Computational Enzyme Design

Thesis by
Bernardo Sosa Padilla Araujo

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

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California
2014
(Defended May 1, 2014)

Abstract

Computational protein design (CPD) is the automated identification of amino acid sequences that will fold into a specified three-dimensional structure. This method has emerged as a promising tool for engineering enzymes. In this thesis, I describe my efforts at improving and applying CPD methods to the design of enzymes.

Chapter II describes the development and benchmark results for a molecular dynamics (MD) protocol to prescreen enzyme designs. Results indicate that the MD protocol is successful in screening enzymes for enzymatic activity. The protocol is general, reproducible, and excels at discarding false positive designs while predicting the most active enzymes correctly.

Conformational changes are part of the repertoire that natural enzymes use to catalyze reactions. Chapter III comprises the computational design and experimental characterization of triosephosphate isomerase (TIM), a model enzyme for the study of catalytic activity in the context of conformational changes. Using a novel multi-state design (MSD) method that can consider multiple states in a single protein sequence optimization calculation, we designed the flexible hinges in TIM's active site loop.

A central challenge in the CPD field is the reliable engineering of an enzyme for any desired reaction. In Chapter IV, I describe the conceptualization and application of a high-throughput computational framework for the *de novo* design of enzymatic activity into inert scaffolds. TIM is used as a model system.

TABLE OF CONTENTS

TABLES AND FIGURES

ABBREVIATIONS

AA	amino acid
CA	catalytic antibodies
CD	circular dichroism
CM	chorismate mutase
CPD	computational protein design
CPU	central processing unit
d-GAP	d-glyceraldehyde-3-phosphate
DC	degenerate codon
DHAP	dihydroxyacetone phosphate
E. coli	*Escherichia coli*
E_{closed}	energy of the closed state
EDTA	ethylenediaminetetraacetic acid
E_{open}	energy of the open state
E_{total}	total energy
EVB	empirical valence bond
fs	femtoseconds
GPDH	α-glycerolphosphate dehydrogenase
GPU	graphics processing unit
k_{cat}	catalytic constant
KE	Kemp elimination
K_M	Michaelis constant
MD	molecular dynamics
MSA	multiple sequence alignment

MSD	multi-state design
NA	not applicable
NADH	reduced nicotinamide adenine dinucleotide
NBZ	5-nitrobenzoxizole
ND	not determined
NMR	nuclear magnetic resonance
NPV	negative predictive value
NPT	constant number of particles, pressure, and temperature
ns	nanoseconds
PBC	periodic boundary conditions
PDB	protein data bank
PME	Particle mesh Ewald
PPV	positive predictive value
ps	picoseconds
RESP	restrained electrostatic potential
RFU	relative florescence units
SDS-PAGE	sodium dodecyl sulfate polyacrylamide gel electrophoresis
SEC	size-exclusion chromatography
SSD	single-state design
TEA	triethanolamine
TEV	tobacco etch virus
TIM	triosephosphate isomerase
T_m	midpoint of thermal denaturation curve
TS	transition state
TSA	transition state analog

V_{after}	reaction rate after sample addition
V_{before}	reaction rate before sample addition
VMD	visual molecular dynamics
V_{obs}	observed reaction rate
w	weighting coefficient in a multi-state design calculation
WT	wild type
Θ	mean residue molar ellipticity

Chapter I

Introduction

Enzymes Are Highly Efficient Catalysts

Life as we know it would not be possible without the existence of enzymes. Enzymes accelerate the rate of biological reactions, which would otherwise not proceed fast enough to sustain life.[1] They are responsible for most of the chemical transformations that occur within a cell. DNA replication, photosynthesis, and glycolysis are some of the innumerable examples of the processes that involve enzymes. Although the origin of the catalytic power of enzymes is not well understood and is still a topic of debate, enzymes work by lowering the activation energy of the reaction.[2; 3]

Even though synthetic catalysts can accelerate chemical transformations to a great extent, enzymes have several advantages over them. Natural enzymes can catalyze very diverse chemical reactions. Most are highly specific for their substrate and can be regio-selective. They are also relatively easy to construct and are biodegradable. More importantly, they can work at normal conditions of pH, temperature, and pressure. Simply put, "enzyme catalysis is not different from chemical catalysis, just better."[4] A priori, these properties make enzymes preferable to other types of catalysts for a broad range of applications.

One limitation of the applicability of enzymes to many medical and industrial problems is the fact that the natural repertoire is not exhaustive; thus, natural enzymes simply do not exist to catalyze many potentially interesting reactions. In addition, our understanding of the physical principles that govern the structure-function relationship of enzymes and proteins in general is limited, impeding the reliable construction of an enzyme for any desired reaction.

Computational Design of Enzymes

Computational protein design (CPD) is the automated identification of amino acid sequences that will fold into a specific three-dimensional structure.[5] CPD methods have been extensively used for the design of proteins for improved stability, solubility, protein-protein binding, novel scaffolds, etc.[6-9] Recent successes in designing novel enzymes have prompted considerable interest in the field.[10-14] To date, designer enzymes have outperformed previously reported catalytic antibodies.[15] The activities of *de novo* designed enzymes are typically lower than those of their natural counterparts; however, *in vitro* evolution methods can be used to improve a moderately active designer enzyme to yield highly active biocatalysts.[16; 17] Despite these successes, it is important to note that each successful example of the application of CPD methods is accompanied by a vast number of false positive results, suggesting a lack of understanding of the physical principles that govern enzyme structure and function.[18] In most cases, a large number of enzyme designs must be tested to find a few active designs, yielding a low signal to noise ratio. The ability to consistently design an enzyme for any desired reaction remains the "holy grail" of the field.[17-19]

In this context, the contributions of my work can be condensed down to three different studies; these are outlined here and are described in more depth in the next chapters.

In silico screening of enzymatic activity

Privett *et al.*[13] reported an iterative approach to computational enzyme design. This approach included the evaluation of initially inactive computational enzyme designs to determine the causes of inactivity. Among other techniques, molecular dynamics (MD) simulations were used to determine structural flaws in the designed enzymes. With this information, changes were made in the design strategy that yielded active catalysts for the Kemp elimination (KE) reaction. This result suggested that MD simulations could serve as a pre-screening step in the CPD framework.

In Chapter II of this thesis, we generalize the idea of using MD simulations for screening enzyme designs. A useful protocol for screening enzymatic activity should meet three criteria: (1) be computationally inexpensive, (2) correctly find the most active designs, and (3) discard most of the false positive designs (Figure 1-1). With these criteria in mind, we tested several MD protocols on a training set of enzymes. The best protocol was subsequently used on two larger and more diverse sets of enzymes that had already been experimentally characterized. Results indicate that the MD protocol is successful in screening enzymes for enzymatic activity. The protocol is general, reproducible, and excels at discarding false positive designs while predicting the most active designs correctly.

Designing conformational changes in enzymes

CPD calculations have typically been used to evaluate and select protein sequences in the context of a single native structure, a method referred to as single-state design (SSD).[6] Although proteins are generally encountered in a single native state,

many must be able to assume two or more different structural or chemical states to carry out their functions.[2; 20-23] For example, some enzymes must undergo conformational changes to be functional. They also need to finely tune their affinities for substrates, transition states, and products to catalyze reactions efficiently; in other words, they must achieve several different chemical states.

Recently, a novel multi-state design (MSD) method was developed that can consider several states of a protein in a single sequence optimization calculation.[24] In Chapter III of this thesis, we describe how this novel algorithm was tested on triosephosphate isomerase (TIM), an enzyme that contains a loop that oscillates between two states, open and closed (Figure 1-2). Loop 6 acts like a lid for the active site of TIM and can assume both open and closed conformations that facilitate substrate binding, product escape, and catalysis.[25-28] The flexibility that allows for the conformational change in loop 6 is provided by two 3-amino acid hinges at the beginning and at the end of the loop.[29] To test the performance of MSD and explore its advantages over SSD, we computationally designed TIM flexible hinges using both methods. Experimental characterization of selected designs revealed that calculations favoring the closed state over the open state were more robust at predicting highly and moderately active mutants. Our results also indicated that MSD could perform better than SSD at recapitulating naturally observed hinge sequences.

The experimental data generated in this study was also used to test the utility of the pre-screening protocol (described in Chapter II of this thesis). MD simulations were run on all the TIM designs. The results suggested that again, as

previously observed for other reactions (Chapter II), the MD protocol can discard false positive designs and correctly predict the most active ones.

Towards the de novo design of enzymatic activity

The most sought-after goal of computational enzyme design is the ability to reliably engineer made-to-order enzymes.[18] In Chapter IV of this thesis, I describe my efforts toward designing catalytic activity into inert protein scaffolds. TIM was used as a model system because it is easy to work with experimentally and there is an incredible amount of mutational and mechanistic data available on it.[4; 29-31] I devised a high-throughput computational framework that could in principle be extended to other reactions (Figure 1-3). The first step in the framework selected protein structures from available databases and identified appropriate sites within each scaffold that could accommodate the TIM active site. Putative catalytic residues and stabilizing mutations were then predicted using conventional enzyme design methods. Finally, by using the MD protocol described in Chapter II of this thesis, candidate designs were screened for binding to the substrate. Out of the starting ~8000 protein structures, 5 novel scaffolds passed the battery of computational tests along with 7 natural TIMs that we included as positive controls. Experimental characterization of the top-ranked designs and degenerate codon (DC) variant libraries based on the 5 novel scaffolds resulted in no active variants.

Despite the lack of success in designing enzymatic activity into inert scaffolds, our methods were able to recapitulate active sites in the wild-type TIM structures that served as positive controls. Moreover, these computationally

redesigned TIMs were able to complement a TIM-deficient *Escherichia coli* strain, demonstrating the overall validity of our computational approach. The general computational framework established in this work is promising and could prove useful for future enzyme design endeavors.

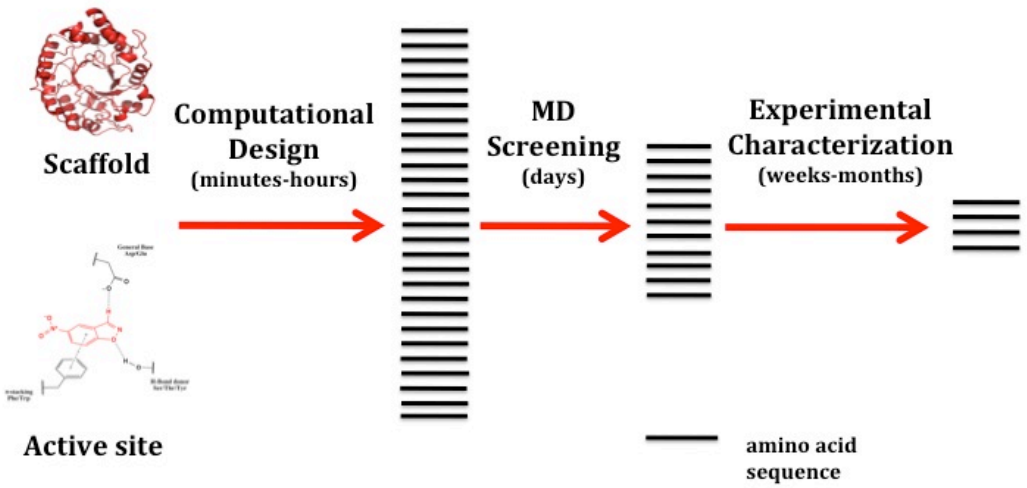


Figure 1-1. *In silico* pre-screening of enzymatic activity applied in the computational protein design framework. The scheme illustrates the design process. The MD protocol helps reduce the number of experimental targets by discarding false-positive sequences.

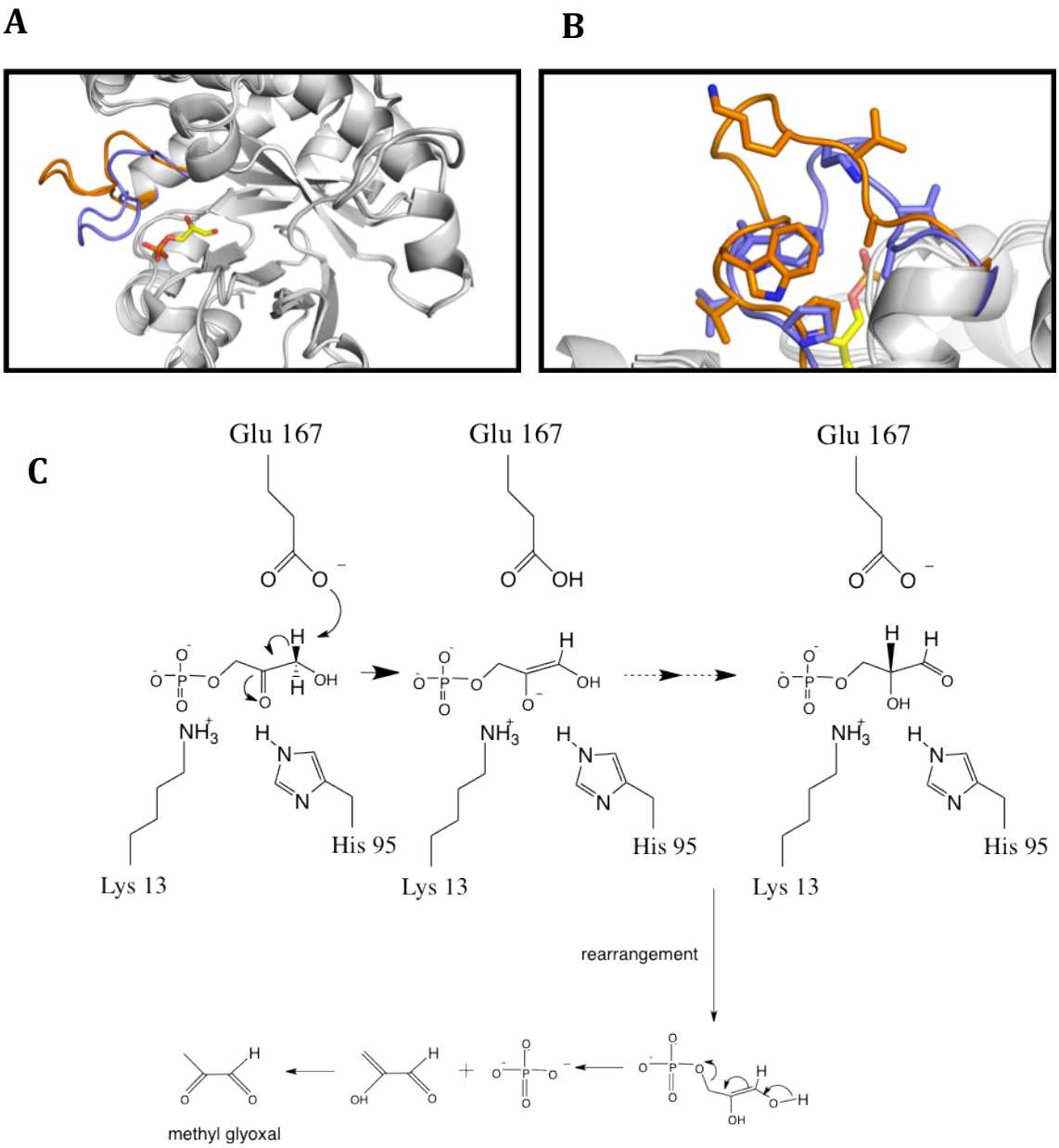


Figure 1-2. Triosephosphate isomerase as a paradigm for conformational changes in enzymatic catalysis. (A) Overall structure of TIM depicting the conformational change between the open (orange loop) and closed (purple loop) states**. (B)** TIM loop 6 emphasizing the 3-residue N-terminal (PVW) and C-terminal (KVA) hinges at the beginning and end of the loop, respectively. The hinges provide the flexibility for the conformational change. **(C)** The TIM reaction scheme. Loop 6's closed conformation is responsible for sequestering the reaction intermediate, which would otherwise react with water and eliminate phosphate producing methyglyoxal, a highly cytotoxic molecule.

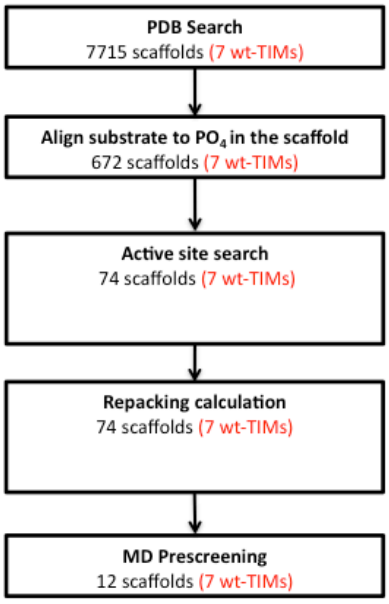


Figure 1-3. Scheme showing the high-throughput computational workflow for designing enzymes into inert scaffolds.

References

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

2. Herschlag, D. (1988). The role of induced fit and conformational-changes of enzymes in specificity and catalysis. *Bioorganic Chemistry* **16**, 62-96.

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

4. Knowles, J. R. (1991). Enzyme catalysis: Not different, just better. *Nature* **350**, 121-124.

5. Dahiyat, B. I. & Mayo, S. L. (1996). Protein design automation. *Protein Science* **5**, 895-903.

6. Alvizo O, Allen BD & Mayo SL. (2007). Computational protein design promises to revolutionize protein engineering. *Biotechniques* **42**, 31–35.

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

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

9. Malakauskas, S. M. & Mayo, S. L. (1998). Design, structure and stability of a hyperthermophilic protein variant. *Nature Structural Biology* **5**, 470-475.

10. Bolon, D. N. & Mayo, S. L. (2001). Enzyme-like proteins by computational design. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 14274-14279.

11. 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. & Baker, D. (2008). Kemp elimination catalysts by computational enzyme design. *Nature* **453**, 190-U4.

12. Siegel, J. B., Zanghellini, A., Lovick, H. M., Kiss, G., Lambert, A. R., Clair, J. L. S., Gallaher, J. L., Hilvert, D., Gelb, M. H., Stoddard, B. L., Houk, K. N., Michael, F. E.

& Baker, D. (2010). Computational design of an enzyme catalyst for a stereoselective bimolecular diels-alder reaction. *Science* **329**, 309-313.

13. Privett, H. K., Kiss, G., Lee, T. M., Blomberg, R., Chica, R. A., Thomas, L. M., Hilvert, D., Houk, K. N. & Mayo, S. L. (2012). Iterative approach to computational enzyme design. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 3790-3795.

14. Khare, S. D., Kipnis, Y., Greisen, P. J., Takeuchi, R., Ashani, Y., Goldsmith, M., Song, Y. F., Gallaher, J. L., Silman, I., Leader, H., Sussman, J. L., Stoddard, B. L., Tawfik, D. S. & Baker, D. (2012). Computational redesign of a mononuclear zinc metalloenzyme for organophosphate hydrolysis. *Nature Chemical Biology* **8**, 294-300.

15. Hilvert, D. (2000). Critical analysis of antibody catalysis. *Annual Review of Biochemistry* **69**, 751-793.

16. Blomberg, R., Kries, H., Pinkas, D. M., Mittl, P. R. E., Grutter, M. G., Privett, H. K., Mayo, S. L. & Hilvert, D. (2013). Precision is essential for efficient catalysis in an evolved kemp eliminase. *Nature* **503**, 418-+.

17. Kries, H., Blomberg, R. & Hilvert, D. (2013). De novo enzymes by computational design. *Current Opinion in Chemical Biology* **17**, 221-228.

18. Baker, D. (2010). An exciting but challenging road ahead for computational enzyme design. *Protein Science* **19**, 1817-1819.

19. Bolon, D. N., Voigt, C. A. & Mayo, S. L. (2002). De novo design of biocatalysts. *Current Opinion in Chemical Biology* **6**, 125-129.

20. Gerstein, M., Lesk, A. M. & Chothia, C. (1994). Structural mechanisms for domain movements in proteins. *Biochemistry* **33**, 6739-6749.

21. Hammes, G. G. (2002). Multiple conformational changes in enzyme catalysis. *Biochemistry* **41**, 8221-8228.

22. Henzler-Wildman, K. & Kern, D. (2007). Dynamic personalities of proteins. *Nature* **450**, 964-972.

23. Henzler-Wildman, K. A., Thai, V., Lei, M., Ott, M., Wolf-Watz, M., Fenn, T., Pozharski, E., Wilson, M. A., Petsko, G. A., Karplus, M., Hubner, C. G. & Kern, D.

(2007). Intrinsic motions along an enzymatic reaction trajectory. *Nature* **450**, 838-U13.

24. Allen, B. D. & Mayo, S. L. (2010). An efficient algorithm for multistate protein design based on faster. *Journal of Computational Chemistry* **31**, 904-916.

25. Williams, J. C. & Mcdermott, A. E. (1995). Dynamics of the flexible loop of triosephosphate isomerase - the loop motion is not ligand-gated. *Biochemistry* **34**, 8309-8319.

26. Sun, J. & Sampson, N. S. (1999). Understanding protein lids: Kinetic analysis of active hinge mutants in triosephosphate isomerase. *Biochemistry* **38**, 11474-11481.

27. Sun, S. H. & Sampson, N. S. (1998). Determination of the amino acid requirements for a protein hinge in triosephosphate isomerase. *Protein Science* **7**, 1495-1505.

28. Xiang, J. Y., Sun, J. H. & Sampson, N. S. (2001). The importance of hinge sequence for loop function and catalytic activity in the reaction catalyzed by triosephosphate isomerase. *Journal of Molecular Biology* **307**, 1103-1112.

29. Wierenga, R. K., Kapetaniou, E. G. & Venkatesan, R. (2010). Triosephosphate isomerase: A highly evolved biocatalyst. *Cellular and Molecular Life Sciences* **67**, 3961-3982.

30. Cui, Q. & Karplus, M. (2003). Catalysis and specificity in enzymes: A study of triosephosphate isomerase and comparison with methyl glyoxal synthase. *Protein Simulations* **66**, 315-+.

31. Knowles, J. R. & Albery, W. J. (1977). Perfection in enzyme catalysis - energetics of triosephosphate isomerase. *Accounts of Chemical Research* **10**, 105-111.

Chapter II

In Silico Screening of Enzymatic Activity: A Molecular Dynamics Approach

Adapted from a manuscript in preparation coauthored with Thomas F. Miller III and Stephen L. Mayo.

Abstract

Current computational protein design (CPD) tools are limited to the reliable prediction of active enzymes and thus produce a large number of false positives. Consequently, in the best scenario, hundreds of designs need to be experimentally screened to find a few active ones. In addition, accurate prediction of enzymatic activity requires inclusion of backbone dynamics and implicit solvation; both factors are generally neglected in common CPD implementations. In this work, we develop a Molecular Dynamics (MD) protocol that provides a reliable and computationally feasible way to pre-screen enzyme designs for active variants. Preliminary work indicated excellent agreement between experimental data and MD predictions. We tested the methodology on a larger and more diverse set of enzymes for which experimental data already exist. Our results suggest that the method is transferable to different reactions and has high sensitivity (0.75) and specificity (0.71). Our protocol also exhibits high negative predictive value (0.94) and always predicts the most active enzymes correctly, providing a useful *in silico* screening tool for the CPD framework.

Introduction

Much attention has recently been paid to the challenge of enzyme design.[1-5]. Because of their high specificity, efficiency, and biodegradability, enzymes are good candidates for industrial and medical applications, and are often preferred to catalysts obtained via chemical synthesis. However, their use in most applications is limited by difficulties in designing an enzyme with the required catalytic activity. The central challenge of the field is thus to consistently design an enzyme for any desired reaction.

Computational protein design (CPD) has proven to be useful in engineering improved stability, novel scaffolds, protein-protein interactions, etc.[6-9] However, enzyme design remains one of the most challenging tasks that this field faces. There are currently only a few examples of successful *de novo* computational enzyme design.[1; 3-5; 10; 11] A major limitation of the methodology is that it cannot reliably predict protein sequences with the desired catalytic function. This problem stems from inaccuracies in the models and force fields used, as well as from an incomplete understanding of the structure-function relationships involved in enzymatic activity. Consequently, computational enzyme design calculations typically predict many amino acid sequences to be active that are not (false positives), yielding a low signal to noise (active to inactive) ratio. In the best scenario, hundreds of designs must be tested to find a few active ones. In addition to the obvious costs of screening in terms of time and resources, experimental screening also presents challenges that may be difficult to overcome. For example, challenges may be encountered in the optimization of heterologous expression, protein purification, or the availability of

high-throughput screening methods. The computational protein design framework would thus benefit from having an *in silico* pre-screening step that is capable of reliably ranking structures according to their catalytic properties.

Computational tools have previously been used to quantify enzymatic catalysis.[12-14] In particular, Warshel *et al.*[13; 15] and Hammes-Schiffer *et al.*[16] independently applied the empirical valence bond (EVB) method to a small set of enzymes. Although excellent correlation between calculated and experimental catalytic rates was reported, neither of these studies analyzed inactive mutants. Therefore they did not assess the method's ability to distinguish active mutants from inactive ones, which is critical in *de novo* enzyme design applications. Additionally Houk *et al.*[17] utilized MD to qualitatively analyze computationally designed enzymes. Although reasonable agreement between experimental results and MD predictions was found, only one reaction was studied and thus the generality of the method was not tested.

Recently, Privett *et al.*[10] reported the successful *de novo* computational design of enzymes for the Kemp elimination (KE)(Figure 2-1A), a reaction for which no natural catalyst exists. In their study, MD simulations combined with simple geometric criteria was used to predict the activity of a small set of enzyme designs. These results illustrated the complementary role of MD simulations in the CPD framework. To keep the calculations computationally affordable, current implementations of CPD do not include backbone dynamics or explicit solvation, both of which play an important role in enzymatic catalysis.[18; 19] MD simulations, on the other hand, can model both.

In the current work, we develop a classical MD protocol aimed at providing a general, reliable, and computationally affordable approach that could analyze many designs and discard most of the false positives. Given that a universal feature of enzyme catalysis is the ability to limit the motions of the substrate in the active site to conformations and geometries that allow chemistry to take place, we decided to use a geometry-based criterion that could be readily generalized to any reaction. In other words, designs that lack structural integrity and do not maintain the putative catalytic contacts during the course of the simulation are predicted to be inactive, and are therefore discarded.

In the following sections, we describe how the MD protocol was optimized and tested for reproducibility, accuracy, and transferability to other reactions. To allow for large-scale screening, the protocol is designed to preserve computational efficiency. We used in total over 170 different structures aimed at catalyzing either of two reactions: the Kemp elimination of 5-nitrobenzisoxazole (NBZ) or the Claisen rearrangement of chorismate. Our results suggest that the MD protocol is general, and excels at discarding false-positive designs while predicting the most active enzymes correctly.

Results and Discussion

Optimization of the protocol

In order to optimize the MD protocol for accuracy and computational performance, different simulation conditions needed to be tested. Four different protocols were

tested (Figure 2-2). We used the 8 Kemp elimination designs reported by Privett *et al.*[10] as a training set (Table 2-1). Given that the key contact in the active site is the hydrogen bond between the general base (Asp or Glu) and the hydrogen on the substrate, this distance was monitored during the simulation and was used to predict activity (Figure 2-1B). Structures that maintained this contact during the simulation (time-average distance ≤ 3.20 Å,[20; 21]) were predicted to be active, whereas those that did not were predicted to be inactive.

In the first version of our screening approach (Figure 2-2), minimization was performed to eliminate clashes with water molecules while keeping all non-water, non-hydrogen atoms fixed. This was followed by a heating phase to bring the system to the target temperature (300 K). An equilibration step was done at NPT conditions (300 K and 1 atm). By monitoring temperature, pressure, radius of gyration, and RMSD, we ensured that the system was equilibrated. Finally, the production phase consisted of 20 ns of NPT MD.

Simulations were run for the 8 Kemp elimination designs in the training set. To illustrate how the simulations were analyzed, we will focus on two designs: HG-1 and HG-2 (Figure 2-1C). Both of these designs use the same protein scaffold even though their active sites are in different locations within the structure. The active site of HG-2 is more buried than HG-1's. In our simulations, HG-1, an experimentally inactive design, exhibits a large average catalytic distance (>20 Å), indicating diffusion of the substrate from the active site early in the simulation (Figure 2-1C). HG-1 is therefore predicted to be inactive. Conversely, HG-2, an experimentally active design, displays an average catalytic distance within canonical hydrogen bond

distances (< 3.2 Å). The MD results indicate that the substrate remains bound to the active site during the entire simulation. Thus, HG-2 is predicted to be active. Despite these successful predictions, the protocol was not accurate enough to be useful. Only three of the eight designs were predicted correctly (Table 2-1). A more detailed analysis of the simulations revealed that the first three steps of Protocol 1 (minimization, heating, and equilibration) were completely disrupting the structure of some of the designs. For instance, the binding and catalytic contacts of other designs in the training set, were lost after equilibration of the system. In light of these observations, we decided to modify the protocol to include restraints on the active site residues in order to keep the structure as close as possible to the design structure, at least until the end of the equilibration.

We therefore devised Protocol 2 (Figure 2-2). The main difference between this procedure and Protocol 1 is that harmonic restraints are applied to the active site residues and substrate during the heating (force constant equal to 50 kcal/mol/A^2) and equilibration steps (force constant starting from 50 kcal/mol/A^2 and gradually releasing to 0 kcal/mol/A^2). Protocol 2 yielded predictions that are more consistent with experimental results, with five of eight designs predicted correctly (Table 2-1). However, 1A53-1 and 1THF-1, both inactive designs (false positives), were again predicted to be active. Further analysis provided some valuable lessons. Using only the first 10 ns of simulation resulted in exactly the same predictions as using 20 ns; therefore, the time of the production phase could be halved. In addition, we hypothesized that the heating step could be discarded and

that just an equilibration at the target temperature might be enough to equilibrate the system.

To explore the impact of these changes on our screening tool, we formulated two more protocols (Protocols 3 and 4, Figure 2-2). The minimization step was the same as before. However, the heating phase was not included, and harmonic restraints were applied on all non-water, non-hydrogen atoms during equilibration. The force constants for these restraints were gradually released from 50 kcal/mol/A^2 using the following loop: $K=50/4^i$, i = 0, 1, 2 … 5. The production phase was also different. Protocol 3 used unrestrained MD on all atoms, whereas Protocol 4 applied restraints on non-active site residues using a low force constant ($K=50/4^5$ kcal/mol/A^2). In both cases, the production phase was only 10 ns.

Both protocols produced results that were in agreement with the experimental data for the training set (Table 2-1). Two of the three false positive designs in the training set were discarded, while all of the active designs were correctly predicted. These results suggested that Protocols 3 and 4 could yield useful predictions with only 10 ns of simulation time.

Simulation and analysis of the KE enzyme designed by Röthlisberger *et al.*[4]

We now apply the screening approach to a larger and more diverse set of KE enzymes that had been computationally designed by Röthlisberger *et al* [4]. In that study, computational methods were used to design candidate sequences. Fifty-nine designs (in seventeen different scaffolds) were selected for experimental characterization. Only eight of them were found to have measurable activity.

Protocols 3 and 4 were run in triplicate, using different random number seeds, for 50 of the KE designs reported by Röthlisberger *et al.,*[4] which included the 8 active enzymes. As seen in Table 2-2 and Figure 2-3, Protocol 3 correctly predicted 6 of the 8 active designs (sensitivity of 0.75), whereas Protocol 4 only did so for 4 designs (sensitivity of 0.50). In both cases, the two most active enzymes were among those predicted correctly. On the other hand, both protocols exhibited high specificity (0.71), which means that 71% of false positives are being discarded. In addition, both protocols showed high (0.94) negative predictive value (NPV), meaning that 94% of the designs that are predicted as inactive are indeed experimentally inactive. Additionally, Protocol 3 exhibits higher positive predictive value (PPV) than Protocol 4. Note also that the small error bars in Figure 2-3 indicate that our protocols are reproducible. Since these results showed that Protocol 3 was the best protocol tested, we decided to use it for the rest of our simulations. . If we were to use Protocol 3, instead of having to do experiments on fifty candidate designs to find eight active ones, we would only have to run experiments on eighteen candidate sequences to find six active designs, which include the two most active ones (Table 2-2).

Our simulations revealed that for most of the designs predicted to be inactive, computationally designed active site geometries are disrupted early in the simulation (within the first 3 ns), evidencing a lack of structural integrity within the binding pocket. For example, one common failure is that the general base or the π-stacking residue (catalytic residues in the KE active site, Figure 2-1B) flips out of the active site, preferring to interact with surrounding water molecules instead of the

substrate. As a result, the substrate moves to a final non-catalytic position within the active site or completely diffuses away from the pocket, rendering a predicted inactive design.

Transferability to another enzymatic reaction: The Claisen rearrangement of chorismate

In a third set of tests, we apply Protocol 3 to the study of chorismate mutases (CMs). CM catalyzes the biologically relevant unimolecular rearrangement of chorismate to prephenate, a Claisen rearrangement (Figure 2-4A). The reaction is part of the biosynthetic pathway of aromatic amino acids such as tryptophan and tyrosine. It has a very well characterized chemistry[22-25] and has been the subject of research in our lab for several years.[26] We based our study on CM mutants that had been previously studied by Lassila *et al.*[27] In that work, exhaustive mutagenesis was carried out on six secondary active site residues, yielding forty active mutants and seventy-four inactive ones (Figure 2-4B).

The geometric criteria appropriate for analysis of MD trajectories depend on the reaction. For the rearrangement of chorismate, we monitored the distance between the two carbons that will form a bond in the product (*i.e.*, the C6-C8 distance) and used a cut-off distance of 3.5 Å as criteria[28](Figure 2-5). Previous studies have shown that the enzyme catalyzes this reaction because the di-axial configuration (reactive configuration) of the substrate is more favored in the active site than in water; in water, the di-equatorial configuration is more stable.[28] Given this, we postulated that the substrate must keep the di-axial conformation for the

enzyme to be active, and therefore decided to also monitor the relevant dihedral angle (*i.e.,* O4-C3-C4-O1) (Figure 2-5).

Wild-type *Escherichia coli* CM (PDB ID: 1ECM) and its 114 single mutants, with chorismate bound to the active site in the di-axial conformation, were subjected to MD simulations. Typical MD results obtained for this set of enzymes are illustrated by wild-type CM (1ECM) and its experimentally inactive mutant 1ECM-D48I (Figure 2-6). For the wild-type enzyme, the substrate C8-C6 distance was very localized at 3.32 Å. Analysis of the dihedral angles showed that the substrate also retained a di-axial configuration (O4-C3-C4-O1 dihedral angle ~ –139°). In contrast, 1ECM-D48I had an averaged C8-C6 distance of 3.9 Å; in addition, a transition at 2.5 ns corresponding to a final di-equatorial conformation was observed. Dihedral angle vs. time plots also exhibit a transition at 2.5 ns to a conformational state consistent with a final di-equatorial configuration of the substrate (O4-C3-C4-O1 dihedral angle ~ –90°).

The results and conclusions for the CM mutants were comparable to those obtained for the KE enzymes (Figure 2-7 and Table 2-4). As in the case of the KE enzymes, the most active mutants are predicted correctly, and the majority of the mutants discarded by the protocol are indeed inactive (high NPV). It is worth noticing that this test was challenging in many ways. It involved a different reaction, but more importantly, the structural differences between the mutants and the wild-type enzyme were minimal—only single mutants were analyzed. These relatively minor structural differences provide a stringent test of the protocol's ability to discriminate between designs with different levels of activity. Together, these

findings suggest that the geometric criteria-based MD protocol can be successfully transferred to a different reaction.

Computational performance

Computational performance of MD simulations depends on many factors, including the size of the system, the cutoff distance, PME, etc.[29] It was therefore important to determine the performance and scalability of our protocol on the type of system we intended to work on. Figure 2-8 shows an assessment of Protocol 3's performance on CPU-only and CPU-GPU configurations. Using GPU-CPU clusters can be an interesting alternative because it provides a 5- to 10-fold improvement in performance.

Conclusions

By making rational-based modifications to our protocol, we developed an explicit solvent molecular dynamics approach to pre-screen enzyme designs. The results suggest that our methodology is general, reproducible, and can discard false positive designs while predicting the most active designs correctly. In addition, the method seems to be readily transferable to other chemistries. The most appropriate geometric criteria to be used with the protocol will depend on the reaction and proposed mechanism, and should be determined based on chemical intuition or previous experimental data.

The MD protocol also exhibits promising computational performance. In the not-so-far future, when a 10 ns simulation is trivial, we envision using it on

thousands of candidate designs for a fast screening. The hits from this initial screening could be further scrutinized by running the protocol for longer time scales, or by setting up simulations in which many instances of the substrate are placed randomly in the simulation box and unbiased binding to the active site is evaluated during the course of the simulation.[30]

Our findings also suggest that future efforts in computational enzyme design should be focused on including backbone flexibility and improving the solvation models used in the design calculations. In particular, a more accurate description of interactions of the protein with the solvent could enhance our ability to predict the positioning of highly charged sidechains within the binding pocket of designer enzymes.

In summary, this *in silico* tool should prove useful by helping researchers focus their experimental efforts on designs that have enough structural integrity to maintain the putative catalytic geometry during the course of a short and computationally affordable simulation.

Materials and Methods

MD simulations and criteria

Simulations typically started with the substrate bound to the active site of the enzyme. The Amber force field was used in all simulations; Amber force field ff99SB[31] was implemented for proteins and the general Amber force field[32] was used for small molecules. For the Kemp elimination substrate (NBZ), we used

charges that had been developed previously[10], and for the chorismate molecule, we calculated charges using the restrained electrostatic potential (RESP) method at HF/6-31G* as implemented in the REDS server.[33; 34] AmberTools 10 was used to prepare the structures for simulations.[35] Hydrogen atoms were added to the structures, after which the systems were completely solvated with water molecules (TIP3P water model). Finally, Na^+ or Cl^- ions were added to neutralize the systems. The cell shape was a rectangular box or a truncated octahedron.

The simulations were carried out using the MD software NAMD 2.7.[29] Pressure was kept constant using the Nose-Hoover Langevin piston method with a damping time of 50 fs and a decay period of 100 fs. Langevin dynamics was used with a target constant temperature of 300 K and a 5 ps^{-1} damping coefficient. We took advantage of the multiple time-stepping capabilities of the code: a 2 fs time step was used, non-bonded interactions (van der Waals and electrostatic) were calculated every two steps, and long range electrostatic interactions [particle mesh Ewald (PME) with periodic boundary conditions (PBC)] were calculated every four steps. Visual molecular dynamics (VMD) was used to visualize and analyze the simulations.[36]

Purely geometric criteria were used to predict enzymatic activity. We hypothesized that structures that maintained the active site-substrate contacts relevant to binding and catalysis would be active. For the Kemp elimination reaction the catalytic contact between the general base (aspartate or glutamate) and the hydrogen on 5-nitrobenzisoxazole was monitored during the simulations and the time average was calculated. Designs that exhibited a time-averaged catalytic

distance ≤ 3.2 Å were predicted to be active, otherwise inactive.[20; 21] Similarly for Claisen rearrangement of chorismate, the distance between the two carbon atoms that would form a bond in the product (C8-C6) was monitored. Designs that exhibited a time-averaged C8-C6 distance ≤ 3.5 Å were predicted to be active, otherwise inactive.[28]

Definitions

True positive: simulation = active AND experiment = active

False positive: simulation = active AND experiment = inactive

True negative: simulation = inactive AND experiment = inactive

False negative: simulation = inactive AND experiment = active

Specificity: proportion of true negatives that are correctly identified:

Specificity = # true negatives/ (# true negatives + # false positives)

Sensitivity*:* proportion of true positives that are correctly identified:

Sensitivity = # true positives/ (# true positives + # false negatives)

Negative predictive value (NPV): proportion of predicted negatives that are true negatives:

NPV = # true negatives / (# true negatives + # false negatives)

Positive predictive value (PPV): proportion of predicted positives that are true positives:

PPV = # true positives / (# true positives + # false positives)

Acknowledgements

The authors thank Marie Ary for help with the manuscript.

Table 2-1. Comparison of experimental results with MD predictions for the 8 KE designs in the training set. Predictions shown shaded are correct. ND = not determined, A = active, and I = inactive. k_{cat}s were reported in Privett *et al.* [10]

Design	Scaffold (PDB ID)	k_{cat} (s^{-1})	Protocol 1	Protocol 2	Protocol 3	Protocol 4
HG-1	1gor	Inactive	I	I	I	I
HG-2	1gor	ND*	A	I	A	A
HG-3	1gor	0.68	I	A	A	A
1A53-1	1a53	Inactive	A	A	I	I
1A53-2	1a53	0.012	I	A	A	A
1A53-3	1a53	0.015	A	A	A	A
1THF-1	1thf	Inactive	A	A	A	A
1THF-2	1thf	ND*	I	A	A	A

* These enzymes are active but their k_{cat} could not be determined.

Table 2-2. Summary of the MD results for the KE designs reported by Röthlisberger *et al.*[4] NA, not applicable**.**

	No MD	**Protocol 3**	**Protocol 4**
# true positive designs	8	6	4
# true negative designs	NA	30	29
# false positive designs	42	12	13
# false negative designs	NA	2	4
# designs to test experimentally	50	18	17
# designs to be discarded	NA	32	33

Table 2-3. Protocol 3 results for the Kemp eliminases by Röthlisberger *et al.*[4]

Design	Scaffold (PDB ID)	k_{cat} (s^{-1})	Catalytic distance d (Å) [a]	Prediction [b]
KE59	1a53	0.29	3.04	A
KE70	1jcl	0.16	2.84	A
KE10	1a53	0.029	2.52	A
KE15	1thf	0.022	2.86	A
KE07	1thf	0.018	2.74	A
KE16	1thf	0.006	3.15	A
KE61	1h61	ND*	5.05	I
KE71	1a53	ND*	6.15	I
KE01	1eix	Inactive	4.97	I
KE02	1eix	Inactive	4.16	I
KE03	1thf	Inactive	5.03	I
KE04	1thf	Inactive	8.34	I
KE05	1lbm	Inactive	3.16	A
KE06	1thf	Inactive	5.11	I
KE08	1thf	Inactive	4.06	I
KE09	1thf	Inactive	3.23	I
KE11	1a53	Inactive	2.53	A
KE12	148l	Inactive	2.92	A
KE13	1thf	Inactive	11.48	I
KE14	1thf	Inactive	4.76	I
KE17	1thf	Inactive	8.45	I
KE18	1a53	Inactive	3.22	I
KE19	1v04	Inactive	7.77	I
KE20	1lbm	Inactive	3.86	I
KE34	1v04	Inactive	5.90	I
KE35	2izj	Inactive	3.79	I
KE36	1tsn	Inactive	3.53	I
KE37	6cpa	Inactive	7.56	I
KE38	1lbm	Inactive	3.36	I
KE39	1a53	Inactive	5.80	I
KE40	1dc9	Inactive	3.44	I
KE42	1igs	Inactive	3.59	I
KE45	1v04	Inactive	5.66	I
KE46	1v05	Inactive	8.13	I
KE47	1dmq	Inactive	2.65	A
KE48	1opy	Inactive	6.61	I
KE49	1dmm	Inactive	6.57	I
KE50	1lbl	Inactive	4.89	I
KE51	1pii	Inactive	2.93	A
KE53	1pii	Inactive	2.99	A

KE54	1jcl	Inactive	2.71	A
KE55	1igs	Inactive	6.42	I
KE58	1wbj	Inactive	2.43	A
KE60	1jul	Inactive	3.08	A
KE62	1ftx	Inactive	2.88	A
KE63	2fpc	Inactive	6.74	I
KE64	1thf	Inactive	3.06	A
KE65	1lbm	Inactive	3.49	I
KE66	1lbl	Inactive	6.39	I
KE67	1s1d	Inactive	2.92	A

ND, not determined.
* These enzymes are active but their k_{cat} could not be determined.
[a] Average catalytic distance measured in a 10 ns MD simulation.
[b] A = active (if d ≤ 3.20 Å), I = inactive (if d > 3.20 Å).[20; 21]

Table 2-4. Protocol 3 results for *E. Coli* chorismate mutase and its 114 single mutants reported by Lassila *et al.*[27]

Mutant	k_{cat}/K_M (M^{-1} s^{-1})	C8-C6 Distance (Å)[b]	O4-C3-C4-O1 Dihedral Angle (°)	Prediction [a]
1ECM (WT)	1.30E+05	3.32	-139.3	A
A32S	2.10E+05	3.31	-140.6	A
A32T	1.60E+05	3.39	-137.1	A
V35I	1.50E+05	3.47	-124.6	A
L7F	1.40E+05	3.49	-123.4	A
L7I	1.20E+05	3.44	-129.2	A
I81M	1.00E+05	3.48	-128.6	A
V85I	1.00E+05	3.48	-122.9	A
I81L	5.50E+04	3.63	-124.6	I
V85C	5.40E+04	3.47	-127.5	A
V35M	5.20E+04	3.50	-121.6	A
V35C	5.20E+04	3.73	-113.1	I
L7V	5.00E+04	3.52	-121.8	I
L7M	4.40E+04	3.34	-140.2	A
V85L	4.30E+04	3.54	-115.9	I
V85R	4.20E+04	3.67	-111.9	I
V35A	3.90E+04	3.98	-96.6	I
V85Y	2.60E+04	3.68	-109.5	I
V85T	2.20E+04	3.67	-113.9	I
V85F	2.10E+04	3.91	-94.7	I
V35T	2.00E+04	3.85	-105.5	I
D48C	1.80E+04	3.39	-133.1	A
V85M	1.50E+04	3.31	-138.7	A
V85K	1.30E+04	3.38	-133.7	A
V35L	1.30E+04	3.38	-141.8	A
L7T	1.30E+04	3.75	-104.4	I
V85N	1.20E+04	3.61	-115.0	I
V85A	1.20E+04	3.32	-137.0	A

L7C	1.00E+04	3.37	-137.9	A
D48Q	9.80E+03	3.37	-135.6	A
I81F	8.70E+03	3.31	-138.4	A
D48N	7.90E+03	3.71	-107.5	I
D48L	7.10E+03	3.38	-135.1	A
A32C	ND*	3.41	-141.9	A
I81C	ND*	3.51	-126.1	I
A32G	ND*	3.73	-108.3	I
A32I	ND*	3.86	-103.8	I
I81W	ND*	3.73	-108.3	I
V85W	ND*	3.52	-127.9	I
A32V	ND*	3.82	-102.2	I
I81V	ND*	3.51	-124.4	I
A32D	Inactive	3.58	-119.2	I
A32E	Inactive	3.39	-137.6	A
A32F	Inactive	4.39	-83.2	I
A32H	Inactive	3.36	-146.1	A
A32K	Inactive	4.03	-93.1	I
A32L	Inactive	3.46	-124.6	A
A32M	Inactive	3.68	-108.7	I
A32N	Inactive	3.43	-129.9	A
A32P	Inactive	3.94	-107.5	I
A32Q	Inactive	3.60	-115.1	I
A32R	Inactive	3.58	-116.5	I
A32W	Inactive	3.88	-101.5	I
A32Y	Inactive	3.54	-122.4	I
D48A	Inactive	3.67	-108.0	I
D48E	Inactive	3.29	-143.2	A
D48F	Inactive	3.38	-137.6	A
D48G	Inactive	3.72	-110.6	I
D48H	Inactive	3.59	-116.4	I
D48I	Inactive	3.97	-90.6	I
D48K	Inactive	3.37	-134.9	A
D48M	Inactive	3.55	-118.9	I

D48P	Inactive	3.72	-105.7	I
D48R	Inactive	3.45	-125.8	A
D48S	Inactive	3.64	-109.1	I
D48T	Inactive	3.43	-137.7	A
D48V	Inactive	3.37	-135.7	A
D48W	Inactive	3.39	-134.0	A
D48Y	Inactive	3.38	-133.0	A
I81A	Inactive	3.78	-99.6	I
I81D	Inactive	4.19	-110.3	I
I81E	Inactive	3.52	-123.6	I
I81G	Inactive	3.43	-135.4	A
I81H	Inactive	3.35	-138.5	A
I81K	Inactive	3.41	-135.8	A
I81N	Inactive	3.32	-142.0	A
I81P	Inactive	4.02	-94.3	I
I81Q	Inactive	3.91	-92.6	I
I81R	Inactive	3.56	-131.6	I
I81S	Inactive	3.48	-129.4	A
I81T	Inactive	3.48	-131.3	A
I81Y	Inactive	3.47	-129.8	A
L7A	Inactive	3.70	-108.8	I
L7D	Inactive	3.64	-113.8	I
L7E	Inactive	3.56	-121.0	I
L7G	Inactive	3.41	-142.0	A
L7H	Inactive	3.58	-125.1	I
L7K	Inactive	3.97	-90.9	I
L7N	Inactive	3.63	-112.6	I
L7P	Inactive	3.66	-114.6	I
L7Q	Inactive	3.44	-126.7	A
L7R	Inactive	3.56	-117.2	I
L7S	Inactive	3.69	-111.5	I
L7W	Inactive	3.52	-120.8	I
L7Y	Inactive	3.40	-136.8	A
V35D	Inactive	3.71	-109.9	I

V35E	Inactive	3.38	-143.4	A
V35F	Inactive	3.57	-110.3	I
V35G	Inactive	4.04	-91.9	I
V35H	Inactive	3.46	-125.0	A
V35K	Inactive	4.29	-104.9	I
V35N	Inactive	3.82	-107.1	I
V35P	Inactive	3.66	-113.7	I
V35Q	Inactive	3.53	-121.2	I
V35R	Inactive	3.33	-146.1	A
V35S	Inactive	3.70	-114.5	I
V35W	Inactive	3.63	-109.6	I
V35Y	Inactive	3.46	-129.7	A
V85D	Inactive	3.43	-130.2	A
V85E	Inactive	3.34	-141.0	A
V85G	Inactive	3.50	-130.1	A
V85H	Inactive	3.68	-107.7	I
V85P	Inactive	3.59	-115.7	I
V85Q	Inactive	3.43	-129.6	A
V85S	Inactive	3.47	-123.9	A

ND, not determined.

* These enzymes are active but their k_{cat} could not be determined.

[b] Average distance measured in a 10 ns MD simulation.

[a] A = active (if C8-C6 distance ≤ 3.50 Å), I = inactive (if C8-C6 distance > 3.50 Å).[28]

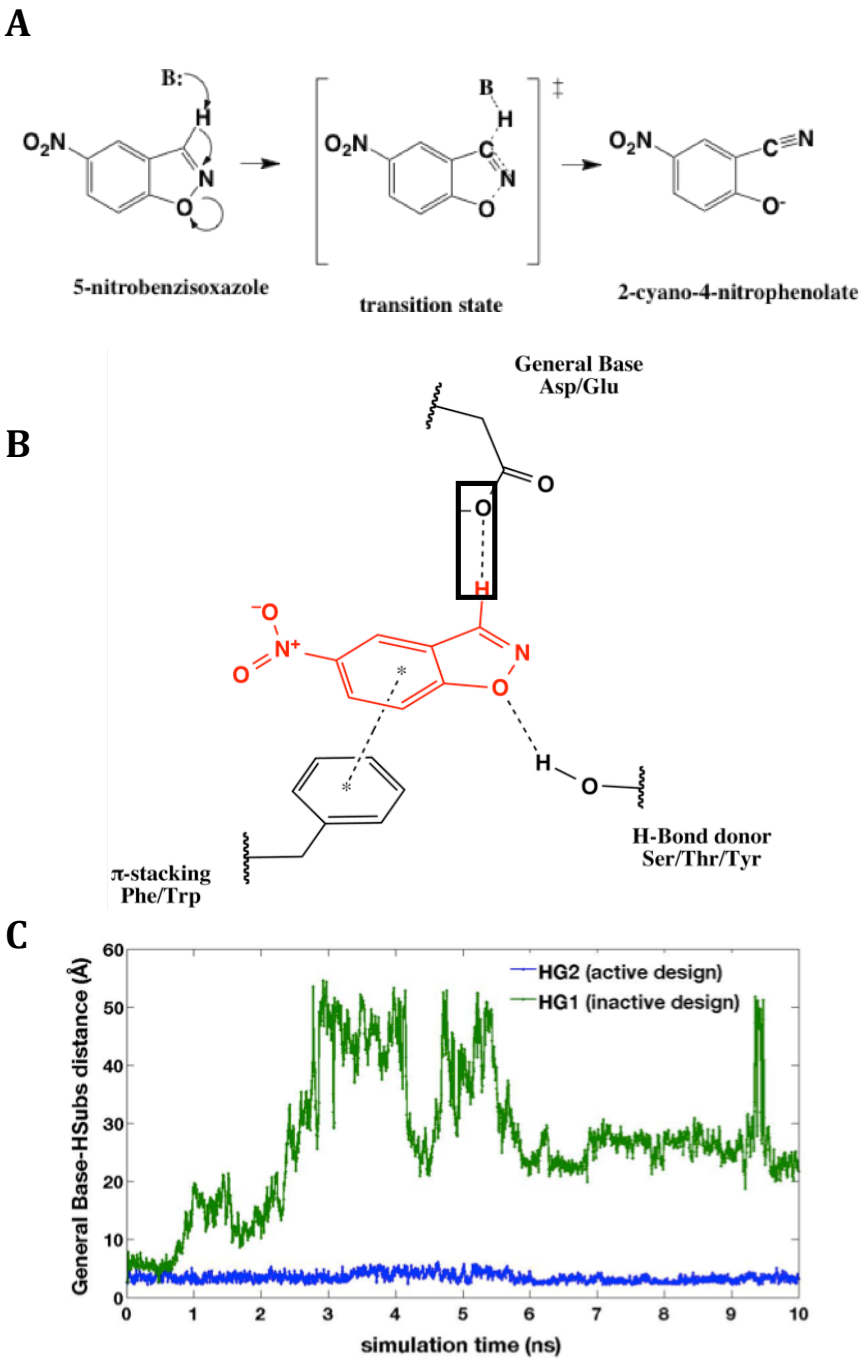


Figure 2-1. Kemp elimination. (A) Kemp elimination reaction scheme. **(B)** Ideal active site for a Kemp eliminase. The substrate is depicted in red. The distance measured in the simulations is indicated with a box. **(C)** Examples of MD results. HG-1 (green), an experimentally inactive design, exhibits diffusion of substrate from the active site early in the simulation. HG-2 (blue), an experimentally active design, keeps the substrate in the designed conformation during the entire simulation.

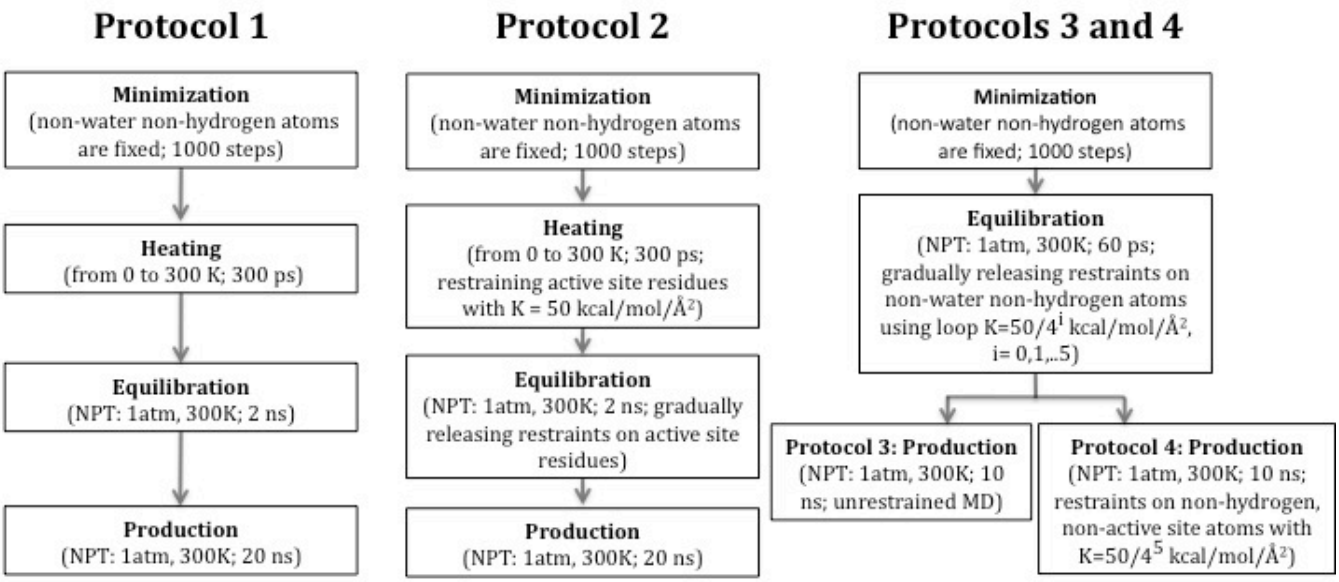


Figure 2-2. MD protocols. Active site residue is defined as any residue with at least one atom within 6.0 Å of the substrate.

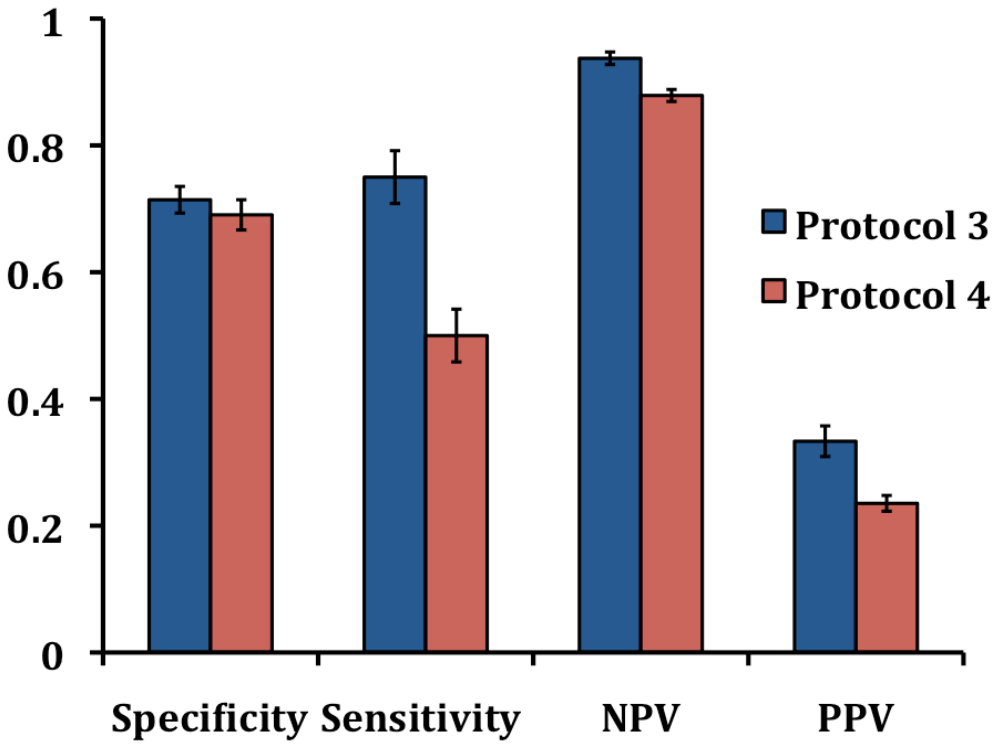


Figure 2-3. Protocols 3 and 4 excel at discarding false-positive designs. Protocol 3 gives the best results. Both protocols are higlhy reproducible. NPV: negative predictive value; PPV: positive predictive value. The statistical parameters are defined in Materials and Methods. Detailed results for each design are available in Table 2-3.

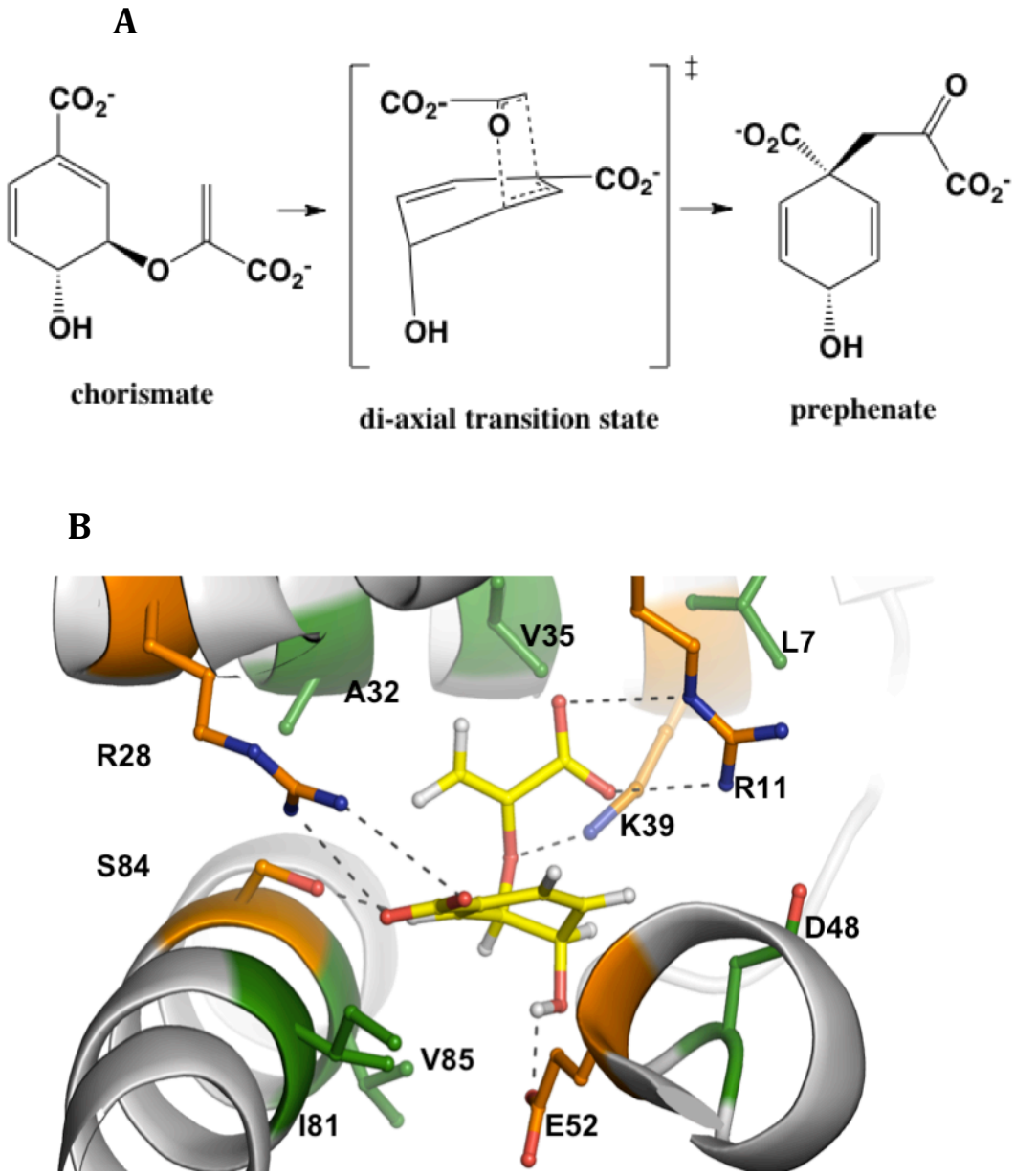


Figure 2-4. Claisen rearrangement of chorismate by chorismate mutase. (A) Reaction scheme showing the Claisen rearrangement of chorismate to prephenate with the di-axial transition state depicted. **(B)** The active site of *E. coli* chorismate mutase. Chorismate is depicted in its di-axial conformation (yellow). Catalytic residues are shown in orange. Lassila *et al.*[27] conducted site-saturation mutagenesis studies on the sites shown in green.

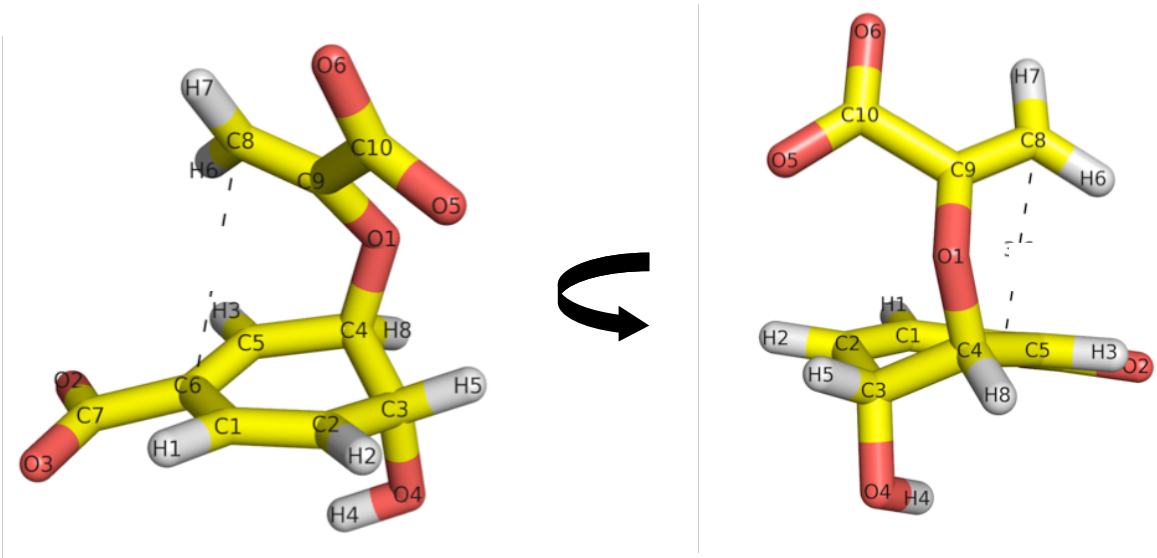


Figure 2-5. Di-axial conformation of chorismate. The substrate is depicted in its di-axial conformation. The atom names are shown. Left: distance between C8 and C6 is shown. Right: distance between C8 and C6 and atoms involved in the O4-C3-C4-O1 dihedral angle are shown.

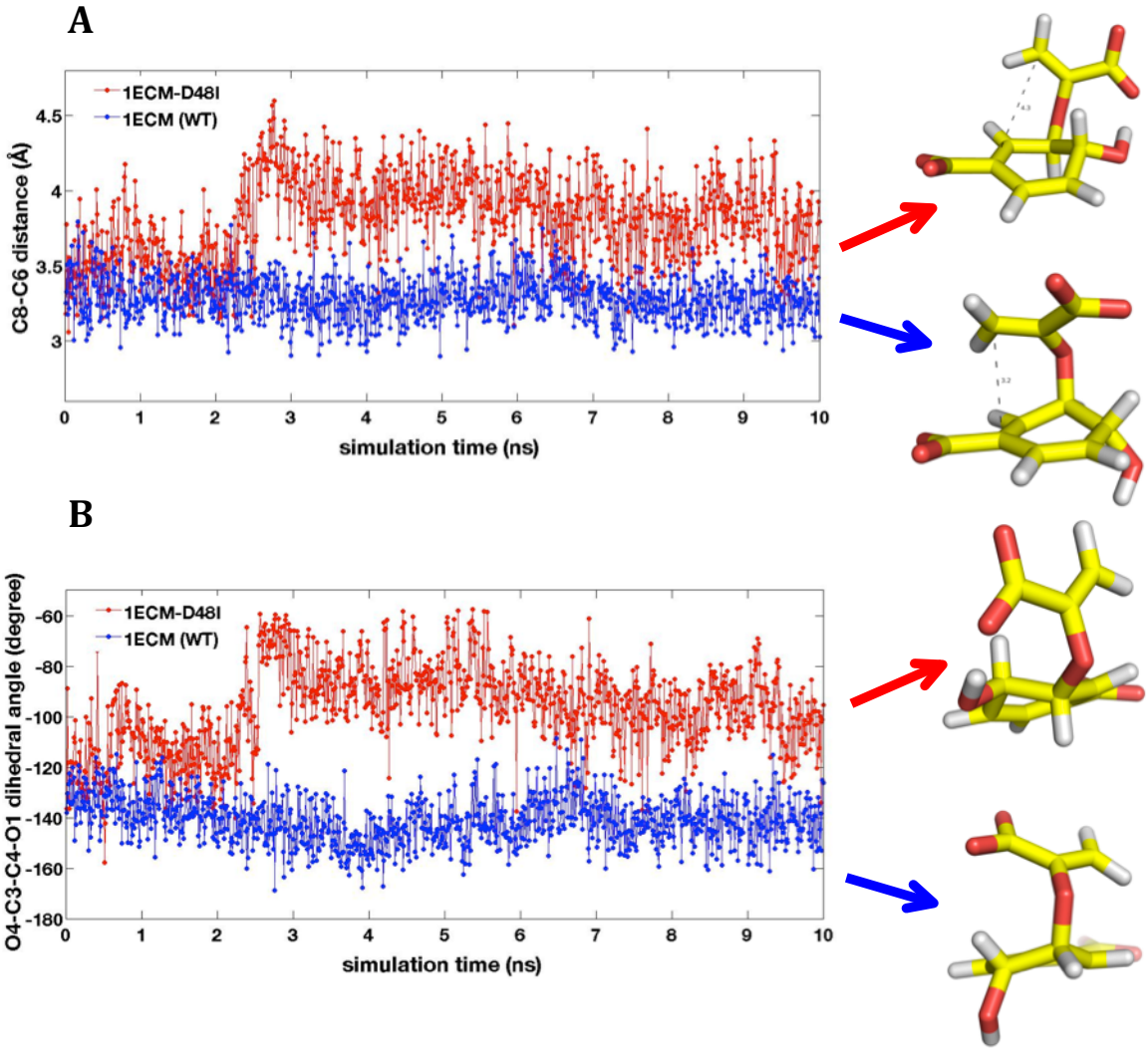


Figure 2-6. Transferability to another reaction: Examples of MD simulations results for WT and its inactive D48I variant. (A) The wild-type enzyme (WT) maintains the C8-C6 contact during the entire simulation. The D48I mutant (inactive) does not keep the C8-C6 contact. To illustrate the different conformations of the chorismate molecule, a representative snapshot of the substrate at the end of the simulation is included. **(B)** The wild-type enzyme keeps the substrate in the di-axial (reactive) conformation. In the D48I mutant, chorismate undergoes a transition to a final di-equatorial conformation (inactive conformation). To illustrate the different conformations of the chorismate molecule, a representative snapshot of the substrate at the end of the simulation is included.

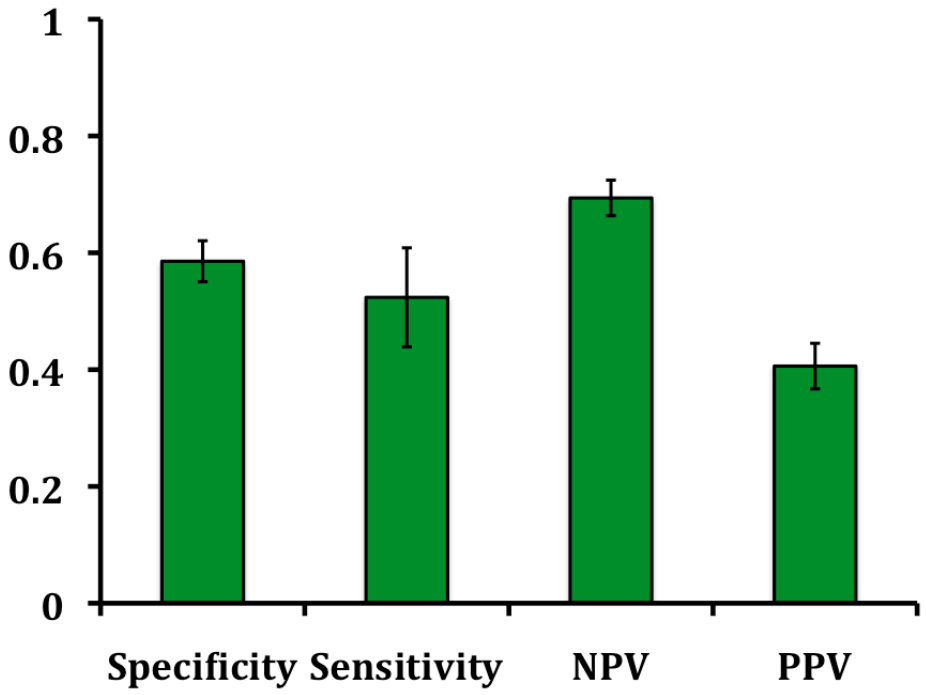


Figure 2-7. Transferability to another reaction: MD simulation results. Analysis of the results for the 114 mutants of *E. coli* chorismate mutase. NPV: negative predictive value; PPV: positive predictive value. The statistical parameters are defined in Materials and Methods. Detailed results for each design are available in the Table 2-4.

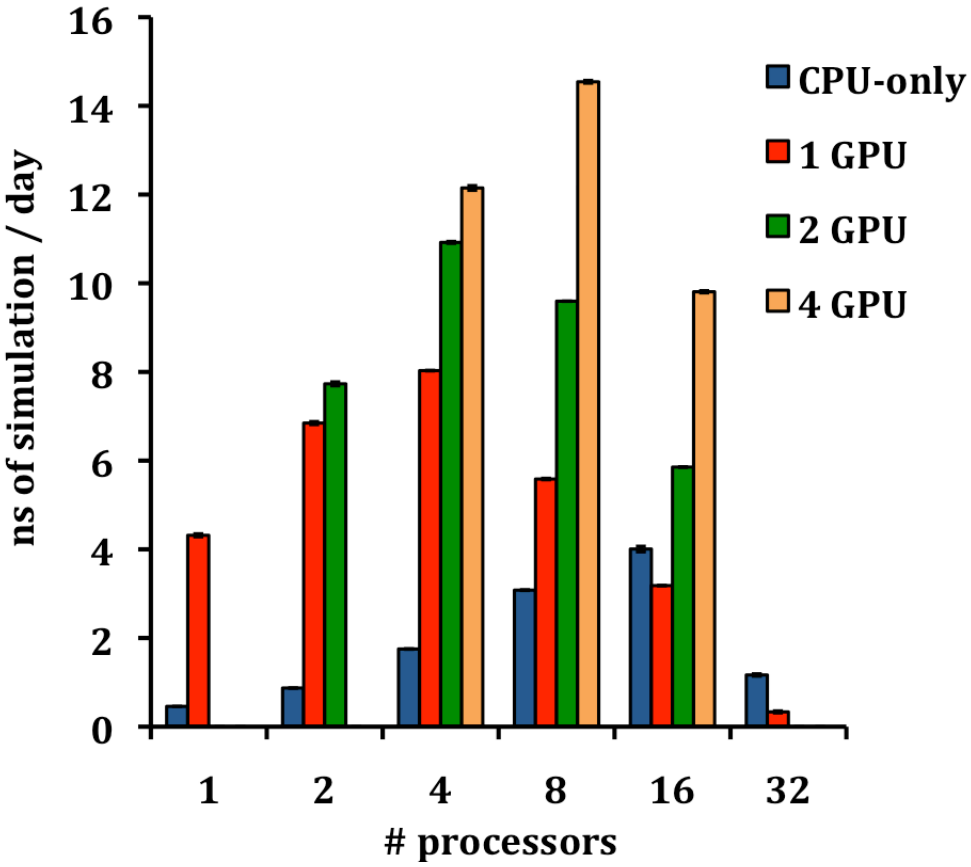


Figure 2-8. Computational performance. The simulation consisted of 32457 atoms, using 12 Å cutoff, 2 fs time step, and PME every 4 steps. NAMD 2.7[29] was used for the simulations. Several combinations of GPUs and CPUs were tested. Specifications of the computer cluster used for benchmarking Protocol 3: Dual 2.4 GHz quad core Xeon, Nvidia M2070 GPUs, and infiniband.

References

1. Bolon, D. N. & Mayo, S. L. (2001). Enzyme-like proteins by computational design. *Proc Natl Acad Sci U S A* **98**, 14274-14279.

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

3. 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., Hilvert, D., Houk, K. N., Stoddard, B. L. & Baker, D. (2008). De novo computational design of retro-aldol enzymes. *Science* **319**, 1387-1391.

4. Röthlisberger, D., Khersonsky, O., Wollacott, A., Jiang, L., DeChancie, J., Betker, J., Gallaher, J., Althoff, E., Zanghellini, A. & Dym, O. (2008). Kemp elimination catalysts by computational enzyme design. *Nature* **453**, 190-195.

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

6. Alvizo O, Allen BD & Mayo SL. (2007). Computational protein design promises to revolutionize protein engineering. *Biotechniques* **42**, 31–35.

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

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

9. Malakauskas, S. M. & Mayo, S. L. (1998). Design, structure and stability of a hyperthermophilic protein variant. *Nature Structural Biology* **5**, 470-475.

10. Privett, H. K., Kiss, G., Lee, T. M., Blomberg, R., Chica, R. A., Thomas, L. M., Hilvert, D., Houk, K. N. & Mayo, S. L. (2012). Iterative approach to computational enzyme design. *Proc Natl Acad Sci U S A* **109**, 3790-5.

11. Khare, S. D., Kipnis, Y., Greisen, P. J., Takeuchi, R., Ashani, Y., Goldsmith, M., Song, Y. F., Gallaher, J. L., Silman, I., Leader, H., Sussman, J. L., Stoddard, B. L., Tawfik, D. S. & Baker, D. (2012). Computational redesign of a mononuclear

zinc metalloenzyme for organophosphate hydrolysis. *Nature Chemical Biology* **8**, 294-300.

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

13. Frushicheva, M. P., Cao, J., Chu, Z. T. & Warshel, A. (2010). Exploring challenges in rational enzyme design by simulating the catalysis in artificial kemp eliminase. *Proc Natl Acad Sci U S A* **107**, 16869-16874.

14. Hammes-Schiffer, S. (1996). Multiconfigurational molecular dynamics with quantum transitions: Multiple proton transfer reactions. *J Chem Phys* **105**, 2236-2246.

15. Roca, M., Vardi-Kilshtain, A. & Warshel, A. (2009). Toward accurate screening in computer-aided enzyme design. *Biochemistry* **48**, 3046-3056.

16. Kumarasiri, M., Baker, G. A., Soudackov, A. V. & Hammes-Schiffer, S. (2009). Computational approach for ranking mutant enzymes according to catalytic reaction rates. *J Phys Chem B* **113**, 3579-3583.

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

18. Benkovic, S., Hammes, G. & Hammes-Schiffer, S. (2008). Free-energy landscape of enzyme catalysis. *Biochemistry* **47**, 3317-3321.

19. Kraut, D., Carroll, K. & Herschlag, D. (2003). Challenges in enzyme mechanism and energetics. *Annu Rev Biochem* **72**, 517-571.

20. Steiner, T. (2002). The hydrogen bond in the solid state. *Angewandte Chemie-International Edition* **41**, 48-76.

21. Jeffrey, G. A. (1997). An introduction to hydrogen bonding. *Oxford: Oxford University Press*.

22. Crespo, A., Scherlis, D., Martí, M., Ordejon, P., Roitberg, A. & Estrin, D. (2003). A dft-based qm-mm approach designed for the treatment of large molecular systems: Application to chorismate mutase. *J Phys Chem B* **107**, 13728-13736.

23. Lyne, P., Mulholland, A. & Richards, W. (1995). Insights into chorismate mutase catalysis from a combined qm/mm simulation of the enzyme reaction. *J Am Chem Soc* **117**, 11345-11350.

24. Wiest, O. & Houk, K. (1994). On the transition-state of the chorismate-prphenate rearrangement. *J Organic Chem* **59**, 7582-7584.

25. Wiest, O. & Houk, K. (1995). Stabilization of the transition-state of the chorismate-prephenate rearrangement: An ab-initio study of enzyme and antibody catalysis. *J Am Chem Soc* **117**, 11628-11639.

26. Lassila, J., Keeffe, J., Oelschlaeger, P. & Mayo, S. (2005). Computationally designed variants of escherichia coli chorismate mutase show altered catalytic activity. *Protein Eng Des Sel*

27. Lassila, J. K., Keeffe, J. R., Kast, P. & 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.

28. Hur, S. & Bruice, T. C. (2003). The near attack conformation approach to the study of the chorismate to prephenate reaction. *Proc Natl Acad Sci U S A* **100**, 12015-12020.

29. Phillips, J., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R., Kale, L. & Schulten, K. (2005). Scalable molecular dynamics with namd. *J Comput Chem* **26**, 1781-1802.

30. Kruse, A. C., Hu, J., Pan, A. C., Arlow, D. H., Rosenbaum, D. M., Rosemond, E., Green, H. F., Liu, T., Chae, P. S., Dror, R. O., Shaw, D. E., Weis, W. I., Wess, J. & Kobilka, B. K. (2012). Structure and dynamics of the m3 muscarinic acetylcholine receptor. *Nature* **482**, 552-6.

31. Hornak, V., Abel, R., Okur, A., Strockbine, B., Roitberg, A. & Simmerling, C. (2006). Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins* **65**, 712-725.

32. Wang, J. M., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. (2004). Development and testing of a general amber force field. *Journal of Computational Chemistry* **25**, 1157-1174.

33. Dupradeau, F. Y., Pigache, A., Zaffran, T., Savineau, C., Lelong, R., Grivel, N., Lelong, D., Rosanski, W. & Cieplak, P. (2010). The r.E.D. Tools: Advances in

resp and esp charge derivation and force field library building. *Phys Chem Chem Phys* **12**, 7821-39.

34. Vanquelef, E., Simon, S., Marquant, G., Garcia, E., Klimerak, G., Delepine, J. C., Cieplak, P. & Dupradeau, F. Y. (2011). R.E.D. Server: A web service for deriving resp and esp charges and building force field libraries for new molecules and molecular fragments. *Nucleic Acids Res* **39**, W511-7.

35. Case DA, T.A. Darden, T.E. Cheatham III, C.L. Simmerling, J. Wang, R.E. Duke, R. Luo, R.C. Walker, W. Zhang, K.M. Merz, B.P. Roberts, B. Wang, S. Hayik, A. Roitberg, G. Seabra, I. Kolossváry, K.F. Wong, F. Paesani, J. Vanicek, X. Wu, S.R. Brozell, T. Steinbrecher, H. Gohlke, Q. Cai, X. Ye, J. Wang, M.-J. Hsieh, G. Cui, D.R. Roe, D.H. Mathews, M.G. Seetin, C. Sagui, V. Babin, T. Luchko, S. Gusarov, Kovalenko, A. & Kollman, a. P. A. (2009). Amber 11. *Univeristy of California, San Francisco*.

36. Humphrey, W., Dalke, A. & Schulten, K. (1996). Vmd: Visual molecular dynamics. *J Mol Graphics* **14**, 33-38.

Chapter III

Computational Design, Molecular Dynamics Screening, and Experimental Characterization of Flexible Hinges in Triosephosphate Isomerase

Adapted from a manuscript in preparation coauthored with Jackson Cahn and Stephen L. Mayo.

Abstract

Traditional computational protein design (CPD) calculations enable the evaluation and selection of protein sequences in the context of a single native structure, single-state design (SSD). However, many proteins natively require multiple distinct conformations for functionality. Recently, a novel multi-state design (MSD) method has been developed, which can consider several states of a protein in a single optimization calculation. We sought to test this novel algorithm on triosephosphate isomerase (TIM), an enzyme that must oscillate between two conformations to be functional. Loop 6 in TIM is a lid with open and closed conformations that facilitate substrate binding, product escape, and catalysis. Two protein hinges provide the flexibility for the open-closed transition. We used both SSD and MSD methods to design the flexible hinges in TIM in the context of the closed and open conformations. Additionally, resulting enzyme designs were subjected to the molecular dynamics (MD) protocol described in Chapter II of this thesis to test the usefulness of MD-based screening. Experimental characterization of computationally designed enzyme libraries revealed that calculations favoring the closed state over the open state were more robust at producing highly and moderately active mutants. In addition, our results suggested that MSD could perform better than SSD at recapitulating naturally observed hinge sequences. Our results clearly indicate that the weighting factors for each of the states can dramatically affect the results of a calculation. Therefore, in future MSD applications, special attention should be paid to this issue. Lastly, the MD-based protocol proved to be a useful complementary tool to the CPD framework.

Introduction

Most computational protein design (CPD) methods involve the optimization of amino acid sequences in the context of a single native structure: a single state.[1; 2] This methodology has been extensively used to design proteins for increased stability, *de novo* design of enzymatic activity, design of novel protein folds, etc.[3-10]

Although most proteins are generally encountered in a single state, many must be able to assume two or more different structural or chemical states to carry out their functions.[11-15] For example, some proteins must undergo conformational changes to be functional.[14] Enzymes need to finely tune their affinities for substrates, transition states, and products to catalyze reactions; i.e., they must achieve several different chemical states.[16] In addition, proteins must not only be able to assume the final native structure but must avoid undesired states (e.g., unfolded or aggregated states).[17] To design proteins that can accommodate multiple states, traditional single-state design (SSD) methodology is not sufficient, as it only considers one state of the system in the calculations.

Recently, a new algorithm was developed for optimizing amino acid sequences in the context of multiple states.[18] Multi-state design (MSD) methods provide a computational tool for selecting protein sequences that are compatible with two or more distinct structural or chemical states. This methodology was successfully used for different design goals, from protein switches to the design of specific protein-protein interfaces.[19-22]

The enzyme triosephosphate isomerase (TIM) is a model system for the study of conformational changes in the context of enzyme catalysis.[23-27] TIM is

present in virtually every cellular organism and plays a central role in the glycolysis pathway. It catalyzes the isomerization of d-glyceraldehyde-3-phosphate (d-GAP) and dihydroxyacetone phosphate (DHAP) by means of acid-base catalysis. Its mechanism has been studied and characterized extensively. TIM loop 6 is considered to act as a lid for the active site of TIM and can assume both open and closed conformations (Figure 3-1A-B). More specifically, loop 6 undergoes a hinge-like conformational change that plays an important role in the catalytic cycle—the loop opens to let the substrate in and the product out of the active site, and also closes to bind tightly to the phosphate moiety of the substrate.[24; 26; 28; 29] Upon substrate binding and closure of loop 6 over the active site, not only the catalytic glutamate (Glu167) becomes optimally positioned for catalysis,[26] but also a glycine residue at the tip of the loop (Gly 173) makes a hydrogen bond with the phosphate moiety of DHAP (Figure 3-1C-D).[24] As a result, the DHAP molecule is tightly bound to the active site. In addition, water molecules are excluded from the vicinity of the reaction, sequestering the reaction intermediate, which would otherwise preferably react with water to eliminate phosphate and produce methyl glyoxal (MG), a highly cytotoxic material (Figure 3-1E).[26; 30] The conformational change of loop 6 is facilitated by two 3-amino acid hinges at the beginning and at the end of the loop (Figure 3-1B). Bioinformatics analysis of genomic databases has shown that TIM hinges are highly conserved in nature. Several functional and structural studies have suggested that the TIM hinges are very sensitive to mutations and are critical for catalysis.[26; 31-33] In addition, the enzyme crystallizes in two distinct conformations, open and closed (Figure 3-1A).[34]

In this work, we applied this novel MSD method to the design of TIM hinges, with the objective of testing its performance against traditional SSD approaches. To achieve our goal, we computationally designed, cloned, and experimentally characterized several libraries of enzyme variants. Additionally, to test the utility of using molecular dynamics (MD) simulations in the CPD framework, the resulting enzyme designs were also screened using the MD protocol described in Chapter II of this thesis.

Results and Discussion

Computationally designed libraries

TIM from *Trypanosome brucei brucei* (PDB ID: 6TIM) was used as the template for our design calculations. This protein crystallizes in both the closed and open conformations. For a MSD calculation, the energies of the different states have to be combined in a scoring function. The MSD scoring function was the following:

$$E_{total} = E_{open} + w\,E_{closed}$$

Where w is the weighting coefficient, E_{open} is the energy of the open state, and E_{closed} is the energy in the close state. We designed five different 24-member libraries: two SSD libraries (closed and open), which used the crystal structure for either the closed or the open conformation as the template, and three MSD libraries (closed biased, even, and open biased) in which different weighting coefficients were applied to the scoring function (Figure 3-2). The three residues comprising each of

the hinges on loop 6 were allowed to sample all 20 natural amino acids. Inspection of the predicted protein sequences reveals some interesting trends (Figure 3-2). First, in all five libraries, prolines are always chosen in the first position of the N-terminal hinge. Second, the closed library is the one that best resembles the amino acid distribution in the multiple sequence alignment (MSA) (Table 3-1); diversity is seen in the 2nd, 4th, and 5th sites, whereas the 3rd and 6th sites are conservative, choosing only the wild-type amino acid. Conversely, the open library is completely different from the MSA, as it introduces diversity in all of the sites except the first. In addition, as we increase the contribution of the open state in the calculation, the predicted sequences diverge from the MSA. Note that as the open state is given more importance, the mutational rate per mutant increases relative to the more closed-state-biased libraries, as does the diversity of previously conserved sites, i.e., sites 3 and 5 (Table 3-2). Also note that there is no overlap between the closed and closed biased libraries, but there is clear overlap between the closed-biased, even, open biased, and open libraries.

TIM activity assays

TIM variants were constructed using a round-the-horn PCR protocol, as described in Materials and Methods. The recently developed TIM-deficient Keio (DE3) strain[35] was used for *in vivo* complementation assays. After transformation and plating on agar plates, one colony was picked and grown overnight in rich liquid media. Cells were then plated onto M63 minimal plates containing L-lactate as the sole carbon source. Colonies will grow on minimal plates only if the plasmid encodes for an

active TIM.

Every member of both the closed and closed biased libraries complemented the knockout strain (Figure 3-3 and Table 3-2). Libraries with higher contributions of the open conformation had fewer mutants that exhibited complementation. Note that the closed biased library has more diversity and a higher mutational rate than the closed library, and all its mutants complemented the knockout strain. However, since the *in vivo* complementation assay read-out is dichotomic (active or inactive) and therefore not sensitive to different degrees of activity, we decided to run *in vitro* activity assays.

Designed enzymes were subsequently expressed, purified, and tested using an *in vitro* activity assay, as detailed in Materials and Methods. Remarkably, all the designs were active (Figure 3-4). Experimental screening of the closed library revealed two kinds of variants: those with wild-type-like activity and those with activity lower than wild-type. Conversely, the closed biased library included only one highly active mutant and others with moderate and low activities. Most of the mutants in the even, open biased, and open libraries were considerably less active than the wild-type enzyme.

Thermofluor stability assays

TIM variants were also subjected to the Thermofluor assay to assess thermal stability.[36; 37] Over all, mutants exhibited very similar melting curves. Apparent melting temperatures calculated from melting curves were comparable to the melting temperature (T_m) of the wild-type enzyme (Figure 3-5). This result suggests

that differences in the activity measurements are due to changes in the mechanism of the reaction and/or the efficiency of catalysis and not to the unfolding, aggregation, misfolding, or lack of structure of the mutants.

All the designs are active and retain wild-type-like stability

The *in vitro* activity assay showed that all the mutants in the five libraries are active (Figure 3-4, and Tables 3-3 to 3-7). Moreover, a high percentage (~87%) are also able to complement the knockout strain, which indicates that even though some mutants are considerably less active than the wild-type enzyme by the *in vitro* assay (variants in the open and open biased libraries), they are still active enough to support complementation.

Remarkably, all the mutants exhibited wild-type-like stability. This held true even for variants with five mutations in a sensitive part of the enzyme (Figure 3-5).

TIM is a highly evolved enzyme

TIM has been identified as one of the most evolved enzymes, and in fact, some natural TIMs exhibit diffusion-controlled kinetics. It is therefore not surprising that none of our computationally designed libraries predicted mutants with activities higher than wild-type. However, twelve mutants exhibited wild-type-like activity (Table 3-8). Bioinformatics analysis of natural TIM sequences shows that only two of our double hinge mutants (PIWVTA and PIWKTA) are found in nature. If the hinges are considered separately, within our twelve highly active mutants, there are three different N-terminal hinges (PIW, PIY and PEW), which are all found in nature.

In contrast, there are 10 different C-terminal hinges, and four of these are not found in nature (YTA, FTA, TTA, and STA).

Designing effective TIM hinges

Previous mutational studies tested the activities of exhaustive degenerate codon libraries for both TIM hinges separately.[32; 33; 38] These results concluded that the N-terminal hinge is very restrictive, with only 13 of 8000 possible sequences exhibiting activity.[38] In contrast, the C-terminal hinge is clearly more permissive, with 180 of 8000 possible sequences retaining activity.[33] In a third set of experiments, the most active N-terminal and C-terminal hinges were combined, but interestingly, the resulting double hinge variants were less active than their parent single hinge mutants.[38] In this work, we allowed both hinges to mutate simultaneously and discovered some new interesting trends. First, we were able to design double hinge mutants with wild-type-like activities, which was not possible by combining highly active single hinge mutants. The N-terminal hinge, previously thought to be restrictive, was found to accept more diverse mutations in the 2nd and 3rd sites (Figures 3-2 and 3-4). Additionally, its 1st site only accepted the amino acid proline, which is in perfect agreement with previous studies (Figures 3-2). With respect to the C-terminal hinge, our results are also in concordance with previously observed trends. The C-terminal hinge can accommodate highly diverse sequences and still produce an active enzyme.[33] Further, all the highly active hinges always exhibit the amino acid alanine in the 6th site (Table 3-8), in good agreement with previously reported mutational data.[39] Conversely, hinges that exhibit amino acids

asparagine and threonine in this site are considerably less active (Tables 3-6 and 3-7).

Comparing our results with previous studies, we conclude that the hinges clearly interact with each other. The best strategy to design active hinges is to consider the entire complexity of the hinges at the same time, instead of carrying out exhaustive mutagenesis of each hinge separately and combining the most active mutations.

Closed conformation is the most important for TIM catalysis

Comparison of the closed and open libraries suggests that CPD calculations are very sensitive to changes in the backbone structure. Relatively slight changes, like the ones observed in the two conformations of TIM, can yield extremely different results.

Comparing the activities of the various libraries shows that the closed state is the most important state for catalysis (Figure 3-4), in good agreement with previous NMR results which suggested that TIM is in the closed conformation ten times more frequently than in the open conformation.[29] Libraries that favored the closed state (closed and closed biased libraries) predicted mutants with higher activities. Conversely, libraries that favored the open state produced mutants with lower activities.

The clear dominance of the closed state prompted us to generate libraries using even stronger closed biases (Figure 3-6 and 3-7). Taking advantage of the experimental data collected in this study and the natural TIM hinge sequences in the

MSA, we counted the number of wild-type (WT), highly active (≥ 70% WT activity), moderately active (10–70% WT activity), and poorly active (< 10% WT activity) hinges within each library. This analysis revealed some interesting trends. As previously noted, MSD libraries favoring the closed state ($w > 1$) produced the best libraries overall (Figure 3-6). Libraries obtained with high weighting coefficients ($w > 100$) were essentially identical to the SSD closed library. Weighting coefficients between 13 and 100 produced more interesting libraries. These libraries not only yielded seven of the highly active hinges found in this study, but also predicted seven WT hinges present in the MSA. In addition, these libraries recovered the WT amino acid lysine in the 4^{th} site (present in 70% of the naturally occurring TIMs) and avoided valine and glutamine at the same site (Figure 3-7). These two amino acids were over-represented in the five libraries characterized in this study, as they are rarely seen at this position in naturally occurring TIM hinges (valine 3.6% and glutamine 0.8%) (Table 3-1 and Figure 3-7).

This analysis indicated that MSD could perform better than SSD at recapitulating naturally observed hinge sequences and prompted interest on two of the new libraries (Figures 3-6 and 3-7): closed biased (50) and closed biased (14). Experiments to characterize these newly designed libraries are ongoing. This analysis also raised the important question of the weighting coefficients for the MSD scoring function.[18] In future applications of MSD, especial attention needs to be paid to choosing an appropriate scoring expression and coefficients for each state, since they dramatically affect the predicted sequences and the outcome of the design calculation.

MD simulations identify the most active mutants

Lastly, to explore the idea that MD simulations can be useful as a screening tool for enzyme activity, simulations were run on every TIM variant. To assess TIM reactivity, the active site glutamate contact to the DHAP substrate was monitored. As a simulation example, we can focus our analysis on PIWVVA, a member of the closed library, and on PDHQVN, a member of the open biased library (Figure 3-8). The glutamate contact to the substrate DHAP is conserved for the highly active PIWVVA (84.5% of wild-type activity). For PDHQVN, a design that is considerably less active (5.6% of wild-type activity) and does not complement the TIM-deficient strain, simulations revealed that the contact was not maintained, evolving to non-catalytic Glu-DHAP distances (> 3.2 Å) (Figure 3-8A). In addition, the hydrogen bond between Gly 173 (loop 6 residue) and the PO_4 of DHAP was monitored. While PIWVVA is capable of keeping the contact during the entire simulation, it is evident that PDHQVN is not (Figure 3-8B). This suggests that a design that was obtained by favoring the open state does not rigidly stay in the closed conformation, but instead samples the open conformation more frequently. The overall results indicate that MD can effectively screen the libraries studied here, predicting the twelve most active designs correctly while discarding almost half of the poorly active hinge mutants (specificity ~ 0.5) (Figure 3-8C and Table 3-9).

Conclusions

In this work, we computationally designed five TIM hinge libraries using SSD and MSD methods. Experimental characterization of the designs showed that the SSD

closed library outperformed the other four libraries, suggesting that the closed conformation of TIM is the most important for catalysis.

Our results indicate that CPD calculations are extremely sensitive to protein backbone alterations. Even though the two conformations of TIM are overall very similar, when used separately as input to design calculations, they can yield dramatically different results.

Although the SSD closed library was the most robust at predicting active sequences, MSD can also produce libraries with active mutants. Our results strongly suggest that MSD calculations are very sensitive to how the states being considered are balanced. The closed biased library was consistently better than the even and open biased libraries, suggesting that in future MSD applications, special care should be taken in choosing the relative importance of the involved states. In addition, our results suggested that MSD could perform better than SSD at recapitulating naturally observed hinge sequences.

Remarkably, all the hinge mutants showed some level of activity and 95% were able to complement the TIM-deficient strain. Stability measurements demonstrated that all the mutants were folded and retained wild-type-like apparent T_ms, indicating that differences in activity cannot be attributed to aggregation or misfolding of the enzyme. Moreover, the twelve most active mutants exhibited wild-type-like activities. Bioinformatics analysis of known TIMs revealed that nine of the highly active mutants are not present in nature.

Finally, screening with MD simulations proved to be a useful complement to CPD by reducing the experimental effort associated with finding the most active

mutants. This suggests that inclusion of MD analysis directly into sequence optimization calculations may improve CPD accuracy.

Materials and Methods

Computational protein design calculations

All computational design was done using the Triad computational protein design platform (Protabit LLC, Pasadena, CA). At each step, care was taken to use commonly used parameters and techniques.

The PDB structure 6TIM[34] contains two monomers of TIM, one in the open conformation and one in the closed conformation. Each monomer was stripped of waters, ligands, and alternate conformations before the addition of hydrogens and the optimization of polar hydrogens. The substrate glycerol-3-phosphate was then replaced into the active site of both monomers (using symmetry for the open form) and the protein atoms were then subjected to complete Cartesian minimization using the Biograf energy function.

Computational protein design of the hinges was carried out using the Rosetta forcefield.[40] The six positions of the hinges were allowed to sample all amino acids and all rotamers, and the other residues of the loop (two residues on each side of the hinges and five residues between the hinges) were allowed to sample repack rotamers but not change residue identity. Changes in residue identity of the hinge residues conserved between the two monomers, but rotamers were repacked independently using the FASTER multistate algorithm.[18] The scoring function for multistate design was:

$$E_{total} = E_{open} + w\,E_{closed}$$

where E_{open} and E_{closed} are the Rosetta energies and w is the weighting factor. Weighting factor for MSD libraries were: closed biased (w = 5), even (w = 1) and open biased (w = 0.2). Design calculations were also done using $E_{total} = E_{open}$ and $E_{total} = E_{closed}$ (SSD) for comparison.

MD simulations and criterion

The Amber force field was used in all simulations; Amber force field ff99SB was implemented for proteins, and the general Amber force field was used for small molecules.[41; 42] For DHAP, we calculated charges using the restrained electrostatic potential (RESP) method at HF/6-31G* as implemented in the REDS server.[43; 44] AmberTools 10 was used to prepare the structures for simulations.[41] Hydrogen atoms were added to the structures after which the systems were completely solvated with water molecules (TIP3P water model). Finally, Na or Cl ions were added to neutralize the systems. The simulation cell shape was a truncated octahedron.

The simulations were carried out using the MD software NAMD 2.7.[45] Pressure was kept constant using the Nose-Hoover Langevin piston method with a damping time of 50 fs and a decay period of 100 fs. Langevin dynamics was used with a target constant temperature of 300 K and a 5 ps^{-1} damping coefficient. We took advantage of the multiple time-stepping capabilities of the code: a 2 fs time step was used, non-bonded interactions (van der Waals and electrostatic) were calculated every two steps, and long range electrostatic interactions [particle mesh

Ewald (PME) with periodic boundary conditions (PBC)] were calculated every four steps. Visual molecular dynamics (VMD) was used to visualize and analyze the simulations.[46]

Minimization (1000 steps) were used to initialize the simulations. Subsequently, the systems were equilibrated for 60 ps at target temperature (300K), while harmonic restraints were applied on all non-water, non-hydrogen atoms during equilibration. The force constants for these restraints were gradually released from 50 kcal/mol/A^2 using the following loop: $K=50/4^i$, i = 0, 1, 2 … 5. Finally, a production step was carried out for 10 ns with no restraints applied to the system.

For the purpose of comparing MD simulations results with experimental results, mutants that exhibited less than 10% of wild-type activity by the *in vitro* assay were assumed to be experimentally inactive. The catalytic hydrogen bond contact of glutamate to the hydroxyl proton of DHAP was monitored during the simulations. The cutoff distance for an effective hydrogen bond is 3.2 Å. Mutants that kept the Glu-DHAP contact below the cut-off distance for 90% of the simulation time were ranked as active, otherwise inactive.

Cloning

Trypanosome brucei brucei TIM was assembled into pET-53-Dest omitting loop 6. This molecule was used as template for later mutagenesis to generate library variants. Forward and reverse primers were designed to add the different hinges and the inner part of the loop. After Round-the-horn PCR reaction with appropriate

primers and Phusion DNA polymerase (Fischer), Dpn I (NEB) treatment for 1 hour at 37˚C was implemented to eliminate the template plasmid. Resulting DNA was purified using the minElute PCR purification kit (Qiagen) and eluted with 15 μL of water. 8 micro liters of the elution were heated at 70 ˚C for 10 minutes to melt the ends of the DNA fragment. Subsequently, 1 μL of T4 Ligase Buffer (NEB) and 1uL T4 Polynucleotide Kinase (NEB) were added and the mixture incubated at 37 ˚C for 45 minutes to Phosphorylate double stranded DNA. Ligation reactions were prepared by mixing 10 μL of the Phosphorylation reaction and 1 μL of T4 Ligase (NEB). After 2 hours at room temperature, the entire ligation reaction volume was transformed into TOP10 chemically competent cells. The transformation product was recovered for 1 hour at 37 ˚ C and plated onto LB plates supplemented with ampicillin (plasmid selection). Colonies were picked and DNA was isolated to confirm identity by sequencing (Beckman Coulter Genomics). The sequence of the complete gene of *Trypanosome brucei brucei* TIM is the following (hinges highlighted in bold):

MHHHHHHGGSGGENLYFQGGGSGGMSKPQPIAAANWKCNGSQQSLSELIDLFNSTSINH
DVQCVVASTFVHLAMTKERLSHPKFVIAAQNAIAKSGAFTGEVSLPILKDFGVNWIVLGHS
ERRAYYGETNEIVADKVAAAVASGFMVIACIGETLQERESGRTAVVVLTQIAAIAKKLKKA
DWAKVVIAYE**PVW**AIGTG**KVA**TPQQAQEAHALIRSWVSSKIGADVAGELRILYGGSVNGK
NARTLYQQRDVNGFLVGGASLKPEFVDIIKATQ*

Protein expression and purification

Proteins were expressed and purified from the TIM-knockout Keio (DE3) cells to prevent the possible contamination with endogenous TIM. Chemically competent

Keio (DE3) cells were transformed with plasmids encoding TIM variants and plated onto LB-agar plates with antibiotics. One colony was picked to start overnight cultures in 2xYT liquid media on 96-well plates. Ten micro liters of saturated culture was used to inoculate 4 mL of auto induce Over Night Express TB media (Novagen) on 24 well plates (Whatman). Cells were grown for 16 hours at 30 °C with continuous shaking after which, they were harvested by centrifugation (25 minutes at 1000 rpm). Cell pellets were resuspended in 400 µL of lysis buffer: 3x Cell Lytic B (Sigma), 300 mM NaCl, 50 mM Tris-HCl, 5 mM imidazole, 5 mM $MgCl_2$, 50 units mL^{-1} benzonase (Sigma), 0.2 mg mL^{-1} lysozyme, and pH 7.8. After 1 hour at room temperature, cellular debris was separated from the soluble part of the lysate by centrifugation (25 minutes at 1000 rpm). Purification was done using a His-select 96-well filter plate (Sigma). Equilibration of the plate resin was done with 300 mM NaCl, 50 mM Tris-HCl, 5 mM imidazole, pH 7.8. After sample loading, two washes with 600 µL of wash buffer (300 mM NaCl, 50 mM Tris-HCl, 20 mM imidazole, pH= 7.8) were done. Finally elution was done into 500 µL of 300 mM NaCl, 50 mM Tris-HCl, 150 mM Imidazole, pH= 7.8. All centrifugation steps were done at 4 °C and 1000 rpm for 2 minutes. Protein purity was confirmed by SDS-PAGE (Figure 3-9). Protein concentration was determined by measuring absorbance at 280 nm on a microplate reader (Tecan) using UV star microplates (Grenier bio-one).

Activity assay

Activity assays were carried out as described previously by Richard and Magliery Labs.[35; 47] Given the number of designs that were tested, the assay was run on liquid

handling robot (Tecan EVO 200) equipped with a microplate reader (Tecan). UV star microplates (Grenier bio-one) were used to be able to monitor confidently at 340 nm. d-GAP (Sigma) was used as substrate and α-glycerolphosphate dehydrogenase (Sigma) was used as coupling enzyme at 3.75 μg/mL. Assay buffer was composed of 0.1 M triethanolamine (TEA) pH 7.4, 5 mM ethylenediaminetetraacetic acid (EDTA), 5 mM nicotinamide adenine dinucleotide (NADH), and 0.5 mM of d-GAP. First, 5 μL of coupling enzyme were placed at each well of plate after which 180 μL of the assay buffer were added. Background absorbance was monitored at 340 nm for ten minutes. Finally, reactions were initiated by addition of 15 μL of enzyme samples, and the entire reaction volume was pipetted three times to ensure mixing. Absorbance at 340 nm was monitored for another 10 minutes. Absorbance rates before and after sample addition were determined using an NADH extinction coefficient was $\varepsilon = 6220\ M^{-1}\ cm^{-1}$. Observed rates were calculated by subtracting the rates before addition of sample from the rate after addition of sample ($V_{obs} = V_{after} - V_{before}$). All assays were run at room temperature. Specific activities of mutants are reported as % of wild type activity.

In vivo complementation assay

The Keio (DE3) cells[35] containing plasmids encoding for TIM variants were grown overnight at 37°C, washed twice with M63 salts (pH 7), and plated on M63 minimal plates. Minimal plates contained kanamycin for strain selection and ampicillin for plasmid selection. Plates were also supplemented with 0.2% w/v lactate as the sole carbon source, 1 mg L^{-1} thiamine, 80 mg L^{-1} histidine, 50 mg L^{-1} uracil, 1mM $MgSO_4$.

Plates were incubated at 37 ˚C for 48 hours. Keio DE3 cells were a generous gift of Tom Magliery.

Protein stability measurements by Thermofluor assay

The Thermofluor assay has been described before.[36; 37] SYPRO® Orange dye (Life Technologies, concentration not disclosed by provider) was diluted 50 times in elution buffer. 5 microliters of diluted dye were added to the 45 μL of protein samples. qPCR equipment (BIO-RAD) was used to run the melting experiments from 25°C to 99°C. Protein concentration was 10 μM for all wells.

Acknowledgments

The authors would like to thank Marie Ary for technical assistance with the manuscript and Tom Magliery for the generous gift of the Keio (DE3) cells.

Table 3-1. Multiple sequence alignment of natural TIMs' loop 6.

N-Terminal Hinge	Inner Loop	C-Terminal Hinge	Number of Sequences
PVW	AIGTG	KTA	92
PIW	AIGTG	KSA	58
PVW	AIGTG	KVA	58
PIW	AIGTG	KTA	57
PVW	AIGTG	KSA	46
PIW	AIGTG	KSS	42
PIW	AIGTG	KTC	19
PLW	AVGTG	KSA	18
PVW	AIGTG	RVA	18
PVW	AIGTG	LTA	15
PVW	AIGTG	RTA	13
PVW	AIGTG	LTP	12
PIW	AIGTG	KAA	11
PVW	AIGTG	LAA	9
PVW	AIGTG	VVA	9
PIW	AIGTG	DTC	8
PIW	AIGTG	RTA	6
PIW	AIGTK	KSA	6
PVW	AIGTG	KVP	6
PIW	AIGTG	KVA	4
PVW	AIGTG	ETA	4
PVW	AIGTG	KNA	4
PVW	AIGTG	RHA	4
PVW	AIGTG	RTP	4
PIW	AIGTG	KNA	3
PIY	SIGTG	VSA	3
PLW	AIGTG	KTA	3
PVW	AIGTG	DTA	3
PVW	AIGTG	MVA	3
PEW	AIGKA	SSA	2
PIW	AIGTD	LEL	2
PIW	AIGTG	CTA	2
PIW	AIGTG	HVP	2
PIW	AIGTG	VSA	2
PSS	AINSG	NCA	2
PVW	AIGTG	HSA	2
PVW	AIGTG	KVC	2
PVW	SIGTG	VVA	2

PEW	AIGKA	NSA	1
PIW	AIATG	KSA	1
PIW	AIGSG	ASA	1
PIW	AIGTG	DTA	1
PIW	AIGTG	ESA	1
PIW	AIGTG	ETA	1
PIW	AIGTG	ETC	1
PIW	AIGTG	HSA	1
PIW	AIGTG	LIP	1
PIW	AIGTG	LTP	1
PIW	AIGTG	LVA	1
PIW	AIGTG	NVP	1
PIW	AIGTG	QTC	1
PIW	AIGTG	RSA	1
PIW	AIGTG	RTP	1
PIW	AIGTG	VTA	1
PIW	AIGTG	VTP	1
PIW	SIGSD	MIP	1
PIW	SIGSG	NSA	1
PIW	SIGTG	VSA	1
PKW	AIGTN	TIP	1
PLW	AIGTG	RTA	1
PRW	AIGTG	KTP	1
PSW	AIGSG	TSA	1
PTW	AIGKK	DSA	1
PVW	AIGKP	QPA	1
PVW	AIGNG	QNA	1
PVW	AIGSG	KAA	1
PVW	AIGSG	LTA	1
PVW	AIGSG	NPA	1
PVW	AIGSG	QAA	1
PVW	AIGSG	SSA	1
PVW	AIGSN	TIP	1
PVW	AIGTG	AVA	1
PVW	AIGTG	EVA	1
PVW	AIGTG	ITA	1
PVW	AIGTG	KAA	1
PVW	AIGTG	KSS	1
PVW	AIGTG	LSA	1
PVW	AIGTG	LVP	1
PVW	AIGTG	NSP	1
PVW	AIGTG	NTA	1
PVW	AIGTG	QTA	1

PVW	AIGTG	RNA	1
PVW	AIGTG	RNC	1
PVW	AIGTG	RSA	1
PVW	AIGTG	RVP	1
PVW	AIGTG	TPA	1
PVW	AIGTG	VAA	1
PVW	AIGTG	VTA	1
PVW	AIGTN	RTA	1
PVW	AIGTN	VTA	1
PVW	AVGTG	NTA	1
PVW	AVGTG	RSA	1
PVW	SIGSC	MVP	1
PVW	SIGTG	ITP	1

Table 3-2. *In vivo* complementation assay results. Every mutant in the closed and closed biased libraries complemented the TIM-deficient strain. Libraries with a higher contribution of the open state (even, open biased, and open) are less effective at complementing the TIM-deficient strain. Mutational rate increases as the open state is given more importance in the design calculation.

Library	Average number of mutations per design	Percent of mutants that complement TIM-deficient strain
Closed	2.88	100
Closed biased	3.67	100
Even	4.00	95.83
Open biased	4.38	66.67
Open	4.33	70.83

Table 3-3. Summary of experimental characterization of closed library. MD metric is the proportion of frames of a 10 ns simulation that exhibit the catalytic contact (Glu-DHAP) ≤ 3.2 Å. Apparent T_ms should be considered approximate since thermal unfolding was not reversible. Specific activity is reported as percent of wt activity.

Hinge sequences	Percent WT activity	*In vivo* complementation assay	Apparent T_m (°C)	MD metric
PIWVTA	96.89	+	50	0.999
PIWITA	75.90	+	50	1.000
PIWKTA	71.57	+	50	0.994
PIWLTA	74.58	+	50.3	0.982
PIWTTA	80.24	+	50	0.996
PIWVVA	84.48	+	50	0.921
PQWVTA	27.73	+	46	0.987
PIWRTA	67.36	+	50	0.656
PIWQTA	94.28	+	50	0.998
PVWVTA	22.39	+	44	0.974
PMWVTA	32.45	+	47	0.205
PLWVTA	29.41	+	47	0.846
PIWETA	49.85	+	51	0.993
PRWVTA	46.51	+	51	0.258
PIWYTA	85.14	+	50.7	1.000
PIWFTA	98.39	+	50	0.995
PEWVTA	76.67	+	50	1.000
PYWVTA	20.91	+	44	0.818
PIWVSA	60.47	+	50	0.867
PAWVTA	45.74	+	50	0.382
PIWVAA	43.91	+	50	0.995
PFWVTA	43.07	+	50	0.820
PIWNTA	47.10	+	50	0.988
PIWSTA	86.76	+	49.7	0.998

Table 3-4. Summary of experimental characterization of closed biased library. MD metric is the proportion of frames of a 10 ns simulation that exhibit the catalytic contact (Glu-DHAP) ≤ 3.2 Å. Apparent T_ms should be considered approximate since thermal unfolding was not reversible. Specific activity is reported as percent of wt activity.

Hinge sequences	Percent WT activity	*In vivo* complementation assay	Apparent T_m (˚C)	MD metric
PIFQTA	58.92	+	49	0.087
PIFKTA	60.77	+	49.3	0.998
PIFRTA	2.38	+	51	0.953
PMFQTA	15.53	+	51	0.967
PIFQVA	29.15	+	48	0.333
PQFQTA	38.36	+	48	0.951
PIFETA	30.10	+	48	0.951
PVFQTA	43.78	+	49	0.054
PIFTTA	44.88	+	48.7	0.164
PIHQTA	26.62	+	48.7	0.988
PLFQTA	53.95	+	50	0.999
PEFQTA	2.38	+	45.7	0.986
PIWKTG	55.08	+	49	0.758
PIFLTA	50.42	+	50	0.994
PRFQTA	26.01	+	50	0.949
PIWQTG	37.68	+	49	0.994
PIYKTA	89.37	+	50	0.964
PMWKTG	35.85	+	47	0.249
PAFQTA	2.38	+	51	0.134
PYFQTA	2.38	+	49	1.000
PVWKTG	43.00	+	50	0.981
PIFQSA	58.21	+	49	0.130
PIFMTA	35.19	+	46	0.999
PLWKTG	2.80	+	47	0.432

Table 3-5. Summary of experimental characterization of even library. MD metric is the proportion of frames of a 10 ns simulation that exhibit the catalytic contact (Glu-DHAP) ≤ 3.2 Å. Apparent T_ms should be considered approximate since thermal unfolding was not reversible. Specific activity is reported as percent of wt activity.

Hinge sequences	Percent WT activity	*In vivo* complementation assay	Apparent T_m (˚C)	MD metric
PMFQTA	15.53	+	51	0.967
PMFQVA	20.39	+	51	0.587
PMFKTA	16.00	+	47	0.999
PMHQTA	24.45	+	48.7	0.993
PEFQTA	2.38	+	45.7	0.986
PIFQTA	58.92	+	49	0.087
PVFQTA	43.78	+	49	0.054
PMFETA	13.94	+	49	0.518
PMHQTN	28.84	+	48.7	0.980
PMFQTS	11.09	+	48	0.998
PMFRTA	11.58	+	48.3	0.730
PQFQTA	38.36	+	48	0.951
PLFQTA	53.95	+	50	0.999
PAFQTA	2.38	+	51	0.134
PRFQTA	26.01	+	50	0.949
PMFQSA	34.25	+	51	1.000
PMFQIA	17.02	+	48.7	0.999
PMHQTD	6.82	-	47.7	0.219
PMFMTA	13.50	+	52	0.137
PDFQTA	15.20	+	49	0.115
PMFQYA	7.52	+	50	0.995
PYFQTA	2.38	+	49	1.000
PMFQAA	34.92	+	50.7	0.977
PKFQTA	19.06	+	50.7	0.989

Table 3-6. Summary of experimental characterization of open biased library. MD metric is the proportion of frames of a 10 ns simulation that exhibit the catalytic contact (Glu-DHAP) ≤ 3.2 Å. Apparent T_ms should be considered approximate since thermal unfolding was not reversible. Specific activity is reported as percent of wt activity.

Hinge sequences	Percent WT activity	*In vivo* complementation assay	Apparent T_m (˚C)	MD metric
PMHQVN	2.38	–	51	0.205
PMFQVA	20.39	+	51	0.587
PMHQTN	28.84	+	48.7	0.980
PMFQVS	10.81	+	48.7	0.996
PMHQIN	2.38	–	47	0.894
PMHQVD	4.32	+	48	0.994
PMHQYN	2.38	+	47.7	0.306
PMHKTN	9.96	+	48	0.991
PEHQVN	2.38	+	45.7	0.651
PMHQFN	2.38	+	47.7	0.951
PMHETN	2.38	+	51	0.989
PMHQSN	8.35	+	48	0.995
PMFQSA	34.25	+	51	1.000
PVHQVN	4.24	–	50	0.978
PMHQQN	2.38	–	46.7	0.993
PAHQVN	2.38	–	50	0.998
PMHQKN	2.38	+	47	0.254
PMHQEN	2.38	–	47	0.511
PMHQNN	2.38	+	47	0.998
PMHQRN	4.79	+	51	0.419
PMHMTN	2.38	+	46.3	0.174
PMFQRA	8.32	+	48	0.358
PDHQVN	5.58	–	47.7	0.407
PMHQVT	2.38	–	51	0.615

Table 3-7. Summary of experimental characterization of open library. MD metric is the proportion of frames of a 10 ns simulation that exhibit the catalytic contact (Glu-DHAP) ≤ 3.2 Å. Apparent T_ms should be considered approximate since thermal unfolding was not reversible. Specific activity is reported as percent of wt activity.

Hinge sequences	Percent WT activity	*In vivo* complementation assay	Apparent T_m (˚C)	MD metric
PMHQVN	2.38	–	51	0.205
PMHQTN	28.84	+	48.7	0.980
PMHQVT	2.38	–	51	0.615
PMHQIN	2.38	–	47	0.894
PMFQVA	20.39	+	51	0.587
PMAQVT	2.38	+	47	0.987
PMSQVT	2.38	+	47.3	0.709
PMHQYN	2.38	+	47.7	0.306
PMFQVS	10.81	+	48.7	0.996
PMAQTT	2.38	+	47.3	0.931
PMHQFN	2.38	+	47.7	0.951
PMHQVD	4.32	+	48	0.994
PMHQLN	2.38	+	47	0.028
PMHQFT	2.45	–	51	0.987
PMEQVT	5.67	+	48	0.998
PMHQLT	2.38	+	47	0.103
PEHQVN	2.38	+	45.7	0.651
PMHKVN	9.28	+	48	0.052
PMHQQN	2.38	–	46.7	0.993
PEHQVT	2.38	+	44.3	0.312
PMHQKN	2.38	+	47	0.254
PMHKVT	2.38	+	48	0.816
PMHQEN	2.38	–	47	0.511
PMAQLT	2.38	–	48	0.999

Table 3-8. Summary of the most active mutants. *In vitro* activity is shown as percent of the wild-type (WT) enzyme activity. The MD protocol predicted all the hinge mutants to be active. Comparison of the most active hinges with the MSA (Table 3-2) indicates that out of the twelve most active hinges, ten are not present in nature. MD = molecular dynamics predictions. A= active.

Most active hinges	Activity	Library	Natural double hinge mutant?	Natural N-terminal hinge?	Natural C-terminal hinge?	MD
PIWFTA	98.4	Closed	no	yes	no	A
PIWVTA	96.9	Closed	yes	yes	yes	A
PIWQTA	94.3	Closed	no	yes	yes	A
PIYKTA	89.4	Closed Biased	no	yes	yes	A
PIWSTA	86.8	Closed	no	yes	no	A
PIWYTA	85.1	Closed	no	yes	no	A
PIWVVA	84.5	Closed	no	yes	yes	A
PIWTTA	80.2	Closed	no	yes	no	A
PEWVTA	76.7	Closed	no	yes	yes	A
PIWITA	75.9	Closed	no	yes	yes	A
PIWLTA	74.6	Closed	no	yes	yes	A
PIWKTA	71.6	Closed	yes	yes	yes	A

Table 3-9. Comparison of MD results with CPD results. NA = not applicable.

	CPD alone	After MD screening
# true positive designs	58	37
# true negative designs	NA	20
# false positive designs	38	18
# false negative designs	NA	21
# designs to characterize	96	55
# designs to discard	NA	41

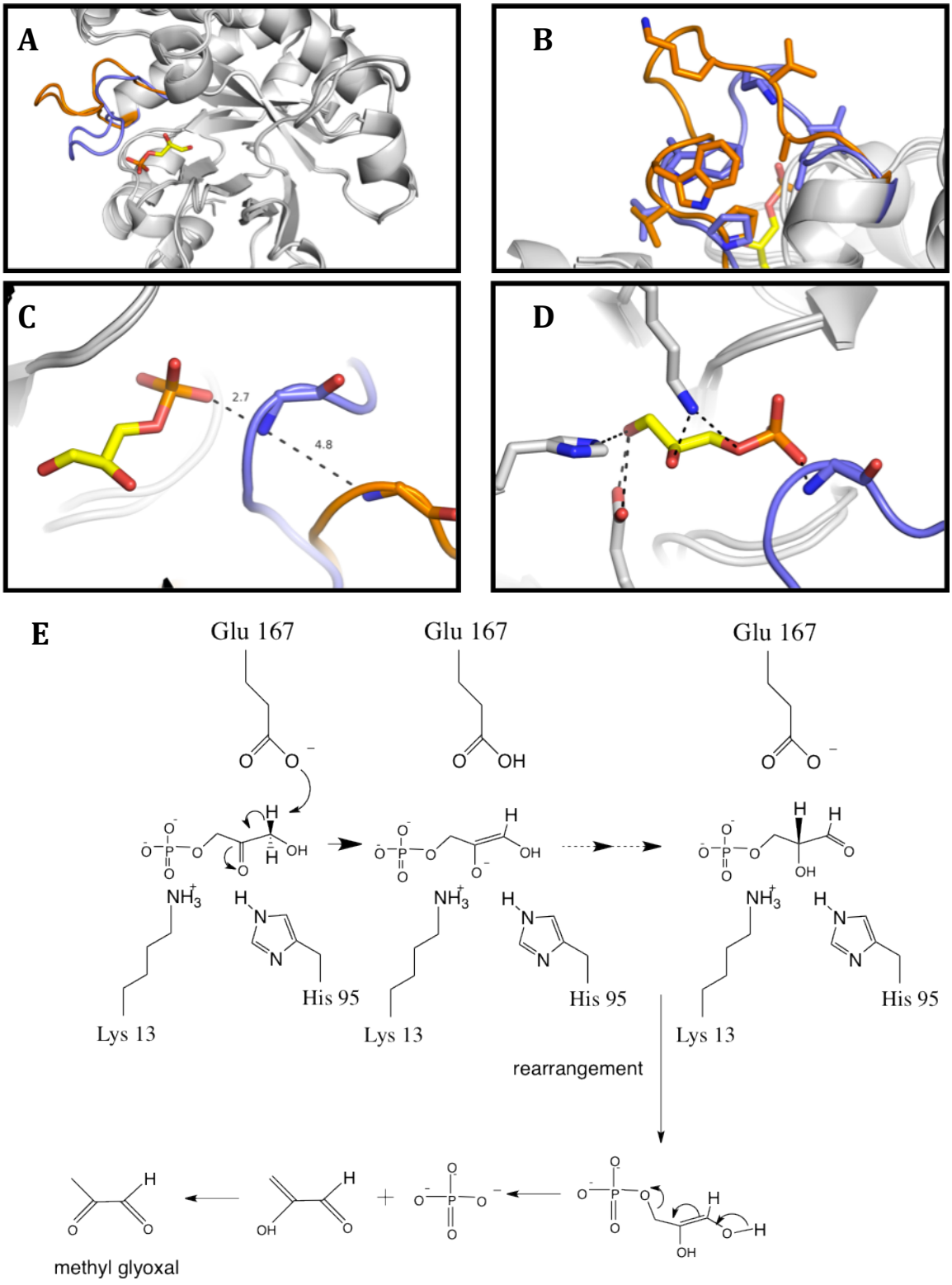


Figure 3-1. Triosephosphate isomerase as a model system for conformational changes in enzymatic catalysis. (A) Overall structure of TIM depicting the conformational change between the open (orange cartoon) and closed (purple cartoon) states**. (B)** TIM loop 6 emphasizing the N-terminal (PVW) and C-terminal

(KVA) hinges at the beginning and end of the loop, respectively. The hinges provide the flexibility for the conformational change. **(C)** Gly-173's backbone hydrogen bond to the DHAP phosphate moiety provides the tight phosphate binding and the sequestration of the reaction intermediate. **(D)** TIM active site depicting catalytic residues (Glu-167, His-95 and Lys-13). **(E)** The TIM reaction scheme. Loop 6's closed conformation is responsible for sequestering the reaction intermediate, which would otherwise react with water and eliminates phosphate producing methyl glyoxal, a highly cytotoxic molecule.

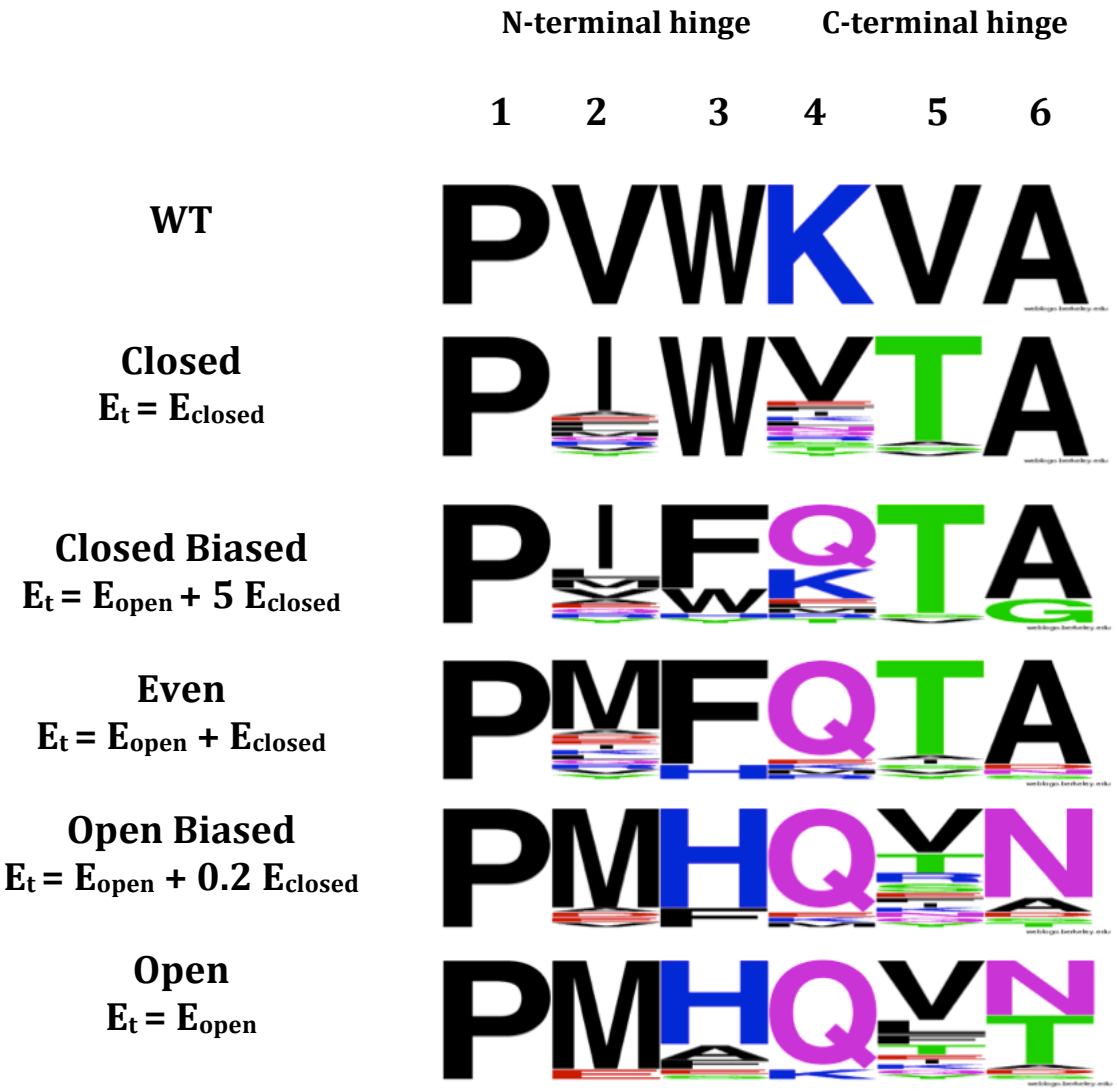


Figure 3-2. Summary of the computationally designed TIM hinge libraries. Mutational frequency logos are shown for the two SSD and the three MSD libraries. The scoring function is shown for each library. The closed library is the one that best resembles the MSA of natural TIMs (Table 3-1), directing diversity to sites 2, 4, and 5 and choosing the wild-type amino acid in the other sites. As the open state is given more importance in the design calculation, diversity is directed to all sites except for the first site, which is always the amino acid proline.

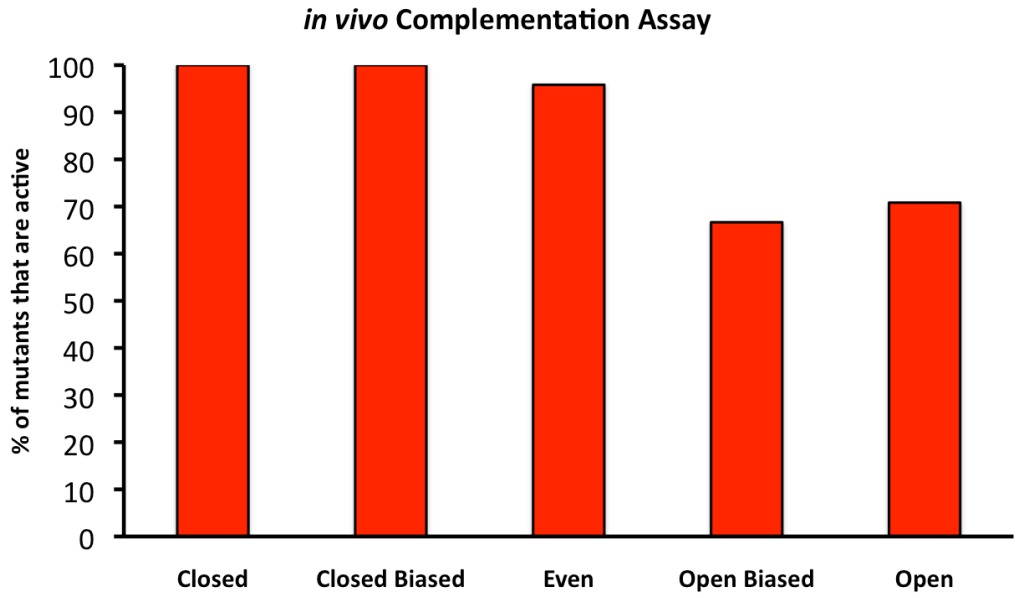


Figure 3-3. *In vivo* complementation assay results for the designed libraries. Every mutant in the closed and closed biased libraries complemented the TIM-deficient strain. Libraries with a higher contribution of the open state (even, open biased, and open) are less effective at complementing the TIM-deficient strain.

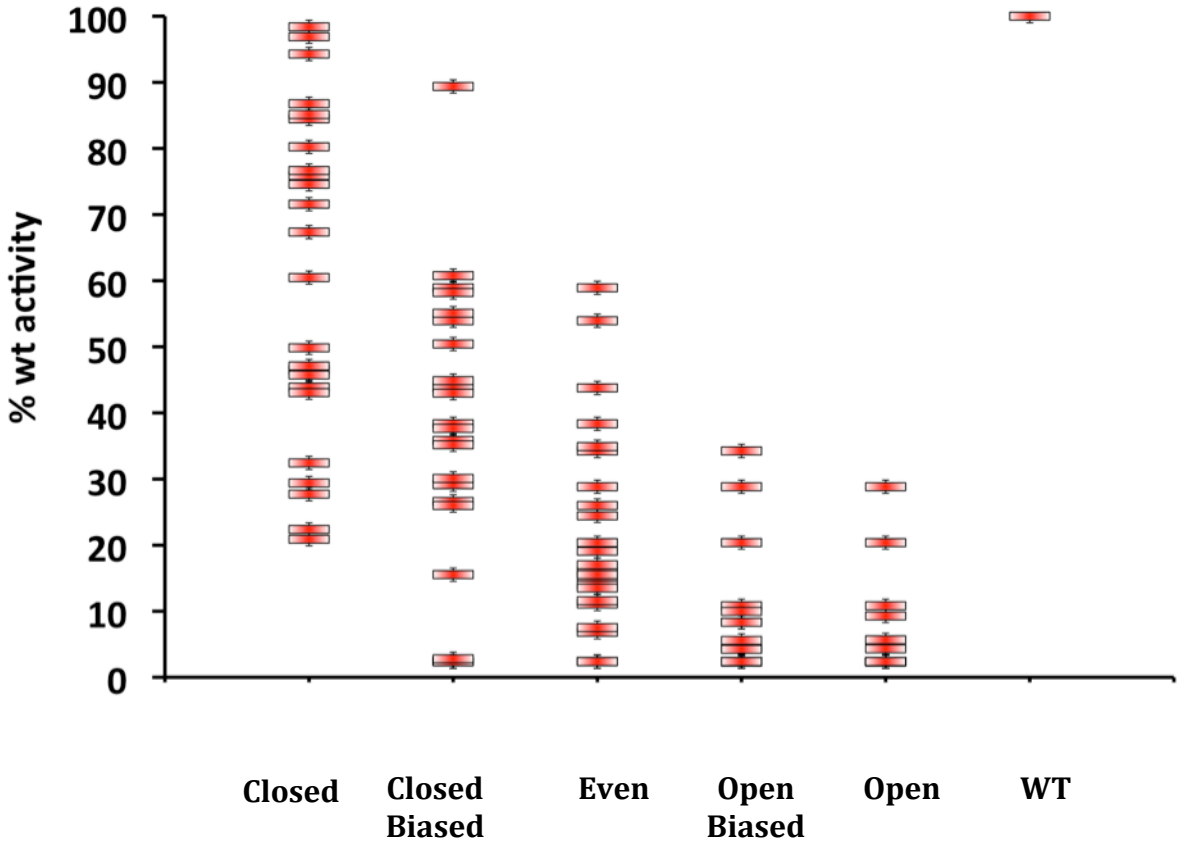


Figure 3-4. *In vitro* activity assay results for each of the libraries. All mutants showed some level of activity. The closed library outperformed the other four libraries.

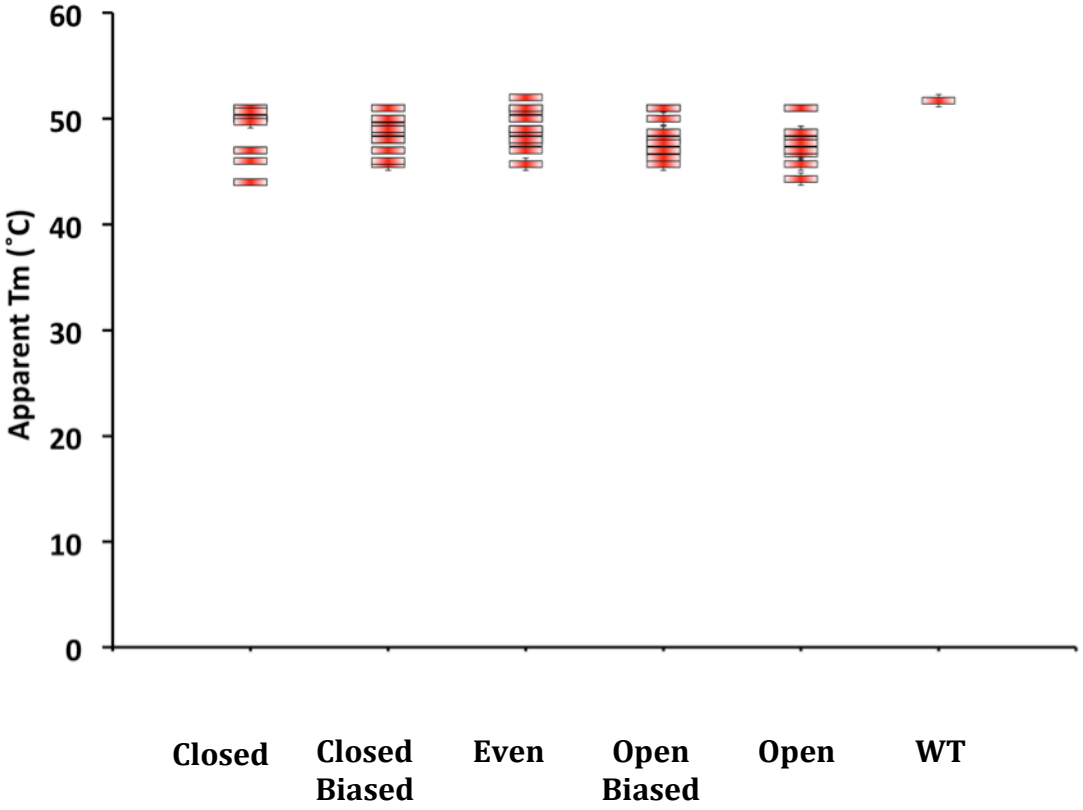


Figure 3-5. Thermofluor assay showed that all mutants exhibit wild-type-like thermal stability. All mutants are folded and stable. Apparent T_ms should be considered approximate given that the thermal denaturation was not reversible.

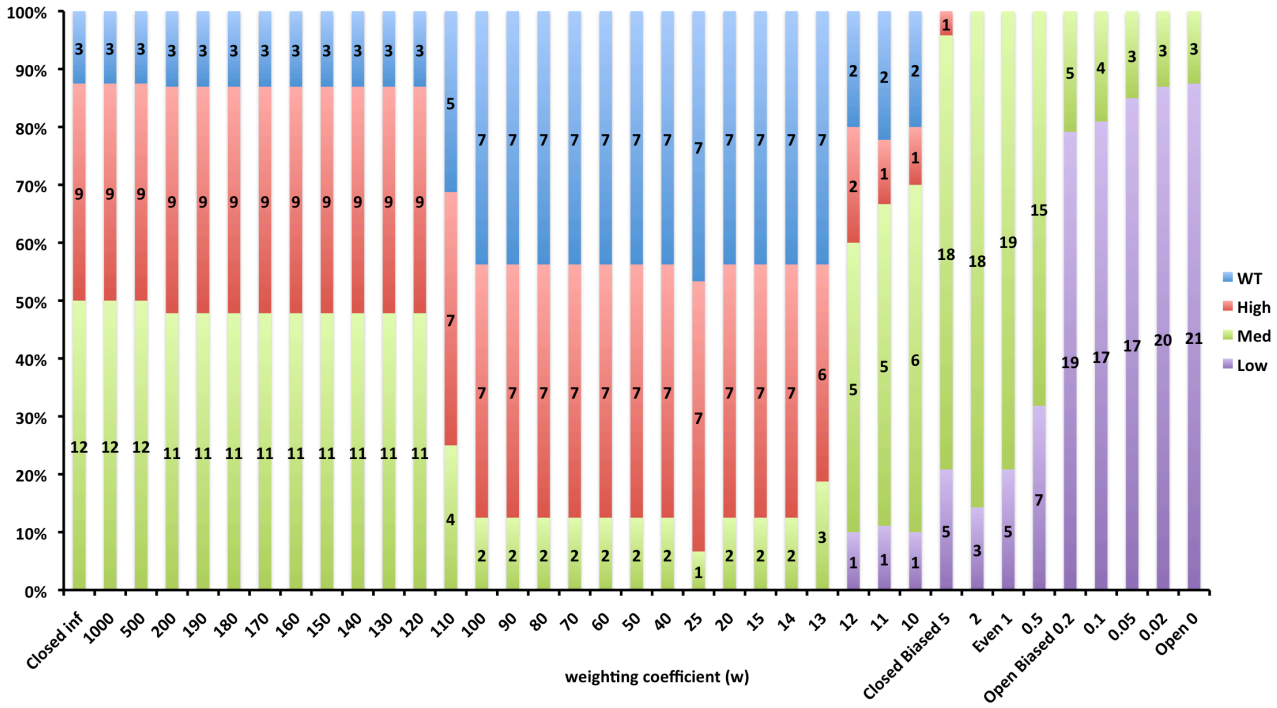


Figure 3-6. Varying the contribution of the open and the closed states in the MSD calculation. Designs within each library are categorized according to their catalytic activity (WT, high activity, medium activity, and low activity). Numbers on the x-axis are the weighting coefficients for each library. On the bars, the number of designs within the top 24 sequences that fall into each category are shown. Libraries with weighting coefficients ranging from 13 to 100 performed the best at recapitulating WT hinge sequences.

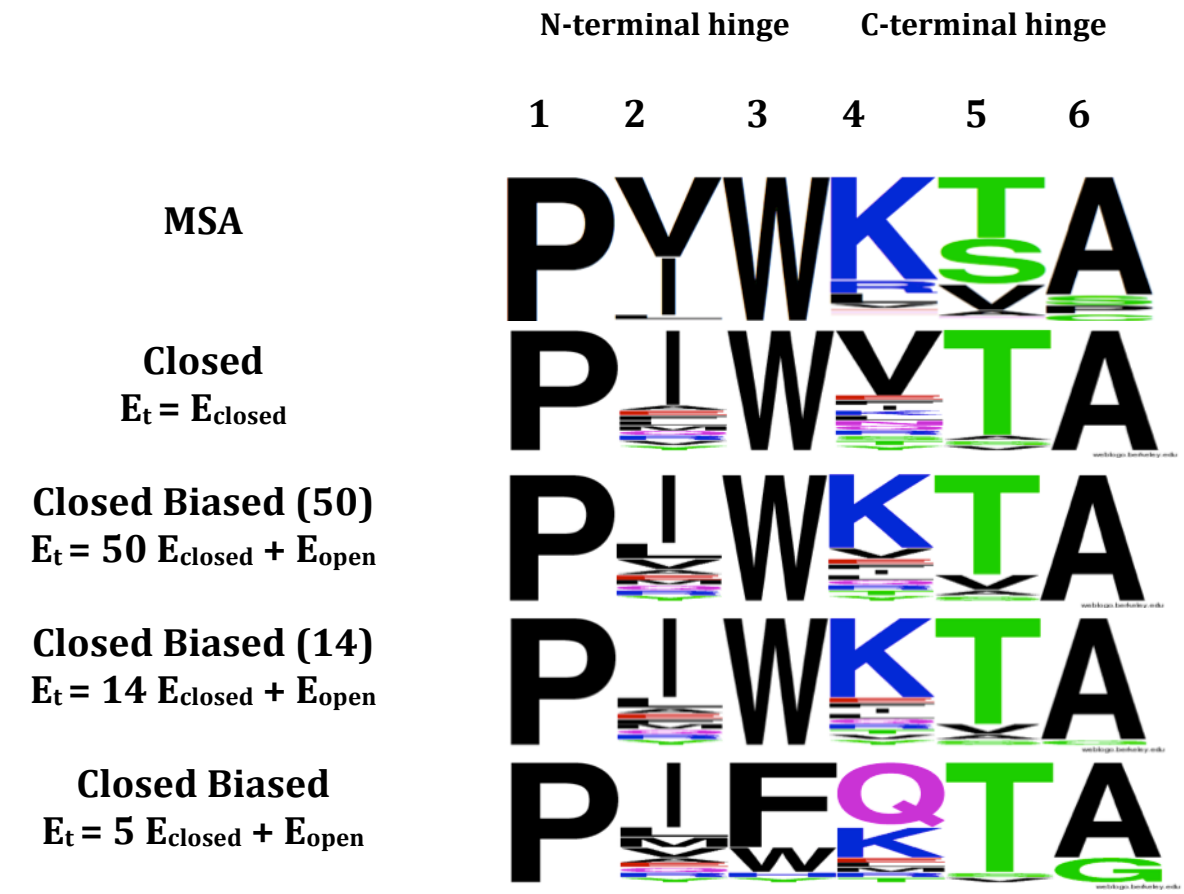


Figure 3-7. Comparison of more closed biased TIM hinge libraries with the MSA, and the original closed and closed biased libraries. Mutational frequency logos are shown for the SSD closed, three MSD libraries, and the MSA. The scoring function is shown for each library. The closed biased (50) and (14) libraries best resemble the MSA of natural TIMs, directing diversity to sites 2, 4, and 5 and choosing the wild-type amino acid in the other sites. Closed biased libraries (50 and 14), recover lysine as the most frequent amino acid at site 4th, mimicking diversity in natural TIMs (MSA).

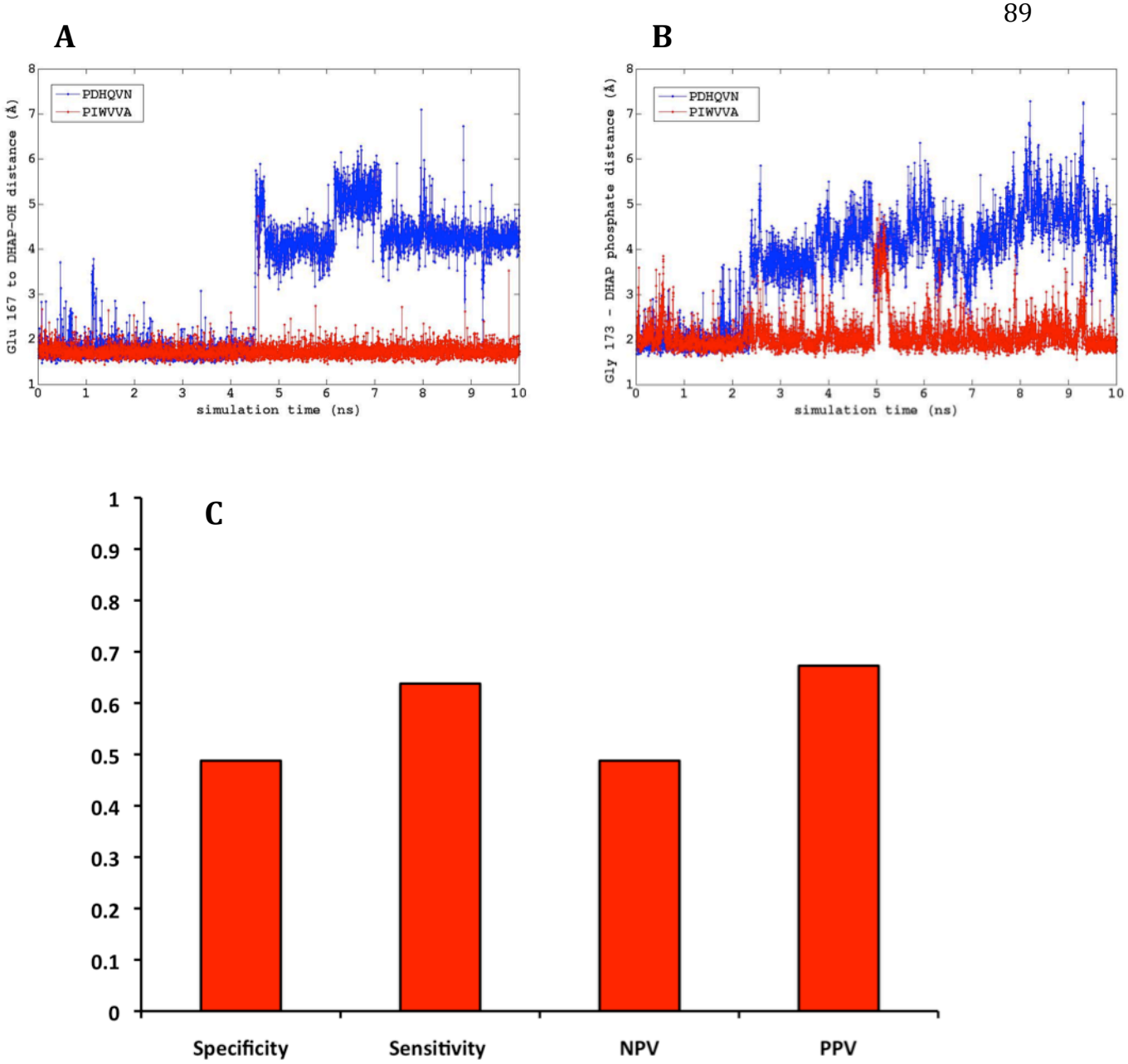


Figure 3-8. MD analysis of TIM hinge mutants. Comparison of simulation results for a highly active hinge mutant (PIWVVA, closed library) and a poorly active mutant (PDHQVN, open biased library) show the utility of MD analysis. **(A)** The catalytic contact of the active site Glutamate to the substrate DHAP is kept constant for the entire simulation of PIWVVA, while the contact is lost for PDHQVN. **(B)** The hydrogen bond of Gly 173, a residue within loop 6 that contacts the PO_4 moiety of DHAP, is monitored. PIWVVA maintains the contact during the entire simulation. Loss of the contact in PDHQVN evidences lack of tight phosphate binding, which suggests that this design could be sampling the open state more often than expected for an highly active enzyme. **(C)** Comparison of MD results with the experimental characterization of all the hinge mutants. NPV = negative predictive value, PPV = positive predictive value.

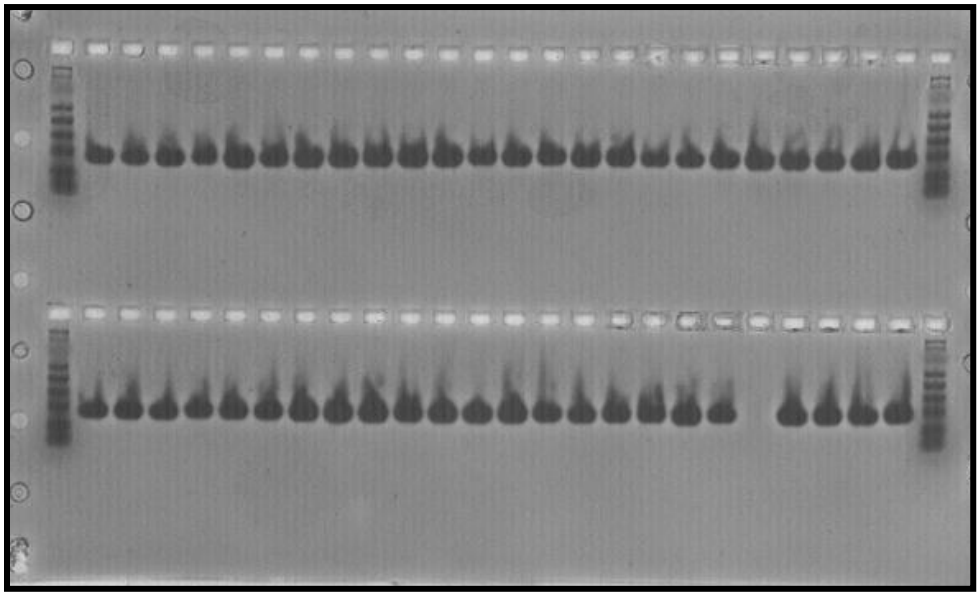


Figure 3-9. SDS-PAGE analysis of enzyme samples. After purification by affinity chromatography (His-select resin), sample purity was confirmed by SDS-PAGE.

References

1. Allen, B. D. & Mayo, S. L. (2006). Dramatic performance enhancements for the faster optimization algorithm. *Journal of Computational Chemistry* **27**, 1071-1075.

2. Alvizo O, Allen BD & Mayo SL. (2007). Computational protein design promises to revolutionize protein engineering. *Biotechniques* **42**, 31–35.

3. Bolon, D. N. & Mayo, S. L. (2001). Enzyme-like proteins by computational design. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 14274-14279.

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

5. Fleishman, S. J., Whitehead, T. A., Ekiert, D. C., Dreyfus, C., Corn, J. E., Strauch, E. M., Wilson, I. A. & Baker, D. (2011). Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. *Science* **332**, 816-821.

6. 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., Hilvert, D., Houk, K. N., Stoddard, B. L. & Baker, D. (2008). De novo computational design of retro-aldol enzymes. *Science* **319**, 1387-1391.

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

8. Privett, H. K., Kiss, G., Lee, T. M., Blomberg, R., Chica, R. A., Thomas, L. M., Hilvert, D., Houk, K. N. & Mayo, S. L. (2012). Iterative approach to computational enzyme design. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 3790-3795.

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. & Baker, D. (2008). Kemp elimination catalysts by computational enzyme design. *Nature* **453**, 190-U4.

10. Siegel, J. B., Zanghellini, A., Lovick, H. M., Kiss, G., Lambert, A. R., Clair, J. L. S., Gallaher, J. L., Hilvert, D., Gelb, M. H., Stoddard, B. L., Houk, K. N., Michael, F. E.

& Baker, D. (2010). Computational design of an enzyme catalyst for a stereoselective bimolecular diels-alder reaction. *Science* **329**, 309-313.

11. Gerstein, M., Lesk, A. M. & Chothia, C. (1994). Structural mechanisms for domain movements in proteins. *Biochemistry* **33**, 6739-6749.

12. Hammes, G. G. (2002). Multiple conformational changes in enzyme catalysis. *Biochemistry* **41**, 8221-8228.

13. Henzler-Wildman, K. & Kern, D. (2007). Dynamic personalities of proteins. *Nature* **450**, 964-972.

14. Henzler-Wildman, K. A., Thai, V., Lei, M., Ott, M., Wolf-Watz, M., Fenn, T., Pozharski, E., Wilson, M. A., Petsko, G. A., Karplus, M., Hubner, C. G. & Kern, D. (2007). Intrinsic motions along an enzymatic reaction trajectory. *Nature* **450**, 838-U13.

15. Herschlag, D. (1988). The role of induced fit and conformational-changes of enzymes in specificity and catalysis. *Bioorganic Chemistry* **16**, 62-96.

16. Richard, J. P. (2012). A paradigm for enzyme-catalyzed proton transfer at carbon: Triosephosphate isomerase. *Biochemistry* **51**, 2652-2661.

17. Davey, J. A. & Chica, R. A. (2012). Multistate approaches in computational protein design. *Protein Science* **21**, 1241-1252.

18. Allen, B. D. & Mayo, S. L. (2010). An efficient algorithm for multistate protein design based on faster. *Journal of Computational Chemistry* **31**, 904-916.

19. Allen, B. D., Nisthal, A. & Mayo, S. L. (2010). Experimental library screening demonstrates the successful application of computational protein design to large structural ensembles. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 19838-19843.

20. Ambroggio, X. I. & Kuhlman, B. (2006). Computational design of a single amino acid sequence that can switch between two distinct protein folds. *Journal of the American Chemical Society* **128**, 1154-1161.

21. Frey, K. M., Georgiev, I., Donald, B. R. & Anderson, A. C. (2010). Predicting resistance mutations using protein design algorithms. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 13707-13712.

22. Havranek, J. J. & Harbury, P. B. (2003). Automated design of specificity in molecular recognition. *Nature Structural Biology* **10**, 45-52.

23. Cui, Q. & Karplus, M. (2003). Catalysis and specificity in enzymes: A study of triosephosphate isomerase and comparison with methyl glyoxal synthase. *Protein Simulations* **66**, 315-+.

24. Knowles, J. R. (1991). Enzyme catalysis: Not different, just better. *Nature* **350**, 121-124.

25. Knowles, J. R. & Albery, W. J. (1977). Perfection in enzyme catalysis - energetics of triosephosphate isomerase. *Accounts of Chemical Research* **10**, 105-111.

26. Wierenga, R. K., Kapetaniou, E. G. & Venkatesan, R. (2010). Triosephosphate isomerase: A highly evolved biocatalyst. *Cellular and Molecular Life Sciences* **67**, 3961-3982.

27. Harris, T. K. (2008). The mechanistic ventures of triosephosphate isomerase. *Iubmb Life* **60**, 195-198.

28. Rozovsky, S. & McDermott, A. E. (2001). The time scale of the catalytic loop motion in triosephosphate isomerase. *Journal of Molecular Biology* **310**, 259-270.

29. Williams, J. C. & Mcdermott, A. E. (1995). Dynamics of the flexible loop of triosephosphate isomerase - the loop motion is not ligand-gated. *Biochemistry* **34**, 8309-8319.

30. Richard, J. P. (1984). Acid-base catalysis of the elimination and isomerization-reactions of triose phosphates. *Journal of the American Chemical Society* **106**, 4926-4936.

31. Kempf, J. G., Jung, J. Y., Ragain, C., Sampson, N. S. & Loria, J. P. (2007). Dynamic requirements for a functional protein hinge. *Journal of Molecular Biology* **368**, 131-149.

32. Sun, J. & Sampson, N. S. (1999). Understanding protein lids: Kinetic analysis of active hinge mutants in triosephosphate isomerase. *Biochemistry* **38**, 11474-11481.

33. Sun, S. H. & Sampson, N. S. (1998). Determination of the amino acid requirements for a protein hinge in triosephosphate isomerase. *Protein Science* **7**, 1495-1505.

34. Noble, M. E. M., Wierenga, R. K., Lambeir, A. M., Opperdoes, F. R., Thunnissen, A. M. W. H., Kalk, K. H., Groendijk, H. & Hol, W. G. J. (1991). The adaptability of the active-site of trypanosomal triosephosphate isomerase as observed in the crystal-structures of 3 different complexes. *Proteins-Structure Function and Genetics* **10**, 50-69.

35. Sullivan, B. J., Durani, V. & Magliery, T. J. (2011). Triosephosphate isomerase by consensus design: Dramatic differences in physical properties and activity of related variants. *Journal of Molecular Biology* **413**, 195-208.

36. Ericsson, U. B., Hallberg, B. M., DeTitta, G. T., Dekker, N. & Nordlund, P. (2006). Thermofluor-based high-throughput stability optimization of proteins for structural studies. *Analytical Biochemistry* **357**, 289-298.

37. Lavinder, J. J., Hari, S. B., Sullivan, B. J. & Magliery, T. J. (2009). High-throughput thermal scanning: A general, rapid dye-binding thermal shift screen for protein engineering. *Journal of the American Chemical Society* **131**, 3794-+.

38. Xiang, J. Y., Sun, J. H. & Sampson, N. S. (2001). The importance of hinge sequence for loop function and catalytic activity in the reaction catalyzed by triosephosphate isomerase. *Journal of Molecular Biology* **307**, 1103-1112.

39. Kursula, I., Salin, M., Sun, J., Norledge, B. V., Haapalainen, A. M., Sampson, N. S. & Wierenga, R. K. (2004). Understanding protein lids: Structural analysis of active hinge mutants in triosephosphate isomerase. *Protein Engineering Design & Selection* **17**, 375-382.

40. Leaver-Fay, A., Tyka, M., Lewis, S. M., Lange, O. F., Thompson, J., Jacak, R., Kaufman, K., Renfrew, P. D., Smith, C. A., Sheffler, W., Davis, I. W., Cooper, S., Treuille, A., Mandell, D. J., Richter, F., Ban, Y. E. A., Fleishman, S. J., Corn, J. E., Kim, D. E., Lyskov, S., Berrondo, M., Mentzer, S., Popovic, Z., Havranek, J. J., Karanicolas, J., Das, R., Meiler, J., Kortemme, T., Gray, J. J., Kuhlman, B., Baker, D. & Bradley, P. (2011). Rosetta3: An object-oriented software suite for the simulation and design of macromolecules. *Methods in Enzymology, Vol 487: Computer Methods, Pt C*, 545-574.

41. Case DA, T.A. Darden, T.E. Cheatham III, C.L. Simmerling, J. Wang, R.E. Duke, R. Luo, R.C. Walker, W. Zhang, K.M. Merz, B.P. Roberts, B. Wang, S. Hayik, A.

Roitberg, G. Seabra, I. Kolossváry, K.F. Wong, F. Paesani, J. Vanicek, X. Wu, S.R. Brozell, T. Steinbrecher, H. Gohlke, Q. Cai, X. Ye, J. Wang, M.-J. Hsieh, G. Cui, D.R. Roe, D.H. Mathews, M.G. Seetin, C. Sagui, V. Babin, T. Luchko, S. Gusarov, Kovalenko, A. & Kollman, a. P. A. (2009). Amber 11. *Univeristy of California, San Francisco*.

42. Wang, J. M., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. (2005). Development and testing of a general amber force field (vol 25, pg 1157, 2004). *Journal of Computational Chemistry* **26**, 114-114.

43. Vanquelef, E., Simon, S., Marquant, G., Garcia, E., Klimerak, G., Delepine, J. C., Cieplak, P. & Dupradeau, F. Y. (2011). R.E.D. Server: A web service for deriving resp and esp charges and building force field libraries for new molecules and molecular fragments. *Nucleic Acids Res* **39**, W511-7.

44. Pigache, A., Cieplak, P. & Dupradeau, F. Y. (2004). Automatic and highly reproducible resp and esp charge derivation: Application to the development of programs red and x red. *Abstracts of Papers of the American Chemical Society* **227**, U1011-U1011.

45. Phillips, J., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R., Kale, L. & Schulten, K. (2005). Scalable molecular dynamics with namd. *J Comput Chem* **26**, 1781-1802.

46. Humphrey, W., Dalke, A. & Schulten, K. (1996). Vmd: Visual molecular dynamics. *J Mol Graphics* **14**, 33-38.

47. Go, M. K., Koudelka, A., Amyes, T. L. & Richard, J. P. (2010). Role of lys-12 in catalysis by triosephosphate isomerase: A two-part substrate approach. *Biochemistry* **49**, 5377-5389.

Chapter IV

A High-Throughput Computational Framework for *De Novo* Enzyme Design and its Application Towards the Design of Triosephosphate Isomerase Activity

Abstract

Recently, computational protein design efforts have focused on a very challenging and extremely useful objective—designing enzymes from scratch. The goal of this study was the *de novo* design of enzymatic activity into an inert scaffold. We chose triosephosphate isomerase (TIM) as our model system, as it is easy to work with and has been well characterized experimentally. We devised and implemented a high-throughput computational framework that selects appropriate structures from protein databases and screens these scaffolds for possible active sites based on the geometric constraints of the reaction. Stabilizing mutations were then predicted using conventional enzyme design methods. Finally, using the molecular dynamics protocol described in Chapter II of this thesis, candidate designs were screened for binding to the substrate. Of the starting ~8000 structures, five novel scaffolds passed the battery of computational tests along with seven natural TIMs that were included as positive controls. Top designs and degenerate codon (DC) libraries based on the five novel scaffolds were subsequently constructed and characterized experimentally. Unfortunately, none of the designs showed measurable activity. In addition, combinatorial DC libraries were designed for two other novel scaffolds, but these again produced no active variants. Despite the lack of success in designing enzymatic activity into inert scaffolds, our methods were able to recapitulate active sites in the wild-type TIM structures that served as positive controls. Moreover, these computationally redesigned TIMs were able to complement a TIM-deficient *Escherichia coli* strain, demonstrating the overall validity of our computational

approach. The general computational framework established in this work is promising and could prove useful for future enzyme design endeavors.

Introduction

Evolution has produced many different types of enzymes that catalyze very diverse chemical reactions.[1] Enzymes have several advantages over other types of catalysts. They are very efficient, have high substrate specificity, and are regio- and stereospecific. Most do not require organic solvents and can function *in vivo* or in aqueous environments under normal conditions of temperature and pressure. They are also relatively easy to construct and are biodegradable. These benefits make enzymes preferable as catalysts for a broad spectrum of medical and industrial applications. One limitation of the applicability of enzymes is the fact that the natural repertoire is limited and therefore there are no natural enzymes for catalyzing many reactions of potential interest. Thus a central goal in chemical biology is to design enzymes with novel functions.

Many protein-engineering approaches have been used to generate novel enzymatic activity. Catalytic antibodies (CA) have been developed to catalyze reactions.[2] Several different chemical transformations have been catalyzed in this way, although with moderate rate enhancements over the uncatalyzed reaction.[3] Despite some successes, there are clear limitations to the generation of CA. Requirements for eliciting CA are: (1) a clear understanding of the reaction transition state (TS), and (2) synthetic methods to produce the transition state analog (TSA). Despite the fact that *in vitro* evolution approaches have been widely used to optimize and change the specificity of enzymes, studies suggest that the use of these techniques alone is not sufficient to create enzymes with novel activities.[4] It appears that the starting protein must at least be minimally active. Recent results,

however, indicate that the intrinsic promiscuity of natural enzymes can be used as a starting point for *in vitro* evolution to yield highly active biocatalysts.[5]

Although catalytic antibodies and directed evolution have both been successfully applied to designing and optimizing enzymatic function, they clearly present limitations that are difficult to overcome.[2; 4] On the other hand, computational protein design (CPD) does not have these limitations.

CPD methods have also been applied to the *de novo* design of enzymatic activity (i.e., the design of novel activity into a previously inert protein scaffold).[3] The catalytic activities of these *de novo* designed enzymes were initially well below those of natural enzymes, but rate enhancements can be achieved by optimizing using directed evolution.[6] Despite these successes, in most cases a shotgun approach was used in which hundreds of designed sequences were experimentally screened, resulting in a few moderately active variants and many inactive ones.[7-12] The ability to consistently design an enzyme for any desired reaction remains the "holy grail" of the field.[3; 13; 14]

The aim of this study is to devise a general high-throughput computational workflow to design enzymes with novel functions. Such a framework must select protein scaffolds from protein structure databases, find (within the scaffold) suitable pockets for harboring the putative active site, suggest active-site-stabilizing mutations, and finally, pre-screen candidate designs (e.g., via molecular dynamics simulations [MD]) to reduce the number of false positive designs. CPD methods already include active site search and active site repacking routines that select

catalytic residues and suggest stabilizing mutations, respectively. In addition, Chapter II of this thesis presents implementation and benchmark results for an MD-based approach to prescreen designs for enzymatic activity. Therefore, it is clear that what remains to be developed is a method for selecting suitable scaffolds. Having this in mind, we devised a general high-throughput computational procedure that can search for scaffolds with specific features that might be required in the design process. The criteria for selecting scaffolds will vary from one system to another. In this work, we used triosephosphate isomerase (TIM) as a model system because it is easy to work with in the laboratory, and extensive experimental and computational studies have been performed on it.[15-18]

TIM is a ubiquitous enzyme, present in nearly all cells. It catalyzes the isomerization of d-glyceraldehyde-3-phosphate (d-GAP) and dihydroxyacetone phosphate (DHAP) and plays a central role in the glycolysis pathway. Studies suggest that TIM accomplishes its stereospecific reaction primarily via three key features (Figure 4-1):[16] (1) tight phosphate binding from backbone hydrogen bonds, (2) traditional acid base catalysis by glutamate and histidine residues in the active site, and (3) a conformational change of an active site loop that closes upon substrate binding and excludes water molecules from the active site. These three mechanisms were taken into account in the application of our computational framework, demonstrating how the computational steps can be tailored for specific reaction requirements.

In the next sections, I describe the implementation details of the high-throughput scaffold search step and the complete computational workflow along

with the experimental characterization of resulting designs. Although none of the designs showed measurable activity, the entire *in silico* framework implemented for this project was successful in recapitulating natural TIM active sites, suggesting that it can be useful for future enzyme design endeavors.

Results and Discussion

Scaffold search

Selection of the appropriate protein scaffold is critical to the success of *de novo* computational enzyme designs. Instead of manually selecting structures by visual inspection, as has been commonly done in previous studies, we devised a computational tool that can search for desired specific geometric features within protein structures in a high-throughout manner. We started by downloading protein structures from the Protein Data Bank (PDB). Initial requirements were: (1) expression in *Escherichia coli*, (2) ≤ 500 residues long, (3) an X-ray crystallography structure with resolution ≤ 2.8 Å, and (4) presence of a phosphate group in the crystal structure. A total of 7715 scaffolds that fit these criteria were downloaded. Inspection of these proteins revealed that seven scaffolds were natural TIMs. We decided to keep them in the computational workflow along with the rest of the scaffolds as positive controls. After this step, the phosphate group of DHAP was aligned to the phosphate group in the crystal structure of the scaffold. To pass this step, a structure had to fulfill two additional requirements (Figure 4-2A-C): (1) to account for the tight phosphate binding present in natural TIMs, at least three hydrogen-bonding contacts had to be made to the phosphate group via the protein

backbone or amino acids R, K, N, Q, H, T and S, and (2) to account for the sequestration of the transition state, the DHAP molecule had to be in a buried pocket that excluded water. Protein structures that fulfilled these requirements would then be ready for processing by our computational protein design software. This step yielded 672 hits in 355 scaffolds. As expected, the seven wild-type TIMs used as positive controls were among the hits.

Active site search

Each scaffold was subjected to an active site search calculation. This procedure entails selecting positions within the scaffold that can accommodate the catalytic residues and achieve the geometries necessary for catalysis.[19] Geometric constraints used in our calculations were inspired and derived from active sites in natural TIMs (Table 4-1). Each of the scaffolds with DHAP docked into the crystallographic phosphate-binding site served as the starting points for these calculations. Pocket residues (within 6 Å of DHAP) were allowed to vary, sampling catalytic amino acids G, E, H, and K. Along with this combinatorial complexity of the search, the substrate was also allowed to move; the translational and rotational degrees of freedom of the substrate were approximated by generating substrate poses.[19] In addition, internal rotational degrees of freedom for the substrate were taken into account by including a substrate rotamer library containing low energy conformations. Since previous mutational studies indicate that the active site lysine in natural TIMs is the less important active site residue (the lysine just provides a positive charge for the active site[20]), scaffolds that could accommodate at least the catalytic glutamate and

histidine passed this step. As a result, only 74 scaffolds were predicted to be able to harbor a TIM active site (this included the seven wild-type TIM positive controls). A comparison of the designed active sites of the seven wild-type TIMs with the actual crystal structures confirmed that our computational methods could recapitulate the active site geometries present in natural TIMs (Figure 4-2D). More details on these calculations can be found in the Materials and Methods section.

Active site repacking calculation

After having found the sites that will accommodate the catalytic amino acids, the rest of the pocket was redesigned to stabilize the overall structure and to compensate for the structural changes introduced by the mutations suggested in the active site search. Given this goal, the rest of the positions within the pocket were allowed to sample all 20 amino acids while the identity of the catalytic residues remained fixed. This active site repacking step typically produces most of the mutations obtained in the entire workflow. For each of the 74 scaffolds, a ranked list of sequences and their approximate computational models are output. More details on the repacking calculation can be found in the Materials and Methods section.

Molecular dynamics screening for enzymatic activity

The top-ranked designs for each of the scaffolds were subjected to analysis by the MD-based screening protocol described in Chapter II of this thesis (Figure 4-2E). Inspection of the simulations revealed that most of the designs were not able to

keep the substrate in the active site and were therefore discarded. Only twelve designs were predicted to be able to bind to DHAP (Figure 4-2A). Further scrutiny of the final designs showed that of the twelve designs that passed all the computational tests, seven were natural TIMs (the positive controls). Given these results, we can confidently state that our methods recapitulated native active sites in wild-type TIM structures. The other five designs were based on scaffolds that had previously been found to catalyze completely unrelated reactions.

Enzyme designs

The five non-native TIM scaffolds that passed all the computational tests are structurally diverse (Figure 4-3). The 2PSN and 3MHG structures include a TIM barrel as one of their domains. The other three scaffolds are completely unrelated folds. 3BE4 and 2C2B were able to accommodate the three canonical catalytic residues, whereas 3H7F, 3MHG, and 2PSN could only accommodate the histidine and glutamate catalytic contacts. The number of mutations in the top designs for the five scaffolds ranges from 5 to 18.

Experimental characterization

The five new scaffold designs were characterized experimentally. Two assays are available to test for TIM activity: an *in vivo* complementation assay using the TIM-knockout Keio (DE3) strain,[21] and a more sensitive *in vitro* assay using d-GAP as substrate and glycerol-3-phosphate dehydrogenase (GPDH) as coupling enzyme,

which consumes the product of the TIM reaction (DHAP) along with NADH.[22] NADH absorbs at 340 nm, which makes it possible to monitor the reaction spectrophotometrically. None of the designs showed measurable activity by either assay (Figure 4-4 and Table 4-2). All the designs could be expressed in *E. coli* except for the one based on the 2C2B scaffold.

To determine why the designs failed to catalyze the reaction, we ran a battery of biophysical experiments. Size-exclusion chromatography (SEC) was performed to confirm the quaternary structure of the designs. In addition, the Thermofluor assay[23] and circular dichroism (CD) experiments were used to determine whether the proteins were folded. The results suggest that two designs were not active, because the mutations predicted by our computational methods destabilized and unfolded the proteins (3BE4 and 3H7F) (Table 4-2). Designs that used scaffolds 3MHG and 2PSN were folded (by Thermofluor assay and CD) (Figure 4-5) and conserved the expected quaternary structure (by SEC). Design 5, which utilized the scaffold 2C2B, did not express in Keio (DE3) strain, and therefore *in vitro* characterization data could not be collected.

In previous computational enzyme design studies, structural analysis of inactive designs was essential for understanding flaws in designs and ultimately for designing active enzymes.[10] Trays with different buffer conditions were set up to crystallize designed proteins that were stable and folded. Unfortunately, no crystal structures were obtained, so a more in-depth structural analysis of design flaws could not be undertaken (Table 4-2).

Recapitulating the natural TIM active site

It is interesting to point out that even though none of the novel scaffold designs were active, the computational framework was successful at indentifying native TIMs as hits. The seven natural TIMs that were included in the first pool of scaffolds were used as positive controls for the battery of computational methods implemented. In addition, the MD screening protocol was able to recognize the wild-type scaffolds as hits as well.

The yeast TIM structure (PDB ID: 7TIM) was among the positive controls. We cloned the 10 top-ranked sequences for that scaffold (Figure 4-6). Nine of these complemented the knockout strain (Table 4-3). These results suggest that the mutations selected by the design software (active site search and repacking calculations) do not compromise the activity of the enzyme *in vivo*.

Second-generation designs: degenerate codon (DC) libraries

Despite the lack of success of the novel scaffold designs, we decided to perform second-generation designs. We hypothesized that sequences similar to the top-ranked sequences could be active. Taking advantage of the fact that a high-throughput *in vivo* selection assay is available for the TIM reaction, DC libraries for the scaffolds 3MHG and 2PSN (previously found to be stable upon mutation) were computationally designed using a previously described algorithm.[24] The algorithm specifies the combination of degenerate codons that can be used to construct a library of CPD-predicted variant sequences (Table 4-4 and Table 4-5).

To test the hypothesis that a scaffold derived from a thermophilic organism would be less likely to unfold or aggregate upon mutation, we also designed a DC library for indole-3-glycerophosphate synthase, an enzyme from the thermophilic organism *Sulfolobus solfataricus* (Figure 4-7 and Table 4-6). Since the crystal structure of this protein (PDB ID: 1A53) also includes a phosphate group, our entire computational approach could be applied to this new scaffold.

To test a third hypothesis, we selected the *Rattus norvegicus* NTPDase2 structure (PDB ID: 3CJ9). This enzyme was included in the starting set of scaffolds but did not pass the MD screening step. It has a feature very similar to the natural TIMs in that the designed catalytic histidine is located at the positive end of the α-helix (Figure 4-8). In addition, histidine N^{∂} is within hydrogen-bonding distance of two main chain –NH groups. Studies suggest that this environment is responsible for lowering the pK_a of the catalytic histidine.[25] To explore the possibility that this feature is critical for catalysis, we cloned and tested another DC library based on this scaffold (Table 4-7).

Experimental characterization of DC libraries

DC libraries were cloned onto pET-53-Dest and transformed into the Keio (DE3) strain. In the four libraries tested, no mutants were found to complement the strain. Selected mutants were expressed and purified to test their activity *in vitro* and to perform biophysical characterization. Although some of the mutants appeared to be folded by the Thermofluor assay (Figure 4-9) and to retain expected quaternary

structure by SEC (Figure 4-10), none showed measurable activity in the *in vitro* assay.

Conclusions

We devised a general computational workflow to design enzymes with novel activity. The scaffold search method proposed here screens thousands of protein scaffolds for possible active sites based on geometric requirements that can be tailored for any reaction. The method searches for the chemical moieties of the substrate within thousands of scaffolds. In addition, a number of chemical criteria can be used during the search: specific protein–substrate contacts, specific amino acids around the pocket, solvent accessibility of the pocket, secondary structure, etc.

Designs based on the seven inert scaffolds showed no measurable activity. Combinatorial DC libraries were also designed on scaffolds that proved amenable to mutations. Again, no active library variants were found. In addition, attempts to crystallize selected designs for more in-depth structural analysis were not fruitful. Despite the lack of success in the *de novo* design of enzymatic activity, it is remarkable that the computational workflow was able to recapitulate 100% of the natural TIMs (positive controls) in the starting pool of scaffolds. In addition, the mutations suggested by CPD calculations and analyzed by the MD protocol for one of the natural TIMs produced active variants in 90% of those tested, demonstrating the overall validity of our computational methods. Thus, although none of the novel designs showed measurable activity, the computational framework implemented here may be useful for future enzyme design endeavors.

Materials and Methods

Scaffold search

Structures were downloaded from the Protein Data Bank. In addition to the criteria described above, proteins with sequence identity ≥ 90% were discarded to eliminate redundant structures. After alignment of DHAP to the PO_4 moiety on the scaffold, the putative active site was considered to be buried if at least 50% of the residues within 4 Å of DHAP had solvent accessible surface area ≤ 15 $Å^2$. The scaffold search script was written in Python and takes advantage of the structure manipulation capabilities of PyMOL (Schrödinger, LLC.).

Design calculations

All design calculations were carried out with the Triad software package (Protabit L LC) using the Rosetta[26] and Dreiding[27] force fields. Structures were prepared for calculations by adding hydrogen atoms and performing minimization to mitigate steric clashes. During active site search and repacking calculations, poses of DHAP were generated by translational and rotational moves, as previously described.[19] Geometry constraints, substrate rotamer library, and partial charges for DHAP were included as previously reported (Table 4-1, Table 4-8 and Figure 4-11).[19]

MD simulations

Please see equivalent section in Chapter III of this thesis.

Gene construction and cloning

Genes were codon optimized for expression in *E. coli*. A His6-tag and a Tobacco Etch Virus (TEV) protease cleavage site were added to the genes. Overlapping oligonucleotides for gene assembly were designed using DNAworks[28] and obtained from Integrated DNA Technologies (IDT). Oligonucleotides were mixed in equal concentrations. The assembly product was amplified using flanking primers. PIPE cloning methods[29] were utilized to put together the inserts with pET-53-Dest. The identity of individual clones was confirmed by sequencing (Laragen). Degenerate codon libraries were constructed using the same cloning protocol, except that some of the original oligonucleotides present in the assembly were replaced with oligonucleotides containing the mutagenic degenerate codons. Random colonies were picked and the fidelity of the cloning protocol was confirmed by sequencing (Laragen).

Protein expression and purification

Proteins were expressed and purified from TIM-knockout Keio (DE3) cells to prevent possible contamination with endogenous TIM. Chemically competent Keio (DE3) cells were transformed with plasmids encoding designs and plated onto LB-agar plates with antibiotics. One colony was picked to start overnight cultures in 2xYT liquid media; 5 mL of saturated culture was used to inoculate 1 L of fresh 2xYT. When the OD_{600} reached 0.6–0.8, protein expression was induced by addition of 1 mM isopropyl β-D-1-thiogalactopyranoside (IPTG). Cells were grown for 16 hours at 30°C with continuous shaking, then harvested by centrifugation (25

minutes at 1000 rpm). Cell pellets were resuspended in 10 mL lysis buffer [3x Cell Lytic B (Sigma), 300 mM NaCl, 50 mM Tris-HCl, 5 mM imidazole, 5 mM $MgCl_2$, 50 units mL^{-1} benzonase (Sigma), 0.2 mg mL^{-1} lysozyme, pH 7.8]. After incubation at room temperature for 1 hour, cellular debris was separated from soluble lysate by centrifugation (25 min at 15000 rpm). Samples were purified using a His-select gel (Sigma) and equilibrated with 300 mM NaCl, 50 mM Tris-HCl, 5 mM imidazole, pH 7.8. After loading, samples were washed twice with 5 mL wash buffer (300 mM NaCl, 50 mM Tris-HCl, 20 mM imidazole, pH 7.8) and eluted with 3 mL elution buffer (300 mM NaCl, 50 mM Tris-HCl, 150 mM imidazole, pH7.8). SDS-PAGE was used to confirm protein purity, and concentration was determined by measuring absorbance at 280 nm on a plate reader (TECAN) using Corstar UV star plates.

Activity assay

Activity assays were carried out as previously described by the Richard and Magliery Labs.[21; 22] The assay was run on a microplate reader (Tecan). UV star microplates (Grenier bio-one) were used to be able to monitor confidently at 340 nm. d-GAP (Sigma) was used as substrate and α-glycerolphosphate dehydrogenase (Sigma) was used as coupling enzyme at 3.75 μg/mL. Assay buffer was composed of 0.1 M triethanolamine (TEA) pH 7.4, 5 mM ethylenediaminetetraacetic acid (EDTA), 5 mM nicotinamide adenine dinucleotide (NADH), and 5 mM of d-GAP. First, 5 μL of coupling enzyme was placed at each well of plate, after which 180 μL of the assay buffer was added. Background absorbance was monitored at 340 nm for 10 minutes. Finally, reactions were initiated by addition of 15 μL of enzyme design

samples and the entire reaction volume was pipetted three times to ensure mixing. Absorbance at 340 nm was monitored for another 10 minutes. Absorbance rates before and after sample addition were determined using an NADH extinction coefficient was $\varepsilon = 6220\ M^{-1}\ cm^{-1}$. Observed rates were calculated by subtracting the rates before addition of sample from the rate after addition of sample ($V_{obs} = V_{after} - V_{before}$).

In vivo complementation assay

The Keio (DE3) cells containing plasmids encoding for enzyme designs were grown overnight at 37°C, washed twice with M63 salts (pH 7), and plated on M63 minimal plates containing kanamycin for strain selection and ampicillin for plasmid selection.[21] Plates were also supplemented with 0.2% w/v lactate as the sole carbon source, 1 mg L^{-1} thiamine, 80 mg L^{-1} histidine, 50 mg L^{-1} uracil, 1 mM $MgSO_4$. Plates were incubated at 37°C for 48 hours.

Size-exclusion chromatography

Size-exclusion chromatography was run with an AKTA system and a Superdex ™ 75 10/300 GL column. All samples were run with buffer containing 50 mM Tris HCl pH 7.4, 300 mM NaCl except for Design #3, which required the following buffer: 20 mM Na_3PO_4 pH 7.5, 10% glycerol, 1 mM $MgSO_4$.

Protein stability measurements by Thermofluor assay

The Thermofluor assay has been described before.[23; 30] SYPRO® Orange dye (Life Technologies, concentration not disclosed by provider) was diluted 50 times with elution buffer. 5 μL diluted dye was added to a small amount of protein (45 μL of 10 μM). Melting temperatures were obtained using a qPCR machine (BIO-RAD) run from 25°C to 99°C with 30 seconds of equilibration time and a temperature step of 1°C.

Circular dichroism

Secondary structure and stability of the designs were analyzed by circular dichroism (CD) spectroscopy with an Aviv 62DS spectrometer. A temperature controller was used to vary temperature for thermal denaturation experiments. All experiments were carried out using a 1 mm cuvette. For the design on scaffold 2PSN, protein concentration was 5 μM in 20 mM Na_3PO_4 pH 7.5, 10% glycerol, 1 mM $MgSO_4$. For the design on scaffold 3MHG, protein concentration was 15 μM in 20 mM Na_3PO_4 pH 7.5, 50 mM NaCl. Wavelength scans were carried out at room temperature from 190 nm to 250 nm with 1 second averaging time at each wavelength. Thermal denaturation was monitored at 220 nm from 4°C to 98°C with 1°C steps, averaging time of 30 sec and 2 minutes temperature equilibration time. Apparent T_m values were obtained; these should not be considered exact given that reversibility of denaturation was not observed.

Crystallization conditions

Using a liquid-handling robot (Gryphon-Art Robbins Instruments), MRC two-well crystallization plates were set up. Vapor-diffusion in sitting-drop format was used at 20 ˚C. Protein samples were mixed with crystallization buffer in a 1:1 ratio to a final volume of 0.4 µL. Different commercially available screens were used: Hampton screens (Crystal HT, MembFac, Peg/Ion, Index and PegRx), Qiagen (JCSG+) and Emerald BioSystems (Wizard 1 & 2). For the design on scaffold 2PSN, protein concentration was 7 mg/mL in 20 mM Na_3PO_4 pH 7.5, 10% glycerol, 1 mM $MgSO_4$. For the design on scaffold 3MHG, protein concentration was 10 mg/mL in 20 mM Na_3PO_4 pH 7, 50 mM NaCl.

Sequences of the top-ranked designs in the five novel scaffolds

Design #1:

MGNSKKHNLILIGAPGSGMGTSCEFIKKEYGLAHLSTGDMLREAIKNGTKIGLEAKSIIESG
NFVGDEIVLGLVKEKFDLGVCVNGFVLSGFPRTIPQAEGLAKILSEIGDSLTSVIYFEIDDSA
VIQKISHRRVHPASGRDYGQKTEPPKQPGIDDVTGEPLVWDDDANAEAVKVLLDVFHKQ
TAPLVKFYEDLGILKRVNAKLPPKEVTEQIKKILGGENLYFQGSSGHHHHHH*

Design #2:

MAHHHHHHMGTLEAQTQGPGSMSAPLAEVDPDIAELLAKELGRQRDTLEMIASENFVPR
AVLQAQGSVLTNKYAEGLPGRRYYGGCEHVDVVENLARDRAKALFGAEFANVQPHSGAQ
AVAAVLHALMSPGERLLGLDLANGGSLEHGMRLNFSGKLYENGFYGVDPATHLIDMDAV
RATALEFRPKVIIAGWSAYPRVLDFAAFRSIADEVGAKLLVHMSHFAGLVAAGLHPSPVPH

ADVVSTSVHKTLGGGRSGLIVGKQQYAKAINSAVFPGQQGGPLMHVIAGKAVALKIAATPE
FADRQRRTLSGARIIADRLMAPDVAKAGVSVVSGGTDVHLVLVDLRDSPLDGQAAEDLLH
EVGITVNRNAVPNDPRPPMVTSGLRIGTPALATRGFGDTEFTEVADIIATALATGSSVDVS
ALKDRATRLARAFPLYDGLEEWSLVGR*

Design #3:

MSILKIHAREIFDSRGNPTVEVDLFTSKGLFRAAVPSGASRGIYEALELRDNDKTRYMGKG
VSKAVEHINKTIAPALVSKKLNVTEQEKIDKLMIEMDGTENKSKFGANAILGVSLAVCKAG
AVEKGVPLYRHIADLAGNSEVILPVPAFGVINGGSNAGNKLAMEHFMILPVGAANFREAM
RIGAEVYHNLKNVIKEKYGKDATNVGDSGGFAPNILENKEGLELLKTAIGKAGYTDKVVIG
MGVAASEFFRSGKYDLDFKSPDDPSRYISPDQLADLYKSFIKDYPVVSIEDPFDQDDWGA
WQKFTASAGIQVVGDDLTVTNPKRIAKAVNEKSCNCLLLKVNQIGSVTESLQACKLAQAN
GWGVMVSHRSGETEDTFIADLVVGLCTGQISTGAPCRSERLAKYNQLLRIEEELGSKAKFA
GRNFRNPLAKGGENLYFQGSSGHHHHHH*

Design #4:

MADGNNIKAPIETAVKPPHRTEDNIRDEARRNRSNAVNPFSAKYVPFNAAPGSTESYSLD
EIVYRSRSGGLLDVEHDMEALKRFDGAYWRDLFDSRVGKSTWPYGSGVWSKKEWVLPEI
DDDDIVSAFEGNSNLFWAERFGKQFLGMNDLWVKHCGISHTGSSSDLGMTVLVSQVNRL
RKMKRPVVGVGCASTGETSAALSAYCASAGIPSIVFLPANKISMAQLVQPIANGAFVLSIDT
DFDGCMKLIREITAELPIYLANALNSLRDEGLKTAAIEILQQFDWQVPDWVIVPGGNLGNA
YAFYKGFKMCQELGLVDRIPRMVTAQAANANPLYLHYKSGWKDFKPMTASTTFASSKQV
GDPVSIDRAVYALKKCNGIVEEATEEELMDAMAQADSTGMFIAPGTGVALTALFKLRNQG

VIAPTDRTVVVSHSHGLKFTQSKIDYHSNAIPDMACRFSNPPVDVKADFGAVMDVLKSYL
GSNTLTSGGENLYFQGSSGHHHHHH*

Design #5:

MAHHHHHHGGENLYFQGSSGTITLTENKRKSMEKLSVDGVISSLGDHQRGALKRMMAQ
HQTKEPTVEQIEELKSLVSEELTPFASSIETDPEYGLPASRVRSEEAGLVLAYEKTGYDATT
TSRLPDCLDVWSAKRIKEAGAEAVGFLLYYDIDGDQDVNEQKKAYIERIGSECRAEDIPFYL
EISTYDEKIADNASPEFAKVKAHKVNEAMKVFSKERFGVDVLEVEVPVNMKFVEGFADGE
VLFTKEEAAQAFRDQEASTDLPYIYASAGVSAKLFQDTLVFAAESGAKFNGTGSGRSTWA
GSVKVYIEEGPQAAREWLRTEGFKNIDELNKVLDKTASPWTEKM*

Acknowledgements

We would like to thank Marie Ary for help with the manuscript and Jan Kostecki for help with the Thermofluor assay and CD measurements. Keio DE3 cells were a generous gift from Tom Magliery.

Table 4-1. Geometric constraint contacts between the catalytic residues and DHAP in the active site search and repacking calculations. Three contacts were required. Distance measurements are in Å. Angle and torsion measurements are in degrees. [O3-O2] is a pseudo atom positioned equidistant between atom O3 and O2 in DHAP.

Contact: GLU to 13P

residue	type	atom1	atom2	atom3	atom4	min	max
GLU	DISTANCE	CD	[O3-O2]			2.6	4
	ANGLE	CD	C3	C2		60	120
	TORSION	CD	OE1	C3	C2	60	120
	TORSION	OE1	C2	C1	O1	160	210
	TORSION	CG	CD	OE1	C2	110	250
GLU	DISTANCE	CD	[O3-O2]			2.6	4
	ANGLE	CD	C3	C2		60	120
	TORSION	CD	OE2	C3	C2	60	120
	TORSION	OE2	C2	C1	O1	160	210
	TORSION	CG	CD	OE2	C2	110	250

Contact: HIE/HID to 13P

residue	type	atom1	atom2	atom3	atom4	min	max
HIE	DISTANCE	NE2	[O3-O2]			2.29	3.1
	ANGLE	NE2	HE2	[O3-O2]		140	180
	TORSION	ND1	CE1	NE2	[O3-O2]	159	201
	ANGLE	NE2	[O3-O2]	O2		70	110
	TORSION	CE1	NE2	[O3-O2]	O2	0	360
	TORSION	NE2	[O3-O2]	O2	C2	159	201
HID	DISTANCE	ND1	[O3-O2]			2.29	3.1
	ANGLE	ND1	HD1	[O3-O2]		140	180
	TORSION	NE2	CE1	ND1	[O3-O2]	159	201
	ANGLE	ND1	[O3-O2]	O2		70	110
	TORSION	CE1	ND1	[O3-O2]	O2	0	360
	TORSION	ND1	[O3-O2]	O2	C2	159	201

Contact: LYS to 13P

residue	type	atom1	atom2	atom3	atom4	min	max
LYS	DISTANCE	NZ	O2			2.4	3.2
	DISTANCE	NZ	O1			2.4	3.2
	ANGLE	NZ	1HZ	O2		120	140
	TORSION	NZ	O2	C2	C1	60	120
LYS	DISTANCE	NZ	O2			2.4	3.2
	DISTANCE	NZ	O1			2.4	3.2
	ANGLE	NZ	2HZ	O2		120	140
	TORSION	NZ	O2	C2	C1	60	120
LYS	DISTANCE	NZ	O2			2.4	3.2
	DISTANCE	NZ	O1			2.4	3.2
	ANGLE	NZ	3HZ	O2		120	140
	TORSION	NZ	O2	C2	C1	60	120

Table 4-2. Experimental characterization of designs on the five novel scaffolds. Melting temperature (T_m) is given in °C. T_m values should be considered approximate given that thermal denaturation was not reversible. *In vitro* characterization of Design #4 was not collected because the protein could not be purified due to poor expression yield. ND = not determined.

Design	Scaffold	Mutations	*in vivo* selection	*in vitro* assay	Size-exclusion chromatography	Tm by ThermoFluor assay	Tm by circular dichrosim	Crystallization
1	3BE4	18	inactive	inactive	soluble aggregate	ND	ND	ND
2	3H7F	5	inactive	inactive	soluble aggregate	ND	ND	ND
3	2PSN	8	inactive	inactive	dimer	65	65	No crystals
4	2C2B	13	inactive	ND	ND	ND	ND	ND
5	3MHG	14	inactive	inactive	dimer	50	50	No crystals

Table 4-3. Top 10 designs on scaffold 1NEY (yeast TIM): molecular dynamics simulations and *in vivo* complementation results.

	10	92	170	209	210	212	230	231	234	# of mutations	MD prediction	*In vivo* Complementation
1NEY_wt	N	I	I	G	G	A	L	V	A	0	Active	Yes
1NEY_0	-	-	N	-	-	V	-	-	S	3	Active	Yes
1NEY_1	-	-	N	-	-	V	-	-	-	2	Active	No
1NEY_2	-	-	-	-	-	V	-	-	S	2	Active	Yes
1NEY_3	-	-	N	-	-	T	-	-	S	3	Active	Yes
1NEY_4	-	-	T	-	-	V	-	-	S	3	Inactive	Yes
1NEY_5	-	V	N	-	-	V	-	-	S	4	Active	Yes
1NEY_6	-	-	N	-	-	V	M	-	S	4	Active	Yes
1NEY_7	-	-	N	-	-	-	-	-	S	2	Active	Yes
1NEY_8	-	-	N	-	-	V	-	-	G	3	Active	Yes
1NEY_9	-	-	V	-	-	V	-	-	S	3	Active	Yes

Table 4-4. Combinatorial degenerate codon library on scaffold 2PSN. Active site residues are highlighted in orange. AAs = amino acids. The library size is 126.

PID	Sampled AAs	Library (126)
D_37	ADGIMNQST	**G**
D_150	ADGILMNQRSTV	**G**
D_157	ADGILMNQRSTVW	**NST**
D_165	**E**	**E**
D_166	**H**	**H**
D_168	ADFGLMNQRSTWY	**M**
D_209	ADGILMNQRSTV	**N**
D_242	ADGMNQRST	**M**
D_244	ADFGILMNQRSTVWY	**G**
D_292	ADFGILMNQRSTVWY	**ILM**
D_317	ADGILMNQRSTVY	**G**
D_340	ADFGILMNQRSTVW	**L**
D_369	ADFGILMNQRSTVY	**AS**
D_370	ADFGLMNQRSTV	**G**
D_393	ADFGILMNQRSTVWY	**AILMPTV**

Table 4-5. Combinatorial degenerate codon library on scaffold 3MHG. Active site residues are highlighted in orange. AAs = amino acids. The library size is 128.

PID	Sampled AAs	Library (128)
D_23	ADGLMNQRSTV	**AS**
D_24	ADGMNQSTV	**G**
D_25	ADGMNQST	**G**
D_27	H	**H**
D_68	ADFGILMNQRSTVWY	**I**
D_88	ADGILMNQRSTV	**IV**
D_125	ADFGILMNQRSTVWY	**LM**
D_163	ADFGILMNQRSTVWY	**LQ**
D_165	ADFGILMNQRSTVWY	**IL**
D_205	E	**E**
D_246	ADFGILNQRSTVY	**V**
D_275	ADGILMNQRSTVW	**DGNS**

Table 4-6. Combinatorial degenerate codon library on scaffold 1A53. Active site residues are highlighted in orange. AAs = amino acids. The library size is 256.

PID	Sampled AAs	Library (256)
A_51	E	**E**
A_53	K	**K**
A_58	DHKLNQRSTW	**N**
A_60	AHLMNQRST	**L**
A_81	G	**G**
A_83	HLMQVWY	**W**
A_89	FY	**F**
A_108	LM	**L**
A_110	HILQVWY	**IV**
A_112	FW	**W**
A_131	IKLMQRTV	**ILMV**
A_133	ILMVWY	**IM**
A_159	ILMV	**ILMV**
A_180	HILMNQRVY	**IV**
A_182	ILMV	**M**
A_184	AHKLMQRWY	**W**
A_210	H	**H**
A_211	H	**H**
A_212	GNS	**N**
A_213	ILV	**I**
A_231	AGS	**AS**
A_232	ILV	**I**
A_233	G	**G**

Table 4-7. Combinatorial degenerate codon library on scaffold 3CJ9. Active site residues are highlighted in orange. AAs = amino acids. The library size is 280.

PID	Sampled AAs	Library (280)
A_125	ADGILMNQSTV	**M**
A_126	ADFGILMNQRSTVY	**GILMRSV**
A_165	E	**E**
A_201	ADFGILMNQRSTVY	**FILMV**
A_202	ADFGILMNQSTVY	**L**
A_203	AGS	**G**
A_208	ADGLMNQSTV	**Q**
A_345	ADGMNQRSTY	**N**
A_434	ADGILMNQSTV	**IV**
A_435	H	**H**
A_436	ADFGIMNQRSTVY	**ILMV**

Table 4-8. Partial charges for dihydroxyacetone phosphate (DHAP).

atom name	charge
1HC1	0.04642
1HC3	0.05164
2HC1	0.02458
2HC3	0.07196
C1	0.08696
C2	0.31944
C3	-0.05002
HO3	0.29821
O1	-0.40376
O1P	-0.47236
O2	-0.63186
O2P	-0.481
O3	-0.86276
O3P	-0.48789
P	0.49041

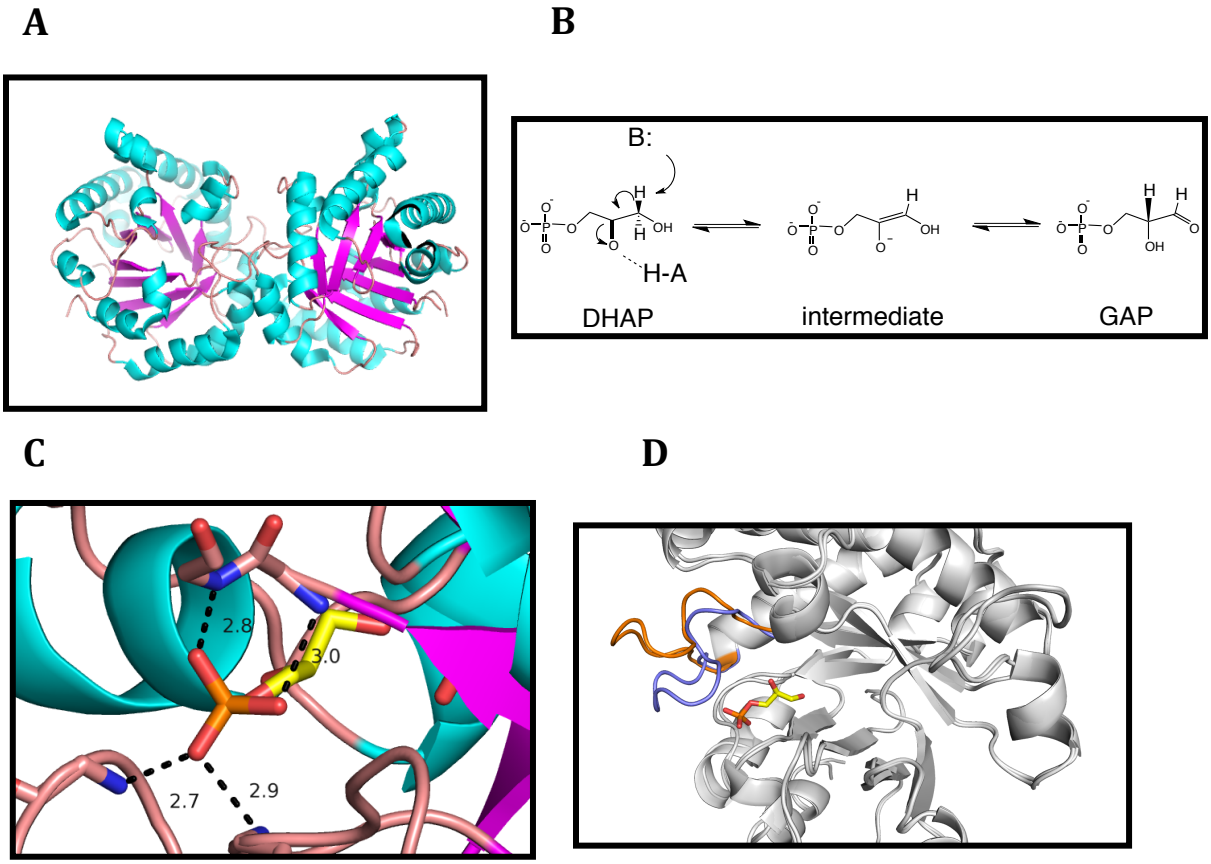


Figure 4-1. Triosephosphate isomerase (TIM) as a model system for enzyme catalysis. (A) Cartoon representation showing structure of TIM dimer. TIM catalyzes the isomerization via three main mechanisms: **(B)** Acid-base catalysis, **(C)** tight phosphate binding using backbone hydrogen bonds, and **(D)** a hinge-like conformational change of active site loop 6 to sequester the enediolate intermediate and avoid side reactions. The open conformation is shown in orange and the closed conformation in purple.

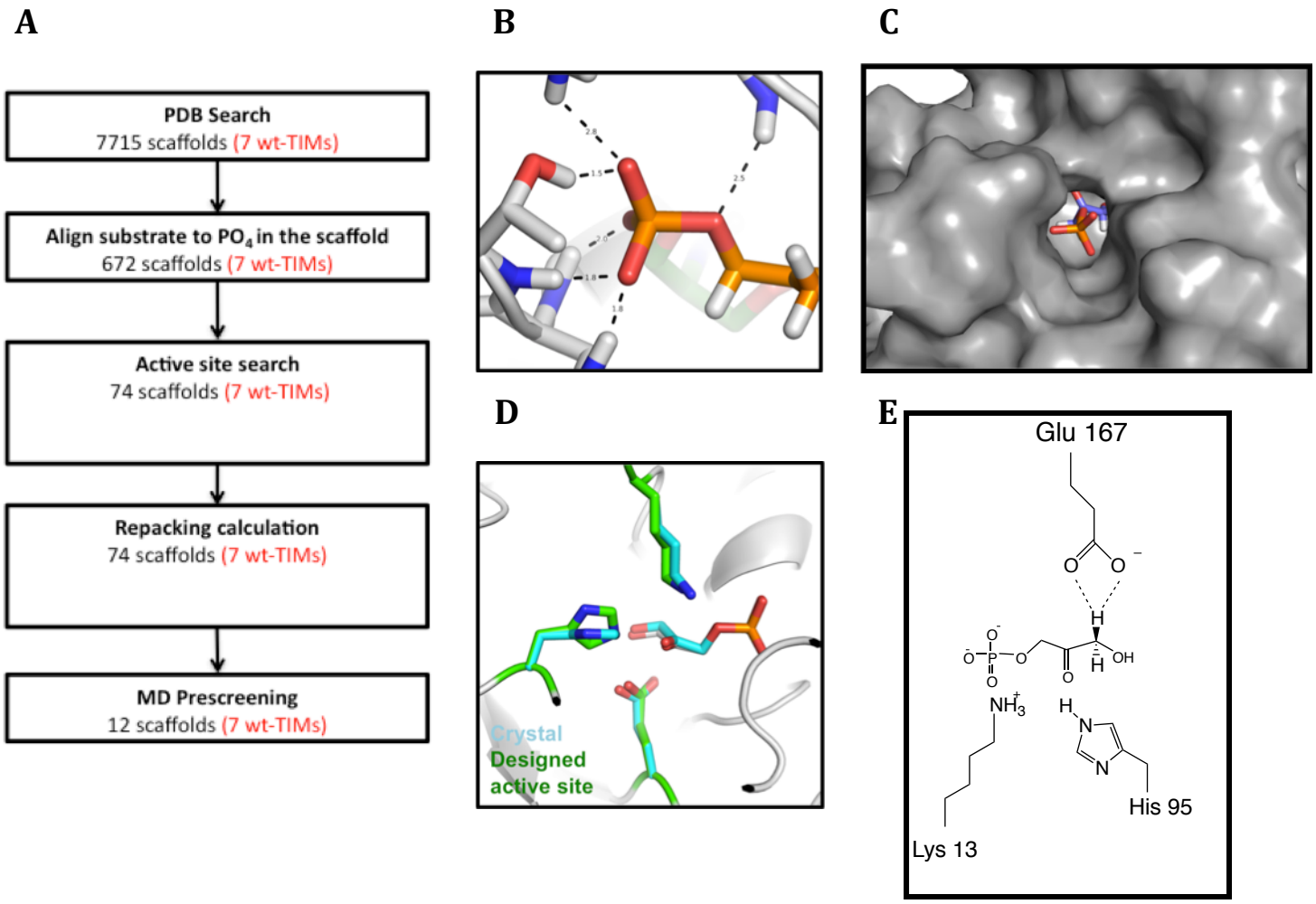


Figure 4-2. High-throughput computational workflow for enzyme design. (A) Scheme of the computational procedure. Structures with PO_4 moieties are downloaded from the PDB. The scaffold must fulfill two requirements: **(B)** native phosphate-binding pockets contacting the PO_4 by hydrogen bonds, and **(C)** buried putative active site. **(D)** The active site search identified the exact active site residues in yeast TIM (PDB ID: 1NEY), a positive control for the computational workflow. The crystal structure active site is shown in cyan, and the designed active site is shown in green. **(E)** Scheme of the TIM active site highlighting the contact that is monitored during the MD simulations (with two dashed lines).

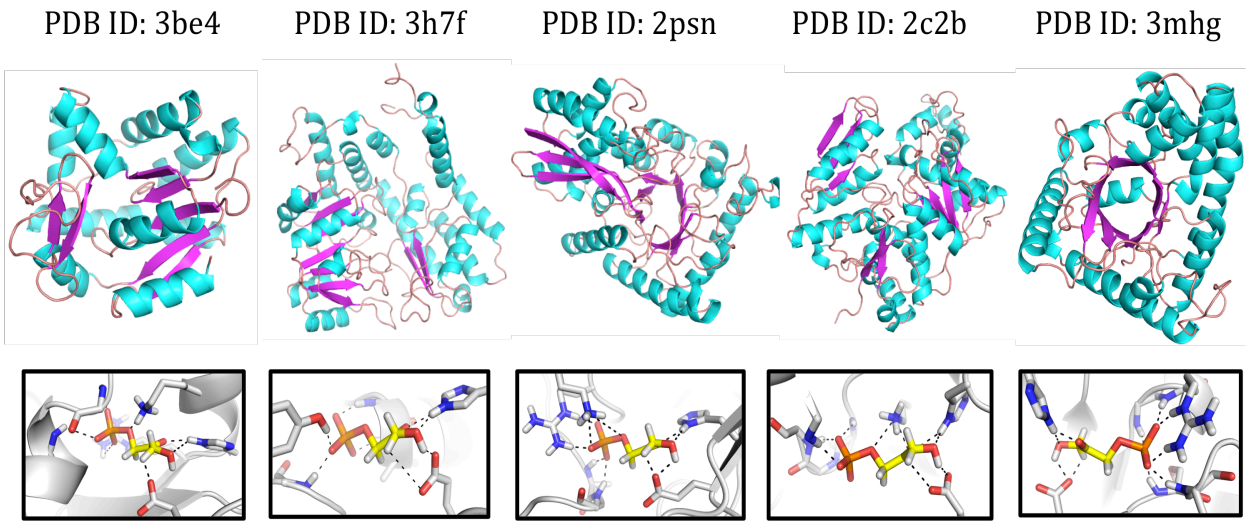


Figure 4-3. Summary of the five inert scaffolds that passed all the computational steps. The overall scaffold is shown for each design. The designed putative active site is shown, emphasizing the phosphate-binding pocket and the designed catalytic residues.

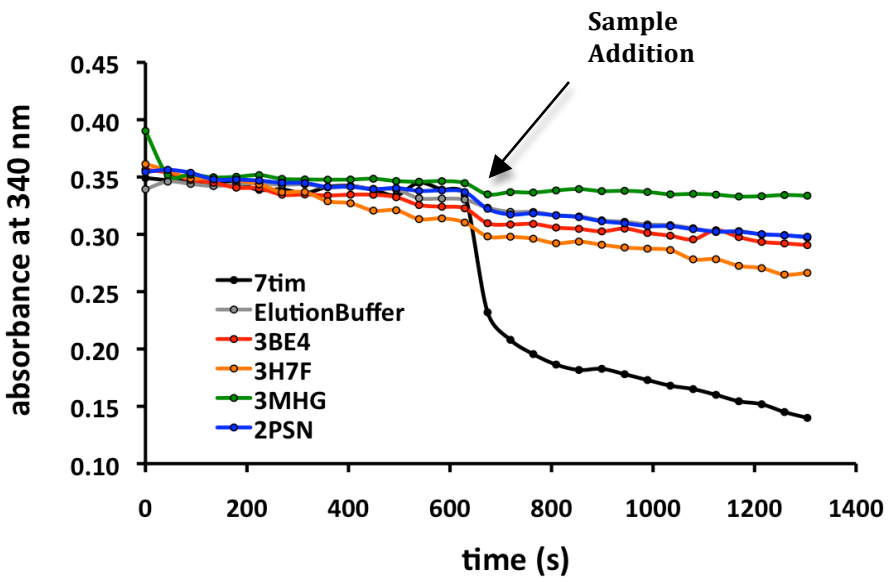


Figure 4-4. *In vitro* activity measurements of designs. Yeast TIM (7tim) is used as the positive control and elution buffer is the negative control. Upon addition of sample, consumption of NADH is apparent as a change in the slope of the curve. None of the designs showed measurable activity.

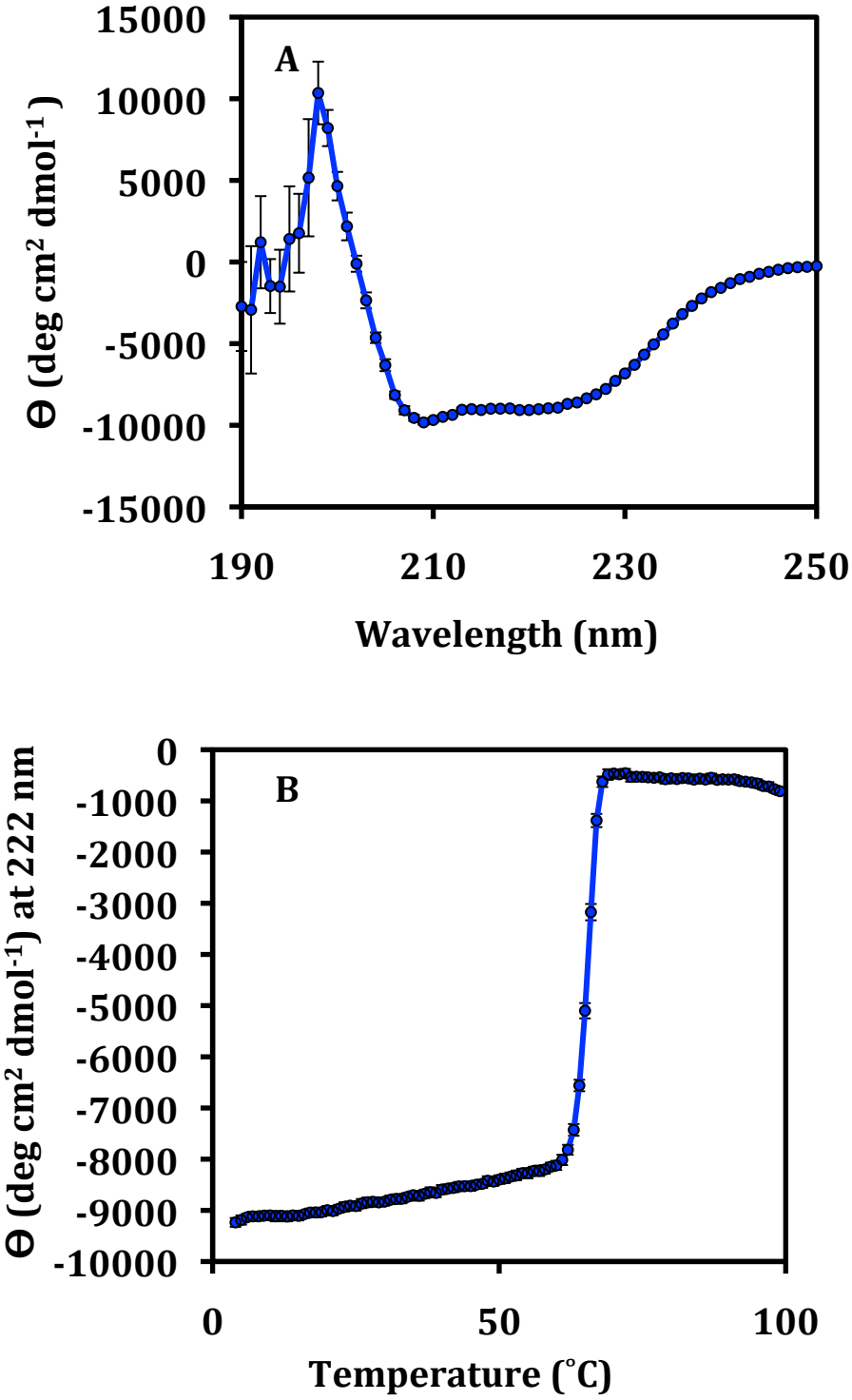


Figure 4-5. Example of circular dichroism measurements. Design #3 retains expected secondary structure and is folded. **(A)** Wavelength scan exhibits a canonical spectrum of a protein with alpha helix and beta strand secondary structure. **(B)** Thermal denaturation confirms that the design retains tertiary structure.

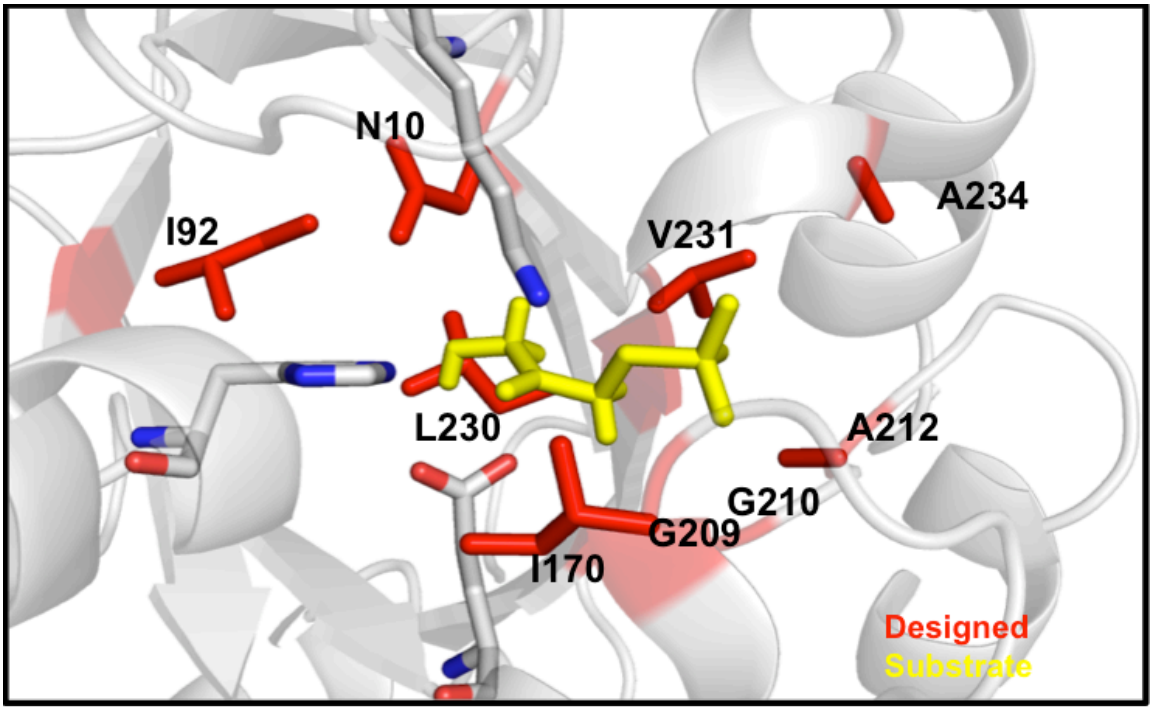
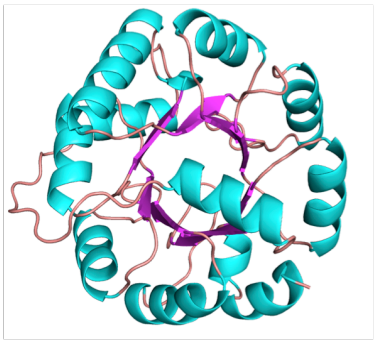
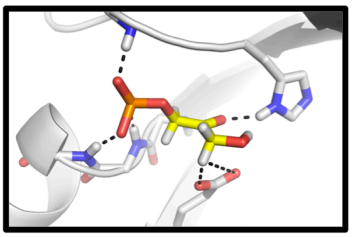


Figure 4-6. Active site of yeast TIM (PDB ID: 1NEY). Sites highlighted in red were allowed to sample all amino acids during the design calculation. Catalytic residues are shown in white and DHAP is shown in yellow.

PDB ID: 1a53





Design #6
(13 mutations)

Figure 4-7. Second-generation designs on a thermophilic scaffold. Structure of indole-3-glycerophosphate synthase from *Sulfolobus solfataricus* (PDB ID: 1A53). Overall structure and designed active sites are shown.

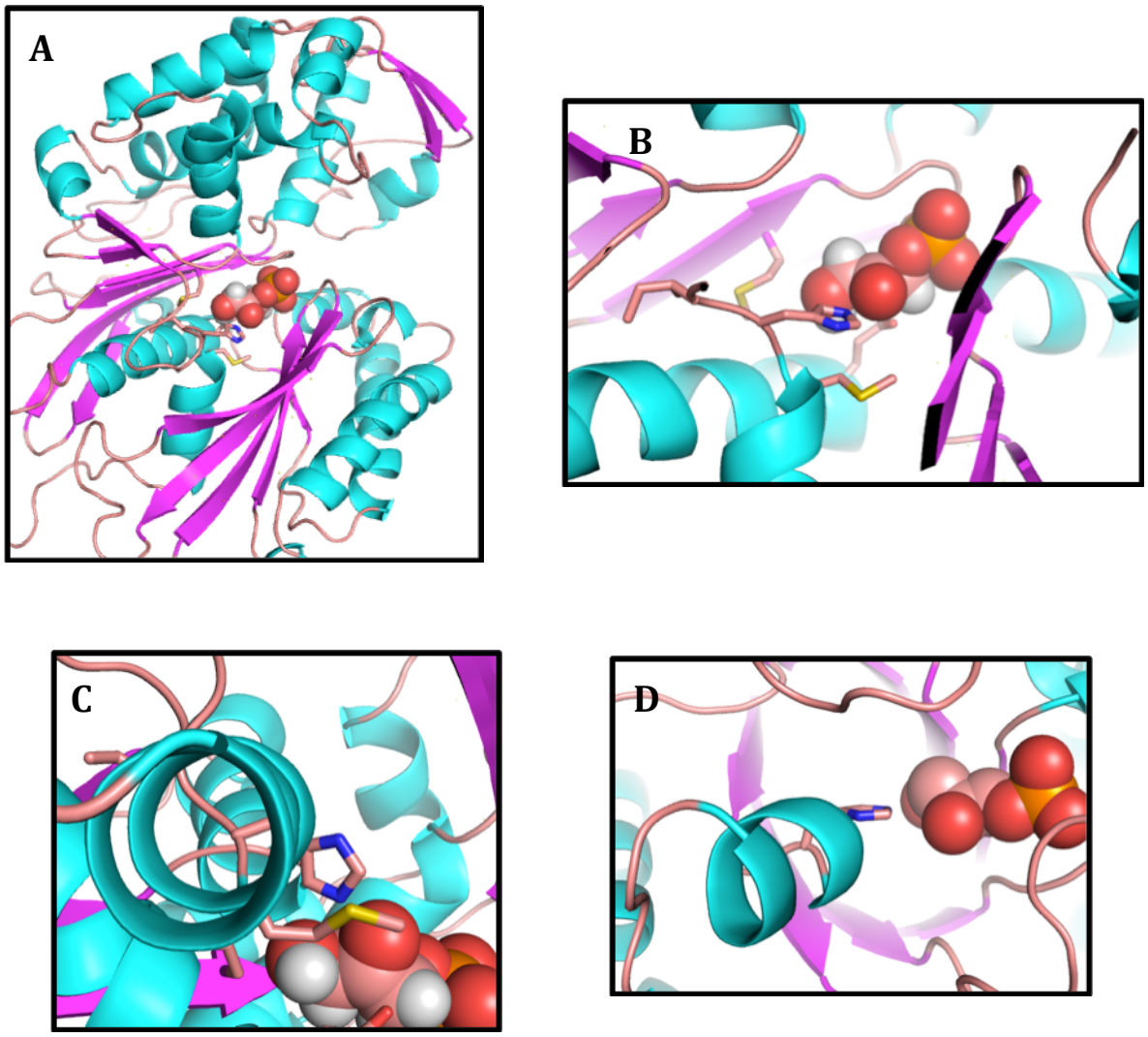


Figure 4-8. Second-generation designs on a scaffold where the designed catalytic histidine lies at the positive end of a short alpha helix, resembling a conserved feature of natural TIMs. (A) Overall structure of *Rattus norvegicus* NTPDase2 (PDB ID: 3CJ9). **(B)** Designed active site. **(C)** Alpha helix in the active site of the top design on scaffold 3CJ9. **(D)** Alpha helix in the active site of yeast TIM.

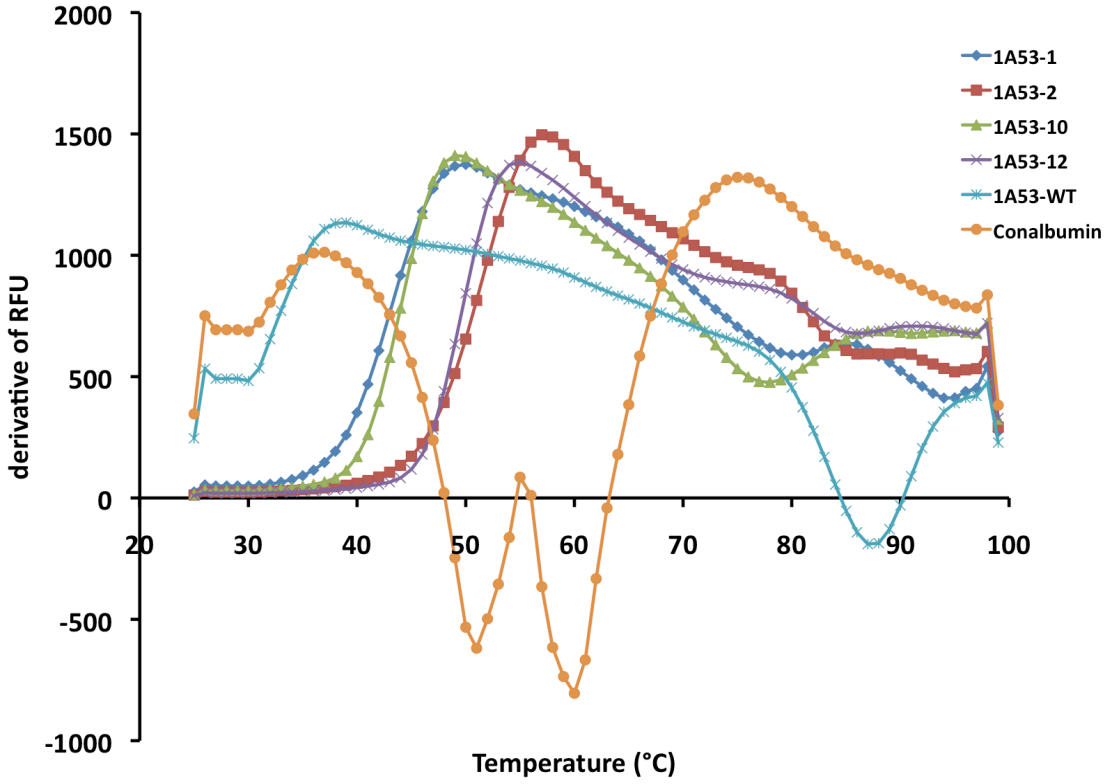


Figure 4-9. Selected designs on scaffold 1A53 are stable. Thermofluor assay for the wild-type scaffold (WT) and selected mutants from the DC library. Conalbumin is used as positive control.

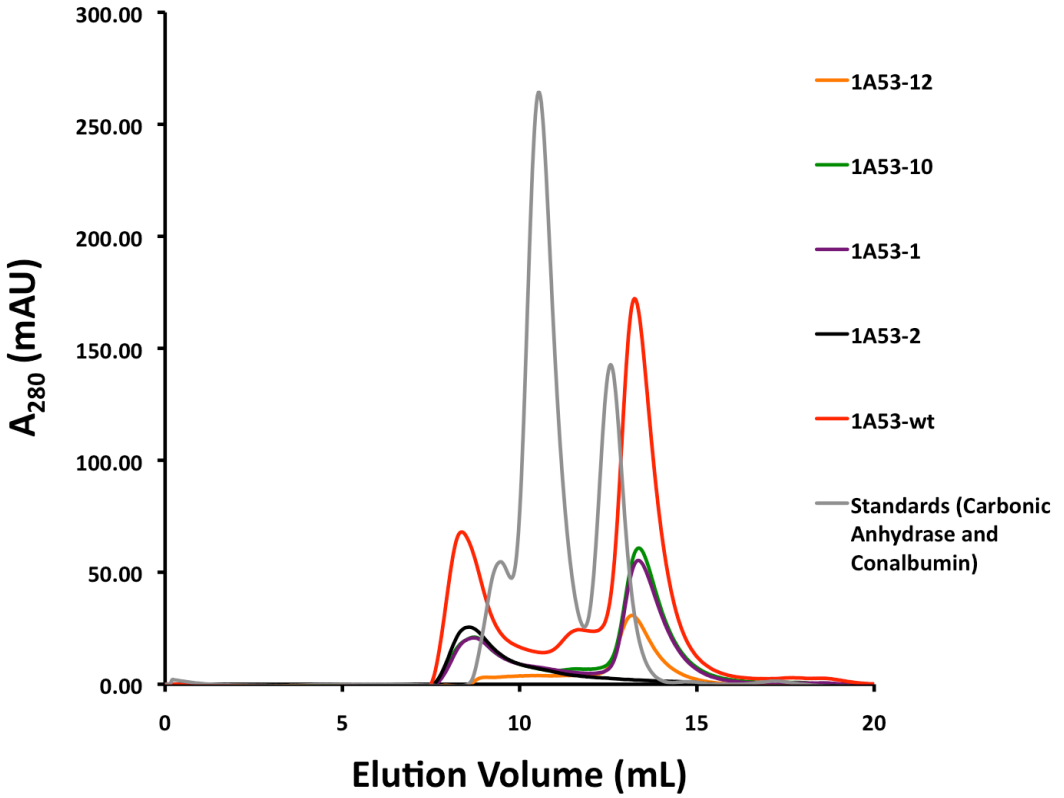


Figure 4-10. Selected designs on scaffold 1A53 elute from SEC column with the expected quaternary structure. Comparison of SEC chromatograms for selected designs and wild-type scaffold (wt). Carbonic anhydrase and conalbumin are used as molecular weight calibration standards.

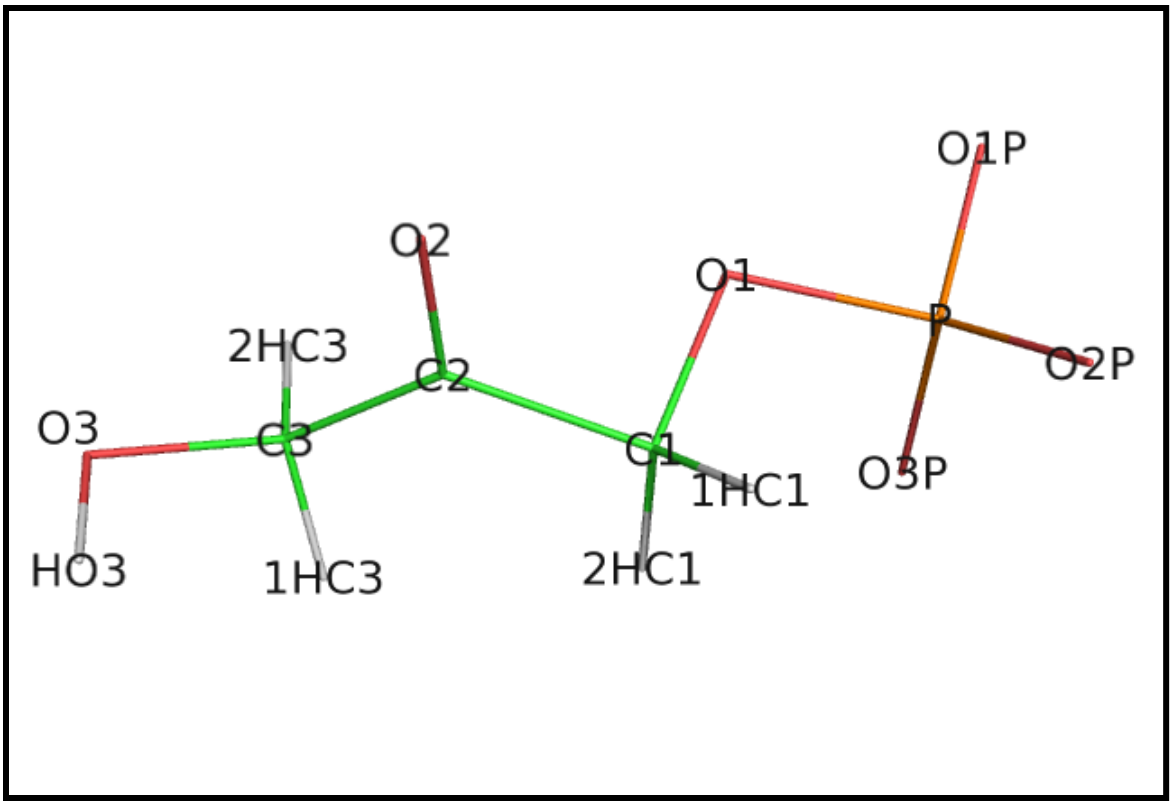


Figure 4-11. DHAP atom names.

References

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

2. Hilvert, D. (2000). Critical analysis of antibody catalysis. *Annual Review of Biochemistry* **69**, 751-793.

3. Kries, H., Blomberg, R. & Hilvert, D. (2013). De novo enzymes by computational design. *Current Opinion in Chemical Biology* **17**, 221-228.

4. Shao, Z. X. & Arnold, F. H. (1996). Engineering new functions and altering existing functions. *Current Opinion in Structural Biology* **6**, 513-518.

5. Coelho, P. S., Brustad, E. M., Kannan, A. & Arnold, F. H. (2013). Olefin cyclopropanation via carbene transfer catalyzed by engineered cytochrome p450 enzymes. *Science* **339**, 307-310.

6. Blomberg, R., Kries, H., Pinkas, D. M., Mittl, P. R. E., Grutter, M. G., Privett, H. K., Mayo, S. L. & Hilvert, D. (2013). Precision is essential for efficient catalysis in an evolved kemp eliminase. *Nature* **503**, 418-+.

7. Bolon, D. N. & Mayo, S. L. (2001). Enzyme-like proteins by computational design. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 14274-14279.

8. 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., Hilvert, D., Houk, K. N., Stoddard, B. L. & Baker, D. (2008). De novo computational design of retro-aldol enzymes. *Science* **319**, 1387-1391.

9. Khare, S. D., Kipnis, Y., Greisen, P. J., Takeuchi, R., Ashani, Y., Goldsmith, M., Song, Y. F., Gallaher, J. L., Silman, I., Leader, H., Sussman, J. L., Stoddard, B. L., Tawfik, D. S. & Baker, D. (2012). Computational redesign of a mononuclear zinc metalloenzyme for organophosphate hydrolysis. *Nature Chemical Biology* **8**, 294-300.

10. Privett, H. K., Kiss, G., Lee, T. M., Blomberg, R., Chica, R. A., Thomas, L. M., Hilvert, D., Houk, K. N. & Mayo, S. L. (2012). Iterative approach to computational enzyme design. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 3790-3795.

11. 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. & Baker, D. (2008). Kemp elimination catalysts by computational enzyme design. *Nature* **453**, 190-U4.

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

13. Baker, D. (2010). An exciting but challenging road ahead for computational enzyme design. *Protein Science* **19**, 1817-1819.

14. Bolon, D. N., Voigt, C. A. & Mayo, S. L. (2002). De novo design of biocatalysts. *Current Opinion in Chemical Biology* **6**, 125-129.

15. Cui, Q. & Karplus, M. (2003). Catalysis and specificity in enzymes: A study of triosephosphate isomerase and comparison with methyl glyoxal synthase. *Protein Simulations* **66**, 315-+.

16. Knowles, J. R. (1991). Enzyme catalysis: Not different, just better. *Nature* **350**, 121-124.

17. Knowles, J. R. & Albery, W. J. (1977). Perfection in enzyme catalysis - energetics of triosephosphate isomerase. *Accounts of Chemical Research* **10**, 105-111.

18. Wierenga, R. K., Kapetaniou, E. G. & Venkatesan, R. (2010). Triosephosphate isomerase: A highly evolved biocatalyst. *Cellular and Molecular Life Sciences* **67**, 3961-3982.

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

20. Lodi, P. J., Chang, L. C., Knowles, J. R. & Komives, E. A. (1994). Triosephosphate isomerase requires a positively charged active-site - the role of lysine-12. *Biochemistry* **33**, 2809-2814.

21. Sullivan, B. J., Durani, V. & Magliery, T. J. (2011). Triosephosphate isomerase by consensus design: Dramatic differences in physical properties and activity of related variants. *Journal of Molecular Biology* **413**, 195-208.

22. Go, M. K., Koudelka, A., Amyes, T. L. & Richard, J. P. (2010). Role of lys-12 in catalysis by triosephosphate isomerase: A two-part substrate approach. *Biochemistry* **49**, 5377-5389.

23. Ericsson, U. B., Hallberg, B. M., DeTitta, G. T., Dekker, N. & Nordlund, P. (2006). Thermofluor-based high-throughput stability optimization of proteins for structural studies. *Analytical Biochemistry* **357**, 289-298.

24. Allen, B. D., Nisthal, A. & Mayo, S. L. (2010). Experimental library screening demonstrates the successful application of computational protein design to large structural ensembles. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 19838-19843.

25. Knowles, J. R. (1991). To build an enzyme. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences* **332**, 115-121.

26. Leaver-Fay, A., Tyka, M., Lewis, S. M., Lange, O. F., Thompson, J., Jacak, R., Kaufman, K., Renfrew, P. D., Smith, C. A., Sheffler, W., Davis, I. W., Cooper, S., Treuille, A., Mandell, D. J., Richter, F., Ban, Y. E. A., Fleishman, S. J., Corn, J. E., Kim, D. E., Lyskov, S., Berrondo, M., Mentzer, S., Popovic, Z., Havranek, J. J., Karanicolas, J., Das, R., Meiler, J., Kortemme, T., Gray, J. J., Kuhlman, B., Baker, D. & Bradley, P. (2011). Rosetta3: An object-oriented software suite for the simulation and design of macromolecules. *Methods in Enzymology, Vol 487: Computer Methods, Pt C*, 545-574.

27. Mayo, S. L., Olafson, B. D. & Goddard, W. A. (1990). Dreiding - a generic force-field for molecular simulations. *Journal of Physical Chemistry* **94**, 8897-8909.

28. Hoover, D. M. & Lubkowski, J. (2002). Dnaworks: An automated method for designing oligonucleotides for pcr-based gene synthesis. *Nucleic Acids Research* **30**.

29. Klock, H. E., Koesema, E. J., Knuth, M. W. & Lesley, S. A. (2008). Combining the polymerase incomplete primer extension method for cloning and mutagenesis with microscreening to accelerate structural genomics efforts. *Proteins-Structure Function and Bioinformatics* **71**, 982-994.

30. Lavinder, J. J., Hari, S. B., Sullivan, B. J. & Magliery, T. J. (2009). High-throughput thermal scanning: A general, rapid dye-binding thermal shift screen for protein engineering. *Journal of the American Chemical Society* **131**, 3794-+.