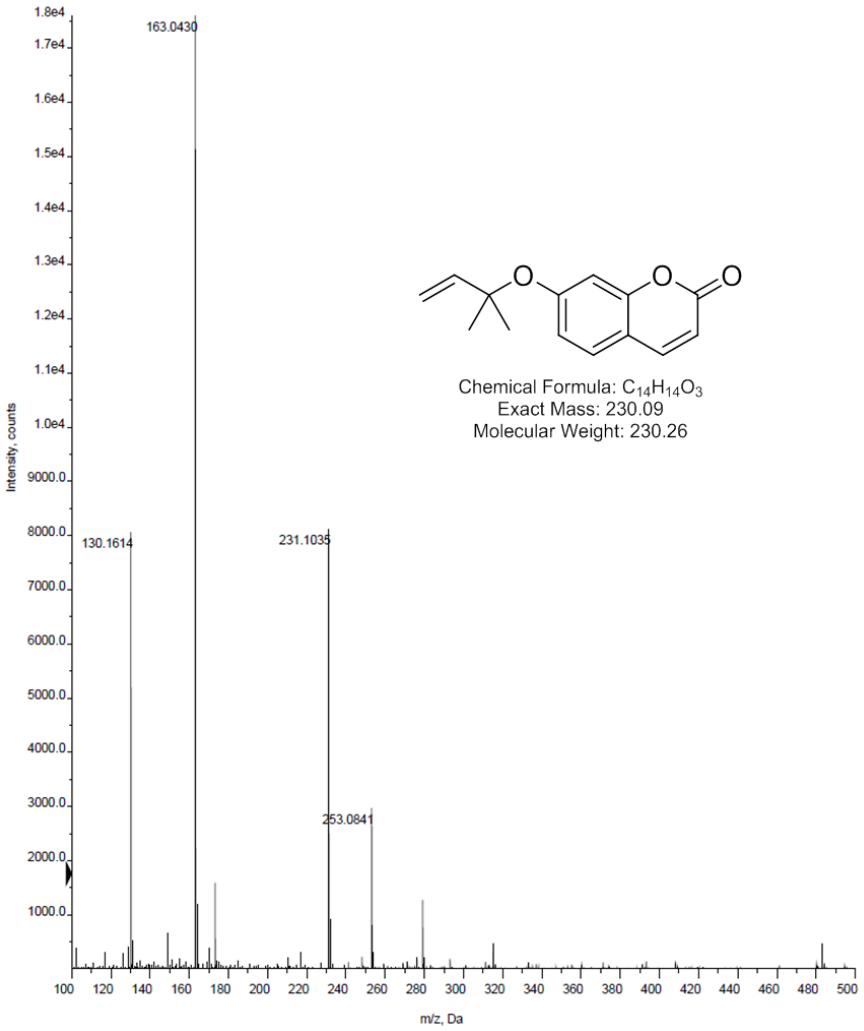


Figure 6.2 Mass spectrum of the gem-dimethyl substrate.

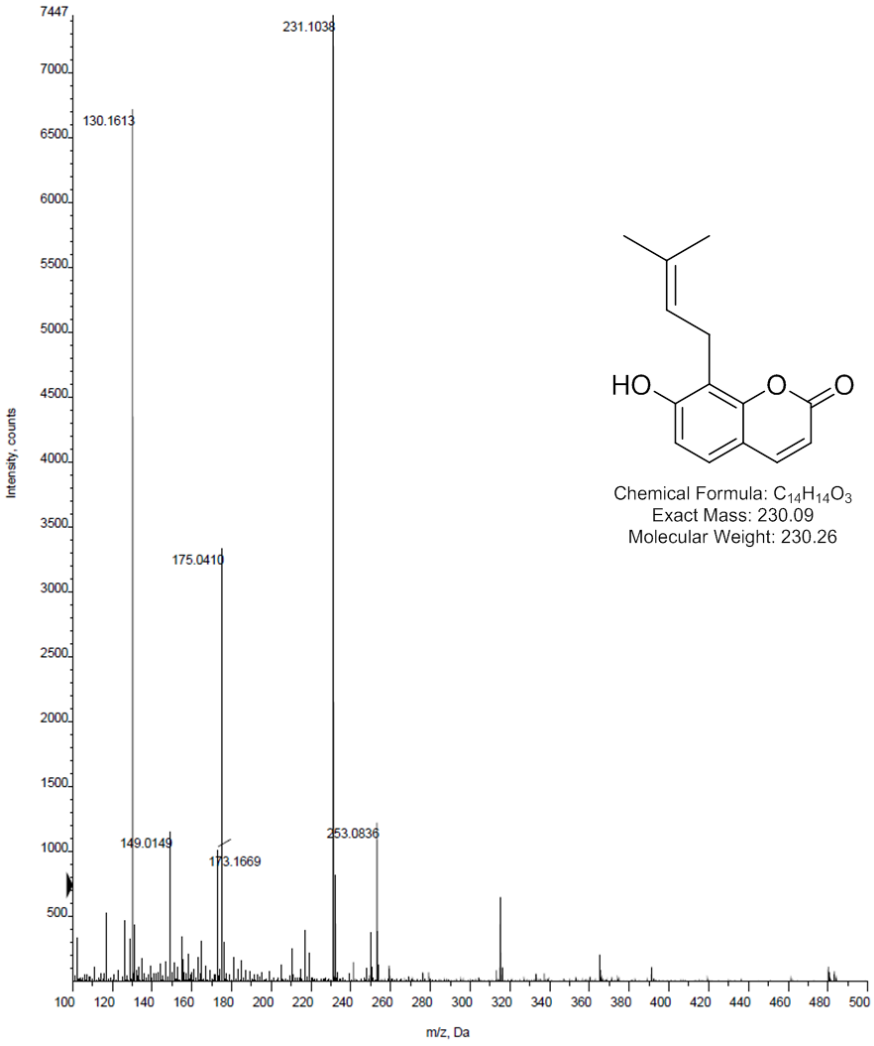


Figure 6.3 Mass spectrum of the gem-dimethyl product.

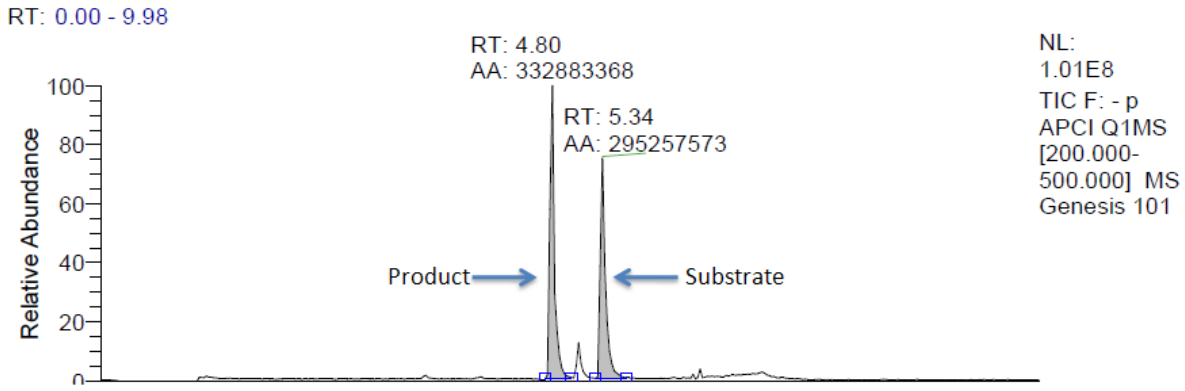


Figure 6.4 HPLC trace showing retention times and resolution of the gem-dimethyl substrate and product for the aromatic Claisen reaction being studied.

6.3.1 Linearity of response for the gem-dimethyl compound

As shown in Figure 6.5, the peak area response in linear within the range of 5-100 μM substrate concentrations. This was determined using both unfiltered solutions and solutions passed through a 0.45 μM filter. This was done to remove any substrate that had not dissolved and might stick to the HPLC syringe during injection, leading to contamination of subsequent injections. As the reaction had been observed to be very rapid, solutions were prepared and analyzed immediately via LC/MS. In both cases, the response was linear within the range studied. The linearity of response of the product was also studied in the same fashion, so as to rule out the possibility of product precipitation as a result of different solubility affecting the rate of reaction. These data are shown in Figure 6.6.

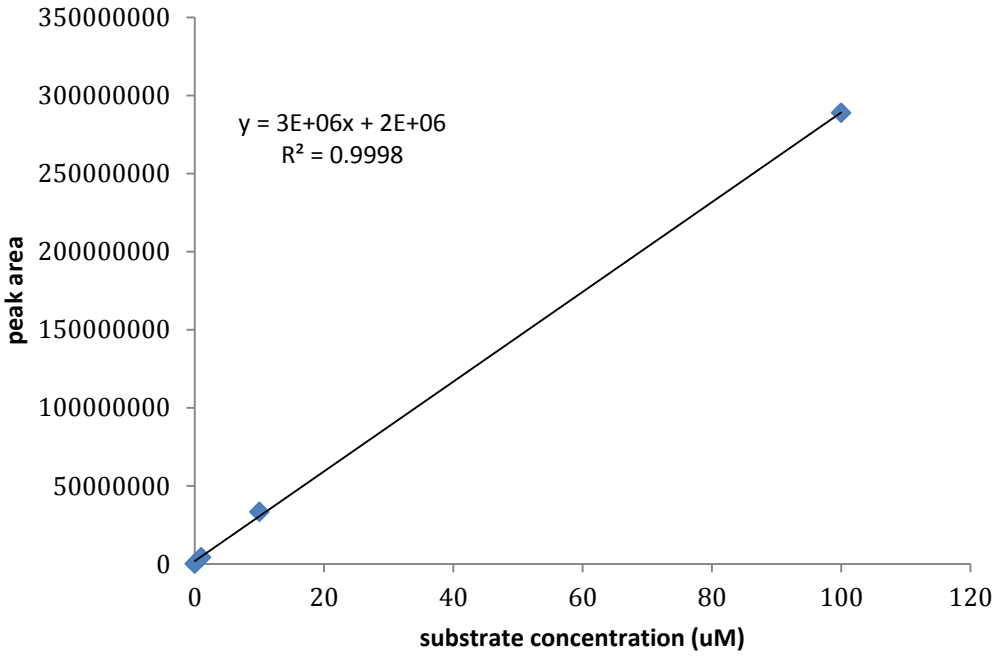


Figure 6.5. Peak area vs. concentration for the gem-dimethyl substrate.

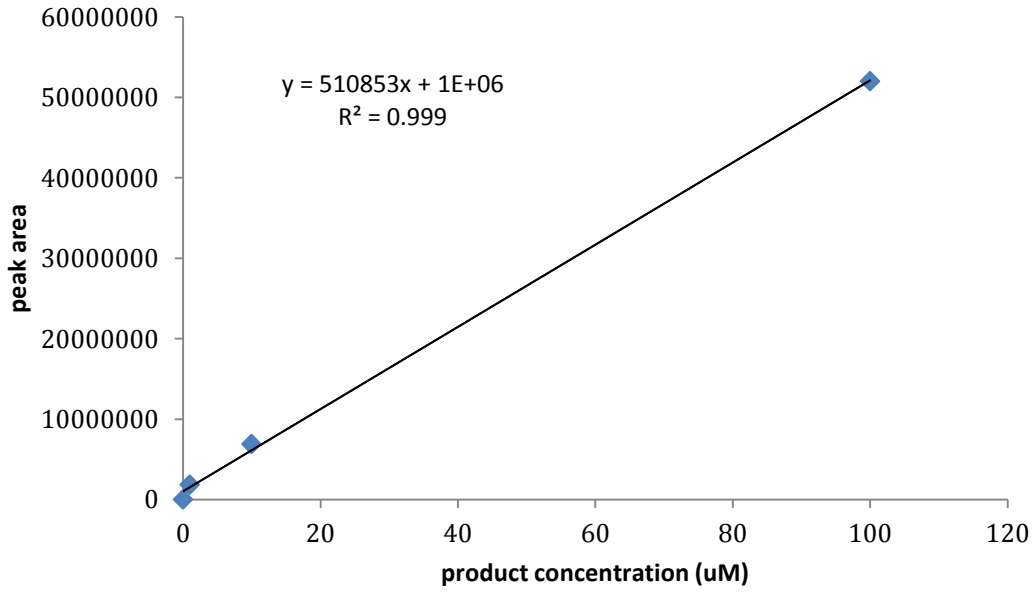


Figure 6.6. Peak area vs. concentration for the gem-dimethyl product.

6.3.2 Effect of substrate concentration on the rate of background reaction

Next, we investigated the effect of substrate concentration on the reaction rate in pure water. The reaction was studied using a timecourse, with measurements taken every 10 minutes for 90 minutes for the gem-dimethyl compound. The proteo compound was also studied in this manner, but due to its very slow rate of reaction, measurements were taken at the initial timepoint, after, 24 hours, and then weekly for 30 days. All samples were analyzed using LC/MS and carried out at 25º C. Concentrations studied were 500 nM, 5 μM, and 50 μM for both compounds. After the experiments were complete, half-life values and $\Delta G^{\ddagger}$ values were calculated for each reaction condition. The results for the gem-dimethyl compound are shown in Table 6.1, and the results for the proteo compound are shown in Table 6.2. Neither compound shows significant variation in half-life within the concentration range studied in these experiments.

Table 6.1. Analysis of reaction rate dependence on concentration for the gem-dimethyl substrate.

Experimental conditions	$\Delta G^{\ddagger}$ (kcal/mol)	Half-life
50 μM, H_2O	22.2	36 min.
5 μM, H_2O	22.2	38 min.
500 μM, H_2O	22.3	42 min.

Table 6.2. Analysis of reaction rate dependence on concentration for the proteo substrate.

Experimental conditions	$\Delta G^{\ddagger}$ (kcal/mol)	Half-life
50 μM, H_2O	> 26	> 30 days
5 μM, H_2O	> 26	> 30 days
500 μM, H_2O	> 26	> 30 days

6.3.3 Effect of solution choice on the rate of background reaction

The choice of solution and its impact on the background reaction rate was also studied for both the gem-dimethyl and proteo compounds. These studies were carried out at 25º C at a substrate concentration of 50 μM. Solution conditions were pure water, phosphate-buffered saline (PBS), pure acetonitrile (ACN), and 50:50 acetonitrile:water. The same timecourse was used as described in the experiments used to study the reaction rate under differing substrate concentrations, except it was extended to 70 days due to the slow reaction rate of the proteo compound. The results of these experiments are shown in Table 6.3.

Table 6.3. The effect of solvent conditions on the uncatalyzed reaction rate for the aromatic Claisen rearrangement being studied.

Compound	Solvent	Conc. (μM)	T (°C)	$\Delta G^{\ddagger}$ (kcal/mol)	Half-life
1	ACN	50	25	> 26.9	> 70 d
1	water	50	25	> 26.9	> 70 d
1	PBS	50	25	> 26.9	> 70 d
2	ACN	50	25	> 26.4	> 30 d
2	water	50	25	22.2	38 m
2	PBS	50	25	22.2	38 m
2	50:50 ACN:H_2O	50	25	25.4	128 h

6.3.4 Effect of temperature on the rate of background reaction

The temperature-dependence of the background reaction rate was also studied. The short half-life of the gem-dimethyl compound in aqueous solution made assays difficult, since it was necessary to immediately begin analysis of samples upon addition of enzyme due to the rapid uncatalyzed conversion of substrate to product. The use of an HPLC autosampler with temperature control allowed the reaction to progress at a fixed temperature of 4° C. The reaction was monitored for 48 hours, with timepoints taken every hour for the first four hours, then every two hours until the timecourse was completed. The results of this experiment are summarized in Table 6.4 (data from the comparable experiment at 25°C is shown for reference).

Table 6.4. The effect of temperature on the uncatalyzed reaction rate for the aromatic Claisen rearrangement being studied.

Compound	Solvent	Conc. (μM)	T (°C)	$\Delta G^{\ddagger}$ (kcal/mol)	Half-life
2	water	50	25	22.2	38 m
2	water	50	4	24.4	24 h

6.4 Results: Enzyme-catalyzed Reaction

Based on the results of the analysis of the background reaction, the designed enzymes were studied using both the gem-dimethyl substrate and the proteo substrate. While the gem-dimethyl compound was likely to be a better candidate for enzyme catalysis, its rapid reaction rate made assays difficult, while the proteo compound was very stable. All three of the designed enzymes were analyzed for activity using the LC/MS assay at both 25° C and 4 °C. The designed retro-aldol enzymes RA45 and RA93 were used as controls. The results of these experiments are summarized in Table 6.5 (25° C) and Table 6.6 (4° C). For substrate 1, samples were assayed after 24 hours, while substrate 2 was assayed every hour for 4 hours.

Table 6.5. Activity analysis of enzymes designed to catalyze the aromatic Claisen rearrangement at 25° C.

Substrate	Enzyme	Temperature (°C)	k_{cat}/k_{uncat}
1	ACR_1GD9	25	1
1	ACR_1HYO	25	1

1	ACR_1E3V	25	1
1	RA45	25	1
1	RA93	25	1
2	ACR_1GD9	25	1
2	ACR_1HYO	25	1
2	ACR_1E3V	25	1
2	RA45	25	1
2	RA93	25	1

Table 6.6. Activity analysis of enzymes designed to catalyze the aromatic Claisen rearrangement at 4° C.

Substrate	Enzyme	Temperature (°C)	k_{cat}/k_{uncat}
1	ACR_1GD9	4	1
1	ACR_1HYO	4	1
1	ACR_1E3V	4	1
1	RA45	4	1
1	RA93	4	1
2	ACR_1GD9	4	1
2	ACR_1HYO	4	1
2	ACR_1E3V	4	1
2	RA45	4	1
2	RA93	4	1

6.5 Conclusions

While none of the enzymes displayed any rate acceleration versus the background reaction, this is not a surprising result. Previous enzyme design efforts have required tens to hundreds of designs to be synthesized in order to find one that has catalytic activity. (*3-5*) This serves to highlight the extreme difficulty of designing a functional enzyme, even with all of the tools that are available.

This work also highlights one of the dangers of simulating hydrogen-bonding based catalysis using an implicit solvent model to represent water. Since water is an excellent hydrogen bonding molecule, it is more appropriate to study this reaction using explicit water molecules when calculating the background rate of reaction. When quantum mechanical calculations are performed on a transition state like the one shown in Figure 6.7, a much more realistic free energy of activation can be obtained.

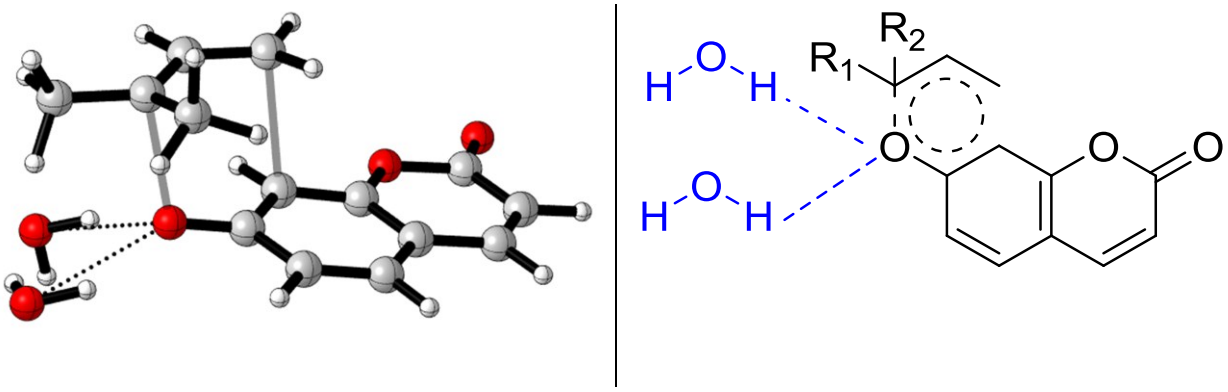


Figure 6.7. (Left) A quantum mechanical model of the gem-dimethyl transition state being stabilized by hydrogen bonds from explicit water molecules. (Right) A Chemdraw representation of this theozyme.

The calculated free energy of activation for the gem-dimethyl compound in the presence and absence of explicit water molecules was studied using B3LYP/6-311++G(d,p) and M06-2X/6-31G(d,p), in both the gas phase and using the CPCM implicit solvent model for water. (*6, 7*) The results of these calculations are shown in Table 6.7.

Table 6.7. Quantum mechanical calculations of the free energy of activation for the gem-dimethyl compound undergoing an aromatic Claisen rearrangement in the presence and absence of explicit water molecules.

Explicit water molecules	$\Delta G^{\ddagger}$ (kcal/mol) Gas phase		$\Delta G^{\ddagger}$ (kcal/mol) CPCM (water)	
	B3LYP/ 6-311++G(d,p)	M06-2X/ 6-31G(d,p)	B3LYP/ 6-311++G(d,p)	M06-2X/ 6-31G(d,p)
0	25.1	32.1	21.7	29.2
1	19.7	27.2	17.0	25.7
2	15.9	24.4	14.2	23.8

The M062X/6-31g(d,p) calculations using two explicit water molecules and the CPCM implicit solvent model for water provided the closest agreement with the experimental results. The average experimental value of $\Delta G^{\ddagger}$ was 22.2 kcal/mol for the gem-dimethyl compound, in reasonable agreement with the calculated value of 23.8 kcal/mol using this method. The calculated activation free energy using this method and two explicit water molecules is 5.4 kcal/mol lower in energy than the activation free

energy calculated in the absence of explicit water molecules. This clearly demonstrates the importance of modeling explicit solvent molecules when calculating background reaction rates of reactions that can be catalyzed by hydrogen bonding.

6.6 References

1. Nosrati, G. R., and Houk, K. N. (2012) SABER: A computational method for identifying active sites for new reactions, *Protein Sci. 21*, 697-706.

2. Bollot, G., and Houk, K. N. Unpublished results.

3. Rothlisberger, 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., and Baker, D. (2008) Kemp elimination catalysts by computational enzyme design, *Nature 453*, 190-195.

4. Jiang, L., Althoff, E. A., Clemente, F. R., Doyle, L., Rothlisberger, D., Zanghellini, A., Gallaher, J. L., Betker, J. L., Tanaka, F., Barbas, C. F., 3rd, Hilvert, D., Houk, K. N., Stoddard, B. L., and Baker, D. (2008) De novo computational design of retro-aldol enzymes, *Science 319*, 1387-1391.

5. Siegel, J. B., Zanghellini, A., Lovick, H. M., Kiss, G., Lambert, A. R., St Clair, J. L., Gallaher, J. L., Hilvert, D., Gelb, M. H., Stoddard, B. L., Houk, K. N., Michael, F. E., and Baker, D. (2010) Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction, *Science 329*, 309-313.

6. Kim, S., and Houk, K. N. Unpublished results.

7. Osuna, S., and Houk, K. N. Unpublished results.