Novel Applications and Methods for the Computer-Aided Understanding and Design of Enzyme Activity

by

Brian M. Bonk

B.S. Chemical Engineering
University of Rhode Island (2011)

Submitted to the Department of Biological Engineering
in partial fulfillment of the requirements for the degree of:

Doctorate of Philosophy in Biological Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2018

Signature redacted

Signature of Author.....
Department of Biological Engineering
October 30th, 2017

Signature redacted

Certified by.....
Bruce Tidor
Professor of Biological Engineering and Computer Science
Thesis Supervisor

Signature redacted

Accepted by.....
Mark Bathe
Chair, Department Committee on Graduate Theses

This thesis has been examined by a committee of the Department of Biological Engineering as follows:

K. Dane Wittrup
C. P. Dubbs Professor of Chemical Engineering and Biological Engineering
Chairman of Thesis Committee

Bruce Tidor
Professor of Biological Engineering and Computer Science
Thesis Supervisor

Amy Keating
Professor of Biology
Thesis Committee Member

Novel Applications and Methods for the Computer-Aided Understanding and Design of Enzyme Activity

by
Brian Bonk
Submitted to the Department of Biological Engineering
on October 30, 2017, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Biological Engineering

Abstract

Despite great progress over the past several decades in the development and application of computer-aided tools for engineering enzymes for a vast array of industrial applications, rational enzyme design remains an ongoing challenge in biotechnology. This thesis presents a set of novel applications and methods for the computer-aided understanding and design of enzyme activity.

In the first part, we apply biophysical modeling approaches in order to design non-native substrate specificity in a key enzymatic step (the thiolase-catalyzed condensation of two acyl-CoA substrates) of an industrially useful de novo metabolic pathway. We present a model-guided, rational design study of ordered substrate binding applied to two biosynthetic thiolases, with the goal of increasing the ratio of C6/C4 products formed by the 3HA pathway, 3-hydroxy-hexanoic acid and 3-hydroxybutyric acid. We identify thiolase mutants that result in nearly ten-fold increases in C6/C4 selectivity. Our findings can extend to other pathways that employ the thiolase for chain elongation, as well as expand our knowledge of sequence-structure-function relationship for this important class of enzymes.

In the second part, we apply methods from machine learning to an ensemble of reactive and non-reactive, but "almost reactive" molecular dynamics trajectories in order to elucidate catalytic drivers in another industrially important model enzyme system, ketol-acid reductoisomerase. Using a small number of molecular features, we show that we can identify conformational states that are highly predictive of reactivity at specific time points relative to the progress of the prospective catalytic event and also that provide mechanistic insight into the reaction catalyzed by this enzyme. We then present a novel theoretical framework for evaluating the contribution to the overall catalytic rate of the conformational states found in the previous part to be predictive of reactivity. Leveraging a computational enhanced sampling technique called transition interface sampling, we show that trajectories sampled in such a manner as to selectively visit the conformations predicted to be characteristic of reactivity exhibit rate constants many orders of magnitude greater than trajectories not required to visit these reactive conformations. The results of this analysis illustrate the importance of incorporating dynamical information into existing frameworks for biocatalyst design.

Acknowledgements

I would like to first acknowledge my thesis advisor, Professor Bruce Tidor, who has challenged me to become a better thinker during my time in his lab. As a direct result of my time working under mentorship of Professor Tidor, I now feel capable of approaching difficult problems with far greater confidence and intellectual rigor.

Next, I would like to acknowledge my thesis committee members, Professors Amy Keating and Dane Wittrup, who have patiently provided insight and support as the work presented in this thesis has evolved over the past six years.

I am indebted to each of the other members of the Tidor lab who I have been fortunate enough to overlap with during my time in graduate school, as they have all shaped the outcome of my graduate studies for the better in some way, be it through sharing code and scripts, maintaining the servers, providing high-level advice, or simply keeping the lab coffee and snack supply refilled. Special shout-outs go to Raja Srinivas and James Weis who have become two of my closest friends and who also have assisted enormously in the technical direction of my thesis. I would also like to thank Ishan Patel for not just providing a great code base with which to start my thesis project, but also for his scientific and personal mentorship, particularly in my early years as a PhD student at MIT. I would also like to thank my experimental collaborator Kat Tarasova for her significant contributions to this thesis.

Although I have learned an enormous amount from my time in spent in the Tidor lab, I have perhaps learned even more from the great friends I have made during my time in Cambridge. I would like to thank Vyas Ramanan, Andrew Warren, Baris Sevinc, Emily Cusick and James Valcourt, and of course Seven, for being awesome and hilarious roommates. I would also like to thank classmates Alec Nielsen, Robert Kimmerling and Nathan Stebbins for simply being part of a great group of boys.

Finally, I acknowledge my family and my girlfriend Barbara, who have been there for me through some very difficult times and who have been responsible for everything worthwhile in my life.

Table of Contents

Chapter 1 : Introduction

Enzymes represent an enormous class of highly efficient catalysts, which have been optimized over billions of years of evolution. While the catalytic power of enzymes has been harnessed extensively for industrial applications in fields ranging from therapeutics to biomaterials to agriculture to commodity chemical production, enzymes must often be fine-tuned and engineered in order to be suitable for a particular human use. The highly specific chemical tasks that enzymes have been optimized over billions of years to perform often do not align perfectly with such industrial applications, and the ability to effectively fine-tune these biocatalysts to suit a particular human-desired task remains an ongoing challenge in biotechnology, despite great progress in the past several decades (Jemli et al. 2016; Church et al. 2014; Baker 2010).

Synthetic metabolic pathways or microbial cell factories, used for the sustainable or otherwise advantageous production of specialty or commodity chemicals and fuels, represent an especially promising application of enzyme engineering (Fisher et al. 2014). Metabolic engineering approaches are particularly attractive for chemical production when traditional chemical synthesis is difficult, and the field of metabolic engineering has seen significant growth due to the numerous advantages conferred by metabolic engineering strategies for chemical production, such as high stereospecificity and mild reaction conditions (Tseng and Prather 2012). Improved catalytic function and thermal or pH stability are typical targets for rational enzyme engineering applications in metabolic pathways (Holland et al. 2012), and often engineered *de novo* pathways for a specific chemical product take advantage of enzymes with naturally

promiscuous substrate selectivity and rely on evolutionary or rational design approaches to optimize and make the pathway fully functional (Fisher et al. 2014).

The decision to undertake a rational design or evolutionary approach to engineering a particular enzyme depends on the amount of structural information available, as well as the throughput of the method that will be used to assay the resulting sequence space (Hicks and Prather 2014), although the two approaches are not necessarily mutually exclusive. Although screening-based approaches such as directed evolution and bioprospecting have been the traditional workhorses for engineering enzymes in metabolic pathways due to the lack of a requirement for a crystal structure of the enzyme under study, ultimately these approaches are limited in their effectiveness by the quality and availability of assays used to perform the selection and screening (Fisher et al. 2014). Rational engineering approaches, although often requiring greater input structural information as well as human intervention and intuition, in theory allow greater fine-tuned control and understanding of the system under study.

Much of the progress in rationally engineering and understanding enzymes in the last several decades has been made possible and driven by advances in computing and computer simulation techniques (Hilvert 2013). Since the first simulations of proteins in the 1960s and the development of hybrid quantum mechanical / molecular mechanics methods, the enormous increases in computing power have allowed these biophysical modeling methods to be applied to increasingly complex systems, at increasing levels of detail. A major challenge in using computer models to design or optimize enzyme catalysts is the vast set of possible configuration states and subtly different potentially reactive pathways. From an algorithmic computer science perspective, due to the vast number of possible amino acid configurations, the design of protein

sequences is an NP-Hard problem (Pierce and Winfree 2002). Accounting for substrate and backbone flexibility, electrostatic interactions and protein conformational changes during the course of the reaction, remain significant challenges in computational enzyme design (Baker 2010; Lippow and Tidor 2007).

The computer-aided rational design of *de novo* enzymes has represented an important milestone in the field of enzyme engineering and a true test of how far understanding in the field has progressed. Recent and exciting milestones in the field of computer-aided enzyme design have included the successful computational design of enzyme catalysts for the Kemp elimination (Röthlisberger et al. 2008) and Diels-Alder reactions (Siegel, et al 2010) – for which no natural catalysts exists. However, the kinetic performance of artificial enzymes remains significantly lower than that of natural enzymes (Kiss et al. 2013; Baker 2010). Often a combination of computer-aided rational design and directed evolution is required to yield highly efficient *de novo* enzymes, a notable example being the iterative rational design and screening approach used to achieve a rate acceleration of 6×10^8 for a *de novo* enzyme designed to catalyze the Kemp elimination (Röthlisberger et al. 2008; Khersonsky et al. 2011; Privett et al. 2012). The significant differences in the rate enhancements between natural enzymes and artificial enzymes, suggest gaps in understanding the complete nature of catalytic events, as well as limitations of existing computer modeling frameworks.

The major conceptual framework governing enzymatic catalysis, and consequently the primary strategy for designing or engineering more active enzymes, has long been transition state theory (Eyring and Stearn 1939). According to transition state theory, enzymes are able to achieve rate enhancements by lowering the activation energy barrier of the catalyzed reaction

relative to the uncatalyzed reaction. This lowering of the activation energy barrier can be accomplished by stabilizing the transition state, debstabilizing the ground state, or a combination of the two (Pauling 1946). Transition state theory alone, however, is often not always sufficient to design enzyme reactivity and an increasing body of literature has pointed to the importance of other effects, such as enzyme dynamics (Kamerlin and Warshel 2010; Ruscio et al. 2009). Further support for the need for new paradigms in biocatalyst design comes from catalytic antibodies, which are also obtained through a transition-state stabilization rationale and have likewise proved catalytically less efficient than their natural counterparts (Richter et al. 2012). Generalized frameworks of transition state theory account for the effect of other factors beyond binding of the transition state through the transmission coefficient, which is a correction term accounting for all the approximations assumed in transition state theory, such that reactant states are in local equilibrium along a reaction coordinate that can be treated by classical mechanics, and the absence of recrossings of the transition state dividing hypersurface (Garcia-Viloca et al. 2004).

The importance of "preorganization" of electrostatic interactions facilitating the formation of the transition state has been proposed to be one of the most significant factors besides transition state stabilization and ground state destabilizing driving enzymatic catalysis (Warshel 1998; Villà and Warshel 2001). A corollary of the electrostatic preorganization theory is that lowering the energetic barrier to facilitate selective formation of subsets of ground state conformations that lie on the path to the transition state, termed "near-attack conformations" (Bruice 2002) can be just as important as lowering the energetic barrier to the transition state itself (Shurki et al. 2002; Štrajbl et al. 2003). Another similar proposed reason for why artificial

enzymes have failed to achieve rate comparable to natural enzymes is the neglect through modeling of transition and ground states as static structures of dynamical effects in certain enzyme systems, such as rapid, subpicosecond "rate-promoting" motions proposed to facilitate the reactive event (Zoi, Antoniou, and Schwartz 2017), although this hypothesis is controversial (Glowacki, Harvey, and Mulholland 2012; Kamerlin and Warshel 2010).

Testing such hypotheses about factors driving enzymatic catalysis is a considerable challenge. Experimentally probing enzyme structure in the transient moments leading up to prospective catalytic event is difficult and consequently most studies of early catalytic drivers have relied on theoretical simulation studies. Accurately modeling bond breaking and bond forming events *in silico* however requires a very computationally expensive quantum mechanical treatment. Another major challenge is the fact that catalytic transitions are rare events, and cannot be efficiently sampled using conventional molecular dynamics simulations. Computational tools for enhanced sampling of transitions such as umbrella sampling (Torrie and Valleau 1974), blue moon sampling (Carter et al. 1989) and metadynamics (Laio and Parrinello 2002) alter the underlying dynamics and thus do not allow the true dynamics of the catalytic trajectory to be studied (Swendsen and Bolhuis 2014).

Transition path sampling (TPS) and transition interface sampling (TIS) are Markov chain Monte Carlo processes designed to sample properly-weighted ensembles of rare event transitions without *a priori* knowledge of the reaction coordinate or relying on the limiting assumptions of transition state theory (van Erp and Bolhuis 2005; Bolhuis et al. 2002). Unlike other enhanced sampling methods, TPS and TIS sample true dynamical events, as if the rare event under study was occurring spontaneously. To illustrate the power of path-sampling methods, consider a

series of molecular dynamics trajectories beginning with a substrate-bound enzyme. Most trajectories beginning from this state would simply remain energetically trapped in the energetic basin of attraction governing the reactant state, and given a 40 kcal/mol activation barrier, the probability of generating a reactive trajectory in this manner at 300 K would be on the order of 10^{-30}, an extremely rare event.

Path-sampling of enzymatic trajectories has been described as "catching a protein in the act" and allows an entire ensemble of rare event trajectories to be compared and analyzed (Hummer 2010). Numerous enzymatic mechanisms have been studied in atomic detail using transition path sampling studies (Hummer 2010; Basner and Schwartz 2005; Quaytman and Schwartz 2007; Crehuet and Field 2007; Saen-Oon, Schramm, and Schwartz 2008, Zoi, Antoniou, and Schwartz 2017; Harijan et al. 2017). Appropriate statistical mechanical theory can then prescribe an analysis to be performed on the resulting trajectories to compute a reaction rate (Dellago et al. 1998; van Erp, Moroni, and Bolhuis 2003). Transition path sampling was designed to overcome the limitations of the traditionally-used Bennett-Chandler approach for computing rate constants (Chandler 1978; Bennett 1977), which has been shown to be extremely sensitive to the choice of reaction coordinate (Bolhuis and Dellago 2015). Since the development of transition path sampling, numerous extensions of the original path sampling idea have been developed to either improve the computational efficiency of the original procedure or to surmount specific challenges posed by certain types of energy landscapes (e.g. diffusive energy barriers). These extensions have included transition interface sampling, forward flux sampling, multiple state transition interface sampling and replica exchange transition interface sampling (Swendsen and Bolhuis 2014; Bolhuis and Dellago 2015).

Transition interface sampling is one path-sampling technique that was designed to overcome several computational limitations of the older transition path sampling method, and provides a computationally efficient procedure for computing a rate constant (van Erp, Moroni, and Bolhuis 2003). The transition interface sampling procedure involves dividing the phase space of interest into a series of non-intersecting interfaces defined along a pre-specified order parameter λ. The only requirement for this order parameter is that it is capable of delineating the reactant basin of attraction from the product basin of attraction, i.e. it does not need to be a true reaction coordinate. Once stable reactant and product basins have been defined along the order parameter, a Monte Carlo procedure can be used to generate a set of transitions according to the statistical mechanical ensemble of interest once the sampling procedure has been bootstrapped by generating an initial "seed" trajectory. In transition interface sampling, this Monte Carlo procedure consists solely of shooting moves, in which candidate trajectories are generated by selecting a time slice of the previous trajectory in the Markov chain at random, making small perturbations to the velocity at the selected time point, and then integrating forward and backward from this point in time. Candidate trajectories are accepted into the Markov chain if they cross the corresponding reactant and product interfaces.

To date, most works applying path-sampling simulations to enzymatic systems have not utilized the formal rate constant computation procedure, but rather have used transition path ensembles as datasets to study mechanistic details such as the importance of "rate-promoting vibrations" (Dametto, Antoniou, and Schwartz 2012). To date, only a few published studies have leveraged insights from path-sampling simulations to propose enzyme variants with improved activity, and none utilized a formal computational path-sampling procedure for computing a rate

constant. Recent examples of transition path sampling studies of enzymes used in design framework have included Zoi et al (2017), which described an aromatic amine dehydrogenase mutant proposed to introduce a rate promoting motion *in silico*. Similar studies by Harijan et al. (2017) and Zoi et al. (2016) used transition path sampling simulations combined with computational protein design tools to identify a purine nucleoside phosphorylase (PNP) variant which does not exhibit a slowed downed rate with heavy isotope ^{2}H, ^{13}C and ^{15}N substitution, as does the wild-type PNP.

A major obstacle preventing the full leverage of rare-event ensembles generated using path-sampling techniques is the high-dimensionality and dynamic nature of the information produced by the simulations. Human intuition is poorly suited to the analysis of such multivariate data, but techniques from artificial intelligence and machine learning can provide powerful tools for analyzing such complex datasets. Machine learning approaches have been used with great success in problems in structural biology such as prediction of protein structure, protein folding pathways, protein-ligand binding affinities and drug design (Wallach, Dzamba, and Heifets 2015; Radivojac et al. 2013; Wu et al. 2017; Ramsundar and Pande 2016), but to date a limited number of studies have applied machine learning approaches to characterize reactive enzymatic trajectories (Zhang et al. 2017; Antoniou and Schwartz 2011).

The application of the aforementioned computer simulation techniques to understand and optimize natural enzymes for industrial uses is the focus of this thesis, which comprises two main parts. In Chapter 2 we apply biophysical modeling approaches in order to design non-native substrate specificity in a key enzyme that is part of an industrially useful *de novo* metabolic pathway. The 3-hydroxyacid (3HA) pathway also referred to as the reverse β-

oxidation or CoA-dependent chain elongation pathway, is a platform pathway for the synthesis of dozens of useful compounds of various chain lengths and functionalities, including acids, alcohols, alkanes and aldehydes, with applications in the pharmaceutical, polymer and flavor and fragrance industries (Clomburg et al., 2015; Kim et al., 2015; Sheppard et al., 2014; Tseng and Prather, 2012). The thiolase enzyme, the first committed step in this pathway, sets the chain length upon which the other downstream enzymes act, and our goal was to obtain a more selective thiolase with high catalytic activity towards the synthesis of longer chain products. We first present a theoretical framework for inducing favorable binding of substrates favoring production of a longer chain (C6) product relative to a shorter chain (C4) product. We apply this framework to the biosynthetic thiolase PhbA from *Z. ramigera* for which there is ample crystallographic data but which exhibits low activity towards longer chain substrates, such as butyryl-CoA, to identify mutants predicted to exhibit higher C6/C4 selectivity compared to wild type. We then applied the same approach to the more active *C. necator* BktB thiolase. Mutants that were computationally predicted to improve the C6/C4 selectivity ratio were initially screened *in vivo* within the context of two heterologous pathways, free HA production and PHA biosynthesis, employing different downstream enzymes (thioesterase vs. PHA polymerase). This process led to the identification of thiolase mutants with up to ten-fold increases in the selectivity ratio. *In vitro* characterization confirmed that one of the most selective mutants had 30-fold lower activity towards formation of the C4 product, whereas the activity towards C6 formation was comparable to wild type. Thiolases represent a large conserved superfamily of enzymes central to many other biological pathways, and lessons learned from this study can help expand

our understanding of the sequence-structure-function relationship for this important class of enzymes.

In Chapter 3, we apply methods from machine learning to a series of molecular dynamics simulations generated using transition interface sampling in order to elucidate catalytic drivers in another industrially important model enzyme system, keto-acid reductoisomerase (KARI). In this study we also explore one of the central ideas underlying the electrostatic preorganization and near-attack conformation theories of reactivity – that successful reactive trajectories selectively visit regions of enzymatic phase space that promote reactivity. We use techniques from artificial intelligence, particularly LASSO (Tibshirani 1996) to systematically identify features that define reactivity, by comparing a series of reactive and non-reactive trajectories at discrete time points prior to the prospective catalytic event. We show that a small number of features can be used to define subtle, but highly predictive conformational differences between the reactive and non-reactive trajectories as early as 150 fs before the prospective catalytic event, with the predictive features pointing to water orientation relative to the active site metal ions, side chain placement and compression of the breaking bond as being key predictors of reactivity. We then present a theoretical framework based on transition interface sampling for evaluating the contribution of these features to the overall catalytic rate and demonstrate that reactive trajectories sampled in a manner that they are forced to visit learned reactive subsets of phase space exhibit rate constants many orders of magnitude greater than trajectories sampled without this constraint.

Chapter 2 : Rational Design of Thiolase Substrate Specificity for Metabolic Engineering Applications

Note: This work performed in close collaboration with Yekaterina Tarasova[1], Michael A. Hicks[2], and Kristala L.J. Prather[1,2]

[1] Microbiology Graduate Program, Massachusetts Institute of Technology, Cambridge, MA

[2] Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA

Author contributions: BB performed all modeling and computations. YT performed all experiments, except for initial PhbA screens, which were performed by MH. BB and YT both contributed equally to data analysis and the final written manuscript.

Abstract

Metabolic engineering efforts require enzymes that are both highly active and specific towards the synthesis of a desired output product in order to be commercially feasible. The 3-hydroxyacid (3HA) or coenzyme-A dependent chain elongation pathway can allow for the synthesis of dozens of useful compounds of various chain lengths and functionalities, but suffers from byproduct formation, which lowers yields of the desired longer chain products, as well as increases downstream separation costs. The thiolase enzyme catalyzes the first reaction in this pathway, and its substrate specificity at each of its two catalytic steps sets the chain length and composition of the chemical scaffold upon which the other downstream enzymes act. However, there have been few attempts reported in the literature to rationally engineer thiolase substrate specificity. In this work, we present a model-guided, rational design study of ordered substrate binding applied to two biosynthetic thiolases, with the goal of increasing the ratio of C6/C4 products formed by the 3HA pathway, 3-hydroxy-hexanoic acid (3HH) and 3-hydroxybutyric acid (3HB). We identify thiolase mutants that result in nearly ten-fold increases in C6/C4 selectivity. Our findings can extend to other pathways that employ the thiolase for chain elongation, as well as expand our knowledge of sequence-structure-function relationship for this important class of enzymes.

Introduction

Microbial fermentation affords many advantages for the synthesis of commodity and specialty chemicals over more traditional methods. These include mild reaction conditions, avoidance of harsh and toxic chemicals, and the ability to utilize renewable feedstocks (Keasling, 2009). Advances in metabolic engineering and synthetic biology now allow for fast construction and manipulation of heterologous pathways in canonical production host strains (Lee et al., 2012). Although a wide variety of useful compounds have been synthesized using biological systems, few of these pathways have been commercialized. For a given pathway to be commercially viable, the process must produce the desired product in high yield, at a high titer and with high productivities.

The 3-hydroxyacid (3HA) pathway (Figure 1A), also referred to as the reverse β-oxidation or CoA-dependent chain elongation pathway, can allow for the synthesis of dozens of useful compounds of various chain lengths and functionalities, including acids, alcohols, alkanes and aldehydes, with applications in the pharmaceutical, polymer and flavor and fragrance industries (Clomburg et al., 2015; Kim et al., 2015; Sheppard et al., 2014; Tseng and Prather, 2012). This is due to the promiscuous activities of pathway enzymes, which on the one hand makes the biological synthesis of these compounds possible, but on the other, always results in a mixture of products at the end of the fermentation (Cheong et al., 2016; Clomburg et al., 2012). Thus, it is imperative to select pathway enzymes with appropriate substrate specificities to maximize yields of the desired product in order to minimize downstream separation costs. The lack of enzymes in the metabolic engineer's toolbox that are both highly active and highly

specific toward the production of a particular product is a major limitation in the construction of commercially feasible metabolic pathways.

In the case of the 3HA pathway, the thiolase enzyme sets the chain length upon which the other downstream enzymes act. Our goal was thus to obtain a more selective thiolase with high catalytic activity towards the synthesis of longer chain products. Previously, we have used the 3HA pathway and demonstrated synthesis of both 3-hydroxy-valeric acid (3HV) and 3-hydroxy-hexanoic acid (3HH). While we achieved 100% conversion of the fed propionate precursor for the synthesis of 3HV, less than 1% of the fed butyrate was converted to 3HH indicating poor specificity and activity of the pathway enzymes towards the longer chain substrates (Martin et al., 2013). As proof of principle, in this work we focused on achieving selective production of the longer chain C6 product, 3-hydroxyhexanoyl-CoA (3HH-CoA), relative to the C4, 3-hydroxybutyryl-CoA (3HB-CoA). Formation of 3HH-CoA results from the initial thiolase catalyzed condensation of a priming butyryl-CoA and extending acetyl-CoA, and subsequent action of a reductase on the condensation product, whereas 3HB-CoA is formed by the condensation of two acetyl-CoA substrates followed by reduction (Figure 1B). We thus sought to increase the thiolase selectivity ratio, which we define here as the ratio of C6 product formed relative to the C4 product. Ideally, this would mean the ratio of the C6 product to the C4 product at the end of the thiolase catalyzed reaction (i.e. 3-oxo-hexanoyl-CoA to acetoacetyl-CoA), but the thermodynamics of this reaction require coupling to a downstream enzyme to enable product formation. Thus as a proxy, we use formation of free 3HH and 3HB, as well as PHAs containing those monomers, which are derived from 3HH-CoA and 3HB-CoA, the condensation products after the reductase step.

To arrive at a more selective thiolase, two general approaches could be considered: bioprospecting or protein engineering, the latter including both rational engineering and directed evolution approaches. The decision to undertake a given approach hinges on the amount of information available at the outset of the study, as well as the throughput of the method that will be used to assay the resulting sequence space (Hicks and Prather 2014). Bioprospecting for more selective thiolases presents several difficulties because very few have been extensively characterized and employed in heterologous pathways despite the fact that thiolase enzymes are ubiquitous in nature, being central to many biochemical processes such as fatty acid biosynthesis and degradation, PHA biosynthesis, and the *Clostridial* ABE fermentative pathway (Haapalainen et al., 2006). Specifically, the BktB thiolase from *Cupravidus necator* (formerly *Ralstonia eutropha)* has been used in the biosynthesis of hydroxyacids and alcohols from C4-C10 in chain length (Cheong et al., 2016; Martin et al., 2013; Sheppard et al., 2014). Interestingly, this organism has 14 other genes in its genome annotated as putative thiolases, but only BktB and one other thiolase, PhaA, have been characterized and explored for metabolic engineering purposes (Reinecke and Steinbüchel, 2008). The catalytic activity of BktB or other thiolases towards >C6 substrates and products has not been studied due to several inherent challenges described later herein, and due to the commercial unavailability of required acyl-CoA substrates. Attempts at rational engineering of the thiolase have been limited due to the lack of *in vitro* data and a poor understanding of the sequence-structure-function relationship of the thiolase. A selection platform or a high-throughput screen would allow for one to assay a large number of variants, however, such methods are not available for the thiolase – the reasons for which

become apparent upon examination of the mechanism of the Claisen condensation reaction catalyzed by the thiolase.

Thiolases catalyze the condensation of a priming acyl-CoA and an extending acyl-CoA using a sequential bi bi ping-pong mechanism (Figure 1C). We were interested in the condensation of butyryl-CoA and acetyl-CoA to form 3-oxo-hexanoyl-CoA with high specificity; however, it is not possible to directly assay for this reaction for several reasons. First, for biosynthetic thiolases, such as BktB from *C. necator* and PhbA from *Zoogloea ramigera*, the condensation direction is thermodynamically unfavorable, requiring the condensation product to be reacted further in order to drive the reaction forward (Thompson et al., 1989). Here, the thiolase is coupled with a kinetically competent dehydrogenase enzyme. Reacting away CoASH, the other product of the condensation reaction, is insufficient to drive the reaction forward because it is released in the first half-step of the overall condensation reaction mechanism. In addition, the self-condensation of two acetyl-CoAs will always occur with some frequency, biasing any measured reaction rate. However, the low yields and high cost of synthesis of these acyl-CoAs precluded the development of a high-throughput activity screen.

It is for the above reasons that engineering of the thiolase has proved to be challenging, with only two examples of such attempts. The first attempt at thiolase engineering described in the literature used directed evolution to arrive at a variant that exhibited robust acetoacetyl-CoA product formation but lower sensitivity to inhibition by CoASH (Mann and Lütke-Eversloh, 2012). Another effort to engineer the thiolase to accommodate α-substituted acyl-CoAs relied on intuition guided rational mutagenesis of just one residue in close proximity of the active site but employed coenzyme-A analogs (Fage et al., 2015).

Limited by a low throughput *in vivo* assay, but armed with extensive crystallographic data, we followed a computationally driven, structure guided rational engineering approach to engineer the biosynthetic thiolase for improved selectivity towards the synthesis of longer chain products. In this work we present a theoretical framework for the design of ordered binding in a sequential bi bi ping-pong reaction. We initially apply this framework to the biosynthetic thiolase PhbA from *Z. ramigera* for which there is ample crystallographic data but which exhibits low activity towards longer chain substrates, such as butyryl-CoA, to identify mutants which we predict will exhibit higher selectivity ratios compared to wild type in order to validate the approach. We then applied this same approach to the more active *C. necator* BktB thiolase. Mutants that were computationally predicted to improve the selectivity ratio were initially screened *in vivo* within the context of two heterologous pathways, free HA production and PHA biosynthesis, which employ different downstream enzymes (thioesterase vs. PHA polymerase).

This process led to the identification of thiolase mutants with up to ten-fold increases in the selectivity ratio. *In vitro* characterization confirmed that one of the most selective mutants had 30-fold lower activity towards formation of the 3HB product, whereas the activity towards 3HH formation was comparable to wild type. Thiolases represent a large conserved superfamily of enzymes central to many other biological pathways, and lessons learned from this study can help expand our understanding of the sequence-structure-function relationship for this important class of enzymes.

Materials and Methods

Chemicals and reagents

All chemicals were obtained from Sigma Aldrich unless stated otherwise. Protein purification reagents were purchased from BioRad Laboratories (Hercules, CA).

Strain and plasmid construction

Escherichia coli MG1655 K12 (DE3) was used as the host for all production experiments. pCDFDuet-*pct-phaC2* was constructed by restriction enzyme cloning. First, *pct* from *M. elsdenii* was amplified using Q5 Polymerase (New England Biolabs, Ipswish, MA) from *M. elsdenii* gDNA. *PhaC2* was synthesized as a codon optimized gBlock from Thermo Fisher and digested with the respective restriction enzymes. Construction of pETDuet-*bktB-phaB* is described in Martin et al. (2013a) . This plasmid served as the template for generating BktB mutants. Primer sequences can be found in Supplementary Table II.

Culture conditions and strain propagation

E. coli DH5α was used for construction and maintenance of all plasmids. For PHA production experiments, *E. coli* MG1655 K12 (DE3) was transformed by electroporation with pCDFDuet-*pct-phaC2* and a pETDuet plasmid with a given thiolase variant and *phaB*. For every production experiment, three individual colonies were picked and grown overnight in LB medium containing carbenicilin (50 μg/mL) and streptomycin (50 μg/mL) at 30°C, 250 rpm. A 250-mL shake flask containing 50 mL of M9 minimal medium with 15 g/L glucose was used for production experiments and inoculated with 1% v/v of the overnight starter culture. Expression of heterologous genes was induced by addition of IPTG to 100 μM final concentration when

OD_{600} was 0.7-1.0. Butyrate was added to 15 mM final concentrations from a neutralized sterIstock solution at induction. Cells were harvested by centrifugation and washed twice with water before freezing at -80°C and lyophilization for polymer extraction and derivatization. For analysis of free acids, cell-free culture supernatants were analyzed directly by HPLC.

Site specific mutagenesis

All point mutants were made using the QuikChange Lightning XL Kit from Agilent Technologies according to the manufacturer's protocols (Agilent Technologies, Santa Clara, CA), except that DH5α cells were used for transformation of QuikChange products. The online primer design tool (http://www.genomics.agilent.com/primerDesignProgram.jsp) was used to generate the mutagenesis primers to be used in the thermal cycling reaction. Primer sequences can be found in the Supplementary Table II. Products of this reaction were used to transform chemically competent *E. coli* DH5α and plated on selective medium after recovery in SOC. Individual colonies were selected and mutations confirmed by sequencing (GeneWiz, Cambridge, MA).

Product analysis

Acidic methanolysis to analyze PHA composition was performed as described by Brandl et al. (1988) and is briefly described below. Cells were harvested by centrifugation and washed twice with water. The cells were then frozen at -80°C. Lyophilized cells were weighed to determine the CDW. Then, 5-20 mg of dried cells was used for methanolysis to determine PHA polymer composition by GC/MS. Hexanoic acid was added as an internal standard to a final concentration of 2.5 mM. In short, 1 mL chloroform, 0.85 mL methanol and 0.15 mL concentrated H_2SO_4 was added to each sample in a screw-capped tube with threads wrapped with

PTFE tape. The samples were then boiled for 1.5-2 hours at 100°C on a heating block with intermittent manual mixing. After boiling, the tubes were cooled and placed on ice, followed by addition of 0.5 mL water and vortexing for 1 minute. Tubes were centrifuged to achieve phase separation. The bottom chloroform layer was then transferred into a glass vial, dried over $MgSO_4$, and filtered through a 0.45 μm PTFE filter into a GC vial. Derivatized 3HAs were analyzed on an Agilent 7890B/5977A GC/MS with a VF-WAX column (30 m x 250 μm x 0.5 μm). The following method parameters were used: inlet temperature of 220°C, initial oven temperature of 80°C and a linear ramp rate of 10°C/min until final oven temperature of 220°C, with a 10:1 split ratio. An FID detector was used for quantification of methyl-3HB and methyl-3HH. Quantification of free acids, 3-hydroxyhexanoic and 3-hydroxybutyric acids, was performed by HPLC. One mL of culture was harvested at induction and at 72 hours post induction and centrifuged at maximum speed for 6 minutes. A sample of the supernatant was then run on an Aminex HP-87x (BioRad, Hercules, CA) column on an Agilent 1200 HPLC instrument equipped with an RID detector. 5 mM sulfuric acid was used as the mobile phase at 0.6 mL/min with column temperature held at 35°C.

Protein purification

Thiolase variants were subcloned into a protein expression vector pTev5 with an N-terminal hexa-histidine tag using CPEC cloning with primers listed in the Supplementary Information. *E. coli* BL21(DE3) was used as the host for protein expression. One liter of culture was grown in TB medium with glycerol at 30°C and induced with 100 μM IPTG when OD_{600} was ~ 0.5. Cells were harvested 15-18 hours post-induction by centrifugation and resuspended in 2.5x vol/wt buffer containing 50 mM Tris-HCl pH 8.0, 500 mM NaCl and 10% vol/vol

glycerol. Lysozyme was added to 1 mg/mL final concentration and cells were lysed by sonication. Protein purification was then performed as described previously (McMahon and Prather 2014). After purification, proteins were exchanged into storage buffer (50 mM Tris pH 8.0, 50 mM NaCl and 10% vol/vol glycerol), flash frozen in small aliquots and stored at -80°C. Protein concentration was determined by a Bradford assay using BSA as standard. PhaB reductase from *C. necator*, which was used as a coupling enzyme in condensation assays, was purified in the same manner as described above.

Enzyme assays

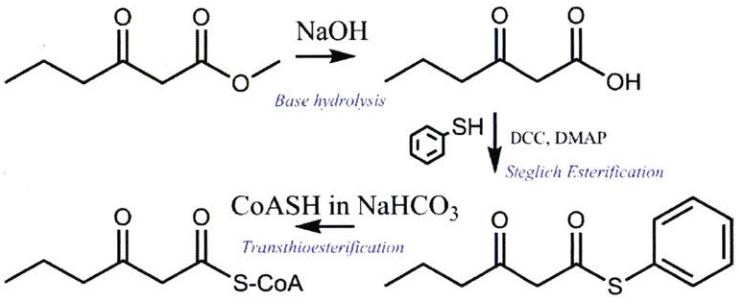
Thiolase variants were assayed in both condensation and thiolysis directions. The condensation assay was performed akin to that described previously (Bond-Watts, Bellerose, and Chang 2011), except at pH 7.0 and coupled to PhaB reductase (from *C. necator)*. Each reaction contained 100 mM Tris pH 7 buffer, 100 μg/mL NADPH, and varying amounts of acetyl-CoA, and reaction progress was monitored by a decrease in A_{340} nm corresponding to NADPH consumption on a Beckman-Coulter DU800 spectrophotometer. Thiolases were also assayed in the thermodynamically favored thiolysis direction with acetoacetyl-CoA and 3-oxo-hexanoyl-CoA. Each assay contained 100 mM Tris pH 7.0, 10 mM $MgCl_2$, 200 μM CoASH, an appropriate amount of enzyme, and varying substrate concentration. A decrease in A_{303}, corresponding to consumption of the Mg-keto-acyl-CoA complex was measured spectrophotometrically. The extinction coefficient for acetoacetyl-CoA was determined to be 4.22 μM^{-1} cm^{-1} under the enzymatic conditions. Concentrations of all enzymes used in the assays were such that the reaction rate was linear for at least 0.5 minutes. Enzymes were diluted in pH 7 dilution buffer (100 mM Tris pH 7, 50 mM NaCl and 10% vol/vol glycerol). Each substrate

concentration was assayed at least in duplicate. Generated concentration vs. initial rate curves were fit to the Michaelis-Menten equation, from which catalytic parameters (k_{cat} and K_m) were determined using the *nlinfit* routine in MATLAB.

Synthesis of 3-oxo-hexanoyl-CoA

The generalized synthesis is outlined in Scheme I below and was inspired from the synthesis of ethylmalonyl-CoA by Dunn et al. (2014) and adapted by M. Blaissey and M.C.Y. Chang (personal communication). 3-oxo-hexanoic acid methyl ester was purchased from Alfa Aesar. 1 mmol of the ester was allowed to react with 1.2 mmol aqueous NaOH at room temperature overnight. The reaction was then neutralized to pH 7.0 and extracted three times with ethyl acetate, dried over $MgSO_4$ and solvent evaporated. 3-oxo-hexanoic acid appeared as a white solid. This crude solid was used in subsequent thioesterification with 1.2 mmol of thiophenol, 1.5 mmol diisopropylcarbodiimide and 2 mg of dimethylaminopyridine in 10 mL of ethyl acetate. The reaction was carried out on ice for 2 hours, followed by 2 hours at room temperature, after which the white precipitate was filtered off and the filtrate extracted with saturated sodium bicarbonate. The organic layer was then dried over $MgSO_4$ and solvent evaporated. Crude thiophenol-coupled product was then re-dissolved in 200 μL acetonitrile and added to 1 mL of 0.5 M $NaHCO_3$ on an ice bath. 25 mg of CoASH was added and the reaction allowed to proceed for 1 hour on ice and then 1 hour at room temperature. The reaction was quenched with 50% formic acid, and extracted with diethyl ether. Final 3-oxo-hexanoyl-CoA product was purified by HPLC with 25 mM ammonium acetate pH 4.5 and 20% acetonitrile in water as the mobile phases using a linear gradient from 1% v/v acetonitrile to 20% over 25 minutes on an Agilent Zorbax Eclipse XDB C18 column. Identity of the compound was verified

by mass spectrometry. Finally, the purified product was desalted on the same column but with only water and acetonitrile as mobile phases.



Scheme I. Synthesis of 3-oxo-hexanoyl-CoA.

Starting X-ray Structures

For PhbA calculations with butyryl-CoA and acetyl-CoA bound with C89 unacylated ("Bind 1"), 1M3Z (C89A mutant with acetyl-CoA bound) was used as the starting crystal structure with the C89 built into the structure using the same dihedral angles as the C89 of the unliganded wild type PhbA thiolase structure 1DLU (Kursula et al. 2002; Y Modis and Wierenga 2000). For PhbA calculations with C89 acylated ("Bind 2"), 1DM3 (wild type enzyme with acetyl-CoA bound and C89 acetylated) was used as the starting structure (Y Modis and Wierenga 2000), and for all BktB calculations, 4NZS (wild type enzyme, unliganded) was used as the starting crystal structure (E. J. Kim et al. 2014a). All crystal structures were prepared for computer modeling with the CHARMM36 force field (Brooks et al. 2009b) using the methodology outlined in Lippow et al. (2007). CHARMM parameters for acetyl-CoA were taken from Aleksandrov and Field (2011).

Computational Methodology

Mathematically, calculations for each of the two binding events sought to optimize:

$$\Delta\Delta\Delta E^{Mut-WT}_{Bind\ B-Bind\ A} = \Delta\Delta E^{Mut-WT}_{Bind\ B} - \Delta\Delta E^{Mut-WT}_{Bind\ A}$$

where the subscript Bind B refers to the structure with a butyryl group in either the first or second binding event and the subscript Bind A refers to the corresponding structure with an acetyl group in either Bind 1 or Bind 2. Note that Bind B refers to the bound conformation leading to BA production in either step and Bind A refers to the bound conformation leading to AA production in either step. For example, the structure optimized for Bind B in the first binding event corresponds to step 2 of Figure 1C where R is a butyryl group, and the structure optimized for Bind A in the first binding event corresponds to step 2 of Figure 1C where R is an acetyl group.

Thus the optimization sought to minimize the difference of the following two terms:

$$\Delta\Delta E^{Mut-WT}_{Bind\ B} = \Delta E^{Mut}_{Bind\ B} - \Delta E^{WT}_{Bind\ B}$$
$$\Delta\Delta E^{Mut-WT}_{Bind\ A} = \Delta E^{Mut}_{Bind\ A} - \Delta E^{WT}_{Bind\ A}$$

where,

$$\Delta E^{Mut}_{Bind\ B} = E^{Mut}_{Complex\ B} - E^{Mut}_{Receptor} - E_{Ligand\ B}$$
$$\Delta E^{WT}_{Bind\ B} = E^{WT}_{Complex\ B} - E^{WT}_{Receptor} - E_{Ligand\ B}$$
$$\Delta E^{Mut}_{Bind\ A} = E^{Mut}_{Complex\ A} - E^{Mut}_{Receptor} - E_{Ligand\ A}$$
$$\Delta E^{WT}_{Bind\ A} = E^{WT}_{Complex\ A} - E^{WT}_{Receptor} - E_{Ligand\ A}$$

subject to

$$\Delta\Delta E^{Mut-WT}_{Fold\ B} < \text{Fold Cutoff}$$
$$\Delta\Delta E^{Mut-WT}_{Fold\ A} < \text{Fold Cutoff}$$

where

$$\Delta\Delta E^{Mut-WT}_{Fold\ B} = \Delta E^{Mut}_{Fold\ B} - \Delta E^{WT}_{Fold\ B}$$
$$\Delta\Delta E^{Mut-WT}_{Fold\ A} = \Delta E^{Mut}_{Fold\ A} - \Delta E^{WT}_{Fold\ A}$$

and

$$\Delta E^{Mut}_{Fold\ B} = E^{Mut}_{Receptor\ B} - E^{Mut}_{Unfolded}$$
$$\Delta E^{WT}_{Fold\ B} = E^{WT}_{Receptor\ B} - E^{WT}_{Unfolded}$$
$$\Delta E^{Mut}_{Fold\ A} = E^{Mut}_{Receptor\ A} - E^{Mut}_{Unfolded}$$
$$\Delta E^{WT}_{Fold\ A} = E^{WT}_{Receptor\ A} - E^{WT}_{Unfolded}$$

To accomplish this, for each mutant and for both binding events, four separate global minimum energy conformations (GMEC) were computed using the Dead End Elimination / A* based approach (Lippow, Wittrup, and Tidor 2007).

$$\Delta E^{Mut}_{Fold+Bind\ B} = E^{Mut}_{Complex\ B} - E^{Mut}_{Unfolded} - E_{Ligand\ B}$$
$$\Delta E^{WT}_{Fold+Bind\ B} = E^{WT}_{Complex\ B} - E^{WT}_{Unfolded} - E_{Ligand\ B}$$
$$\Delta E^{Mut}_{Fold+Bind\ A} = E^{Mut}_{Complex\ A} - E^{Mut}_{Unfolded} - E_{Ligand\ A}$$
$$\Delta E^{WT}_{Fold+Bind\ A} = E^{WT}_{Complex\ A} - E^{WT}_{Unfolded} - E_{Ligand\ A}$$

Note that the difference of the four above terms is equivalent to $\Delta\Delta\Delta E^{Mut-WT}_{Bind\ B-Bind\ A}$:

$$\Delta\Delta\Delta E^{Mut-WT}_{Bind\ B-A} = (\Delta E^{Mut}_{Fold+Bind\ B} - \Delta E^{WT}_{Fold+Bind\ B}) - (\Delta E^{Mut}_{Fold+Bind\ A} - \Delta E^{WT}_{Fold+Bind\ A})$$

Mutants were then sorted on $\Delta\Delta\Delta E^{Mut-WT}_{Bind\ B-Bind\ A}$ and filtered with a fold cutoff of 15 kcal/mol in order to identify thiolase sequences with predicted differential selectivity as compared to wild type toward accommodating the butyryl group as opposed to the acetyl group. Sequences were then sorted on $\Delta\Delta E^{Mut-WT}_{Bind\ B}$ to allow identification of mutants for testing that were also predicted to accommodate the butyryl group more favorably than wild type, and not

just resulting in improved differential specificity by accommodating both acetyl and butyryl groups more poorly than wild type, but with the acetyl worse than butyryl.

After optimized structures of mutants were computed and sequences were sorted, the dominant energetic interactions contributing to $\Delta\Delta\Delta E_{Bind\ B-Bind\ A}^{Mut-WT}$ for the top mutants were computed and analyzed. The top sets of four structures for each mutant for each binding event were manually inspected using this information. An example of the energetic breakdowns and four GMEC structures for one experimentally tested mutant in Bind 1 and Bind 2 is presented in Supplementary Figures 3-4 and Supplementary Tables III-IV.

The 17 positions allowed to mutate in the PhbA design calculations were: V57, Q87, L88, S91, L93, D146, L148, T149, D150, M157, M288, N316, I350, S353, L377, I379 and Q64. The 20 positions allowed to mutate in BktB design calculations were: V57, R88, L89, S92, L94, A148, L149, H150, D151, M158, M290, N318, A320, F321, I352, T355, M379, I381, I387 and Y66.

Rotamer Library

For each mutant calculation, a dead-end elimination / A* based rotameric search was performed following the methodology of (Lippow, Wittrup, and Tidor 2007), with all residues within 4.75 Å of the mutated residue allowed to relax. For the rotameric search, a modified version of the Dunbrack rotamer library (Dunbrack and Cohen, 1997) was used in which χ_1 and χ_2 were expanded by ± 10° from the crystal structure rotamers. The substrates and acylated cysteines were rotamerized in a manner to allow the acyl groups to rotate while keeping the rest of the atoms' positions fixed to that of the crystal structure. Using the atomic nomenclature introduced in Figure 1D, for acetyl-CoA, acetyl-C, butyryl-CoA and butyryl-C the dihedrals Sγ–Cδ–Cε–Oδ and Cβ–Sγ–Cδ–Cε were allowed to rotate by ± 10° from the corresponding crystal structure rotamers. For acetyl-CoA and acetyl-C, all atoms except Cδ, Oδ and Cε were held fixed. For butyryl-CoA and butyryl-C, the dihedrals Sγ–Cδ–Cε–Cζ and Cδ–Cε–Cζ–Cη were allowed to rotate in 30° increments, with atoms except Sγ, Cβ, Cδ, Oδ, Cε, and Cζ held fixed.

Results

Computational design of mutants predicted to exhibit increased selectivity in *Z. ramigera* PhbA thiolase

The structure of *Z. ramigera* PhbA thiolase has been well studied, with crystal structures representing each step of the catalytic cycle, a total of 22 structures, including the following mutants: C89A, N316A/H/D, H348A/N, N316H-H348N and Q64A (Meriläinen et al. 2009, 2008; Y Modis and Wierenga 2000; Yorgo Modis and Wierenga 1999; Kursula et al. 2002). Especially relevant to this study were the structures of the C89A mutant with acetyl-CoA bound (1M3Z; 1.7 Å), the wild-type thiolase with acetyl-CoA bound and C89 acetylated (1DM3; 2.0 Å), and unliganded wild-type thiolase with C89 butyrylated (1M4T; 1.77 Å), as these provided a structural basis for examining acyl group specificity at each binding event (Meriläinen et al. 2009, 2008; Y Modis and Wierenga 2000; Yorgo Modis and Wierenga 1999; Kursula et al. 2002). Due to this wealth of available crystallographic data, PhbA was chosen as the starting point for structure-based design calculations. Note that no published BktB crystal structures existed at the start of this study.

In the case of the 3HA pathway, we were interested in improving the overall pathway selectivity ratio, i.e. the production of the longer chain C6, BA product relative to AA, using the nomenclature in Figure 2. At the thiolase level, this ratio could be improved either by increasing the formation of 3-oxohexanoyl-CoA (BA), by decreasing the formation of acetoacetyl-CoA (AA), or a combination of the two approaches. There are several possible steps in the thiolase catalytic cycle where BA production might be limited compared to AA production. For example, steric constraints might lead the thiolase active site to preferentially accommodate acetyl-CoA relative to butyryl-CoA during the priming CoA binding step ("Bind 1" in Figure 1C). Similarly, steric constraints could also prevent the thiolase active site from accommodating butyrylated-

C89 relative to acetylated-C89 in a conformation favorable to nucleophilic attack by the acetyl carbon of acetyl-CoA ("Bind 2" in Figure 1C). Effectively, the butyryl group must be accommodated in at least two orientations in the active site: on the bound priming butyryl-CoA, and on the butyrylated catalytic C89.

While it is possible that one of the catalytic steps (e.g. proton abstraction, breakdown of acyl-enzyme intermediate) might limit BA production, crystallographic studies by Kursula et al (2002) suggest that butyrylation of C89 inhibits catalytically productive binding of the extending acetyl-CoA. Kursula et al (2002) report that soaking experiments with butyryl-CoA and wild-type PhbA crystals result in butyrylation of C89 with no detectable CoA bound, indicating that butyryl-CoA is able to act as the priming acyl-CoA, but not as the extending acyl-CoA once the enzyme is butyrylated. We observe very low levels of 3HH formation *in vivo* with wildtype PhbA, suggesting poor PhbA activity with butyryl-CoA as the priming acyl-CoA and acetyl-CoA as the extending acyl-CoA. It should also be noted that on studies of ketosynthase domains in polyketide synthetases, which exhibit a similar bi bi ping-pong mechanism, it has been reported that the extending step is more often the bottleneck for acceptance of alternative substrates than the priming/acylation step (Jenner et al., 2015).

Superimposing the butyrylated C89 structure (which corresponds to step 4 of the catalytic cycle in Figure 1C; PDB: 1M4T) upon the acetylated structure with acetyl-CoA bound as the extending acyl-CoA (representing step 5 in Figure 1C; PDB: 1M3Z) reveals that the butyryl group of C89 lies directly over the sulfur atom of the bound extending acetyl-CoA, with the butyryl group pointing directly into the hydrophobic pocket formed by conserved active site residues M157 and M288, and the thioester oxygen pointing into the oxyanion hole formed by N(C89) and N(G380) (Kursula et al. 2002). Modis and Wierenga (1999) suggest that M157 and

M288 in the PhbA active site prevent the accommodation of larger acyl-CoA substrates, although they did not test these observations experimentally (Modis and Wierenga, 1999b). We sought to develop an *in silico* model that would allow prediction of the ability of mutations at these positions, as well as additional positions to allow the butyrylated C89 to take on a conformation more favorable to catalytically productive binding of acetyl-CoA as the extending acyl-CoA.

Rather than building models of the transition state for each step of the condensation reaction and optimizing the active site binding of the transition state associated with each catalytic step leading to BA production, we chose to build on the published crystal structures of acetyl-CoA bound as the priming acyl-CoA (1M3Z) and acetyl-CoA bound as the extending acyl-CoA with C89 acetylated (1DM3), assume these crystal structures represent catalytically productive binding modes at each step, and identify mutations that could accommodate a butyryl group in the appropriate place while keeping the rest of the crystal structure fixed outside of a defined radius (4.75 Å) of the residue to be mutated (see **Computational Methodology**).

Although poor binding affinity of the extending acetyl-CoA due to the native conformation of butyryl-C89 was likely the primary driver for poor BA production, it was nonetheless important to consider the effect of active site mutations on the ability to accommodate butyryl-CoA as the priming acyl-CoA. If a thiolase mutant was able to accommodate the butyryl group in Bind 2, but as a result of the mutation was unable to accommodate the butyryl group in Bind 1, then this would likely lead to poor BA production. Although they observe that butyryl-CoA is capable of acting as the priming acyl-CoA with wild type PhbA, Kursula, et al (2002) also report poor (mM) affinity of PhbA for butyryl-CoA. It was

critical that designed mutants did not further decrease this affinity, or butyryl-CoA may no longer be capable of acting as the priming acyl-CoA.

We thus performed design calculations on conformations representing both Bind 1 and Bind 2. We chose to focus structure-based design calculations on identifying mutations with the potential to improve the energy of bound conformations leading to BA relative to those leading to AA, at either the first or second Michaelis complex (steps 2 and 5 in Figure 1C, respectively). Design calculations were performed as described in Methods, and Table I lists the PhbA mutants chosen for experimental testing along with their corresponding values of $\Delta\Delta E^{Mut-WT}_{Bind\ B}$, $\Delta\Delta E^{Mut-WT}_{Bind\ A}$, and $\Delta\Delta\Delta E^{Mut-WT}_{Bind\ B-Bind\ A}$ for both Bind 1 and Bind 2.

Note that mutants presented in Table I and chosen for experimental validation involve paring down of a bulky hydrophobic (L88, M157, M288, L377) residue to a smaller residue, such as serine, alanine or glycine. Also note that all mutants except M288A and M288G have negative values of $\Delta\Delta\Delta E^{Mut-WT}_{Bind\ B-Bind\ A}$ in Bind 1, Bind 2 or both Bind1 and Bind 2. All mutants chosen for experimental testing also have negative values of $\Delta\Delta E^{Mut-WT}_{Bind\ B}$ in either Bind 1 or Bind 2. All mutants also have positive values of $\Delta\Delta E^{Mut-WT}_{Bind\ A}$ in both steps, indicating decreased binding preference for accommodating for the acetyl group in both binding events.

Of all positions, M157 was judged the most promising candidate due to its negative values of $\Delta\Delta\Delta E^{Mut-WT}_{Bind\ B-Bind\ A}$ in both binding events, the high magnitude of $\Delta\Delta\Delta E^{Mut-WT}_{Bind\ B-Bind\ A}$ relative to the other mutants, and the fact that the same trend of $\Delta\Delta\Delta E^{Mut-WT}_{Bind\ B-Bind\ A}$ was exhibited for the similar mutations of M157S/A/G.

Because according to energetic calculations and upon inspection M288S appeared to be a promising candidate for improving selectivity in Bind 2 but not M288A and M288G, M288A and M288G were also chosen for testing to account for the possibility that the model might not be able to accurately distinguish the small chemical differences between serine, alanine and glycine. This position was also included because previous crystallographic studies of PhbA hypothesized that the bulky hydrophobic group of M288 (along with M157) prevents the accommodation of larger acyl-CoA substrates (Y Modis and Wierenga 1999). Figures 3A-D show the location of the residues chosen for PhbA mutagenesis relative to the active site catalytic residues in both the Bind 1 and Bind 2 orientations.

Initial screening of *Z. ramigera* PhbA mutants identifies several improved enzyme variants

Z. ramigera PhbA thiolase mutants were initially assayed *in vivo* in the context of a previously established pathway for 3-hydroxyalkanoic acid (3HA) (Martin et al. 2013b) production. This pathway consists of an activator enzyme (Pct, *M. elsdenii*), a thiolase (BktB from *C. necator* or PhbA from *Z. ramigera*), an NADPH dependent reductase (PhaB from *C. necator*), and a thioesterase (TesB from *E. coli*), which generates the final 3HA product. Specifically, when the cultures are supplied with butyrate and grown on glucose, the cells produce 3HB and 3HH. Examining the amount of 3HH produced relative to 3HB provides a measure of thiolase selectivity.

Of the twelve tested thiolase variants, several resulted in increased selectivity ratios *in vivo* (Figure 4A). This higher selectivity ratio is mostly due to decreased production of 3HB by the pathway, and not increased 3HH titers (Figure 4B). Specifically, five mutants: M157A/G and M288S/A/G resulted in an approximately 30-fold higher ratio of 3HH relative to the

undesired 3HB by-product, with a roughly 80-fold decrease in their sum. Motivated by these results, we wanted to further characterize the most selective mutants. Because the extent to which the enzymes downstream of the thiolase could affect final product distribution was unknown, we also wanted to assay the mutants within the context of another pathway. Thioesterases exhibit varying levels of activity towards different acyl-CoA substrates, depending on the carbon chain length and functional group of the substrate (McMahon and Prather 2014). The PHA biosynthesis pathway was thus subsequently used to screen the thiolase mutants because it is known that over 100 different 3HA monomers can be incorporated into PHAs, suggesting a broad substrate range for the PHA synthase (Agnew and Pfleger, 2013). We chose to use the PhaC2 polymerase enzyme from *R. aetherivorans* I24 because it has been previously employed to synthesize PHA polymers with large amounts of the longer chain C6 monomer, 3-hydroxyhexanoate, 3HHx (Budde et al. 2011). Using the polymerase as the terminal enzyme removes any possible limitation or specificity imposed by the thioesterase, providing further evidence for thiolase imposed selectivity on the distribution of observed products.

When the most selective PhbA thiolase variants were profiled using the PHA assay, M157 mutants resulted in an 18-fold higher 3HHx:3HB selectivity ratio (Figure 4C). The resulting PHA polymers synthesized by M158A/G/S thiolase mutants contained 83-85 mol% of the 3HHx monomer, as compared to wild type, which only resulted in a 22 mol% of the 3HHx monomer (Table III). To eliminate the possibility that native *E. coli* thiolases or reductases could influence final PHA composition, we performed several control experiments; no PHA accumulation was observed without plasmid-based overexpression of all four genes of the pathway (data not shown). This assay was consistent with our previous observations and supports our initial hypothesis of these thiolases exhibiting reduced activity towards the

condensation of two acetyl-CoAs, while maintaining similar or better activity towards the condensation of butyryl-CoA and acetyl-CoA compared to the wild type enzyme.

Validated computational approach applied to more active BktB thiolase from *C. necator*

Having successfully identified mutants with increased C6/C4 (BA/AA) selectivity in the PhbA thiolase, we applied the newly validated modeling framework to identify mutants that might increase selectivity of the more active *C. necator* BktB thiolase. Although the BktB thiolase only exhibits 51% sequence identity with PhbA, the active site is highly similar, with 86% of the residues within 10 Å of the PhbA acetyl-CoA carbonyl center conserved between PhbA and BktB (Supplementary Table I). Two unliganded crystal structures were available in the PDB for BktB (E. J. Kim et al. 2014b; Fage, Meinke, and Keatinge-Clay 2015), and due to the active-site similarity, the *Z. ramigera* PhbA structures,1M3Z and 1DM3 were used as templates to build structures of BktB with acetyl-CoA and butyryl-CoA bound.

The results of the computational model applied to the BktB thiolase are shown in Table II. Given the active-site similarity, it was not surprising that two BktB residues with analogous PhbA positions (M157/M158, M288/M290) were also predicted to improve BA/AA selectivity. Additionally, a position unique to BktB was predicted, Y66, which is part of a loop that comprises the major structural difference between the PhbA and BktB active sites. The positions of the BktB residues chosen for mutagenesis relative to the Bind 1 and Bind 2 conformations are shown in Figures 5A-D.

BktB thiolases enable synthesis of PHAs enriched in 3HHx

The BktB thiolase has been previously used by us and other groups to achieve synthesis of longer (>C4) and branched chain acids, aldehydes and alcohols by the same CoA dependent

pathway (Dhamankar et al. 2014; Hsien-Chung Tseng et al. 2009; S. Kim, Clomburg, and Gonzalez 2015; Cheong, Clomburg, and Gonzalez 2016a). Of the M158, M290 and Y66 mutants assayed, the M158 mutants resulted in the highest selectivity ratios, with M158G and M158S exhibiting selectivity ratios 10-fold greater than wild type for 3HHx in PHAs (Figure 6A). Based on previous reports of their activities, it was not surprising that wild type baseline selectivity was higher for BktB at 3.45 compared to PhbA at 0.292 (Slater et al. 1998). The PHA polymers isolated from *E. coli* strains expressing these mutants varied from 77 to 97 mol% 3HHx (Table III), with BktB M158A mutant resulting in the highest yields of 3HHx as a percentage of the CDW (Figure 6B). Protein gels of lysates of strains expressing wild type vs. mutant enzymes showed no significant difference in the soluble expression level of the thiolase enzymes, pointing to differences in the activities of these enzymes (Supplementary Figure 1). It was surprising that the M290 mutants resulted in very low yields of PHAs *in vivo*. Although it is possible that BktB expression or solubility was affected as a result of this mutation, soluble expression was detected via a protein gel.

In vitro characterization of BktB thiolase mutants with highest selectivity ratios

Having achieved increased selectivity ratios of the 3HHx:3HB in the PHA polymers with our computationally designed mutants, we next studied the effects of the M158 mutations on thiolase activity. Our *in vivo* data suggested that we were able to obtain increased selectivity ratios due to decreased activity of these mutant thiolases for the formation of the AA condensation product (and subsequently 3HB), and not due to increased activity towards the condensation of butyryl-CoA with acetyl-CoA, which results in the formation of 3-oxo-hexanoyl-CoA (and 3HH-CoA upon reduction). We thus sought to purify and assay both the mutant and wild type BktB thiolases *in vitro* to remove many of the confounding variables

present *in vivo*. For example, differences in stability of the enzymes as well as fluctuating pools of substrates and coenzymes could influence thiolase activity. Further, activities of the downstream enzymes could also influence the final product distribution. Each wild type and mutant enzyme was purified as a His-tagged fusion protein to homogeneity and assayed in the condensation direction with acetyl-CoA, and thiolysis directions with acetoacetyl-CoA (AA) and 3-oxo-hexanoyl-CoA (BA). *In* vitro characterization of the BktB WT and M158A enzymes reveal a 10-fold lower catalytic activity of the mutant towards the condensation of two acetyl-CoA molecules (Table IV). This result is consistent with *in vivo* observations of reduced 3HB product formation which arises from the condensation of two acetyl-CoAs. From the *in vitro* kinetic parameters it can be concluded that the M158 mutants do indeed have lower catalytic efficiencies towards the formation and degradation of AA, the C4 product (1.52×10^4 vs. 1.46×10^3 $M^{-1}sec^{-1}$, WT vs mutant), whereas the thiolysis k_{cat}/K_m value towards the degradation of BA, the C6 product, is 3-fold higher as compared to wild type ($2.67x10^5$ vs. $9.82x10^5$, WT vs mutant, Table IV). In all, there is an 80-fold improvement in the selectivity ratio of the M158A thiolase as compared to wild type. Further, activity measurements of the BktB M290A mutant revealed a very low k_{cat} for the condensation of two acetyl-CoAs consistent with *in vivo* observations (data not shown).

Using *C. necator* BktB M158A mutant allows for biosynthesis of PHAs enriched in 3HHx from glucose as sole carbon source

Having demonstrated increased selectivity for the BktB M158A mutant *in vivo* while supplying both butyrate and glucose, we next wanted to determine if this mutant could allow for more selective synthesis of longer chain product using glucose as the sole source of carbon. We sought to use the same model system as before, except that now we had to overexpress additional enzymes that would allow for conversion of 3-hydroxybutyryl-CoA to butyryl-CoA. Trans-

enoyl-CoA reductase, Ter from *Treponema denticola* was cloned into the first MCS of pCDFDuet and enoyl-CoA hydratase, PhaJ4b from *C. necator* was cloned into an operon with PhaC2 generating pCDFDuet(ter_{Td})-(phaC2-phaJ4). This vector, along with pETDuet(BktB WT or M158A)-(phaB) was used to transform *E. coli* MG1655(DE3) and the strain was grown in M9 medium with glucose as a sole carbon source. Figure 7 shows that the residual activity of BktB M158A towards the condensation of two acetyl-CoAs was sufficient to allow for formation of butyryl-CoA and subsequently 3-hydroxybutyryl-CoA. Using the BktB M158A mutant led to an almost 2-fold increase in selectivity for the 3HHx monomer as compared to using wild type BktB, though the overall yield of PHAs was low. However, we used an almost wild type *E. coli* for all production experiments in this work, and it is likely that strain engineering to increase precursor supply and elimination of competing pathways will lead to increased product yields.

Discussion

In this work we present a rational design framework for increasing the thiolase selectivity ratio, which we define as the ratio of C6 to C4 condensation products. We then apply this framework to two related biosynthetic thiolases, PhbA from *Z. ramigera* and BktB from *C. necator*. *In vivo*, we observe the synthesis of PHAs that are highly enriched for 3HHx (C6) when our rationally selected mutants are employed. *In vitro* characterization of one of the most selective mutants (M158A) revealed a 10-fold reduction in activity for formation and breakdown of the C4 product with uncompromised thiolysis activity toward the C6 substrate as compared to the wild type enzyme.

Although designed thiolase mutants exhibited nearly 10-fold improvements in the selectivity ratio, this increase was primarily driven by the reduced ability of the thiolase to synthesize AA (acetoacetyl-CoA, C4) and not improved ability to synthesize BA (3-oxo-hexanoyl-CoA, C6). The decreased synthesis of C4 products by the mutants we tested is consistent with *in silico* predictions, as all mutants tested exhibited positive computed values of $\Delta\Delta E_{Bind\ A}^{Mut-WT}$ for both Bind 1 and Bind 2. From the *in vitro* kinetic characterization of the BktB M158A mutant, the reduced rate of condensation of two acetyl-CoA substrates is consistent with reduced 3HB production within both pathway contexts (3HA and PHA).

The fact that all mutants tested failed to significantly increase 3HH titers was not consistent with *in silico* predictions however. With the exception of M288A/G, all mutants tested had negative computed values of $\Delta\Delta E_{Bind\ B}^{Mut-WT}$ for either Bind 1 or Bind 2, meaning that each of the tested mutants were expected to preferentially accommodate the butyryl group in either the first (PhbA L377S/G and BktB M290S/A/G), second (PhbA L88A/G, PhbA M288S, BktB

Y66N/V/T/A/G) or both (PhbA L88S, PhbA M157S/A/G, PhbA L377A, BktB M158S/A/G, and Y66Q/S) binding events.

It remains entirely possible that activities of downstream enzymes on the longer chain substrates limited 3HH production *in vivo*. Activities of all downstream enzymes (3-ketoacyl-CoA reductase and/or thioesterase and PHA polymerase) with the pathway acyl-CoA intermediates must be examined to rule this out. This is challenging due to the commercial unavailability of required substrates as well as lack of robust assays as is the case with PHA polymerase, in which the observed *in vitro* and *in vivo* substrate specificities differ (Stubbe and Tian 2003; Yuan et al. 2001).

In vitro, we were unable to directly assay the rate of condensation between acetyl-CoA and butyryl-CoA to test whether mutants exhibited increased production of 3-oxo-hexanoyl-CoA (and subsequently 3-HH-CoA) in the absence of any potential confounding factors *in vivo*. While we can assay the thiolase in the thermodynamically favored direction, thiolysis, and observe higher activity of the M158A mutant with the C6 substrate, we cannot conclude that a similar rate enhancement results for the forward condensation direction. One might expect the observed increase in the C6 thiolysis rate to lead to decreased overall titers in the biosynthetic direction; however, one must keep in mind that the reductase which is present in both *in vivo* and *in vitro* contexts is necessary to allow for formation and detection of condensation products. For example, when PHB synthesis was modeled *in vitro*, inclusion of the reductase enzyme was necessary to observe accumulation of the 3-ketoacyl-CoA condensation product (Burns et al. 2007). Put another way, both *in vivo* and *in vitro*, the thiolase must always be coupled to the reductase and the substrate specificity and activity of the reductase enzyme will influence the behavior of the overall system. Indeed, a similar approach has been used to model the kinetics of

in vivo PHB accumulation (Leaf and Srienc, 1998; van Wegen et al, 2001). For this reason, the system must necessarily be examined whilst considering, at a minimum, the thiolase and reductase enzymes in combination.

Given that the butyryl group must be accommodated at two distinct locations within the active site, it is possible that multiple active-site mutations, rather than the point mutations tested in this study, are required to improve C6 product titers beyond that of wild type. When the butyryl group is built onto the acetyl-CoA Cε carbon in structure 1M3Z (representing Bind 1) in its minimum energy planar zigzag conformation, the Cη atom of the butyryl group clashes directly with the backbone atoms of I379 and C378. The four non-polar, non-charged residues within 5 Å of the Cε carbon of acetyl-CoA in this structure, and the non-hydrogen atoms closest to the Cε carbon (side chain atom followed by distance in parenthesis) include M157 (S; 5.3 Å), M288 (S; 3.1 Å), A318 (Cß; 4.5 Å), I379 (Cß; 5.3 Å). Of these, M288 is the only residue that is directly in a position to clash with the butyryl group in a non-planar conformation, as the other three side chains are either too far away, or point away from the substrate. Similarly, if a butyryl group is built onto acetyl-C89 from structure 1DM3 (representing Bind 1) it clashes directly with G380. The three nonpolar, non-charged side chains within 5 Å of the acetyl carbon on acetylated C89 are L88 (Cß; 3.2 Å), M157 (S; 4.9 Å away) and I379 (Cγ2; 4.5 Å). Of these, L88 and M157 are potentially positioned to clash with the C89 butyryl group in a non-planar conformation.

M157 is thus the only PhbA active site residue that satisfied the energetic filtering criteria and was positioned to directly relieve a steric clash imposed by a bulky butyryl group in both binding events within the fixed-backbone context of this work. Downstream enzyme specificities aside, from a modeling standpoint its failure to improve 3HH production may potentially be due to a number of factors including the fixed backbone modeling approach, the fact that the catalytic

residues were fixed to their crystal structure conformations, failure to consider the correct transition state conformation, or a decrease in stability as a result of the mutations.

Applying the same analysis to BktB, If the butyryl group is added to the acetyl-CoA in its planar zigzag conformation (representing Bind 1), it clashes directly with the backbone atoms of C380 and I381. The nonpolar, non charged residues within 5 Å and their corresponding closest atoms are M158 (S; 4.6 Å), M290 (S; 3.1 Å), I381 (Cß; 5.6 Å), A320 (Cß; 4.2 Å), F321 (Cε2; 5.4 Å). Of these only M158 and M290 are positioned to directly clash with a non-planar butyryl C90. For Bktb Bind 2, the butyryl group in its planar zigzag conformation clashes with the sidechains Y66 (the Cη of the butyryl group as lies 2.3 Å from the Cε2 of Y66), L89 (the C4 of the butyryl group as lies 2.8 Å from the Cδ2 of Le89), and G382. The nonpolar, non-charged residues within 5 Å are M158 (S; 5.0 Å), I352 (Cγ2; 4.6 Å), L89 (Cß; 3.0 Å), Ile381 (Cγ2; 4.7 Å). Of these, L89, Y66, M158 and I352 are in orientations that could potentially clash with a butyryl C90. Only residues M290, M158, and Y66 met the energetic filtering criteria. Again, analogous to the PhbA case, M158 is the only side chain positioned to directly relieve a steric clash imposed by a butyryl group at both binding events.

The relative success of M157/M158 may be related to its location and orientation between the acyl group of the CoA substrate and the catalytic residue C89/C90. The fact that degradative thiolases, which are known to be able to accommodate >C6 substrates also have methionines at positions 290 and 158 suggests that mutations at other positions play a role in accommodating >C6 substrates (Fage et al., 2015; Modis and Wierenga, 1999b). Mutations at multiple positions, although beyond the scope of this work, may well prove to be a fruitful starting point for future thiolase engineering attempts.

From a practical metabolic engineering standpoint, the thiolase mutants identified in this study, specifically the BktB M158A thiolase, should be useful in other pathways where the condensation of acetyl-CoA and different acyl-CoA species is required (Sheppard et al. 2014; Cheong, Clomburg, and Gonzalez 2016b). In addition, we have shown that the thiolase can be used to modulate PHA polymer composition, resulting in PHAs that are highly enriched for medium-chain length monomers. Typically, PHA composition is modulated by process engineering such as novel feeding strategies and choice of feedstock, as wells as various strain engineering strategies to remove endogenous competing enzymes from native PHA synthesizing microbes. Using the thiolase to control PHA monomer composition opens up a new avenue for achieving the synthesis of PHAs with specific, desired properties for diverse applications.

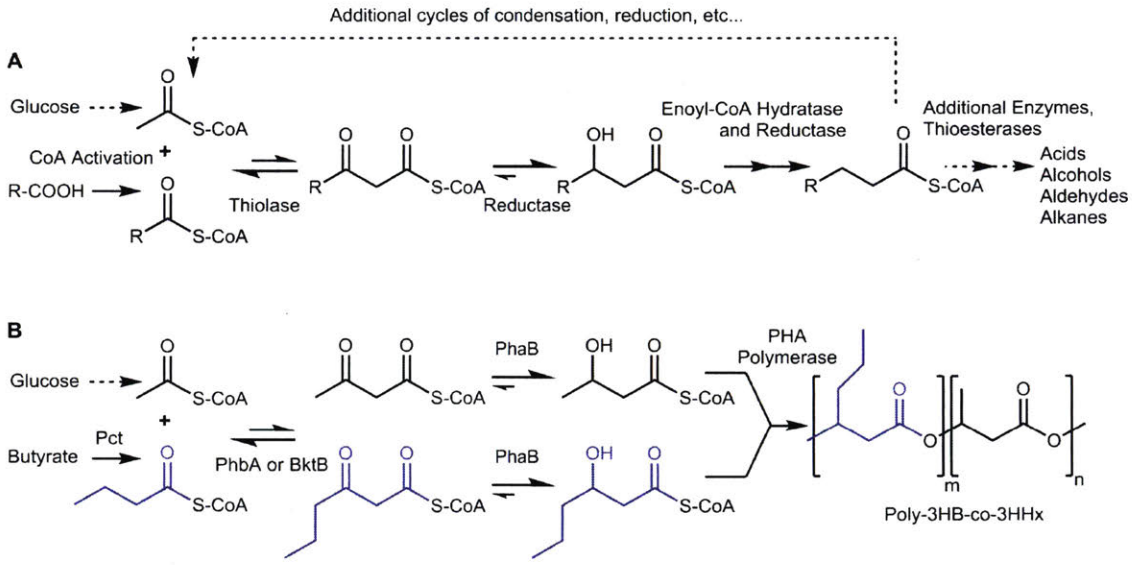
Acknowledgements

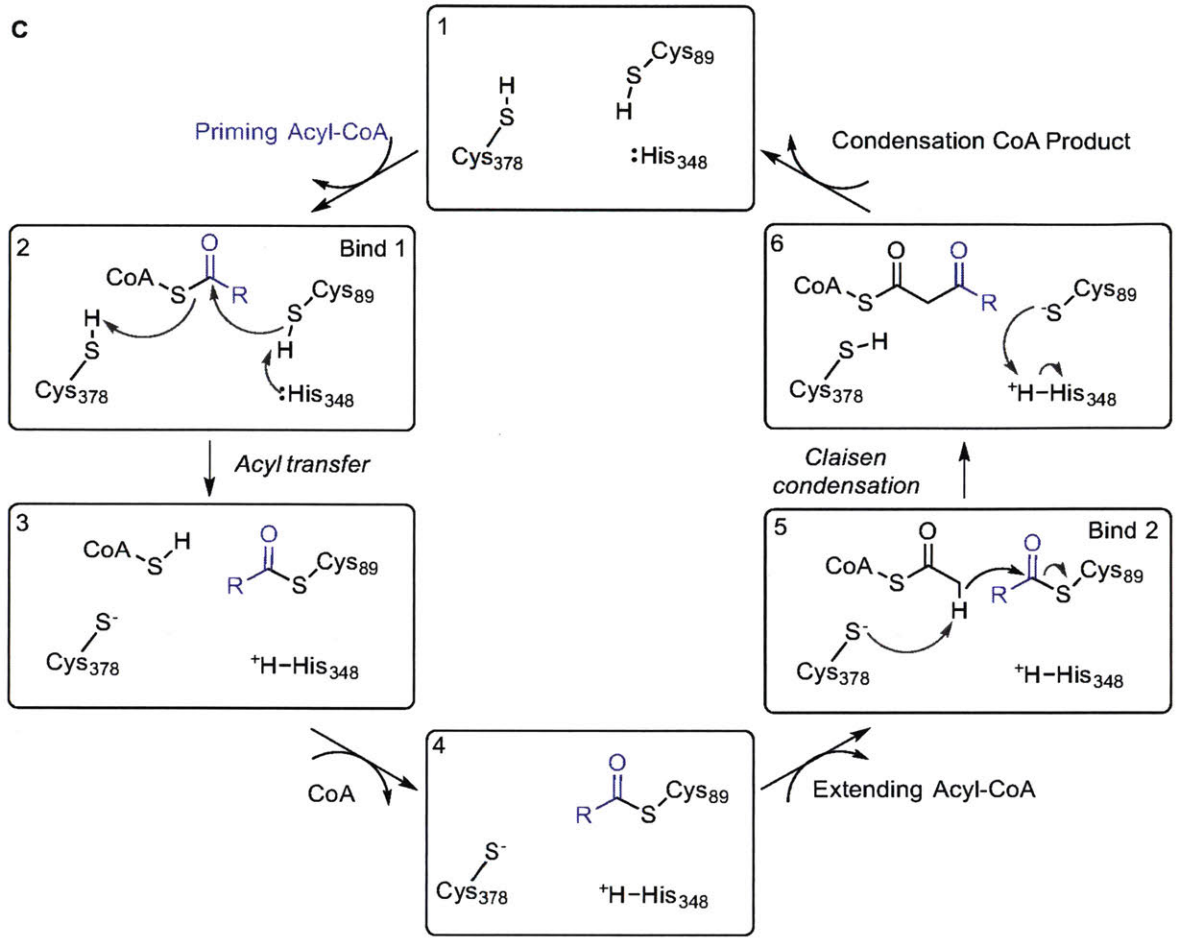
We thank Michael Blaisse and Prof. Michelle C.Y. Chang (University of California, Berkeley USA Department of Chemistry) for sharing the protocol for the synthesis of 3-oxo-acyl-CoA compounds and Dr. Denyce Wicht (Suffolk University) for her assistance with organic synthesis.

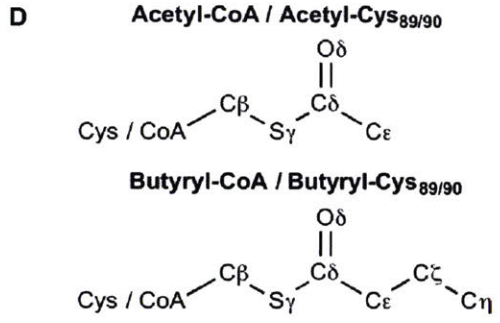
This work was supported under the Cooperative Agreement between the Masdar Institute of Science and Technology, Abu Dhabi, UAE and the Massachusetts Institute of Technology, Cambridge, MA, USA, Reference Number 02/MI/MIT/CP/11/07633/GEN/G/00 (Tarasova), as well as the NDSEG Fellowship and NIH Grants R01 GM082209 and R01 GM065418 (Bonk).

Figures

Figure 1







A Generalized 3HA pathway, which is also referred to as CoA-dependent chain elongation or reverse β-oxidation. This pathway consists of four core enzymes – a coenzyme-A (CoA) activating enzyme which converts a small acid precursor to a CoA thioester, a thiolase which brings about the condensation of the CoA activated acid and acetyl-CoA, a reductase which reduces the β-carbonyl of the resulting longer chain intermediate, and finally a thioesterase

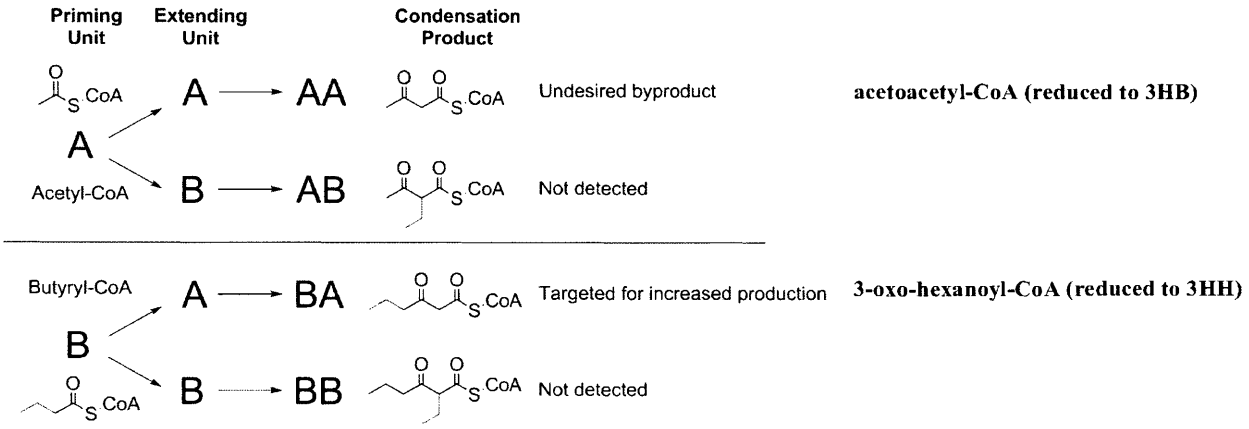
which cleaves the thioester bond of the 3-hydroxyacyl-CoA, releasing free CoASH and the free 3-hydroxyacid. A wide variety of other compounds can be produced by addition of other enzymes that can act on the 3-hydroxyacyl-CoA intermediates, such as enoyl-CoA dehydratases and reductases, and alcohol and aldehyde dehydrogenases. Biosynthesis of longer chain 3HAs and carboxylic acids, as well as ω-carboxylic acids, and longer chain alcohols has been demonstrated (Cheong, Clomburg, and Gonzalez 2016c; Sheppard et al. 2014). However, a mix of products of variable chain lengths always results.

B In this study we employ a four-enzyme pathway for the synthesis of poly-3HB-co-3HHx as a readout of thiolase selectivity. The cells are grown on glucose and supplied with butyrate. Activation of butyrate by the action of Pct (*M. elsdenii)*, leads to butyryl-CoA which is then condensed with acetyl-CoA by a thiolase, either BktB (*C. necator*) or PhbA (*Z. ramigera*), to produce 3-oxohexanoyl-CoA. This intermediate is then reduced to 3HH-CoA by an acetoacetyl-CoA reductase PhaB (*C. necator*). The thiolase is also capable of condensing two acetyl-CoA molecules which leads to production of 3HB-CoA upon reduction by PhaB. The 3HA-CoA intermediates are then polymerized into PHAs by PhaC2 (*R. aetherivorans* I24).

C Reaction mechanism of the thiolase occurs by a biological Claisen condensation reaction though a sequential bi bi ping-pong mechanism. In addition to other thiolases, this mechanism is also similar to those utilized by acetyltransferase and ketosynthase domains of polyketide synthetases. Panel 2 corresponds to Bind 1 and Panel 5 corresponds to Bind 2 on which structure based design calculations were performed.

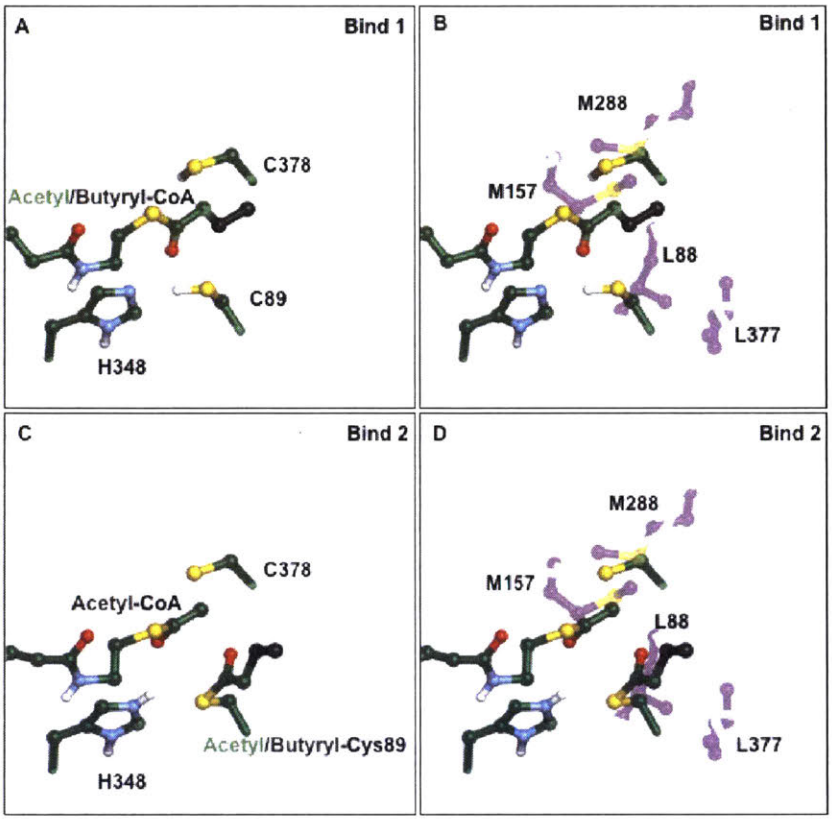
D Atomic nomenclature used throughout the rest of the paper.

Figure 2



Four different products can result from the condensation reaction of acetyl-CoA (A) and butyryl-CoA (B) catalyzed by the thiolase. The product formed depends on the order of addition of the acyl-CoAs into the active site of the enzyme. The priming acyl-CoA serves as an electrophile at the carbonyl carbon and forms an acyl-enzyme intermediate. The extending acyl-CoA in this case acts as a nucleophile after abstraction of an α proton and formation of a carbanion. Self-condensation of two acetyl-CoA molecules results in formation of acetoacetyl-CoA, which we term the AA condensation product, and subsequent reduction by PhaB leads to the formation of 3HB. Condensation with butyryl-CoA as the priming acyl-CoA and acetyl-CoA as the extending acyl-CoA forms 3-oxo-hexanoyl-CoA which we term the BA condensation product, and subsequent reduction by PhaB leads to the formation of 3HH. In this study we sought to increase the ratio of 3HH to 3HB by increasing the ratio of the BA condensation product relative to the AA condensation product.

Figure 3



A Structure of *Z. ramigera* PhbA thiolase active site during the first binding event (Bind 1, corresponding to step 2 in Figure 1C). The atoms colored black show the extra atoms of the butyryl group compared to the acetyl group that must be accommodated in order to preferentially produce 3-oxo-hexanoyl-CoA rather than acetoacetyl-CoA.

B Structure of *Z. ramigera* PhbA thiolase active site during first binding event with residues selected for mutation colored purple.

C Structure of *Z. ramigera* PhbA thiolase active site during second binding event (Bind 2, corresponding to step 5 in Figure 1C). Atoms colored black show the extra atoms that must be accommodated in order to preferentially produce 3-oxo-hexanoyl-CoA.

D Structure of *Z. ramigera* PhbA thiolase active site during second binding event (corresponding to step 5 in Figure 1C) with residues selected for mutation colored purple.

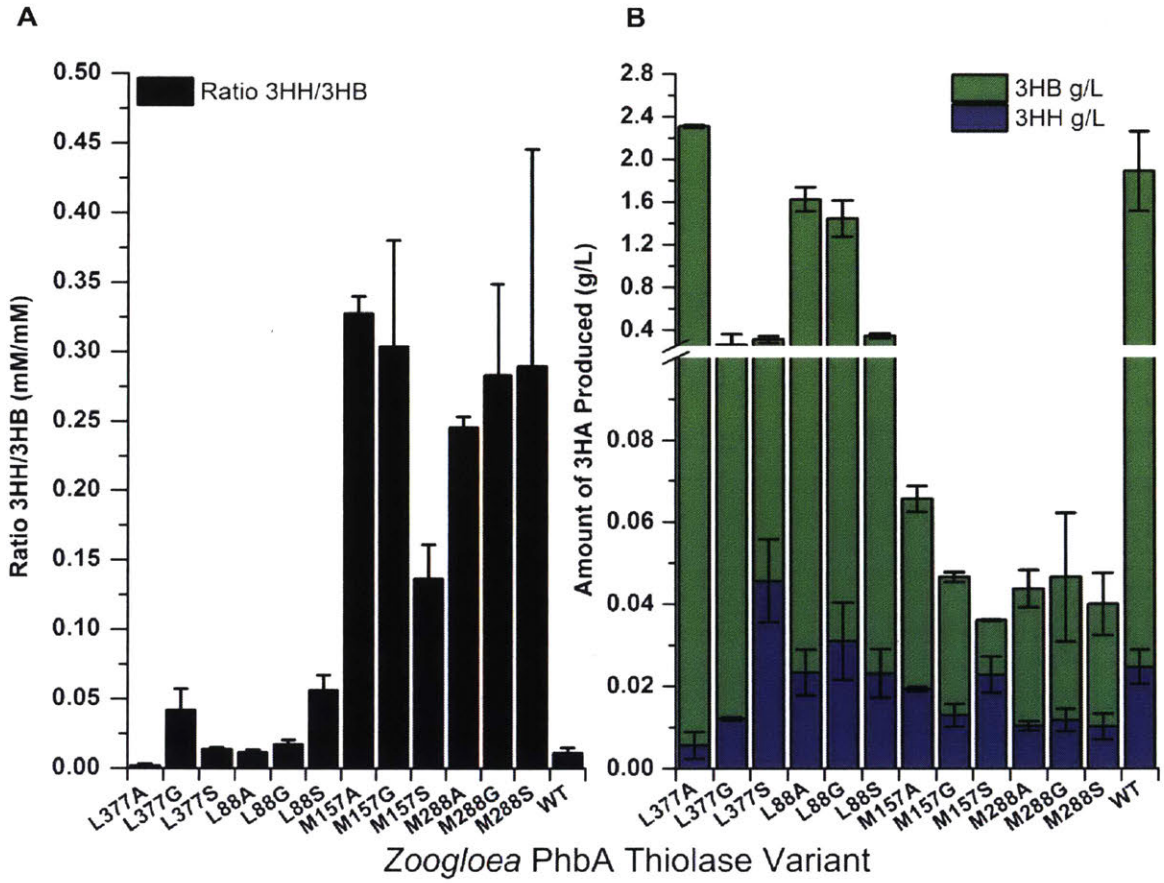
Table I. Energetic calculations of *Z. ramigera* PhbA mutants selected for experimental testing

Mutant	Bind 1[(a)]			Bind 2[(b)]		
	$\Delta\Delta E_{Bind\ B}^{Mut-WT}$	$\Delta\Delta E_{Bind\ A}^{Mut-WT}$	$\Delta\Delta\Delta E_{Bind\ B-A}^{Mut-WT}$	$\Delta\Delta E_{Bind\ B}^{Mut-WT}$	$\Delta\Delta E_{Bind\ A}^{Mut-WT}$	$\Delta\Delta\Delta E_{Bind\ B-A}^{Mut-WT}$
L88S	-0.09	0.09	-0.18	-1.05	0.09	-1.14
L88A	0.02	0.09	-0.07	-1.33	0.09	-1.42
L88G	0.53	0.15	0.38	-1.90	-0.26	-1.64
M157S	-18.09	0.83	-18.92	-3.98	1.32	-5.30
M157A	-17.36	1.38	-18.74	-2.76	1.88	-4.64
M157G	-16.34	2.20	-18.54	-1.34	2.37	-3.71
M288S	1.60	1.18	0.43	-0.03	0.48	-0.52
M288A	1.85	1.25	0.60	1.55	0.57	0.98
M288G	2.29	1.34	0.95	1.08	0.65	0.43
L377S	-0.65	0.02	-0.68	0.50	0.02	0.47
L377A	-1.01	0.07	-1.09	-0.30	0.08	-0.37
L377G	-1.14	0.11	-1.26	0.35	0.12	0.22

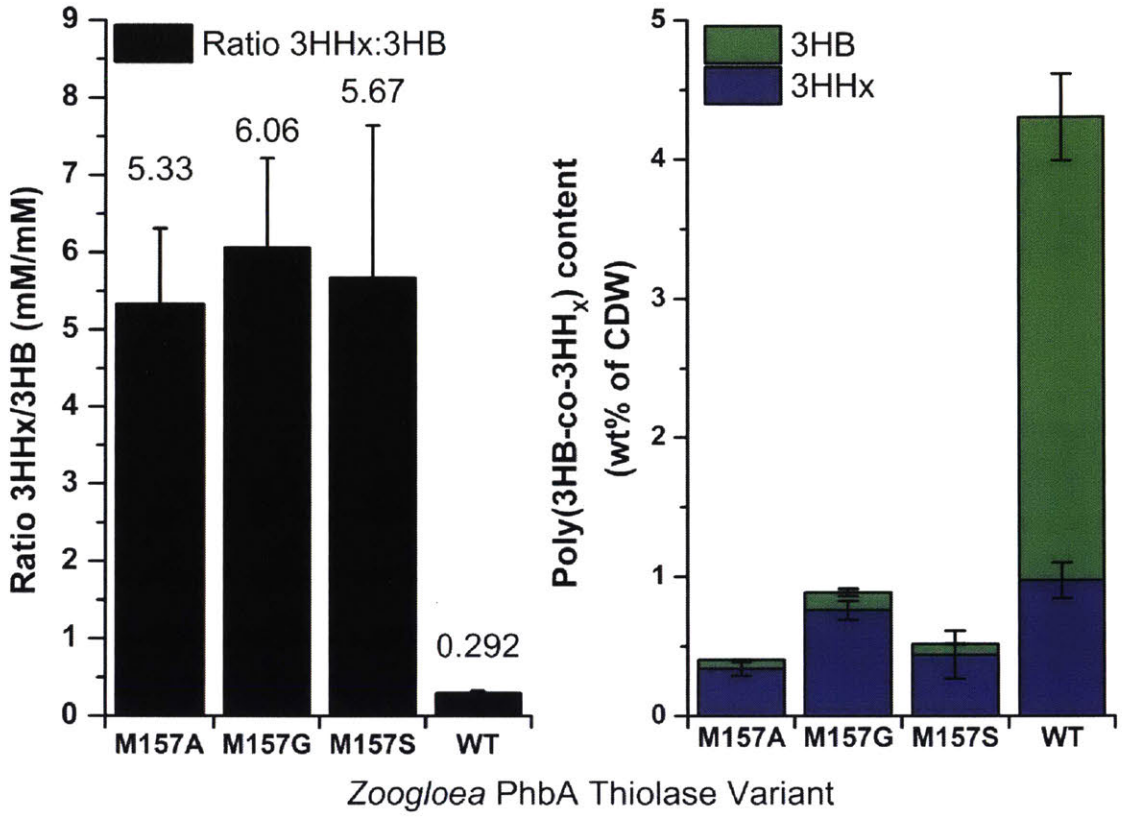
(a) In the Bind 1 column $\Delta\Delta E_{Bind\ B}^{Mut-WT}$ is the difference in binding energies between mutant and wild type bound to butyryl-CoA with free C89, $\Delta\Delta E_{Bind\ A}^{Mut-WT}$ is the difference in binding energies between mutant and wild type bound to acetyl-CoA with free C89, and $\Delta\Delta\Delta E_{Bind\ B-A}^{Mut-WT}$ is the difference between $\Delta\Delta E_{Bind\ B}^{Mut-WT}$ and $\Delta\Delta E_{Bind\ A}^{Mut-WT}$ corresponding to the differential specificities for mutant versus wild type for binding butyryl-CoA versus acetyl-CoA as the priming acyl-CoA.

(b) In the Bind 2 column $\Delta\Delta E_{Bind\ B}^{Mut-WT}$ is the difference in binding energies between mutant and wild type bound to acetyl-CoA with butyrylated C89, $\Delta\Delta E_{Bind\ A}^{Mut-WT}$ is the difference in binding energies between mutant and wild type bound to acetyl-CoA with C89 acetylated, and $\Delta\Delta\Delta E_{Bind\ B-A}^{Mut-WT}$ is the difference between $\Delta\Delta E_{Bind\ B}^{Mut-WT}$ and $\Delta\Delta E_{Bind\ A}^{Mut-WT}$ corresponding to the differential specificities for mutant versus wild type for binding acetyl-coA as the extending CoA with C89 either butyrylated or acetylated. Negative energies are highlighted with a green background, while positive energies are highlighted with a red background. All energies are reported in kcal/mol.

Figure 4


A
B
Ratio 3HH/3HB
Ratio 3HH/3HB (mM/mM)
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
3HB g/L
3HH g/L
Amount of 3HA Produced (g/L)
2.8
2.4
2.0
1.6
1.2
0.8
0.4
0.08
0.06
0.04
0.02
0.00
L377A
L377G
L377S
L88A
L88G
L88S
M157A
M157G
M157S
M288A
M288G
M288S
WT
Zoogloea PhbA Thiolase Variant

C

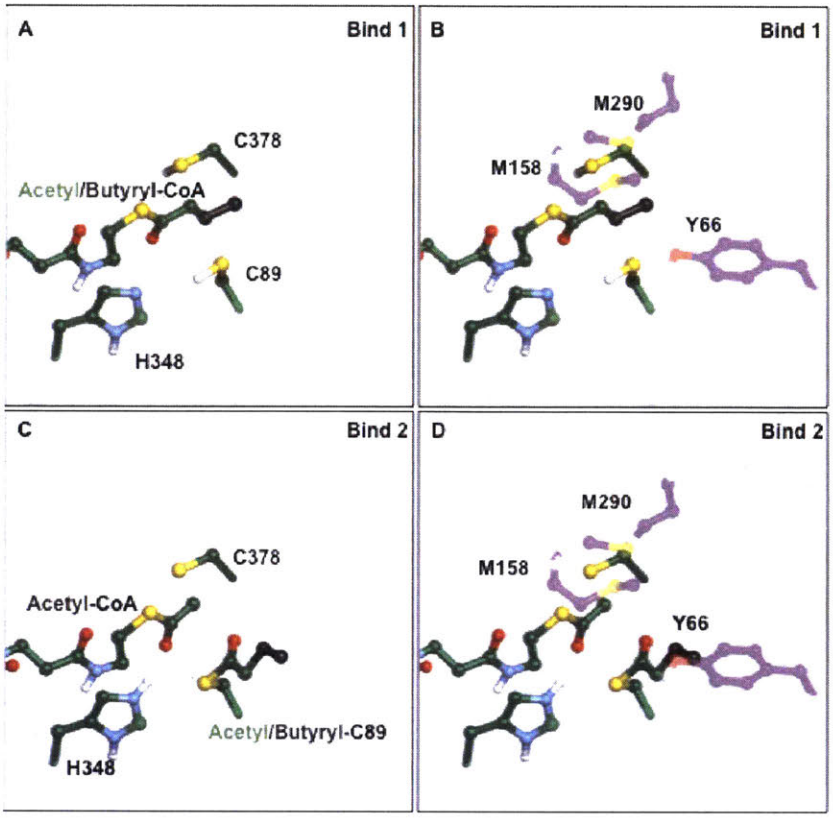


A Initial screening of *Z. ramigera* PhbA thiolase variants as designed by our computational method, using the previously established 3HA pathway, which results in production of free 3HAs which are detected in the supernatant. Products were analyzed from cell-free culture supernatants 72 hours post induction by HPLC and ratios calculated on a molar basis.

B Final concentrations of 3HB and 3HH acids 72 hours post induction as analyzed by HPLC.

C *Z. ramigera* mutant thiolases profiled within the context of PHA biosynthesis. Ratios represent the composition of the PHA polymer as measured by GC after methanolysis.

Figure 5



A Structure of *C. necator* BktB thiolase active site during the first binding event (corresponding to step 2 in figure 1C). The atoms colored black show the extra atoms of the butyryl group that must be accommodated compared to the acetyl group in order to preferentially produce 3-oxo-hexanoyl-CoA rather than acetoacetyl-CoA.

B Structure of *C. necator* BktB thiolase active site during first binding event with residues selected for mutation colored purple.

C Structure of *C. necator* BktB thiolase active site during the second binding event (corresponding to step 5 in figure 1C). Atoms colored black show the extra atoms that must be accommodated in order to preferentially produce 3-oxo-hexanoyl-CoA.

D Structure of *C. necator* BktB thiolase active site during second binding event (corresponding to step 5 in figure 1C) with residues selected for mutation colored purple.

Table II. Energetic calculations of *C. necator* BktB mutants selected for experimental testing

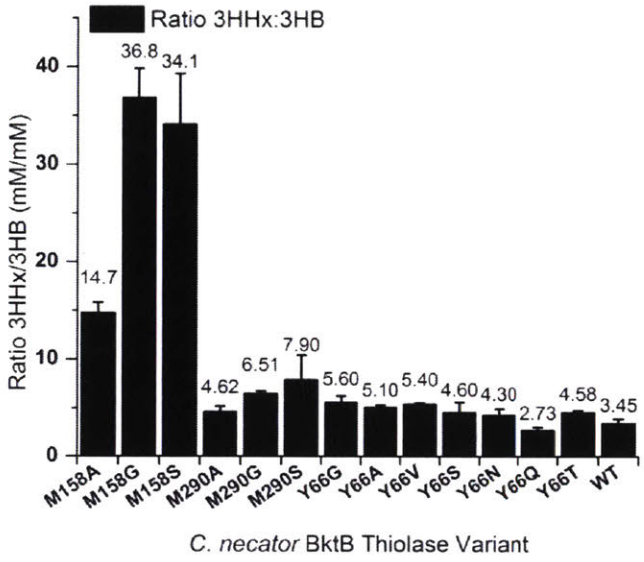
Mutant	Bind 1			Bind 2		
	$\Delta\Delta E_{Bind\ B}^{Mut-WT}$	$\Delta\Delta E_{Bind\ A}^{Mut-WT}$	$\Delta\Delta\Delta E_{Bind\ B-A}^{Mut-WT}$	$\Delta\Delta E_{Bind\ B}^{Mut-WT}$	$\Delta\Delta E_{Bind\ A}^{Mut-WT}$	$\Delta\Delta\Delta E_{Bind\ B-A}^{Mut-WT}$
M158S	-2.18	1.27	-3.45	-1.33	1.03	-2.36
M158A	-1.86	1.58	-3.44	-1.54	1.49	-3.03
M158G	-1.54	1.61	-3.15	-0.16	1.99	-2.15
M290S	-21.81	0.45	-22.26	0.13	0.31	-0.18
M290A	-22.03	0.55	-22.59	0.18	0.42	-0.23
M290G	-21.52	0.64	-22.16	0.29	0.51	-0.22
Y66Q	-0.36	0.12	-0.48	-2.52	0.08	-2.60
Y66N	0.10	0.09	0.01	-2.49	0.12	-2.61
Y66V	0.26	0.12	0.14	-2.49	0.11	-2.60
Y66T	0.09	0.12	-0.04	-2.46	0.12	-2.58
Y66S	-0.32	0.11	-0.43	-2.48	0.13	-2.61
Y66A	0.08	0.12	-0.04	-2.44	0.14	-2.58
Y66G	0.18	0.12	0.05	-2.45	0.15	-2.60

In the Bind 1 column, $\Delta\Delta E_{Bind\ B}^{Mut-WT}$ is the difference in binding energies between mutant and wild type bound to butyryl-CoA with free C90, $\Delta\Delta E_{Bind\ A}^{Mut-WT}$ is the difference in binding energies between mutant and wild type bound to acetyl-CoA with C90 unacylated, and $\Delta\Delta\Delta E_{Bind\ B-A}^{Mut-WT}$ is the difference between $\Delta\Delta E_{Bind\ B}^{Mut-WT}$ and $\Delta\Delta E_{Bind\ A}^{Mut-WT}$ corresponding to the differential specificities for mutant versus wild type for binding butyryl-CoA versus acetyl-CoA as the priming acyl-CoA.

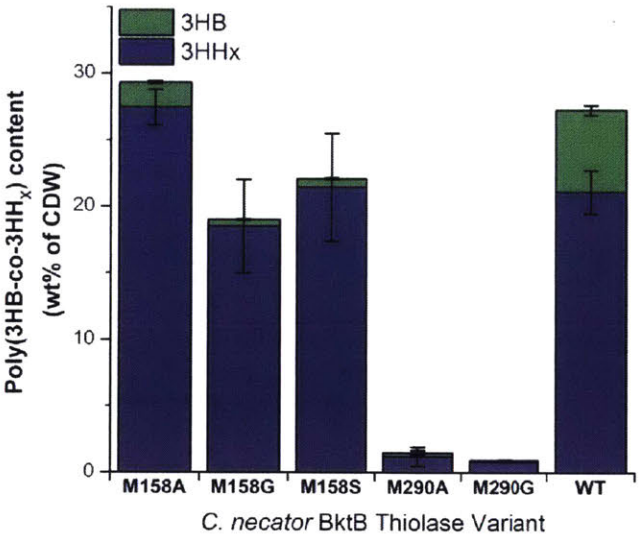
In the Bind 2 column $\Delta\Delta E_{Bind\ B}^{Mut-WT}$ is the difference in binding energies between mutant and wild type bound to acetyl-CoA with butyrylated C90, $\Delta\Delta E_{Bind\ A}^{Mut-WT}$ is the difference in binding energies between mutant and wild type bound to acetyl-CoA with acetylated C90, and $\Delta\Delta\Delta E_{Bind\ B-A}^{Mut-WT}$ is the difference between $\Delta\Delta E_{Bind\ B}^{Mut-WT}$ and $\Delta\Delta E_{Bind\ A}^{Mut-WT}$ corresponding to the differential specificities for mutant versus wild type for binding acetyl-coA as the extending CoA with C90 either butyrylated or acetylated. Negative energies are highlighted with a green background, while positive energies are highlighted with a red background. All energies are reported in kcal/mol.

Figure 6

A



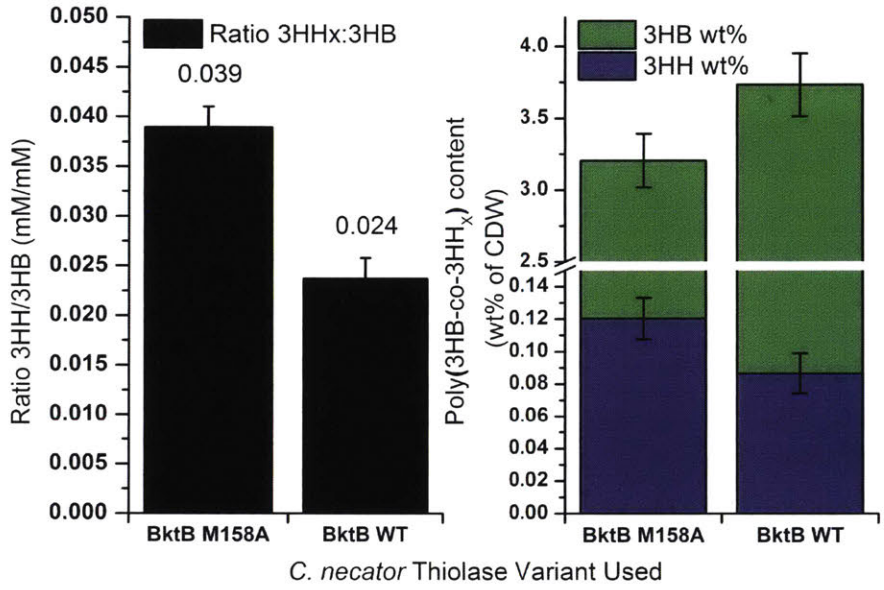
B



A Computationally predicted mutant BktB thiolases assayed within the context of PHA biosynthesis. Ratios represent the composition of the PHA polymer as measured by GC after methanolysis.

B PHA content as a weight percentage of the CDW of *E. coli* overexpressing a given BktB thiolase variant.

Figure 7



Overexpression of a trans-enoyl-CoA reductase (ter_{Td}) and reductase ($PhaJ4b_{Cn}$), in addition to a thiolase, and acetoacetyl-CoA reductase and PHA polymerase, allows for synthesis of C6 products solely from glucose. Increased selectivity for longer chain products is achieved with BktB M158A in place of wild-type BktB.

Table III Composition of PHAs extracted from engineered *E. coli* strains overexpressing different thiolases and grown on glucose with fed butyrate.

Thiolase	CDW (g/L)	PHA Content (wt%)	Mol% 3HHx
Z. ramigera PhbA WT	0.52 ± 0.022	4.3 ± 0.41	22.6 ± 1.81
Z. ramigera PhbA M158A	1.26 ± 0.073	0.41 ± 0.046	83.9 ± 2.7
Z. ramigera PhbA M158G	0.72 ± 0.34	0.083 ± 0.08	85.6 ± 2.15
Z. ramigera PhbA M158S	0.96 ± 0.42	0.51 ± 0.17	84.2 ± 4.02
C. necator BktB WT	1.04 ± 0.13	27.2 ± 1.2	77.3 ± 2.3
C. necator BktB M158A	0.81 ± 0.16	29.3 ± 1.4	93.6 ± 0.42
C. necator BktB M158G	0.71 ± 0.17	18.9 ± 3.55	97.3 ± 0.43
C. necator BktB M158S	0.88 ± 0.16	22.1 ± 4.08	97.3 ± 0.22
C. necator BktB M290A	0.65 ± 0.23	1.47 ± 0.90	81.7 ± 1.4
C. necator BktB M290G	0.54 ± 0.04	0.90 ± 0.05	86.7 ± 0.46

Table IV *In vitro* kinetic characterization of thiolase variants

Reaction	k_{cat} (sec^{-1})	K_m (μM)	k_{cat}/K_m ($M^{-1}sec^{-1}$)	C6/C4 Selectivity
BktB WT C4 Condensation	14.1	919	1.52×10^4	N/A
BktB M158A C4 Condensation	1.33	913	1.46×10^3	
BktB WT C4 Thiolysis	148	17.5	8.45×10^6	0.032
BktB WT C6 Thiolysis	4.06	15.2	2.67×10^5	
BktB M158A C4 Thiolysis	4.63	14.1	3.28×10^5	2.99
BktB M158A C6 Thiolysis	16.7	17	9.82×10^5	

In vitro characterization of *C. necator* BktB wild type and mutant thiolases in the forward direction (condensation) with C4, and reverse direction (thiolysis) with both C4 and C6 substrates. Catalytic parameters were computed from fits to the Michaelis-Menten equation (Supplementary Figure 2).

Supplementary Information

Supplementary Table I Differences between PhbA and BktB active sites

Distance from Acetyl-CoA Carbonyl Center	PhbA/BktB Chain	PhbA Residue ID	PhbA Residue Name	BktB Residue ID	BktB Residue Name	Cumulative BktB Residue Differences
3.89	**A**	**348**	**H**	**350**	**H**	**0**
3.93	**A**	**378**	**C**	**380**	**C**	**0**
3.97	A	318	A	320	A	0
4.01	**A**	**288**	**M**	**290**	**M**	**0**
4.39	**A**	**89**	**C**	**90**	**C**	**0**
5.79	A	316	N	318	N	0
6.30	A	319	F	321	F	0
6.53	A	147	G	148	A	1
6.57	B	64	Q	65	M	2
6.57	A	148	L	149	L	2
6.77	A	350	I	352	I	2
6.77	A	247	S	249	S	2
7.01	**A**	**157**	**M**	**158**	**M**	**2**
7.31	A	380	G	382	G	2
7.42	A	322	Q	324	Q	2
7.47	**A**	**377**	**L**	**379**	**M**	**3**
8.11	A	353	S	355	T	4
8.24	**A**	**88**	**L**	**89**	**L**	**4**
8.33	A	289	G	291	G	4
8.35	A	379	I	381	I	4
8.57	A	160	T	161	T	4
8.85	A	161	A	162	A	4
8.94	A	292	P	294	P	4
8.94	A	248	G	250	G	4
9.03	A	343	A	345	G	5
9.26	A	119	M	120	M	5
9.48	A	57	V	57	V	5
9.66	A	156	H	157	H	5
9.85	A	246	A	248	A	5
10.06	A	249	L	251	L	5
10.33	A	347	G	349	G	5
10.41	A	158	G	159	G	5
10.41	A	383	M	385	Q	6

10.44	A	91	S	92	S	6
10.56	A	349	P	351	P	6
10.61	A	164	V	165	V	6
10.65	A	381	G	383	G	6
10.66	A	90	G	91	G	6
10.66	A	283	V	285	V	6
10.75	A	235	F	236	F	6
10.90	A	382	G	384	G	6
11.00	A	326	V	328	V	6
11.01	A	323	A	325	A	6
11.08	A	357	I	359	I	6
11.29	A	351	G	353	G	6
11.34	A	291	G	293	G	6
11.49	A	384	G	386	G	6
11.56	A	321	A	323	A	6
11.58	B	65	N	66	Y	7
11.81	A	344	I	346	I	7
11.88	A	150	D	151	D	7
11.93	A	290	T	292	I	8
11.96	A	87	Q	88	R	9
12.07	A	315	A	317	A	9
12.10	A	356	R	358	L	10
12.14	A	376	T	378	T	10
12.25	A	58	L	58	I	11
12.28	A	86	N	87	N	11
12.41	A	241	V	243	V	11
12.55	A	144	I	146	L	12
12.60	A	149	T	150	H	13
12.62	A	385	V	387	I	14

Differences between PhbA and BktB active sites ordered by distance from PhbA priming acetyl-CoA acetyl carbonyl carbon. Note that overall the two thiolases share only 52% sequence identity, however their active sites are highly conserved, with only 5 amino acid differences within a 10 Å shell from the catalytic Cys89/90 residues. This table was generated by aligning Chains A and B of the PhbA structure (PDB: 1M3Z) to Chains A and B of the BktB crystal structure (PDB: 4NZS) using the *super* command in Pymol. Residues in **bold** indicate catalytic residues (C89/C90, H348/H350, C378/C380). Residues in **green** indicate residues mutated in both BktB and PhbA (M157/M158, M288/M290). Residues in **purple** indicate those mutated only in PhbA (L377, L88). Residues in blue indicate those mutated in BktB only (Y66).

Supplementary Table II Primers and Strains Used in this Study

Primer Name	Sequence	Primer Name	Sequence
bktB_m158a_F	ggcggtcacgcccgcgtggatgcgatgg	phbA_L378A_F	gggtctcgccacggcctgcatcggcggc
bktB_m158a_R	ccatcgcatccacgcgggcgtgaccgcc	phbA_L378A_R	gccgccgatgcaggccgtggcgagaccc
bktB_m158g_F	ccatcgcatccacgggggcgtgaccgcc	phbA_L378G_F	gggtctcgccacgggctgcatcggcggc
bktB_m158g_R	ggcggtcacgccccgtggatgcgatgg	phbA_L378G_R	gccgccgatgcagcccgtggcgagaccc
bktB_m158s_F	cccttccatcgcatccacagcggcgtgaccg	phbA_L378S_F	gggtctcgccacgagctgcatcggcggc
bktB_m158s_R	cggtcacgccgctgtggatgcgatggaaggg	phbA_L378S_R	gccgccgatgcagctcgtggcgagaccc
bktB_m290a_F	ggacccgaaggcggcgggcatcggcccg	phbA_L89A_F	gcccgagccgcaagcctggttcatgccc
bktB_m290a_R	cgggccgatgcccgccgccttcgggtcc	phbA_L89A_R	gggcatgaaccaggcttgcggctcgggc
bktB_m290g_F	ggacccgaaggcgggggcatcggcccg	phbA_L89G_F	gcccgagccgcaaccctggttcatgccc
bktB_m290g_R	cgggccgatgccccccgccttcgggtcc	phbA_L89G_R	gggcatgaaccagggttgcggctcgggc
bktB_m290s_F	tggacccgaaggcgagcggcatcggccc	phbA_L89S_F	ggcccgagccgcaactctggttcatgcccc
bktB_m290s_R	gggccgatgccgctcgccttcgggtcca	phbA_L89S_R	ggggcatgaaccagagttgcggctcgggcc
bktB_y66a_F	gccgcgcgacatggctctgggccgcgtc	phbA_M158A_F	ctacggctaccacgcgggcacgaccgcc
bktB_y66a_R	gacgcggcccagagccatgtcgcgcggc	phbA_M158A_R	ggcggtcgtgcccgcgtggtagccgtag
bktB_y66g_F	gccgcgcgacatgggtctgggccgcgtc	phbA_M158G_F	ctacggctaccacgggggcacgaccgcc
bktB_y66g_R	gacgcggcccagacccatgtcgcgcggc	phbA_M158G_R	ggcggtcgtgcccccgtggtagccgtag
bktB_y66n_F	ccgcgcgacatgaatctgggccgcg	phbA_M158S_F	gccttctacggctaccacagcggcacgaccg
bktB_y66n_R	cgcggcccagattcatgtcgcgcgg	phbA_M158S_R	cggtcgtgccgctgtggtagccgtagaaggc
bktB_y66q_F	gccgcgcgacatgcagctgggccgcgtcg	phbA_M289A_F	cgatcccaaggtcgcgggcaccggcccg
bktB_y66q_R	cgacgcggcccagctgcatgtcgcgcggc	phbA_M289A_R	cgggccggtgcccgcgaccttgggatcg
bktB_y66s_F	gccgcgcgacatgagtctgggccgcgtc	phbA_M289G_F	cgatcccaaggtcgggggcaccggcccg
bktB_y66s_R	gacgcggcccagactcatgtcgcgcggc	phbA_M289G_R	cgggccggtgcccccgaccttgggatcg
bktB_y66t_F	gccgcgcgacatgactctgggccgcgtc	phbA_M289S_F	cgtcgatcccaaggtcagcggcaccggc
bktB_y66t_R	gacgcggcccagagtcatgtcgcgcggc	phbA_M289S_R	gccggtgccgctgaccttgggatcgacg
bktB_y66v_F	gccgcgcgacatggttctgggccgcgtc		
bktB_y66v_R	gacgcggcccagaaccatgtcgcgcggc		
BktB_SeqF	gcagcctgaaggatgtg		
BktB_SeqR	gcgccttcaccgtgatc		
PhaC2_F	gctagcgaattcctacaaggagaacaaac		
PhaC2_R	ctagaggatccccgggctcgag		

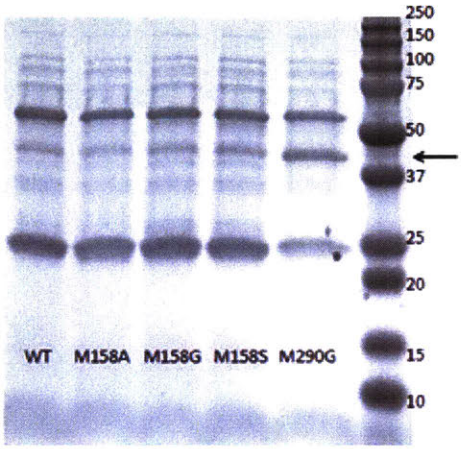
ptev5_BktB_F	Gaaaacctgtattttcagggcatgacgcgtgaagtggtag
ptev5_BktB_R	Gctcgagaattccatggtcagatacgctcgaagatgg
phaB_F	aacctgtattttcagggcatgactcagcgcatt
phab_R	gctcgagaattccatggctcagcccatatgcag

Codon Optimized Sequence of PhaC2 from R. aetherivorans I24

This sequence was cloned into the second MCS of pCDF_pct$_{M.elsdenii}$ using the NdeI and XhoI restriction sites which are underlined in the sequence below.

GATATACATATGGCACAGGCACGTACCGTTATTGGTGAAAGCGTTGAAGAAAGCATTGGTG
GTGGTGAAGATGTTGCACCGCCTCGTCTGGGTCCGGCAGTTGGTGCACTGGCAGATGTTTTT
GGTCATGGTCGTGCAGTTGCACGTCATGGTGTTAGCTTTGGTCGTGAACTGGCAAAAATTGC
AGTTGGTCGTAGCACCGTTGCACCGGCAAAAGGTGATCGTCGTTTTGCAGATAGCGCATGGT
CAGCAAATCCGGCATATCGTCGCCTGGGTCAGACCTATCTGGCAGCAACCGAAGCAGTTGAT
GGTGTTGTTGATGAAGTGGGTCGTGCAATTGGTCCGCGTCGTACCGCAGAAGCACGTTTTGC
CGCAGATATTCTGACCGCAGCACTGGCACCGACCAATTATCTGTGGACCAATCCGGCAGCAC
TGAAAGAAGCATTTGATACCGCAGGTCTGAGCCTGGCACGTGGCACCAAACATTTTGTTAGC
GATCTGATTGAAAATCGTGGTATGCCGAGCATGGTTCAGCGTGGTGCATTTACCGTTGGTAA
AGATCTGGCAGTTACACCGGGTGCAGTTATTAGCCGTGATGAAGTTGCCGAAGTTCTGCAGT
ATACCCCGACCACCGAAACCGTTCGTCGTCGTCCGGTTCTGGTTGTTCCGCCTCCGATTGGTC
GTTATTACTTTCTGGATCTGCGTCCGGGTCGTAGCTTTGTTGAATATAGTGTTGGTCGTGGCC
TGCAGACCTTTCTGCTGAGCTGGCGTAATCCGACCGCAGAACAGGGTGATTGGGATTTTGAT
ACCTATGCAGGTCGTGTTATTCGTGCAATCGATGAAGTTCGTGAAATCACCGGTAGTGATGA
TGTTAATCTGATTGGTTTTTGTGCCGGTGGTATTATTGCAACCACCGTTCTGAATCACCTGGC
AGCCCAGGGTGATACCCGTGTTCATAGCATGGCCTATGCAGTTACCATGCTGGATTTTGGTG
ATCCGGCACTGCTGGGTGCATTTGCCCGTCCTGGTCTGATTCGTTTTGCCAAAGGTCGTAGCC
GTCGTAAAGGTATTATTAGCGCACGTGATATGGGTAGCGCATTTACCTGGATGCGTCCGAAT
GATCTGGTTTTTAACTATGTGGTGAACAACTATCTGATGGGTCGTACCCCTCCTGCCTTTGAT
ATTCTGGCATGGAATGATGATGGTACAAATCTGCCTGGTGCCCTGCATGGCCAGTTTCTGGA
TATTTTTCGTGATAATGTTCTGGTGGAACCGGGTCGTCTGGCCGTTCTGGGTACACCGGTTGA
TCTGAAAAGCATTACCGTTCCGACCTTTGTGAGCGGTGCAATTGCCGATCATCTGACCGCGT
GGCGTAATTGTTATCGTACCACACAGCTGCTGGGAGGTGAAACCGAATTTGCACTGAGCTTT
AGCGGTCATATTGCAAGCCTGGTTAATCCTCCGGGTAATCCGAAAGCACATTATTGGACCGG
TGGCACACCGGGTCCGGATCCTGATGCATGGCTGGAAAATGCAGAACGTCAGCAGGGTAGT
TGGTGGCAGGCCTGGTCAGATTGGGTTCTGGCACGCGGTGGCGAAGAAACAGCAGCACCGG
ATGCACCGGGTAGTGCACAGCATCCTGCACTGGATGCCGCACCGGGTCGCTATGTTCGTGAT
CTGCCTGCAGGTTAACTCGAGTCTGGT

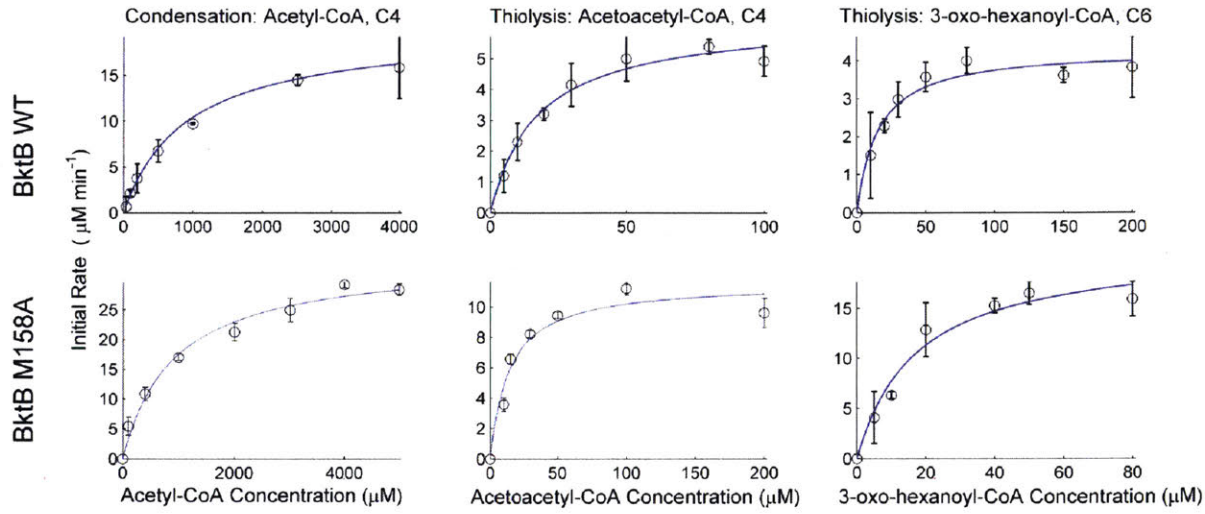
Supplementary Figure 1



Crude cell lysate of strains expressing the PHA pathway with either wild-type or mutant BktB enzymes. Briefly, 1 ml of each culture at 48 hours post-induction was collected and supernatant removed after centrifugation. Cells were resuspended in 0.4 mL His buffer and lysed by bead-beating. Protein concentration was determined by a Bradford assay and 5 µg total protein/BSA equivalent was loaded onto an AnykD Bio-Rad mini Protean gel. No significant differences in expression are observed between wild type and mutant thiolases.

BktB = 40.9 kDa, Pct = 55.6 kDa, PhaB = 26 kDa, and PhaC2 = 60.3 kDa.

Supplementary Figure 2



Michaelis-Menten curve fits of assayed wild type and M158 BktB thiolases, in both condensation direction with acetyl-CoA, and thiolytic direction with acetoacetyl-CoA and 3-oxohexanoyl-CoA substrates.

Structural Analysis of M158A Mutation – Bind 1

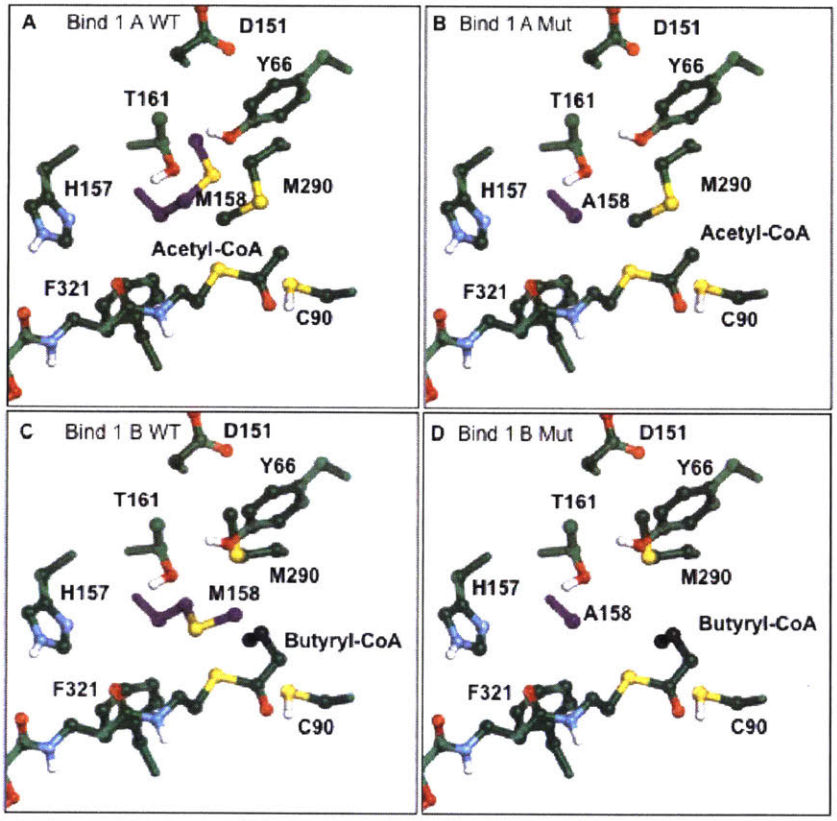
Supplementary Figure 3 shows the four global minimum energy structures that comprise the $\Delta\Delta\Delta E^{Mut-WT}_{Bind\ B-A}$ calculation for Bind 1 for the *C. necator* mutant we chose for in vitro characterization, M158A. Supplementary Table III shows the detailed pairwise energetic breakdowns of each of the terms comprising $\Delta\Delta\Delta E^{Mut-WT}_{Bind\ B-A}$. Note that according to Table III, M158A is predicted to improve butyryl-CoA binding in the first binding event, hurt acetyl-CoA binding in the first step, improve accommodation of butyryl-Cys90 with acetyl-CoA bound in the second binding event, and disfavor accommodation of acetyl-Cys90 with acetyl-CoA bound in the first binding event. According to Supplementary Table III, the bulk of the improvement in butyryl-CoA binding comes from the van der Waals (vdW) energy, particularly from the interaction of residue 158 with the mercapto group of acetyl/butyryl-CoA. Note that although the butyryl group between wild type and mutant takes on the same conformation, the M158 side chain takes on a conformation that clashes with the mercapto group. Residues shown with a ball-and-stick model are those included in the mobile region. Panels A-D of Supplementary Figure 3 show that the majority of the mobile region does not locally rearrange in response to mutation or substrate binding. Only residues 290, 158, and 90 change conformation across the four panels. Nothing except the mutated residue changes conformation between panels C and D, which represent the wild type and mutant bound to acetyl-CoA. The loss of favorable van der Waals contacts between the substrate and M158 as a result of paring M158 to a smaller alanine explains the positive value of 1.58 kcal/mol for $\Delta\Delta E^{Mut-WT}_{Bind\ A}$, which represents the difference in binding energy between mutant and wild type bound to acetyl-coA. Note that to accommodate the bulkier butyryl group, the side chains of M290 and M158 take on conformations different from those in Panels C and D. In order to accommodate the butyryl group, the side chain of M158

takes on a less favorable conformation to butyryl binding. Mutating M158 to an alanine relieves this unfavorable interaction, explaining the favorable -1.86 kcal/mol value of $\Delta\Delta E_{Bind\ B}^{Mut-WT}$.

Structural Analysis of M158A Mutation – Bind 2

Supplementary Figure 4 shows the four global minimum energy conformations that comprise the $\Delta\Delta\Delta E_{Bind\ B-A}^{Mut-WT}$ calculation for Bind 2 for the M158A structure. Compared to Bind 1, even fewer residues included in the mobile region change conformations between the four structures. In panels C and D, as for Bind 1, the M158A mutation causes a loss of favorable van der Waals interactions between the M158 residue and the substrate, which explains the -1.49 kcal/mol value of $\Delta\Delta E_{Bind\ A}^{Mut-WT}$ for this binding event. Panels A and B represent the conformations of the mutant and wild type bound to acetyl-CoA with butyryl-C90. In Panel A, the butyryl group of butyryl-C90 takes on a conformation that has an unfavorable interaction with the mercapto group of acetyl-CoA. Mutating M158 to alanine allows the butyryl group to relax to a more favorable conformation, which is worth -1.33 kcal/mol as shown in Supplementary Table IV.

Supplementary Figure 3



A Minimum energy structure of *C. necator* BktB thiolase wild type active site with acetyl-CoA bound as priming acyl-CoA (Bind 1). In Panels A-D all residues shown as ball and stick models are included in the mobile region of the conformational search.

B Minimum energy structure of *C. necator* BktB thiolase M158A active site with acetyl-CoA bound as priming acyl-CoA (Bind 1). Note that $\Delta\Delta E_{Bind\ A}^{Mut-WT}$ for Bind 1 is computed as the difference in binding energies between Panel A and B.

C Minimum energy structure of *C. necator* BktB thiolase wild type active site with butyryl-CoA bound as priming acyl-CoA (Bind 1). Note that residues M158 and M290 adopt conformations different from the wild type structure with acetyl-CoA bound as priming acyl-CoA in panel A.

D Minimum energy structure of *C. necator* BktB thiolase M158A active site with butyryl-CoA bound as priming acyl-CoA (Bind 1). Note that $\Delta\Delta E_{Bind\ B}^{Mut-WT}$ for Bind 1 is computed as the difference in binding energies between Panel C and D.

Supplementary Table III. Dominant pairwise energetic interactions comprising $\Delta\Delta\Delta E^{Mut-WT}_{Bind\ B-A\ Total}$ for Bind 1 and corresponding to the four minimum energy structures shown in Supplementary Figure 3

Interaction[e]	$\Delta\Delta E^{Mut-WT}_{Bind\ B\ vdW}$[a]	$\Delta\Delta E^{Mut-WT}_{Bind\ A\ vdW}$[b]	$\Delta\Delta\Delta E^{Mut-WT}_{Bind\ B-A\ vdW}$[c]	$\Delta\Delta\Delta E^{Mut-WT}_{Bind\ B-A\ Total}$[d]
A-158-Side chain – C-1-Mercapto	-2.65	0.73	-3.37	-3.32
Non-mobile – C-Acyl	-0.19	0.00	-0.19	-0.19
A-158-Side chain – C-1-PantoADP	0.45	0.17	0.29	0.20
A-158-Side chain – C-1-CoACO	0.49	0.21	0.29	0.26
Total	**-1.73**	**1.49**	**-3.23**	**-3.44**

(a) $\Delta\Delta E^{Mut-WT}_{Bind\ B\ vdW}$ is the van der Waals (vdW) energy comprising the $\Delta\Delta E^{Mut-WT}_{Bind\ B}$ term in the above table.

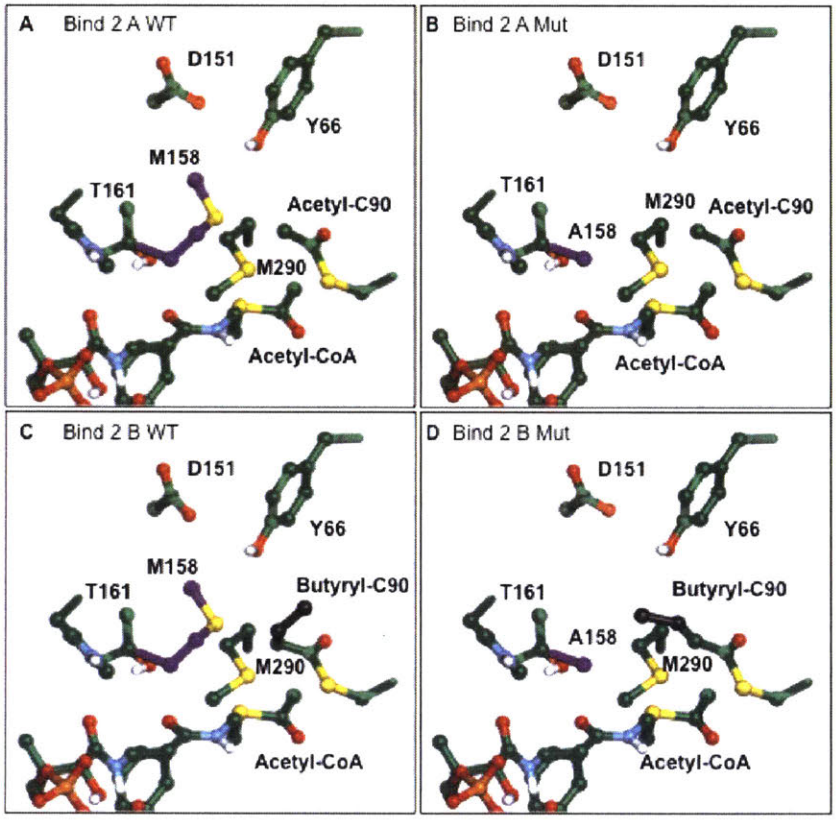
(b) $\Delta\Delta E^{Mut-WT}_{Bind\ A\ vdW}$ is the van der Waals energy comprising the $\Delta\Delta E^{Mut-WT}_{Bind\ A}$ term in the above table.

(c) $\Delta\Delta\Delta E^{Mut-WT}_{Bind\ B-A\ vdW}$ is the van der Waals energy comprising the $\Delta\Delta\Delta E^{Mut-WT}_{Bind\ B-A\ Total}$ term in the above table.

(d) $\Delta\Delta\Delta E^{Mut-WT}_{Bind\ B-A\ Total}$ is the total energy of each pairwise interaction comprising the sum of vdW, geometric and electrostatic interactions (latter two not displayed due to negligible contribution). Note the the total in this column represents the $\Delta\Delta\Delta E^{Mut-WT}_{Bind\ B-A}$ term for M158A in the Bind 1 column of Table II. Note that the four pairwise interactions listed are the only four interactions with $|\Delta\Delta\Delta E^{Mut-WT}_{Bind\ B-A\ vdW}| > 0.1$ kcal/mol.

(e) In this Table, A-158-Side chain refers to the non-backbone atoms in residue 158, C-1-Mercapto refers to the atoms in the β-mercaptoethylamine group of the acetyl-CoA, C-1-PantoADP refers to the atoms in the pantothenic acid moiety of the acyl-CoA, C-1-CoACO refers to the two atoms in the acyl group carbonyl moiety of the acyl-CoA. Non-mobile refers to the atoms not included in the mobile region in the calculation (everything not shown as a ball and stick model in Supplementary Figure 3).

Supplementary Figure 4



A Minimum energy structure of *C. necator* BktB thiolase wild type active site with acetyl-CoA bound as the extending acyl-CoA and C90 acetylated (Bind 2). In Panels A-D all residues shown as ball and stick models are included in the mobile region of the co3nformational search.

B Minimum energy structure of *C. necator* BktB thiolase M158A active site with acetyl-CoA bound as the extending acyl-CoA and C90 acetylated (Bind 2). Note that $\Delta\Delta E_{Bind\ A}^{Mut-WT}$ for Bind 2 is computed as the difference in binding energies between Panel A and B.

C Minimum energy structure of *C. necator* BktB thiolase wild type active site with acetyl-CoA bound as the extending acyl-CoA and C90 butyrylated (Bind 2).

D Minimum energy structure of *C. necator* BktB thiolase M158A active site with butyryl-CoA bound as priming acyl-CoA (Bind 1). Note that $\Delta\Delta E_{Bind\ B}^{Mut-WT}$ for Bind 1 is computed as the difference in binding energies between Panel C and D.

Supplementary Table IV: Dominant Energetic Interactions Comprising $\Delta\Delta\Delta E^{Mut-WT}_{Bind\ B-A}$ in M158A Bind 2 and corresponding to the four minimum energy structures shown in Supplementary Figure 4

Interaction[e]	$\Delta\Delta E^{Mut-WT}_{Bind\ B\ vdW}$[a]	$\Delta\Delta E^{Mut-WT}_{Bind\ A\ vdW}$[b]	$\Delta\Delta\Delta E^{Mut-WT}_{Bind\ B-A\ vdW}$[c]	$\Delta\Delta\Delta E^{Mut-WT}_{Bind\ B-A\ Total}$[d]
A-90-Acyl – C-1-Mercapto	-1.33	0.00	-1.34	-1.30
A-90-CO – C-1-Acyl	-1.05	-0.02	-1.03	-1.00
A-90-CO – C-1-CoACO	-0.21	0.00	-0.21	-0.27
Non-mobile – C-Acyl	-0.27	0.00	-0.27	-0.26
A-90-Acyl – C-1-CoACO	-0.18	0.00	-0.19	-0.21
A-90-Cys – C-1-CoACO	-0.19	0.00	-0.18	-0.16
A-90-Cys – C-Acyl	0.20	0.00	0.20	0.20
Total	**-1.69**	**1.38**	**-3.07**	**-3.03**

(a) $\Delta\Delta E^{Mut-WT}_{Bind\ B\ vdW}$ is the van der Waals (vdW) energy comprising the $\Delta\Delta E^{Mut-WT}_{Bind\ B}$ term in the above table.

(b) $\Delta\Delta E^{Mut-WT}_{Bind\ A\ vdW}$ is the van der Waals energy comprising the $\Delta\Delta E^{Mut-WT}_{Bind\ A}$ term in the above table.

(c) $\Delta\Delta\Delta E^{Mut-WT}_{Bind\ B-A\ vdW}$ is the van der Waals energy comprising the $\Delta\Delta\Delta E^{Mut-WT}_{Bind\ B-A\ Total}$ term in the above table.

(d) $\Delta\Delta\Delta E^{Mut-WT}_{Bind\ B-A\ Total}$ is the total energy of each pairwise interaction comprising the sum of vdW, geometric and electrostatic interactions (latter two not displayed due to negligible contribution). Note the the total in this column represents the $\Delta\Delta\Delta E^{Mut-WT}_{Bind\ B-A}$ term for M158A in the Bind 2 column of Table II. Note that the seven pairwise interactions listed are the only seven interactions with $|\Delta\Delta\Delta E^{Mut-WT}_{Bind\ B-A\ vdW}| > 0.1$ kcal/mol.

(e) In this table, A-90-Acyl refers to the aliphatic (non-carbonyl) portion of the acyl group on C90, A-90-CO, refers to the two atoms in the carbonyl portion of the acyl group on C90, A-90-Cys refers to the non-acyl, non-backbone portion of C90, C-1-Mercapto refers to the atoms in the β-mercaptoethylamine group of the acetyl-CoA, C-1-CoACO refers to the two atoms in the acetyl group carbonyl moiety of the acetyl-CoA, C-1-acyl refers to the non-carbonyl atoms in the acetyl group of acetyl-CoA. Non-mobile refers to the atoms not included in the mobile region in the calculation (everything not shown as a ball and stick model in Supplementary Figure 4).

Chapter 3 : Machine Learning Identifies Chemical Characteristics that Promote Enzyme Catalysis

Brian M. Bonk[1,2], James W. Weis[3,4], Bruce Tidor*[1,2,3,4]

[1] Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA

[2] Computer Science and Artificial Intelligence Laboratory (CSAIL), Massachusetts Institute of Technology, Cambridge, MA

[3] Computational and Systems Biology, Massachusetts Institute of Technology, Cambridge, MA

[4] Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA

*Corresponding author (BT: tidor@mit.edu)

Author contributions: BMB, JWW, and BT conceived of the overall project and developed the approach and plan. BMB performed the simulations, generated the data sets, led the data analysis, and wrote the initial manuscript draft. JWW explored methods for feature selection and implemented the method used here. BMB, JWW, and BT contributed to the analysis of the data and developed the final manuscript.

Abstract

Despite tremendous progress in understanding and engineering enzymes, knowledge of how enzyme structures and their dynamics induce observed catalytic properties is incomplete, and capabilities to engineer enzymes fall far short of industrial needs. Here we investigate the structural and dynamic drivers of enzyme catalysis for the rate-limiting step of the industrially important enzyme ketol-acid reductoisomerase (KARI) and identify a portion of the conformational space of the bound enzyme–substrate complex that, when populated, leads to large increases in reactivity. We apply computational statistical mechanical methods that implement transition interface sampling to simulate the kinetics of the reaction and combine this with machine learning techniques from artificial intelligence to select features relevant to reactivity and to build predictive models for reactive trajectories. We find that conformational descriptors alone, without the need for dynamic ones, are sufficient to predict reactivity with greater than 85% accuracy (90% AUC). Key descriptors distinguishing reactive from almost-reactive trajectories quantify substrate conformation, substrate bond polarization, and metal coordination geometry and suggest their role in promoting substrate reactivity. Moreover, trajectories constrained to visit a portion of the reactant well separated from the rest by a simple hyperplane defined by ten conformational parameters show increases in computed reactivity by many orders of magnitude. This study provides evidence for the existence of reactivity hot spots within the conformational space of the enzyme–substrate complex and develops methodology for identifying and validating these particularly reactive regions. We suggest that identification of reactivity hot spots and re-engineering enzymes to preferentially populate them, can lead to significant rate enhancements.

Introduction

Enzymes are remarkable catalysts that produce substantial rate enhancements, often accompanied by high substrate and product selectivity as well as regio- and stereo-specificity. They are increasingly important for industrial-scale chemical and bioengineering applications, not only because of the chemistry they accomplish but also because they can do so sustainably in mild, aqueous conditions. The interest and need for custom enzymes developed for specific purposes continues to motivate computational and experimental research in catalytic biochemistry.

Despite substantial progress made, more is still required along two principal avenues in order to advance enzyme engineering to meet industrial needs. We need a better understanding of the drivers of chemical reactivity promoted by enzymes, some of which have been hypothesized to be dynamic (Basner and Schwartz 2005; Ruscio et al. 2009; Kamerlin and Warshel 2010) rather than structural, together with a richer collection of tools to probe and potentially manipulate the active-site catalytic environment. We also need even more powerful and robust methods of designing and engineering enzyme function. Current approaches include directed evolution (Porter, Rusli, and Ollis 2016; Hammer, Knight, and Arnold 2017; Molina-Espeja et al. 2016), catalytic antibodies (Lerner, Benkovic, and Schultz 1991; Nevinsky and Buneva 2004; Maeda et al. 2016) and computational enzyme design (Kiss et al. 2013; Baker 2010), the latter two of which focus on tight-binding of transition states. While these approaches have produced tremendous successes in the hands of developers, they have not yet become general-purpose tools. The need for directed evolution to improve designs obtained by other methods and our inability to fully understand the improvements accumulated through evolution suggest our understanding may be incomplete, perhaps in some fundamental way, and might

require us to incorporate other factors beyond transition-state binding and transition-state stabilization (relative to the bound ground state).

Here we investigate two fundamental questions of enzyme function motivated by the larger goal of enzyme engineering; we focus on the enzyme–substrate complex without specific reference to the transition state. First, can we gain insight into the nature of the drivers of chemical reactivity, and to what extent are these drivers apparent in the behavior of the bound enzyme–substrate complex? And second, based on previous work of ourselves and others (Silver 2011; Hur and Bruice 2003b; Sadiq and Coveney 2014; Zhang et al. 2017; van Erp et al. 2016) can we identify regions of the conformational space of the enzyme–substrate complex that are inherently more reactive than others? These questions are addressed using a new approach that combines machine learning with kinetic transition sampling techniques, applied to the rate-limiting step for the industrially important enzyme ketol-acid reductoisomerase (KARI).

There are a number of approaches for studying and analyzing enzyme reactivity that do not focus on the transition state per se, although it may enter implicitly. These include the literature investigating near-attack conformations, which has suggested that lowering the energetic barrier to selectively facilitate formation of subsets of ground state conformations that lie on the path to the transition state, can be just as important as lowering the energetic barrier to the transition state itself (Lau and Bruice 1998; Bruice and Lightstone 1999; Bruice 2002; Sadiq and Coveney 2014) and the computational methodology of kinetic transition sampling, embodied in the transition path sampling (Dellago et al. 1998) and transition interface sampling (van Erp, Moroni, and Bolhuis 2003) approaches. Here we use transition interface sampling (TIS), which is computationally more efficient. Both sampling approaches are statistical mechanical techniques for directly computing the rate of a chemical reaction without reliance on transition-

state theory or knowledge of either the transition state or a valid reaction coordinate connecting the reactant well with the product well on the free energy surface. TIS uses Monte Carlo sampling to construct an ensemble of trajectories that all start in the reactant well and pass through an interface on the way toward the product well. Ensembles are collected in a prescribed order such that successive ensembles progress further towards the product well, with trajectories from the final ensemble reaching the product well. Appropriate statistical methods exist to compute the progressive probability that a trajectory starting in the reactant well will reach each interface, a rapidly diminishing probability can drop tens of orders of magnitude on its way to the product, and to convert the probability into a reaction rate, corresponding to the specific activity, k_{cat}, for enzymes. While a valid reaction coordinate is not a requirement, the method uses an order parameter that cleanly distinguishes reactant from product to track progress between the two wells (van Erp, Moroni, and Bolhuis 2003). (The placement of interfaces is shown schematically in Figure 1A and their progression in Figure 2, with λ representing the order parameter.)

KARI is a natural enzyme required for branched-chain amino-acid synthesis, found broadly across plant and microbial species (Dumas et al. 2001). It also now has an important role in industrial processes for the microbial production of isobutanol, and, due to its role as the rate-limiting step, improvements in its specific activity would improve processes for large-scale isobutanol production (Chen and Liao 2016). Natural KARIs show two principal activities, converting *(2S)*-acetolactate (AL) to *(2R)*-2,3-dihyroxy-3-isovalerate (leading to valine or leucine) and converting *(2S)*-2-aceto-2-hydroxybutyrate (AHB) to *(2R,3R)*-2,3-dihydroxy-3-methylvalerate (leading to isoleucine). The enzyme carries out two enzymatic steps in sequence, first a rate-limiting isomerization consisting of an alkyl migration (a methyl migration for AL

and ethyl for AHB) and then a faster reduction carried out by a nucleotide cofactor. Our studies have focused on the homodimeric enzyme from spinach (*Spinacia oleracea*), due largely to the availability of appropriate crystal structures, and we have studied the industrially relevant, rate-limiting reaction step involving isomerization of AL through methyl migration (Chen and Liao 2016; Bastian et al. 2011; Tadrowski et al. 2016) (Figure 1B).

The natural spinach enzyme exhibits a strong preference for NADPH as a cofactor and has two divalent magnesium cations bound at the active site, in intimate contact with substrate (Figure 1C; Biou et al. 1997). Models show AL coordinates both magnesium ions, with magnesium M16 coordinated to hydroxyl O6 and a carboxylate oxygen and magnesium M17 coordinated to hydroxyl O6 and carbonyl O8 (note that here we adopt atom naming and numbering from Figure 1C). M16 is additionally coordinated by an Oε of Glu 319, an Oδ of Asp 315, and two water molecules to make it hexacoordinate. M17 is also hexacoordinate, with additional coordination to the other Oδ of Asp 315 and three water molecules. Thus, both Asp 315 and hydroxyl O6 bridge the magnesium ions. An additional polar contact to the substrate is made by the hydrogen bond donated by the protonated form of the side chain of Glu 496 to carbonyl O8. Note that the C5 (and its associated hydrogen atoms) is the methyl group that migrates from C4 to C7.

The current study is based on previous work we carried out on KARI, which identified a "pump-and-push" mechanism for the rate-limiting isomerization reaction, whereby the local environment vibrationally excites the breaking C4–C5 bond and the side chain of Glu 319 helps direct and potentially stabilize the migrating methyl group (C5 and its hydrogens) towards its destination, bound to C7 (Silver 2011). Moreover, the work suggested that some portions of the conformational and motional space of the bound enzyme–substrate complex (the reactant well)

led to trajectories that have a greater probability of reacting than those that do not pass through or spend as much time in those same portions of the reactant well. The term "more reactive" portions of the reactant well is applied to represent this idea.

Here we carried out TIS simulations of wild-type spinach KARI and performed detailed comparative analysis on two sets of ensembles of trajectories—one set that reacted and another set that approached the barrier but did not react (termed "almost-reactive"). We tabulated data on 68 different geometric measurements in the active site that represent elements of the local conformation in the form of distances between pairs of atoms (whether or not the atoms are bonded to each other), planar angles across triplets of atoms (again, whether or not there is a true bond angle involving them), and dihedral angles across quadruplets of atoms. The set was selected based on mechanistic hypotheses of others and ourselves, and includes internal metrics within the substrate; measures of the position and orientation of substrate relative to the environment, particularly for groups that might stabilize the bound substrate or transition state; and measures of conformation of the environment (Table 1 and Supplementary Figure 1). Machine learning techniques were applied to identify subsets of this feature list and build predictive models that accurately distinguished reactive from almost-reactive trajectories, based only on data tabulated from before trajectories departed the reactant well. We reasoned that these reduced feature sets and models describe key features sufficient to drive reactivity. We analyzed these feature sets in the context of the reactive and almost-reactive trajectories to understand in more detail these drivers and to gain insight into mechanism. We found key descriptors capable of identifying reactive conformations included those that quantify substrate conformation, substrate bond polarization, and metal coordination geometry and suggest their role in promoting substrate reactivity. To test the notion that these drivers are sufficient and that they define

inherently reactive portions of the reactant well, we compared the computed specific activity of the wild-type enzyme when trajectories were constrained to visit these regions with those that were not. We found that ten features alone were sufficient to describe a portion of the reactant well that led to very large rate increases, demonstrating it as a highly reactive portion of the well.

Figure 1: (A) Schematic of interface placements used to generate reactive and almost-reactive trajectories, where λ_A denotes the reactant interface, λ_{AR} indicates the product interface used to generate the almost-reactive trajectory ensembles and λ_R indicates the product interface used to generate the reactive trajectory ensembles. (B) Reaction catalyzed by KARI with states 2 and 3 indicating initial and final states used for the specific rate-limiting step of the isomerization studied (C) Atoms and residues included in QM region (non-polar hydrogens not shown) Residue AC6 refers to the deprotonated acetolactate substrate, residue NDP refers to the NADPH cofactor and residue MG6 refers to the two magnesium ions and the five coordinating active site waters (D) Distribution of λ values for reactive (red) and almost-reactive (blue) trajectories time-shifted such that last trough before prospective catalytic event occurs at the 0 fs time point. Vertical black lines indicate the location of time points where features were computed.

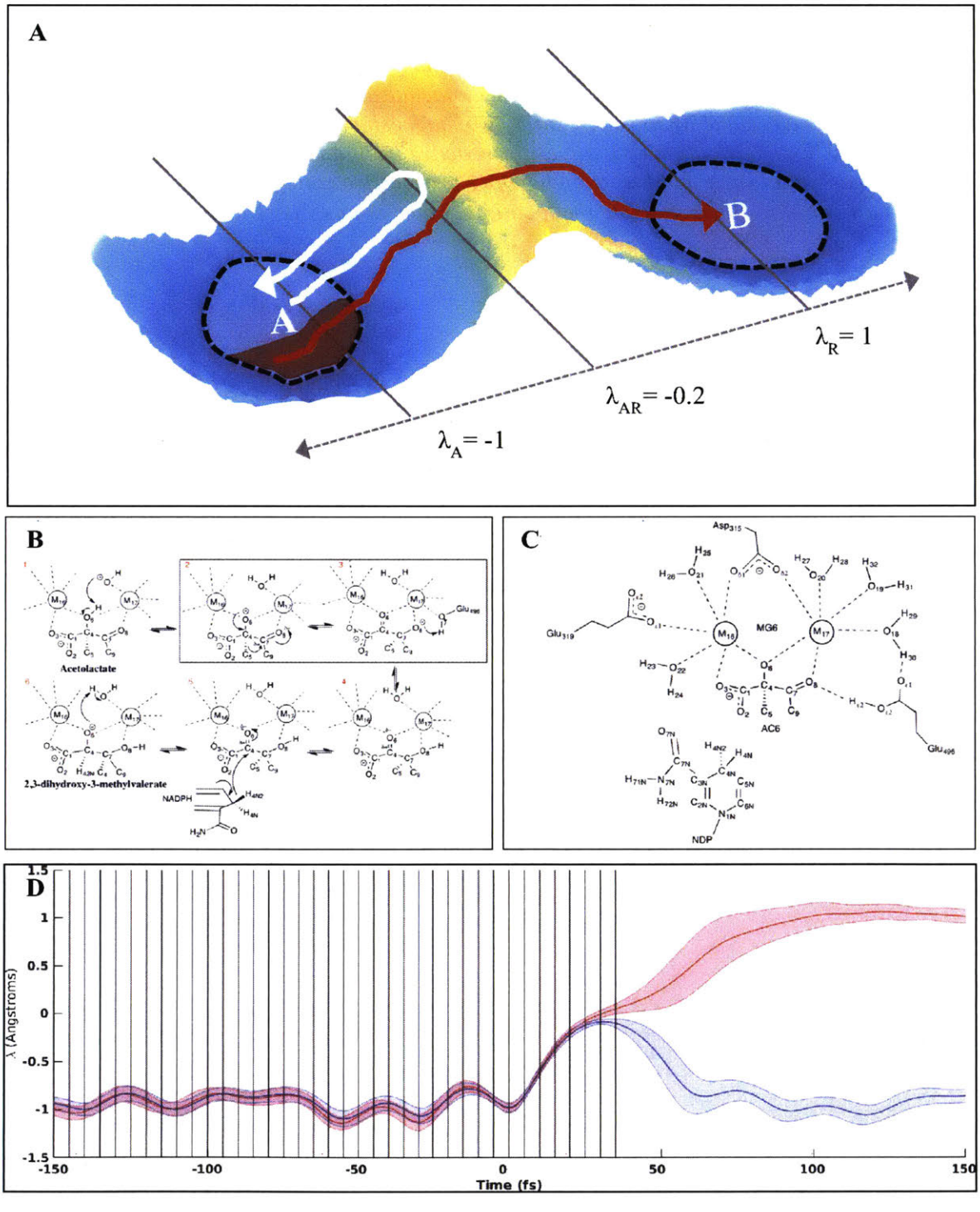
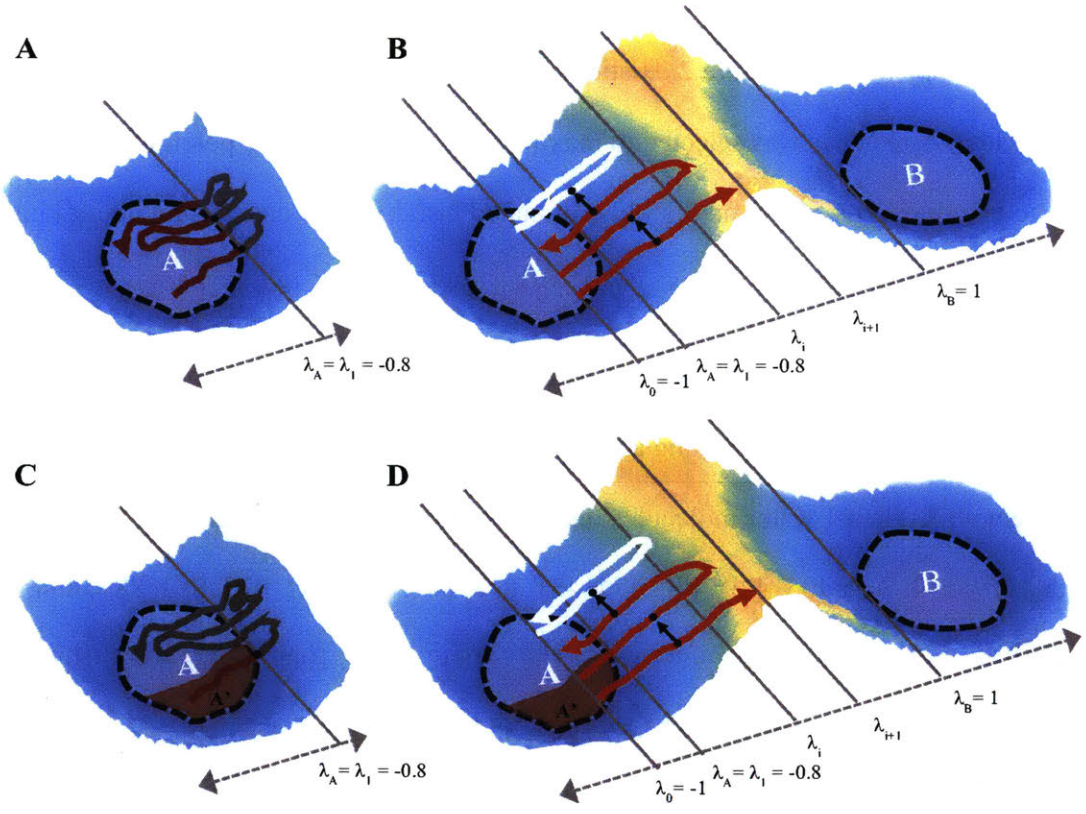

A
B
A
B
$\lambda_R = 1$
$\lambda_{AR} = -0.2$
$\lambda_A = -1$
B
Acetolactate
2,3-dihydroxy-3-methylvalerate
NADPH
C
Asp315
Glu319
MG6
AC6
Glu496
NDP
D
λ (Angstroms)
Time (fs)
1.5
1
0.5
0
-0.5
-1
-1.5
-150
-100
-50
0
50
100
150

Figure 2: (A) Illustration of computation of the TIS flux factor. The red and gray line represents a long molecular dynamics trajectory originating in region A. Portions of the trajectory in red indicate the time points in region A used to normalize the flux factor. Black dots represent effective crossings of the λ_A interface. (B) Illustration of computation of a $P(\lambda_{i+1}|\ \lambda_i)$ ensemble. Each red and white line indicates an attempted shooting move. Black dots indicate shooting points. Red lines indicate accepted shooting moves, while white lines indicate rejected shooting moves. (C) Illustration of procedure used to compute the constrained flux factor. The dark red region indicates the reactive subregion A' identified using machine learning. Portions of the trajectory in red indicate the time point in either region A' used to compute . Black dots represent effective crossings of the λ_A interface. (D) Illustration of a constrained $P(\lambda_{i+1}|\ \lambda_i)$ ensemble. The dark red region indicates the reactive subregion A' identified using machine learning.



Methods

Structure Preparation. The crystal structure of spinach (*Spinacia oleracea)* KARI bound to the transition-state analog N-hydroxy-N-isopropyloxamate was obtained from the Protein Data Bank (Bernstein et al. 1977; Berman et al. 2000; Rose et al. 2017) with the accession code 1YVE (Biou, Dumas, Cohen-Addad, et al. 1997) and prepared as described previously by Silver (2011). Although the enzyme crystallizes as a homodimer with two identical active sites (Biou, Dumas, Cohen-Addad, et al. 1997), only the chain A monomer was used for all simulations in order to improve computational efficiency. This choice was justified by the significant separation between the active sites of the two monomers (Biou, Dumas, Cohen-Addad, et al. 1997) which is illustrated in Supplementary Figure 2. Histidine side-chain orientation and protonation for the following chain A residues was selected to maximize hydrogen-bonding potential, resulting in no changes to histidine orientation, no doubly protonated histidine side chains, and neutral histidine protonation as indicated: 103-δ, 215-δ, 226-δ, 232-δ, 280-ε, 328-ε, 484-δ, 506-ε, and 564-ε. Crystallographic water molecules that were neither in the active site nor made least three hydrogen bonds with the protein (using a maximum heavy-atom hydrogen-bond distance of 3.33 Å) were removed. The 61 water molecules remaining had residue identifiers of 72, 75, 87, 93, 106, 109, 179, 194, 379, 405, 429, 440, 474, 481, 838–841, 852, 862, 878, 883, 887, 894, 895, 941–949, 965, 967–969, 975, 998, 999, 1023–1025, 1032, 1072, 1089, 1093–1095, 1097, 1105, 1108, 1206, 1250, 1252, 1253, 1257, 1304, 1305, and 1779.

A model of the substrate-bound enzyme was then constructed by running an *in vacuo* QM ground-state minimization of the substrate, two magnesium centers, five magnesium-coordinating water molecules, as well as the side chains of three surrounding active-site residues,

Asp 315, Glu 319, and Glu 496. The Glu 496 residue was protonated, consistent with previous studies indicating its importance in stabilizing the transition and product state by forming a hydrogen bond with the substrate O8 (Proust-De Martin, 2000). The GAUSSIAN03 computer program (Frisch et al. 2003) was used to perform *in vacuo* QM calculations at the rhf/3-21g* level of theory, using ground-state energy minimization (keyword OPT) to obtain reactant and product structures and a saddle-point search (keyword QST3) to obtain the transition-state structure. Both types of optimization were performed using the Berny algorithm (Peng and Bernhard Schlegel 1993; Peng et al. 1996, Li and Frisch 2006). To ensure low-energy pathways to the reactant and product state of isomerization, the resulting transition state was validated by following the vibrational eigenmode corresponding to the single negative eigenvalue.

Each of the optimized and validated QM-derived structures was combined with the prepared crystallographic structure for the rest of the enzyme by alignment of the carbon atoms of the QM-optimized substrate to the crystallographic transition-state analog, followed by ten rounds of sliding, constrained minimization. During this minimization, which consisted of 100 steps of steepest descent minimization followed by 100 steps of adopted basis Newton-Raphson minimization, all substrate, magnesium ion, and coordinating aspartate and glutamate oxygen atoms were held fixed, and the remaining active-site residues were harmonically constrained using a force constant of 50 kcal/(mol · Å). Harmonic constraints were reset after each round of minimization.

Simulation Methodology. CHARMM version 41 (Bernard R. Brooks et al. 1983; B R Brooks et al. 2009) compiled with the SQUANTUM option was used to perform all molecular dynamics simulations. The QM portion of the energy function was calculated with the AM1 semi-empirical quantum mechanical force field (Dewar et al. 1985); the MM portion of the

energy function was computed using the CHARMM36 all-atom force field (Huang and MacKerell 2013). Additional AM1 parameters published by Stewart (2004) were used for the magnesiums. The following atoms made up the QM region: substrate (acetolactate), both magnesium centers, five magnesium-coordinating active site water molecules, the side chains of Asp 315, Glu 319, and Glu 496, and the nicotinamide group of NADPH (illustrated schematically in Figure 1C). The Generalized Hybrid Orbital method (Gao et al. 1998) was used to treat the QM/MM boundary atoms, included the Cα atoms of residues Asp 315, Glu 319, and Glu 496, as well as the C5' atom of the ribose ring in NADPH linking to the nicotinamide group. The substrate O6 was deprotonated and the coordinating Glu 496 was protonated, paralleling previous QM/MM studies of KARI (Proust-De Martin, 2000). All molecular dynamics simulations were performed *in vacuo* with a distance dependent dielectric ($4r$) using the leapfrog integrator at 300 K with a 1-fs integration time step.

Seed Trajectory Generation. The initial reactive trajectories used to bootstrap the TIS simulations were found by computing a potential of mean force (PMF) along the order parameter λ, defined as the difference of the distance between the substrate breaking bond (C4–C5) and the forming bond (C5–C7), which has units of ångstroms. This PMF was computed using umbrella sampling and the weighted histogram analysis method (Kumar et al.1992). The umbrella sampling was performed in CHARMM41 using the RXNCOR module with windows 0.05 Å in width and harmonic constraints of 300 kcal/(mol · Å). The resulting PMF provided an estimate of the location of the transition state along the order parameter λ, roughly within the $-0.05 < \lambda < +0.05$ region. Candidate seed trajectories were then generated by integrating forward and backward for 2,000 fs starting from a randomly chosen frame from the umbrella sampling window ensembles with centers at λ values of –0.05, 0.00, and +0.05. Trajectories were selected

as successful seed trajectories if they connected the reactant basin ($\lambda < -1$) and product basin ($\lambda > +1$).

Training Data Set Generation and Time Point Selection. Three randomly-selected connecting seed trajectories from the resulting collection described above were used as starting trajectories for the generation of a larger ensemble of reactive and almost-reactive trajectories. Each seed was used to generate 9 reactive ensembles and 9 almost-reactive ensembles of 20,000 trajectories each. The combined data set contained 461,422 almost-reactive and 618,578 reactive trajectories. The greater number of reactive trajectories resulted because the sampling process for almost-reactive trajectories also could also generate some reactive ones, but the sampling process for reactive trajectories could not generate almost-reactive ones. When the almost-reactive process produced a reactive trajectory, it was removed from that set and added to the reactive data set. To ensure a balanced number of reactive and almost-reactive trajectories in each training and testing data set, the reactive trajectories were randomly sampled without replacement to produce a set of 461,422 reactive trajectories.

For the reactive ensembles, as shown in Figure 1A the product interface was defined as $\lambda_R = +1.00$, and for the almost-reactive ensembles, the product interface was defined as $\lambda_{AR} = -0.20$. In both ensembles, the reactant interface was defined as $\lambda = -1.00$. In order to collect time points early in the reactant basin for analysis, integration was not stopped once a trajectory reached the reactant and product interface (and had been accepted into the Markov chain), but continued forward and backward for a total of 200 fs in each direction. A MATLAB wrapper that launched individual CHARMM41 trajectory runs was used to perform all TIS computations.

To ensure that candidate features (see below) were computed at analogous time points between reactive and almost-reactive trajectory ensembles, in a post-processing step, all almost-

reactive and reactive trajectories from all 27 pairs of ensembles were time-shifted and aligned such that the 0-fs time point corresponded to the bottom of the last "trough" in λ (when plotted vs. time) before the prospective alkyl migration event, a geometric feature that all the collected trajectories shared. Aligned reactive and almost-reactive trajectory ensembles are illustrated in Figure 1D. Chemically, the last trough represents the point at which the C4–C5 bond is most compressed, before, like a spring, launching into the prospective bond-breaking event (whether or not that event occurred). This trough was found by first finding the point in the trajectory closest to the transition region at $\lambda = 0$, then scanning along the trajectory backward from this point until the first change in sign of the derivative of λ with respect to time was found with a value of λ less than 0 (i.e., was located in the reactant basin). All other time points were defined relative to this first trough at time 0. Cartesian coordinate frames of atomic positions were collected in 5-fs increments from the 0-fs time point, going backward to –150 fs and forward to +35 fs from the t=0-fs point, for a total of 38 total time points. This collection of sub-sampled time points was used for all subsequent analysis.

Feature Computation. At each of the 38 time points between –150 and +35 fs, the set of 68 structural features in Table 1 (see Introduction), were computed for each of the trajectories in each of the 27 reactive and 27 almost-reactive ensembles. The 68 features are illustrated structurally in *Supplementary Figure 1*A (distances), *Supplementary Figure 1*B (angles), and *Supplementary Figure 1*C (dihedrals). These data were pooled across ensembles to produce one combined reactive and one combined almost-reactive data set at each of the 38 time points, which were used in machine learning and subsequent analysis described below and stored as a row in a data matrix. A separate data matrix was constructed for each time point by augmenting

the 68 computed features with the trajectory outcome (1 for reactive or 0 for almost-reactive), as well as the ensemble and trajectory indices.

Note that for the remainder of this study, the residue name AC6 will be used to refer to the reactant state of the substrate shown in Figure 1C. The residue name NDP refers to the NADPH cofactor and the residue name MG6 refers to the five quantum mechanically-treated waters and two magnesium ions in the active site.

Machine Learning. For feature regularization and discovery, the LASSO method (Tibshirani 1996) was used with the *lassoglm* implementation in MATLAB. For an intercept β_0 and predictor coefficients β_j, LASSO solves the general problem,

$$\min_{\beta_0,\beta}\left(\frac{1}{N}\sum_{i=1}^{N}\rho_{\beta_0,\beta}(X_i,Y_i)+\lambda\sum_{j=1}^{p}|\beta_j|\right)$$

where p is the number of input predictor features, N is the number of observables (the number of reactive and almost-reactive trajectories used in a given LASSO training set), the X_i are each a p-dimensional vector of predictor features (generally interatomic distances, angles, and dihedrals), the Y_i are scalar outcomes (1 for a trajectory that was reactive and 0 for one that was almost-reactive), λ is a non-negative regularization (penalty strength) parameter, and an underlying logistic learning model was composed of an intercept β_0, a set of p feature coefficients β_j, and the loss function $\rho_{\beta_0,\beta}(X_i,Y_i)$.

Due to the binary nature of the response variables, a logistic loss function was used,

$$\rho_{\beta_0,\beta}(x,y)=-y(\beta_0+\sum_{j=1}^{p}\beta_j x_j)+\log\left(1+e^{(\beta_0+\sum_{j=1}^{p}\beta_j x_j)}\right)$$

where x and y denote individual observations of X_i and Y_i. Note that only the values of the predictor coefficients β_j were penalized using LASSO, and not the value of the intercept β_0. In

order to select a given number of features with LASSO, the regularization parameter λ was adjusted until a specific number m (1, 5, 10, 15, 20, 25 or 30) of non-zero coefficients β_j remained (using a tolerance of 1.0×10^{-4}). These m LASSO-selected predictor features with non-zero coefficients were then fit using the *fitglm* function in MATLAB to a logistic classifier of the form,

$$\mu = \frac{e^{\left(\beta_0 + \sum_{j=1}^{m} \beta_j x_j\right)}}{1 + e^{\left(\beta_0 + \sum_{j=1}^{m} \beta_j x_j\right)}}$$

where μ is the probability of evaluating to 1 (reactive) given a specific linear combination of predictor features x_j. Trajectories were considered reactive if this probability evaluated to greater than 0.5 ($\beta_0 + \sum_{j=1}^{p} \beta_j x_j > 0$) and non-reactive if this probability evaluated to less than or equal to 0.5 ($\beta_0 + \sum_{j=1}^{p} \beta_j x_j \leq 0$). The logistic classifier essentially defines a hyperplane with the equation $\beta_0 + \sum_{j=1}^{p} \beta_j x_j = 0$ that partitions the reactant well in two, with reactive predictions on one side and non-reactive on the other.

After fitting predictor coefficients, the area under the receiver operating characteristic (AUC) was computed for each logistic classifier using the *perfcurve* function in MATLAB to vary the classifier threshold β_0 in order to generate a receiver operating characteristic, and subsequently compute the area under the resulting curve. Other classifier performance metrics were computed using the *classperf* function in MATLAB, where accuracy was defined as the number of correctly classified trajectories divided by the total number of trajectories, sensitivity was defined as the number of correctly classified reactive trajectories divided by the total number of reactive trajectories, and specificity was defined as the number of correctly classified almost-reactive trajectories divided by the total number of almost-reactive trajectories.

Cluster Assignment. Reactive clusters were assigned by k-means clustering, with the *kmeans* function in MATLAB using $k = 5$ applied to the matrix of consensus feature Z-scores weighted by their corresponding logistic coefficient β_j for all correctly classified reactive trajectories. The number of clusters (5) was chosen based on a hierarachical clustering analysis also performed in MATLAB (data not shown). The Euclidian distance of the consensus feature set from each almost-reactive trajectory to each of the five k-means centers was computed, and each almost-reactive trajectory was then assigned to the cluster with the shortest Euclidian distance to its respective centroid.

Rate Constant Computations. The TIS rate constant was computed as the product of two terms—a flux term denoted $\frac{\langle \Phi_{A,\lambda_1} \rangle}{\langle h_A \rangle}$ and a probability term denoted $P(\lambda_B|\lambda_1)$ (van Erp, Moroni, and Bolhuis 2003). The flux term represents the number of crossings through interface λ_1 having come directly from state A (also referred to as the reactant basin, defined as all points for which $\lambda \leq \lambda_A = -0.8$), normalized by the total time spent in state A. The probability term represents the probability a trajectory of reaching interface λ_B given that it has also crossed interface λ_1, and for computational efficiency can be decomposed into a series of conditional probabilities $\mathcal{P}(\lambda_B|\lambda_1) = \prod_{i=1}^{n-1} \mathcal{P}(\lambda_{i+1}|\lambda_i)\mathcal{P}(\lambda_B|\lambda_n)$ representing the probability of a trajectory reaching the next successive interface λ_{i+1}, given it has also reached interface λ_i.

For the flux factor calculations, a total of 10 independent 1-nanosecond molecular dynamics simulations were performed starting from reactant structures derived from each of five randomly selected seed trajectories generated as described above. The λ_A interface was set equal to the λ_1 interface at $\lambda = -0.8$. For the control flux factor computations (as illustrated in Figure 2A), the effective positive flux was computed as the number of times the trajectory crossed the

λ_A=–0.8 interface, having come from the region below the λ_A interface, divided by the total amount of time spent below the λ_A interface. For the constrained test flux factor computations (as illustrated in Figure 2C), the top 10 LASSO-selected features at the t=0 time point were written out during the dynamics run, and the effective positive flux was computed as the number of times the trajectory crossed the λ_1=-0.8 interface, having come from the region A', where region A' refers to all points in phase space which lie on the last trough (i.e., the first point at which $\frac{d\lambda}{dt} = 0$ and $\frac{d^2\lambda}{dt^2} > 0$) before crossing λ_A=–0.8, having first crossed λ_0=–1, and for which the logistic classifier with coefficients and features listed in Table 4 evaluated to true. Derivatives of λ with respect to time were computed using finite differences.

For the probability factor calculations, a total of 29 $P(\lambda_{i+1}|\ \lambda_i)$ interface ensembles from each of the five seed trajectories were computed, with the λ_i interfaces spaced between λ = –0.8 and λ = 0. The placement of these interfaces relative to the potential of mean force surface used to generate initial seed is shown in Supplementary Figure 3. To ensure sufficient sampling, interfaces between λ = –0.8 and λ = –0.15 were spaced in 0.025 Å increments and the remaining interfaces between –0.15 and 0 spaced in 0.05 Å increments. For each interface ensemble, a total of 5000 shooting moves was attempted. In each λ_i ensemble, candidate trajectories were generated using full shooting moves and accepted if they both crossed the λ_A= –0.8 interface and crossed the $\lambda = \lambda_i$ interface having first come from crossing interface λ_A. For the unconstrained control ensembles (as illustrated in Figure 2B), no further acceptance rules were applied.

For constrained ensembles (as illustrated in Figure 2D), once the trajectory connected the λ_A = –0.8 and λ_{i+1} interfaces, the trajectory was only included in the ensemble if the logistic classifier evaluated with features and coefficients described in Table 4 evaluated to true at the

first point at which $\frac{d\lambda}{dt} = 0$ and $\frac{d^2\lambda}{dt^2} > 0$ before crossing λ_A=-0.8, having first crossed interface λ_0=–1. Derivatives in λ with respect to time were computed using finite differences.

Integration was stopped when the candidate trajectories crossed its respective $\lambda = \lambda_{i+1}$ interface or the λ_0 interface, which was accomplished by modifying the RXNCOR module of CHARMM41 (Brooks et al. 1983, Brooks et al. 2009) All shooting moves and acceptance criteria were implemented using a MATLAB wrapper around CHARMM41, i.e. CHARMM was only used for the actual molecular dynamics integration. The number of accepted trajectories varied between the interface ensembles, seed trajectories and whether or not the additional sampling constraint was applied, and ranged between 10 and 95%.

Table 1: Feature names, feature indices and feature types computed at each time point. Residue name AC6 refers to the substrate, residue name NDP refers to the NADPH cofactor, and the residue name MG6 refers to the 5 active site waters and two magnesium ions. Structural representations of features are shown in Supplementary Figure 1.

Feature Index	Feature Name	Feature Type
1	Dist AC6/O2,NDP/N7N	Substrate-environment
2	Dist AC6/O2,NDP/O7N	Substrate-environment
3	Dist AC6/O3,MG6/H24	Substrate-environment
4	Dist AC6/O6,MG6/M16	Substrate-environment
5	Dist AC6/O8,GLU496/Hε2	Substrate-environment
6	Dist AC6/O8,MG6/M17	Substrate-environment
7	Dist GLU319/Oε1,AC6/C5	Substrate-environment
8	Dist MG6/H25,AC6/O6	Substrate-environment
9	Dist MG6/H26,AC6/O6	Substrate-environment
10	Dist MG6/H27,AC6/O6	Substrate-environment
11	Dist MG6/H28,AC6/O6	Substrate-environment
12	Dist MG6/H31,AC6/O6	Substrate-environment
13	Dist MG6/H32,AC6/O6	Substrate-environment
14	Dist MG6/M16,AC6/O3	Substrate-environment
15	Dist MG6/M17,AC6/O6	Substrate-environment
16	Dist NDP/H4N2,AC6/C4	Substrate-environment
17	Ang AC6/O6,MG6/M16,AC6/O3	Substrate-environment
18	Ang AC6/O8,MG6/M17,AC6/O6	Substrate-environment
19	Ang MG6/M17,AC6/O6,MG6/M16	Substrate-environment
20	Dist AC6/C1,AC6/C4	Intra-substrate
21	Dist AC6/C1,AC6/O2	Intra-substrate
22	Dist AC6/C1,AC6/O3	Intra-substrate
23	Dist AC6/C4,AC6/C7	Intra-substrate
24	Dist AC6/C4,AC6/O6	Intra-substrate
25	Dist AC6/C5,AC6/C4	Intra-substrate
26	Dist AC6/C5,AC6/C7	Intra-substrate
27	Dist AC6/C7,AC6/C9	Intra-substrate
28	Dist AC6/C7,AC6/O8	Intra-substrate
29	Ang AC6/C1,AC6/C4,AC6/C7	Intra-substrate
30	Ang AC6/C4,AC6/C7,AC6/C5	Intra-substrate
31	Ang AC6/C4,AC6/C7,AC6/C9	Intra-substrate
32	Ang AC6/C5,AC6/C4,AC6/C1	Intra-substrate
33	Ang AC6/C5,AC6/C7,AC6/C9	Intra-substrate
34	Dihe AC6/C1,AC6/C5,AC6/C7,AC6/C4	Intra-substrate
35	Dihe AC6/C5,AC6/C4,AC6/C7,AC6/C9	Intra-substrate
36	Dist NDP/H4N2,NDP/C4N	Intra-cofactor
37	Dist NDP/N7N,NDP/O2N	Intra-cofactor
38	Ang NDP/C4N,NDP/N1N,NDP/C1NQ	Intra-cofactor
39	Ang NDP/C6N,NDP/C3N,NDP/C7N	Intra-cofactor
40	Ang NDP/N7N,NDP/H72N,NDP/O2N	Intra-cofactor
41	Dihe NDP/C2N,NDP/C3N,NDP/C7N,NDP/N7N	Intra-cofactor
42	Dihe NDP/C2NQ,NDP/C1NQ,NDP/N1N,NDP/C6N	Intra-cofactor
43	Dihe NDP/C4N,NDP/C3N,NDP/C7N,NDP/O7N	Intra-cofactor
44	Dihe NDP/H1NQ,NDP/C1NQ,NDP/N1N,NDP/C2N	Intra-cofactor
45	Dist MG6/O18,MG6/M17	Water-metal
46	Dist MG6/O19,MG6/M17	Water-metal
47	Dist MG6/O20,MG6/M17	Water-metal
48	Dist MG6/O21,MG6/M16	Water-metal
49	Dist MG6/O22,MG6/M16	Water-metal
50	Ang MG6/H23,MG6/O22,MG6/M16	Water-metal
51	Ang MG6/H24,MG6/O22,MG6/M16	Water-metal
52	Ang MG6/H25,MG6/O21,MG6/M16	Water-metal
53	Ang MG6/H26,MG6/O21,MG6/M16	Water-metal
54	Ang MG6/H27,MG6/O20,MG6/M17	Water-metal
55	Ang MG6/H28,MG6/O20,MG6/M17	Water-metal
56	Ang MG6/H29,MG6/O18,MG6/M17	Water-metal
57	Ang MG6/H30,MG6/O18,MG6/M17	Water-metal
58	Ang MG6/H31,MG6/O19,MG6/M17	Water-metal
59	Ang MG6/H32,MG6/O19,MG6/M17	Water-metal
60	Dist GLU496/Oε2,GLU496/Hε2	Other environment
61	Dist GLN136/Nε2,NDP/O7N	Other environment
62	Dist MG6/H25,MG6/O21	Other environment
63	Dist MG6/H26,MG6/O21	Other environment
64	Dist MG6/H27,MG6/O20	Other environment
65	Dist MG6/H28,MG6/O20	Other environment
66	Dist MG6/H31,MG6/O19	Other environment
67	Dist MG6/H32,MG6/O19	Other environment
68	Ang GL136/Nε2,GLN136/Hε22,NDP/O7N	Other environment

Results

Data Collection and Preparation. Transition interface sampling simulations were carried out using a combined QM/MM approach to collect 27 ensembles of reactive trajectories (that each start as reactant and end as product, with the reaction progress order parameter λ reaching at least +1). A parallel method was used to collect 27 corresponding ensembles of almost-reactive trajectories (that also start as reactant but are only required to reach a λ value of –0.2 and that all returned to reactant [λ value less than –1] rather than continuing on to product). For both the reactive and almost-reactive ensembles, nine Markov chains containing 20,000 trajectories were initiated from each of three seed trajectories. All 54 of the resulting trajectories were then aligned by time shifting such that at t=0.0 fs, λ describing reaction progress was at the bottom of its last trough before attempting to cross the reaction barrier. Averages of the time traces for λ are illustrated in Figure 1D for reactive and almost-reactive trajectories.

The combined data set contained 461,422 almost-reactive and 618,578 reactive trajectories. There were a greater number of reactive trajectories because the sampling process for almost-reactive trajectories also could also generate some reactive ones, but the sampling process for reactive trajectories was not used to generate almost-reactive ones. When the almost-reactive process produced a reactive trajectory, it was moved to the reactive data set. To ensure a balanced number of reactive and almost-reactive trajectories in each training and testing data set, the reactive trajectories were randomly sampled without replacement to produce a set of 461,422 reactive trajectories.

At each of 38 time points between –150 and +35 fs (5-fs spacing and shown in Figure 1D), the 68 features listed in

Table 1 and illustrated structurally in *Supplementary Figure 1* were computed for each of the trajectories in each of the 27 reactive and 27 almost-reactive ensembles. These data were pooled across ensembles to produce one combined reactive and one combined almost-reactive data set at each of the 38 time points, which were used in machine learning and subsequent analysis described below.

Machine Learning. The prepared data sets were analyzed with machine learning to identify features with the ability to distinguish reactive from almost-reactive trajectories for each of the 38 time points. To assess individual feature performance, AUC (area under the curve of the receiver operating characteristic) was computed for all single features at the 0-fs time point. The results are shown in Figure 3A. The single feature with the maximum AUC performance was the distance between Glu 319 Oε1 and substrate C5 (AUC of 0.73). Only two features (distance Glu 319/Oε1–AC6/C5 and distance AC6/C4–AC6/C5) produced models with individual AUCs above 0.70, and 18 features produced models with AUCs above 0.60.

To find highly predictive groups of features, the LASSO method (Tibshirani 1996) was applied iteratively with different penalty strengths to identify an ordered set of features for each trajectory time point, optimized to distinguish reactive from almost-reactive conformations (see Methods). That is, for each time point a collection of separate machine-learning classifiers was built, trained, and tested, enabling comparisons of the useful sets of features across time points as well as the performance benefits for increased numbers of features at each time point. For model training, the data matrix at each time point was randomly sampled without replacement to produce 5 equal partitions containing 73,827 trajectories each, and for model testing, the remaining trajectories were randomly sampled to produce five equal partitions containing 18,456 trajectories each. Figure 3B shows the machine learning results for four classifier performance

statistics computed from each model constructed from data at each time point. The four statistics are AUC, accuracy, sensitivity, and specificity. Results for models constructed with optimized sets of 1, 5, 10, 15, and 20 features selected by LASSO are shown. Uncertainty was computed as the standard error over the 5 separate partitions of the data and is roughly equivalent to the thickness of each line. The results show progressively improved performance as the number of features was increased, with not insignificant performance with just one feature (generally 0.65–0.75 AUC) that rose to excellent performance with 10, 15, and 20 features (generally 0.85–0.95 AUC). Note that the performance of the LASSO-selected 1-feature models, being the "best" feature for each time point, was significantly better than the average AUC of all possible 1-feature models shown in Figure 3A, which was 57.18%. The similarity in performance between 15- and 20-feature models suggests near convergence with this number of features. The models developed were well balanced between false positives and false negatives as judged by similar values for the sensitivity and specificity metrics of individual classifiers, as well as the AUC values. Machine learning models performed similarly (for the same number of features) for time points between –150 and +20 fs, and then became substantially better (approaching an AUC of 1.00) for time points after +20 fs, which corresponds to times when the reactive and almost-reactive trajectories began to separate based on the order parameter λ (Figure 1D).

To assess the effect of LASSO-optimized feature selection for use in machine learning models, a control was carried out in which a classifier was trained similarly but using feature sets randomly chosen from the original 68 features. That is, each control classifier was optimally trained for the best performance possible with the random (and not optimized) features it was assigned. Analogous performance statistics for these control classifiers are shown in Figure 3C, with error bars indicating standard errors of classifier performance statistics across 100 randomly

selected feature groups. The results showed improved performance with additional features randomly selected from a chemically plausible set, together with large error bars, which is consistent with the notion that at any given time point some features or combinations of features were much better able than others to create predictive models, and the performance of models depended greatly on the features making up that model. Models with any given number of features performed much better on average when those features were selected by LASSO based on predictive ability than when selected randomly, demonstrating the value of the LASSO-selected features in distinguishing reactive from almost-reactive trajectories; for example, many of the one-feature models with LASSO-selected features had AUCs of about 0.70, whereas the random models had average AUCs of 0.57. The random models showed improved average performance after t=+20 fs, consistent with the notion that many features report on the fact that the reaction has largely begun.

Analysis of Consensus Feature Set Predictive Throughout Pre-Launch Time Window. The union of the complete 20-feature sets predictive at all 31 time points between –150 and 0 fs is depicted in Figure 3E. Features are listed in decreasing order of frequency of appearance, and the colored bars indicate the time points for which each feature appears as one of the 20 LASSO-selected features. (The time range –150 to 0 fs will be called the "pre-launch time window" for shorthand, as the 0-fs time point represents the last compression before the ultimate expansion of the putative breaking bond.) The results show that 17 of the features were used throughout at least half the window, 31 features were used at 10 or more time points, nearly all of the original features were used at least once (54 from the collection of 68), and 8 were used at five or fewer time points. The results suggest a commonality amongst the geometric descriptors that, broadly across the pre-launch window, were predictive. The names and feature types of the top 30

consistently predictive, consensus features are presented in *Table 2* along with the number of occurrences in the top 20 LASSO-selected sets within the pre-launch window. Figure 3D shows the classification performance of models trained using the top 1, 5, 10, 15, 20, 25, and 30 consensus features across the 31 time points between –150 and 0 fs. With the 30 consensus features, classification performance was nearly equivalent to or better throughout the pre-launch window (approximately 0.90 AUC) than the performance obtained from 20 LASSO-selected features optimized for each of the individual time points. That is, 30 shared features performed as well as 20 custom features across the range, which is strong evidence that the fundamental determinants of reactivity do not change during the pre-launch window. Because the classifiers were each trained separately at each time point to produce models with different learned coefficients, these fundamental determinants of reactivity can (and do) play somewhat different roles at different times.

A structural representation of the set of 30 consensus predictive features is shown in Figure 4A (17 distances) and Figure 4B (12 planar angles and 1 dihedral angle). Half of the features (15) represent interactions between the substrate and its environment (nearby water molecules, the two magnesium ions, and the side chain of Glu 319), 7 represent intra-substrate conformational metrics, 7 represent water–metal interactions, 1 represents an intra-co-factor orientation, and 2 represent other intra-environment interactions. A full third of the features (10) represent distances or angles describing the relationship of a single atom, the substrate hydroxyl oxygen (O6), to its environment—the coordinating magnesium ions and water molecules interacting with the metal ions. The largest number of intermolecular features involving any other substrate atom is 2, for both a substrate carboxylate oxygen (O3) and the substrate carbonyl oxygen (O8), whose carbon receives the migrating methyl group. Only one intermolecular

interaction involves the migrating methyl itself. We note two additional characteristics of the feature set: (1) the substrate intramolecular features involve the geometry local to the C4–C7 covalent bond, which is parallel to the path of the migrating methyl group (it departs from C4 and arrives at C7), and (2) 8 of the 10 intermolecular angle features describe the orientation of groups coordinating the metal ions—either their ligated water molecules or oxygen atoms of the substrate (carboxylate, hydroxyl, or carbonyl). It must be recognized that the composition of the initial 68 features had some effect on the composition of the selected features. Nevertheless, the composition of this consensus feature set suggests important roles for substrate conformation, substrate bond polarization, and metal coordination in the reaction mechanism.

Table 3 and Figure 5 describe the contribution of individual features to the classifiers trained at representative time points (–150, –100, –50, and 0 fs) using the consensus feature set. Table 3 presents standardized logistic regression coefficients β_j for each classifier, with a positive (negative) coefficient β_j indicating that increasing the value of feature j (corresponding to a bond length, bond angle, or dihedral angle), tends to increase (decrease) the likelihood of classifying a trajectory being as reactive. A higher absolute value for coefficient β_j indicates that feature j has a greater contribution to the probability of a trajectory evaluating as reactive relative to the other features at the same time point (note that in this standardized representation the features themselves are input to the classifier as Z-scores, so different scales for bond lengths and bond angles, for instance, don't contribute to coefficient values).

Average reactive and almost-reactive time traces for the consensus feature set are presented in Figure 5A. The four time points for which coefficients are presented in Table 3 are shown as vertical black dashed lines for each of the 30 features in Figure 5A. Bonded intra-substrate distances and angles, such as features 9, 14, 15, 20, and 25 tend to exhibit oscillatory

behavior consistent with the vibrations of the breaking bond, other inter-atomic distances and angles, such as features 2, 5, 8, and 22, tend to progress monotonically, and still others tend to remain relatively constant in the pre-launch window (features 7, 21, 26, and 32), sometimes with a few characteristic deviations. The results follow the general trend that features with lower magnitude coefficients in Table 3 tend to exhibit more closely overlapping reactive and almost-reactive distributions at corresponding time points in Figure 5A, and that greater magnitude coefficients correspond to more distinct trajectory distributions, but there are exceptions as well. For example, at the –150 and –100-fs time points, β_1 exhibits values of –0.361 and –0.527, respectively, indicating that for feature 1 (distance Glu 319/Oε1–AC6/C5) increases in the distance tend to decrease reactivity; the observation from the simulations, shown in Figure 5A, matches in the sense that this distance is smaller, on average, for the reactive compared to the almost-reactive trajectories at these time points (the average reactive trace (red) is lower than the average almost-reactive trace (blue) for these time points), but the difference is larger for the smaller absolute-value coefficient. Interestingly, this feature has the opposite effect at the later time point of 0 fs, where higher values of feature 1 are more predictive of reactivity, (β_1 coefficient of 0.470) and the average reactive trace for feature 1 is greater than the average almost-reactive trace for at the 0-fs time point in Figure 5A. Thus, the same features in the consensus set are generally predictive across all time points, but they can be used somewhat differently at different times.

The closely overlapping distributions of most features in Figure 5A suggest the need for multiple features in combination to make usefully accurate predictions. Histograms of reactive and almost-reactive trajectories for feature pairs and triplets (2D and 3D histograms; data not shown) show somewhat greater separation than that seen in Figure 5A, but still considerable

overlap between reactive and almost-reactive distributions at individual time points, consistent with the relatively poor classification performance of models with fewer than 10 features.

Subtle variations distinguish different reactive channels. We examined the question of whether there were single or multiple channels through which reaction proceeded. Clustering was used to organize the correctly predicted reactive and almost-reactive trajectories into related sets, and the magnitude of the differences between the sets was examined. This also allowed a more fine-grained analysis of the determinants of reactivity as identified by the machine learning. Specifically, all correctly predicted reactive trajectories were clustered based on the 0-fs time point using the 30 consensus features, each weighted by its β_j value (we refer to this as the feature weight; see Methods; results for five clusters are shown in Figure 5B). The results show at least five different modes of reacting, with each cluster distinguished by which features contribute most and least to the classifier outcome. In Figure 5B, the thirty columns represent the contribution from each of the 30 consensus features and the rows each represent one trajectory. Red bars correspond to a positive (more reactive) contribution to the classifier and blue bars correspond to a negative (less reactive) contribution to the classifier, with a darker color corresponding to a stronger contribution. For example, cluster 1 is distinguished by a dark blue band for distance MG6/H32–AC6/O6 (feature 2; indicating a contribution that disfavors reactivity), offset by the dark red bands for distance MG6/H31–AC6/O6 and distance MG6/M16–AC6/O3 (features 22 and 27, respectively; favoring reactivity). Conversely, cluster 3 exhibits partial reversal of that pattern—a strong red band for distance MG6/H32–AC6/O6 (feature 2; favoring reactivity) and a strong blue band for distance MG6/M16–AC6/O3 (feature 27; disfavoring reactivity), which correspond to higher than average values for distance MG6/H32–AC6/O6 (feature 2 has a positive coefficient in the 0-fs model) and also higher values

for distance MG6/H31–AC6/O6 (feature 22 has a negative coefficient) associated with this cluster. Figure 5B shows that at the 0-fs time point, roughly half of the 30 features contribute very little to the decision as indicated by white bands in each cluster. Further confirmation is seen by the observation that features that appear as white bands usually do not occur in the top 20 LASSO selected set at this time point (see Figure 3E; distance AC6/O6–MG6/M16 and distance MG6/O19–MG6/M17 are exceptions and rank 15 and 18, respectively, in the top 20 LASSO selected set).

A corresponding set of almost-reactive clusters was constructed such that each correctly predicted almost-reactive trajectory was associated with a cluster shown in Figure 5B (feature contributions at 0 fs were computed for correctly predicted almost-reactive trajectories, and each was assigned to the nearest reactive cluster). The almost-reactive trajectory feature contributions are shown in Figure 5C grouped into the five clusters. The almost-reactive weighted features fall into groups that approximate the distinguishing features of each reactive cluster. For example, cluster 1 in both Figure 5B and its corresponding almost-reactive cluster in Figure 5C is characterized by a dark blue band for distance MG6/H32–AC6/O6 (feature 2) and dark red bands for distances MG6/H31–AC6/O6 and MG6/M16–AC6/O3 (features 22 and 27, respectively). If each cluster is viewed as a somewhat distinct channel by which reactive and almost-reactive trajectories approach the barrier, these results suggest that each channel can accommodate both reactive and almost-reactive trajectories, and that a comparison of Figure 5B and Figure 5C might be helpful in identifying subtle differences contributing to relative reactivity (see next paragraph). Moreover, the differences in the numbers of trajectories in each cluster between reactive and almost-reactive sets indicate that some of the channels (clusters 1, 2, and 5) led to a greater fraction of reactive as compared to almost-reactive trajectories than others (clusters 3 and

4; note that these observations remain true when one does a comparison that includes all reactive and almost-reactive trajectories, not just those correctly predicted; results not shown).

Grouping the weighted features into reactive clusters and corresponding almost-reactive clusters allows the subtle differences that define reactivity for each of these subgroups to be more closely examined. To this end, the mean feature contribution for each almost-reactive cluster in Figure 5C was subtracted from each of the weighted features from the corresponding cluster of reactive trajectories from Figure 5B to obtain a mapping of how each feature in each reactive trajectory differs from its mean in the corresponding almost-reactive cluster (Figure 5D); the results show several common features that distinguish correctly predicted reactive from correctly predicted almost-reactive clusters. For example, across all five clusters shown in Figure 5D, the darkest red bands appear for distances AC6/C5–AC6/C4 and MG6/M16–AC6/O3 (features 10 and 27, respectively), indicating that these features are critical in driving the reactive/almost-reactive decision. However, there are other differences that are cluster-specific; for example, differences in the distance AC6/O8–Glu 496/Hε2 (feature 6) are responsible for distinguishing reactive from nearly-reactive more for cluster 3 than for any of the others, on average.

Distributions of feature values with the strongest contributions to differences in reactivity amongst the clusters (i.e., the darkest bands in Figure 5D), are shown, per cluster, in Figure 6. Although there is often considerable overlap in the individual feature distributions between each reactive and almost-reactive cluster, the set of 5 features alone, when re-trained on each cluster alone, achieved AUCs of 1.00, 1.00, 0.94, 0.91 and 1.00, in classifying trajectories from clusters 1 though 5, respectively, as reactive or almost-reactive. These very high scores suggest that the more general classifiers presented earlier somehow carry out the dual tasks of determining which

reaction channel the trajectory is headed toward, as well as whether the trajectory will successfully react through that channel. The high AUCs suggest that determining which channel is being approached may be the harder portion of the task, although this effect is convolved with the fact that these clusters are composed of trajectories that were correctly classified previously. When all (including incorrectly classified) data points are used, the intercluster AUCs using the same set of features are 0.92, 0.93, 0.80, 0.88 and 1.00 respectively, supporting the interpretation that predicting reactivity within a cluster is easier than in the absence of knowledge of the cluster for most of the clusters.

Figure 6 shows that across all five clusters, some general trends exist for the five features and their relative distribution between reactive and almost-reactive trajectories. The strongest observation is that in almost every instance, each significant feature has a much narrower distribution in the reactive than the almost-reactive set of trajectories. This is consistent with the notion that there are many ways of not reacting, but fewer modalities for successfully traversing the reaction barrier. Across most of the five clusters, in general, reactivity is associated with a shorter AC6/C5–AC6/C4 bond length (column 2; feature 9; clusters 1, 2, 4, and 5), a longer AC6/C1–AC6/C4 bond length (column 4; feature 25; clusters 2, 3, and 5), a longer GLU319/Oε1–AC6/C5 distance (column 1; feature 1; clusters 1, 2, 4, and 5), and a shorter MG6/M16–AC6/O3 distance (column 5; feature 27; clusters 1, 2, and 5). The value of the MG6/H29–MG6/O18–MG6/M17 angle (column 3; feature 20) is associated with reactivity for small values in cluster 1 but large values in cluster 5. Nevertheless, the absolute values associated with reactivity for some of the features varies greatly between clusters (column 3 for clusters 1 and 5, and column 5 for clusters 1 and 2, for example). Taken together, these results reinforce the notion that a common set of fundamental reaction-promoting mechanisms are

deployed in somewhat different combinations in the different clusters. The structural implications of these observations were explored. Here we illustrate the trends through a comparison of the reactive and almost-reactive structures closest to the centroid of each respective cluster (Figure 7B–F), indicated by the colored or grey open circles, respectively, in Figure 6. Additionally, a single structure closest to the centroid of each reactive cluster is shown in Figure 7A.

Figure 7A shows a number of interesting structural variations, particularly considering that they represent the final bond compression of reactive trajectories. The five water molecules that coordinate to one or the other magnesium ion each show significant variability across the clusters. Some of these are differences in water molecule positioning are represented in the features significant for cluster identify (e.g, features 2 and 22) and others in those significant for reactivity within a cluster (e.g., feature 20). Additionally, there is substantial variability in the internal substrate conformation across the different reactive clusters, with cluster 2 being an especially unusual outlier.

Illustrated in Figure 6 (column 5; feature 27), clusters 1, 2, and 5 exhibit significantly shorter values for the distance MG6/M16–AC6/O3 for the reactive than the almost-reactive trajectories. Structurally, Figure 7B, C, and F show that this corresponds to a different conformation of the substrate carboxylate group and a different engagement of magnesium ion M16 between reactive and almost-reactive trajectories. This shorter distance corresponds to a somewhat different orientation for the entire substrate relative to the two magnesium ions that also affects substrate hydroxyl O6 and the metal coordination environment. By contrast, clusters 3 and 4 show much less difference in the distribution of MG6/M16–AC6/O3 (feature 27)

between reactive and almost-reactive sets (Figure 6) and this can also be seen structurally in Figure 7D and E.

Also illustrated in Figure 6 (column 2; feature 9), all five clusters show that the length of the breaking bond, AC6/C5–AC6/C4, spans a wider range of values for the nearly-reactive trajectories and is on the shorter side of that distribution for the reactive ones. Keeping in mind that these conformations are for the 0-fs time point, when the bond is fully compressed before launching toward the barrier, this represents the notion that reactive trajectories require substantial potential energy by stored in the bond that is not always seen for almost-reactive trajectories (that is, this extra compression is necessary but not sufficient).

Figure 6 indicates that the adjacent substrate bond, AC6/C1–AC6/C4 (column 4; feature 25), is distributed somewhat longer in reactive than almost-reactive trajectories for clusters 2, 3, and 5; examining the corresponding structures in Figures 7C, D, and F doesn't show a clear effect of this on conformation. Figure 6 also indicates that a water molecule orientation, angle MG6/H29–MG6/O18–MG6/M17 (column 3; feature 20), is distributed substantially larger for reactive than almost-reactive trajectories in cluster 5, and much more so than in any of the other clusters. Figure 7F seems to indicate that this allows engagement of a lone pair from O18 to interact much more favorably with magnesium, and perhaps affect the polarization of the substrate, in a typical reactive rather than almost-reactive trajectory. Some of the other clusters appear to show a difference in the interaction between that water molecule and magnesium ion, although it may not show up in the angle indicated. Finally, Figure 6 indicates that the distance from Glu 319 Oε1 to the substrate's migrating methyl group C5 (column 1; feature 1) is distributed longer in reactive than almost-reactive trajectories for clusters 1, 2, and 5 (and partially for clusters 3 and 4). Figure 7B–F indicates the interaction, but it is unclear how

much of it is steric (the Glu side chain must be far enough away from the methyl at compression to be adequately poised to push it toward product upon bond expansion) and how much is stabilizing of the methyl during the transition.

In summary, a comparison of feature histograms and representative structures shows that features distinguishing reactive from almost-reactive trajectories include internal conformational degrees of freedom of the substrate, which may provide distortion toward the transition state and ground-state destabilization; subtle changes to polar interactions of the two magnesium ions with the substrate and with their ligating water molecules and side chains, which could have important effects in polarizing the substrate toward reactivity; and interactions of the side chain of Glu 319 with the migrating methyl group, which could be important for steric, kinetic, and electronic reasons. It is anticipated that more detailed molecular orbital analyses will contribute to an understanding of how these structural differences are responsible for changes in relative reactivity.

Predictive Features Direct Reactivity. Machine-learning analysis was used here to develop predictive models capable of distinguishing reactive from nearly reactive trajectories. Predictions of reactivity were successful, even when applied to trajectories not used in training the models, further supporting the notion that model features represent characteristics of reactivity. We reasoned that these characteristics could be useful not only to predict reactivity, but also to direct it. That is, if the features identify characteristics that are largely sufficient for reactivity, rather than just indicative of it, then trajectories constrained to possess reactive characteristics should show markedly increased reactivity. We tested that notion, described below, and our findings confirm the directive power of the machine learning features and their associated models.

The LASSO-selected, ten-feature model at the 0-fs time point was used, with testing performance AUC of 89.03% and accuracy of 81.57%. Model features and the corresponding logistic-regression coefficients are listed in Table 4. Eight of the ten features occur in the 30-feature consensus set, with the exceptions being distance AC6/C4–AC6/O6 and distance AC6/O8–MG6/M17. Of the five features shown in Figure 6, four appear in the ten-feature model, with the exception being angle MG6/H29–MG6/O18–MG6/M17 (feature 3). Thus, the ten-feature model achieves very good predictive performance and is composed of many of the consensus features found to be important at other time points.

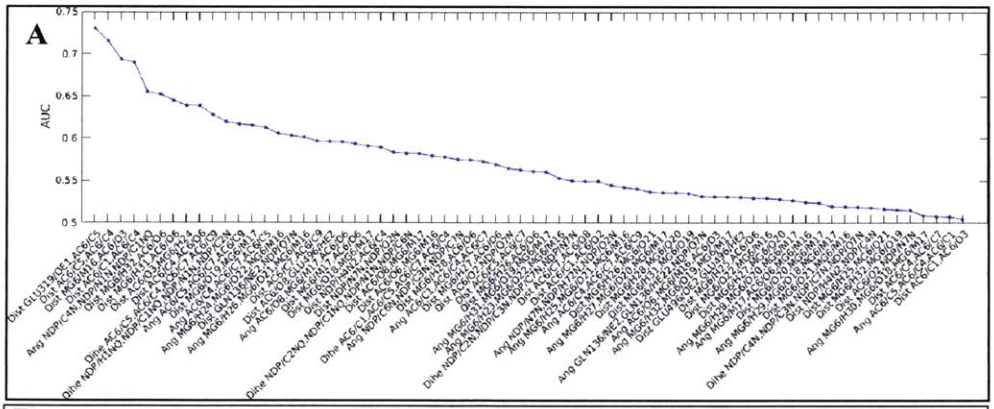
The logistic regression machine learning models used here effectively create a dividing surface in the reactant well (the hyperplane defined by the β_j coefficients; see Methods), and make successful predictions of reactivity based on whether the trajectory is in the "reactive portion" of the well at the appropriate time. We modified the statistical mechanical TIS sampling procedure used here to compute reaction rates, so that we could require all trajectories to be on the reactive side of the hyperplane encoded in the ten-feature model (Table 4) during a rate calculation (see Methods). Calculations of the reaction rate were performed with ("test") and without ("control") this constraint applied only at the 0-fs time point from five different starting seeds (three were used previously to train the model, and two were new). The expectation was that the test simulations would show greater reactivity (larger computed k_{cat}) than the controls, as the test simulations satisfied the reactivity conditions in every trajectory (by constraint), whereas on average only 8.03% of control trajectories satisfied them through ordinary statistical sampling.

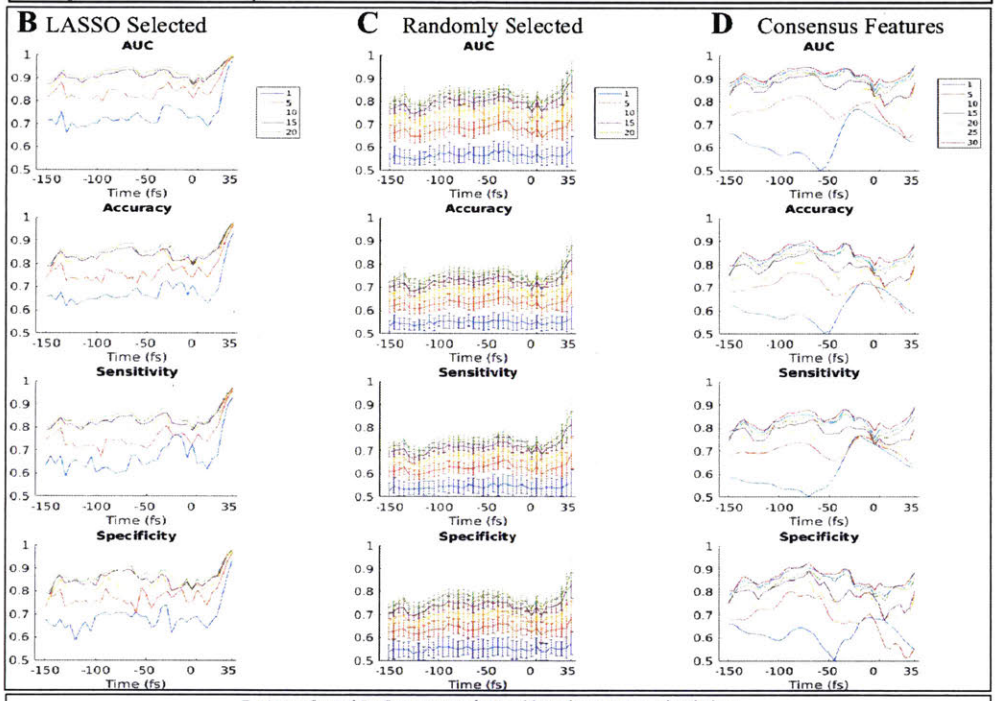
The observed relative differences in rate constants in all five sets of simulations was consistent with this expectation and quite large, on the order of 10^{16} to 10^{19}, depending on the

initial seed trajectory (Table 5). The computed rate is a product of a factor representing the rate of reactant starting toward the barrier and a probability factor representing the likelihood of progress toward and over the barrier. Here the rate enhancement was driven by both factors, but with a significantly larger effect from the probability factor and with contributions across much of the approach to the barrier, which suggests that greater reactivity was due to increased productivity at multiple stages of the reaction, including those after the reactant left the reactant well (see additional results below).

Contributions to the probability factor were examined in more detail. Figure 8A shows the cumulative logarithm of the probability factor as a function of reaction progress for test (red) and control (blue) simulations (essentially the probability that a trajectory that started toward the barrier will reach this value of λ, the reaction progress variable). Figure 8B shows the individual multiplicative contribution to the probability factor at each progress window (essentially the probability that a trajectory that made it through the previous window will continue through this window). The test simulations show much smaller decreases in the reaction probability (Figure 8A) and much larger contributions to reactivity (Figure 8B) than the control simulations earlier in the reaction (below λ=–0.4) but show similar behavior beyond that point (between λ=–0.4 and 0.0). These data indicate a strong reactivity advantage of the constrained simulations (which was applied at the 0-fs time point, corresponding to a λ value of approximately –0.9 and well before the barrier) across the whole region from λ=–0.9 through –0.4 but not past this point, noting that by λ=–0.2 the reaction has essentially already occurred. This is consistent with a picture in which the constraint achieved its large gains in reactivity not by giving those simulations a local, near-term boost in reaction progress, but by directing them into channels that retained a continuous reactivity advantage.

Figure 3: (A) AUC performance for all 68 individual features. AUC is the area under the curve for the receiver operating characteristic for the machine learning model. Values of AUC shown represent the mean computed across 5 equal cross-validation training and testing partitions. (B) AUC, accuracy, sensitivity, and specificity for models with LASSO-selected features (C) AUC, accuracy, sensitivity, and specificity are plotted for models with randomly-selected features. Models with 1, 5, 10, 15, and 20 features are shown for each of 38 time points across the time range –150 to +35 fs. Error bars in (B) correspond to standard error of the mean across 100 randomly-selected feature sets. Accuracy is the number of correctly classified trajectories divided by the total number of classified trajectories. Sensitivity is the number of correctly classified reactive trajectories divided by the total number of reactive trajectories. Specificity is the number of correctly classified almost-reactive trajectories divided by the total number of nearly-reactive trajectories. (E) Top 20 features selected by LASSO at each time point before the last trough. Features are colored by feature type and sorted by the total number of occurrences in the top 20 between -150 and 0 fs before the last trough.


A
AUC
0.75
0.7
0.65
0.6
0.55
0.5


B LASSO Selected
C Randomly Selected
D Consensus Features
AUC
Accuracy
Sensitivity
Specificity
Time (fs)
1
5
10
15
20
25
30

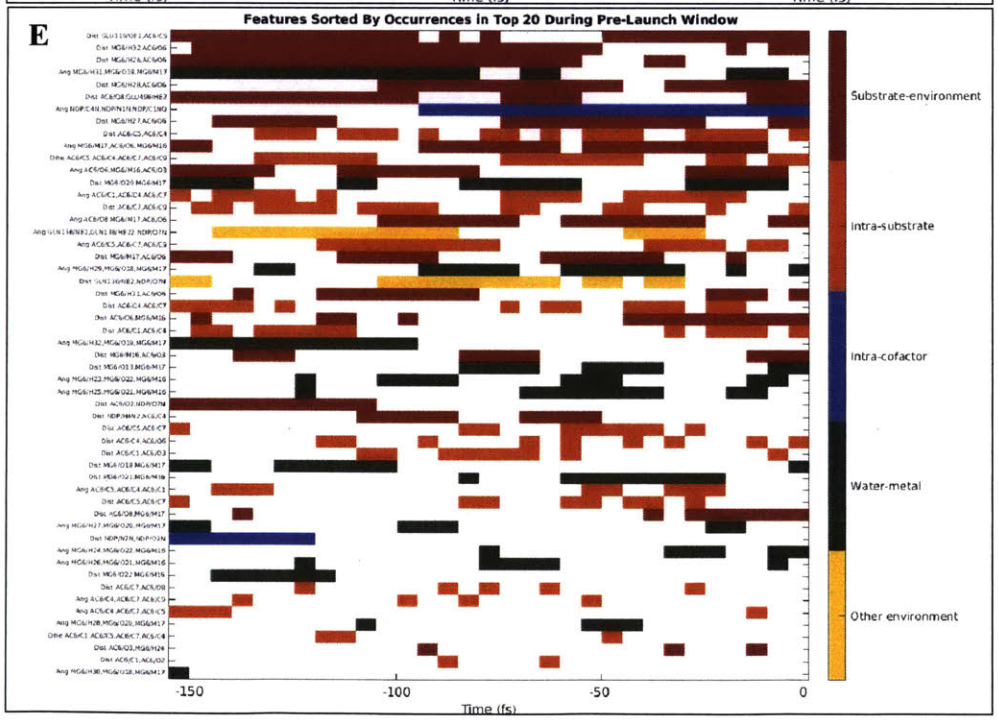

E
Features Sorted By Occurrences in Top 20 During Pre-Launch Window
Substrate-environment
Intra-substrate
Intra-cofactor
Water-metal
Other environment
Time (fs)

Table 2: Top 30 consensus features for the –150 to 0 fs time window. Feature rank indicates ranking according to the number of occurrences in the 20 LASSO-selected feature sets.

Feature Rank	Feature Name	Feature Type	Occurrences in Top 20 Between -150 and 0 fs
1	Dist GLU319/Oε1,AC6/C5	Substrate-environment	24
2	Dist MG6/H32,AC6/O6	Substrate-environment	23
3	Dist MG6/H26,AC6/O6	Substrate-environment	22
4	Ang MG6/H31,MG6/O19,MG6/M17	Water-metal	20
5	Dist MG6/H28,AC6/O6	Substrate-environment	19
6	Dist AC6/O8,GLU496/Hε2	Substrate-environment	19
7	Ang NDP/C4N,NDP/N1N,NDP/C1NQ	Intra-cofactor	19
8	Dist MG6/H27,AC6/O6	Substrate-environment	18
9	Dist AC6/C5,AC6/C4	Intra-substrate	18
10	Ang MG6/M17,AC6/O6,MG6/M16	Substrate-environment	18
11	Dihe AC6/C5,AC6/C4,AC6/C7,AC6/C9	Intra-substrate	17
12	Ang AC6/O6,MG6/M16,AC6/O3	Substrate-environment	17
13	Dist MG6/O20,MG6/M17	Water-metal	17
14	Ang AC6/C1,AC6/C4,AC6/C7	Intra-substrate	16
15	Dist AC6/C7,AC6/C9	Intra-substrate	16
16	Ang AC6/O8,MG6/M17,AC6/O6	Substrate-environment	16
17	Ang GLN136/Nε2,GLN136/Hε22,NDP/O7N	Other environment	16
18	Ang AC6/C5,AC6/C7,AC6/C9	Intra-substrate	15
19	Dist MG6/M17,AC6/O6	Substrate-environment	15
20	Ang MG6/H29,MG6/O18,MG6/M17	Water-metal	15
21	Dist GLN136/Nε2,NDP/O7N	Other environment	15
22	Dist MG6/H31,AC6/O6	Substrate-environment	13
23	Dist AC6/C4,AC6/C7	Intra-substrate	13
24	Dist AC6/O6,MG6/M16	Substrate-environment	13
25	Dist AC6/C1,AC6/C4	Intra-substrate	12
26	Ang MG6/H32,MG6/O19,MG6/M17	Water-metal	12
27	Dist MG6/M16,AC6/O3	Substrate-environment	10
28	Dist MG6/O19,MG6/M17	Water-metal	10
29	Ang MG6/H23,MG6/O22,MG6/M16	Water-metal	10
30	Ang MG6/H25,MG6/O21,MG6/M16	Water-metal	10

Figure 4: Structural representations of top 30 most consistently predictive (A) distances and (B) angles and dihedrals during the -150 to 0 fs time window. Labeling of features corresponds to ranking in Table 2. Coloring of features corresponds to the feature type with red indicating substrate-environment interactions, orange indicating intra-substrate conformations, blue indicating intra-cofactor conformations, green indicating water-metal interactions and gold indicating other environment interactions.

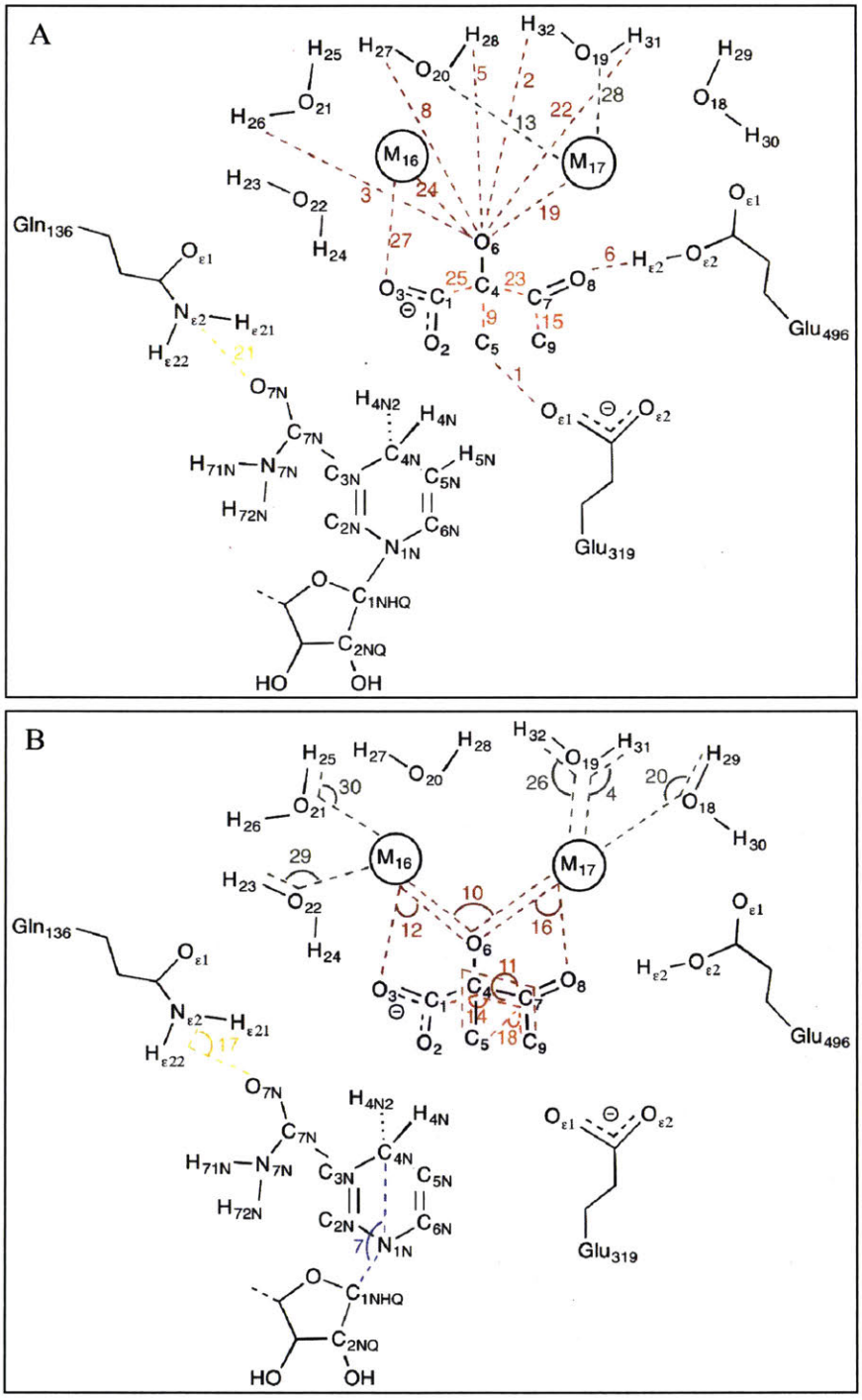


Table 3: Mean standardized logistic regression coefficients fit to classifier trained using the top 30 most consistently predictive features between -150 and 0 fs (listed in Table 2 and illustrated structurally in Figure 4) at the -150, -100, -50 and 0 fs time points relative to the last trough in the order parameter prior to the prospective catalytic event. Coefficients shown represent the mean values across 5 cross-validation partitions.

Standardized Regression Coefficient	Time Before Last Trough			
	-150 fs	-100 fs	-50 fs	0 fs
β_0	-0.059	-0.195	-0.269	-0.094
β_1	-0.361	-0.527	0.354	0.470
β_2	-0.198	-0.569	-0.374	0.874
β_3	-0.303	-0.957	-0.497	-0.035
β_4	0.615	0.706	0.117	0.453
β_5	0.094	0.069	0.401	-0.477
β_6	-0.365	-0.265	-0.147	-0.613
β_7	0.273	-0.251	-0.423	-0.397
β_8	0.293	-0.428	-1.134	-0.356
β_9	0.068	0.446	0.533	-1.030
β_{10}	0.318	-1.060	-0.666	-0.025
β_{11}	-0.307	0.058	-1.379	-0.289
β_{12}	-0.723	0.414	0.179	-0.510
β_{13}	0.236	-0.129	0.610	0.050
β_{14}	-0.256	0.214	-0.348	-0.107
β_{15}	-0.132	-0.460	-0.227	0.269
β_{16}	-0.330	-0.237	1.049	0.106
β_{17}	0.065	0.302	0.137	0.039
β_{18}	0.193	-0.704	0.665	0.026
β_{19}	-0.426	0.252	-0.425	0.007
β_{20}	0.033	0.141	0.477	0.704
β_{21}	0.319	-0.327	-0.471	-0.013
β_{22}	-0.135	-0.630	-0.162	-1.100
β_{23}	0.790	0.281	-0.089	0.200
β_{24}	-0.048	-0.179	-0.127	-0.014
β_{25}	0.083	-0.047	-0.182	0.504
β_{26}	0.592	0.592	0.434	-0.244
β_{27}	0.142	0.093	0.241	-0.944
β_{28}	-0.208	0.477	0.437	-0.083
β_{29}	-0.148	-0.370	-0.327	-0.061
β_{30}	0.183	0.151	0.338	-0.174

Figure 5: (A) Average time traces of consensus features across -150 to +100 fs time points with red indicating average reactive traces and blue indicating average almost-reactive traces. Error bars indicate 2 standard errors of the mean at each time point. Vertical black lines indicate time points at -150,-100,-50 and 0 fs where coefficients listed in Table 3 were fit. (B) Z-scores for consensus features (listed in Table 2 and illustrated structurally in Figure 4) evaluated at the 0 fs time point and weighted by their corresponding standardized logistic regression coefficient for all correctly classified reactive trajectories in data set. Dark lines indicate cluster boundaries assigned using k-means clustering with k=5. Within each cluster, features are sorted by distance from the centroid of the respective cluster (closest to centroid at top). (C) Z-scores for the consensus features evaluated at the 0 fs time point and multiplied by their corresponding standardized logistic regression coefficient for all correctly classified almost-reactive trajectories in data set. Dark lines indicate cluster assignments, based on the closest centroid to the five centroids learned on the reactive features shown in (B). (D) Z-scores differences between reactive features in each cluster and the mean almost-reactive feature set of the corresponding almost-reactive cluster. In (B), (C) and (D), blue lines indicate negative values, red lines indicate positive values, and white lines indicate zero values.

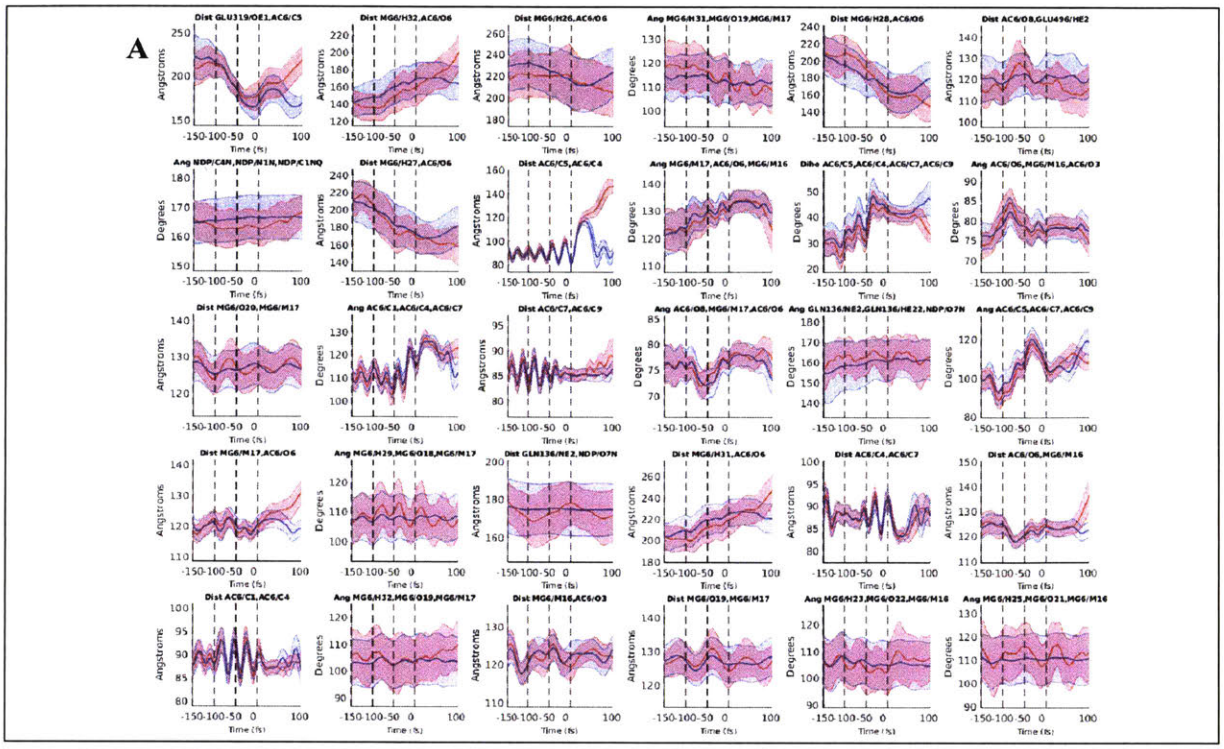

A
Dist GLU319/OE1,AC6/C5
Dist MG6/H32,AC6/O6
Dist MG6/H26,AC6/O6
Ang MG6/H31,MG6/O19,MG6/M17
Dist MG6/H28,AC6/O6
Dist AC6/O8,GLU496/HE2
Ang NDP/C4N,NDP/N1N,NDP/C1NQ
Dist MG6/H27,AC6/O6
Dist AC6/C5,AC6/C4
Ang MG6/M17,AC6/O6,MG6/M16
Dihe AC6/C5,AC6/C4,AC6/C7,AC6/C9
Ang AC6/O6,MG6/M16,AC6/O3
Dist MG6/O20,MG6/M17
Ang AC6/C1,AC6/C4,AC6/C7
Dist AC6/C7,AC6/C9
Ang AC6/O8,MG6/M17,AC6/O6
Ang GLN136/NE2,GLN136/HE22,NDP/O7N
Ang AC6/C5,AC6/C7,AC6/C9
Dist MG6/M17,AC6/O6
Ang MG6/H29,MG6/O18,MG6/M17
Dist GLN136/NE2,NDP/O7N
Dist MG6/H31,AC6/O6
Dist AC6/C4,AC6/C7
Dist AC6/O6,MG6/M16
Dist AC6/C1,AC6/C4
Ang MG6/H32,MG6/O19,MG6/M17
Dist MG6/M16,AC6/O3
Dist MG6/O19,MG6/M17
Ang MG6/H23,MG6/O22,MG6/M16
Ang MG6/H25,MG6/O21,MG6/M16
Angstroms
Degrees
Time (fs)
B Correctly Classified Reactive
C Correctly Classified Almost-Reactive
D Reactive /Almost-Reactive Difference

Figure 6: Histograms of weighted feature weight differences across each of the five reactive clusters and corresponding almost-reactive clusters. The set of five features shown was determined by computing the top three weighted feature differences by absolute value for each cluster shown in Figure 5D, then taking the union of the resulting set, which led to the five unique features listed. Magenta corresponds to cluster 1, cyan corresponds to cluster 2, green corresponds to cluster 3, yellow corresponds to cluster 4, orange corresponds to cluster 5 and gray corresponds to the corresponding almost-reactive cluster for the reactive cluster shown in each histogram. Dots in Figure 6B indicate representative structures (the reactive or almost-reactive structures closest to the mean of the centroid for each respective cluster) which are shown in Figure 7B-F.

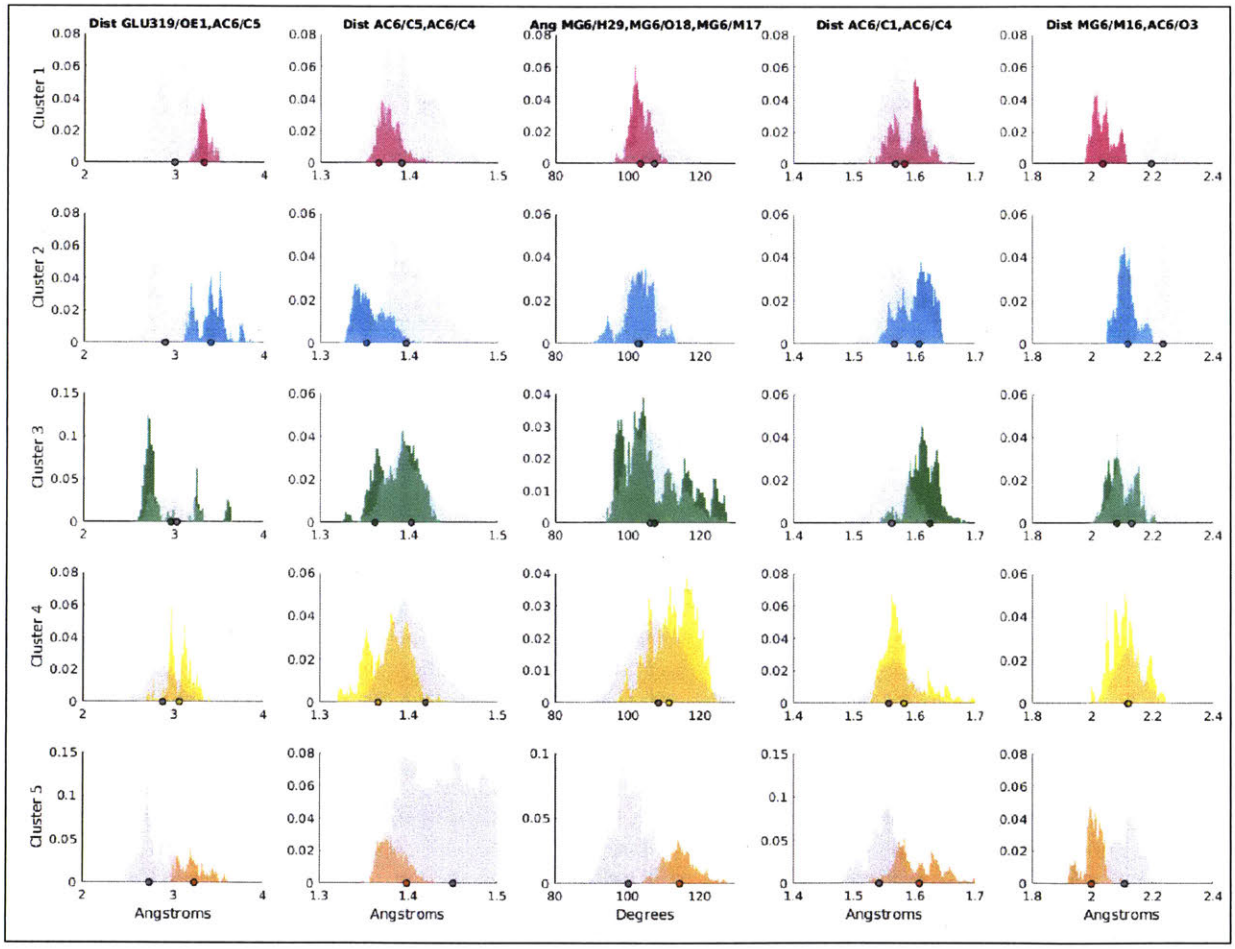


Figure 7: Representative structures for the reactive cluster and corresponding almost-reactive clusters described in Figure 5B,C and D and Figure 6. Feature numbering corresponds to that of Table 2. (A) Representative structures from all five reactive clusters. (B) Representative structures from cluster 1 and its corresponding almost-reactive cluster. (C) Representative structures from cluster 2 and its corresponding almost-reactive cluster. (D) Representative structures from cluster 3 and its corresponding almost-reactive cluster. (E) Representative structures from cluster 4 and its corresponding almost-reactive cluster. (F) Representative structures from cluster 5 and its corresponding almost-reactive cluster. In all panels, magenta corresponds to cluster 1, cyan corresponds to cluster 2, green corresponds to cluster 3, yellow corresponds to cluster 4, orange corresponds to cluster 5 and gray corresponds to the corresponding almost-reactive cluster for the reactive cluster shown in each histogram. In all panels, structures were aligned to minimize the root mean square difference between the two magnesium centers.

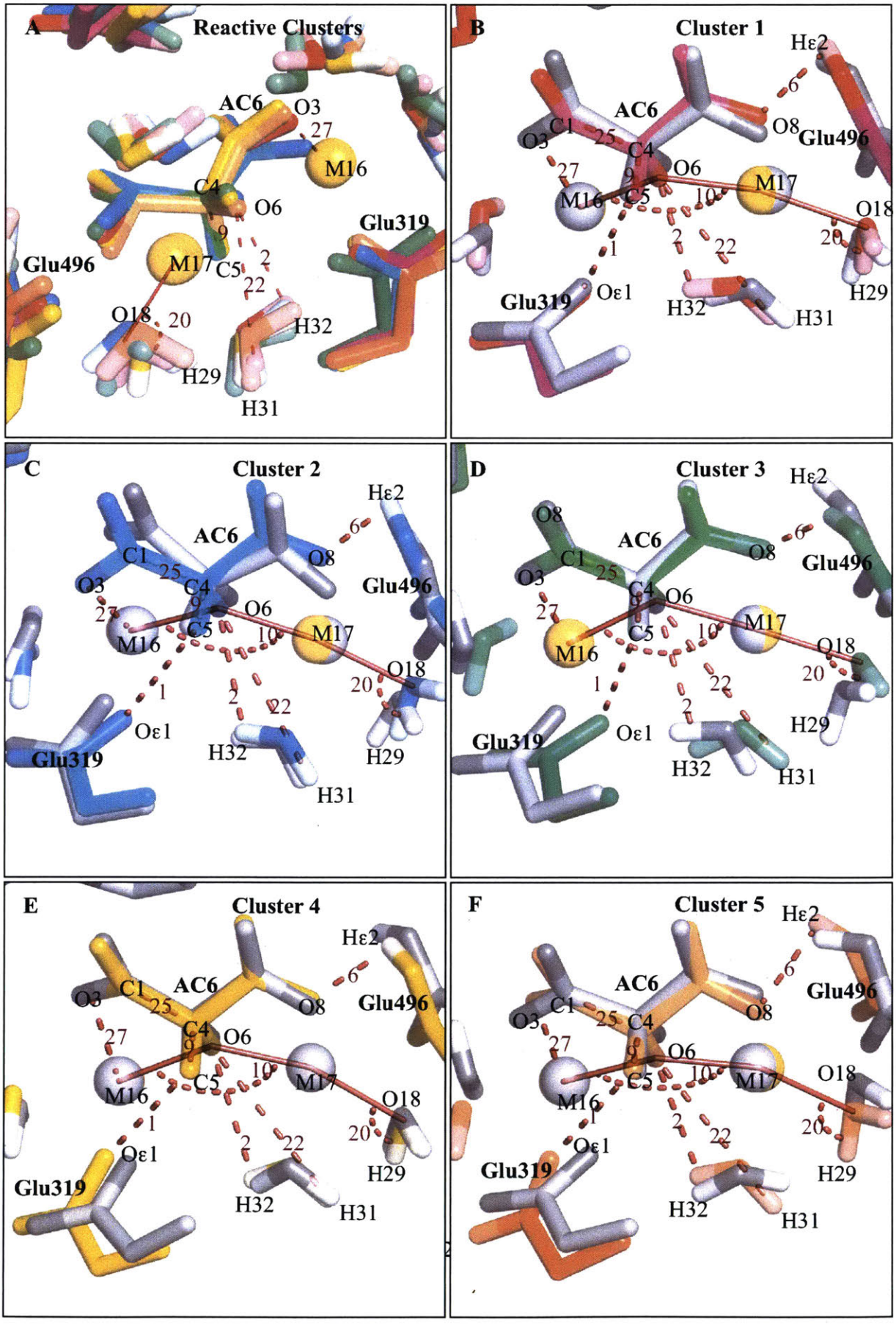

A
Reactive Clusters
AC6
O3
27
M16
C4
O6
9
Glu319
Glu496
M17
C5
2
22
O18
20
H32
H29
H31
B
Cluster 1
Hε2
6
AC6
C1
O8
Glu496
O3
25
C4
27
9
O6
M16
C5
10
M17
O18
2
20
1
22
Glu319
Oε1
H32
H31
H29
C
Cluster 2
Hε2
AC6
6
C1
O8
25
Glu496
O3
C4
9
O6
27
M16
C5
10
M17
O18
20
1
2
22
Oε1
Glu319
H32
H29
H31
D
Cluster 3
Hε2
O8
AC6
6
C1
O8
Glu496
25
O3
C4
9
O6
27
M16
C5
10
M17
O18
20
1
22
2
Oε1
H29
Glu319
H32
H31
E
Cluster 4
Hε2
AC6
6
C1
Glu496
O3
25
O8
C4
O6
27
9
M16
C5
10
M17
O18
20
1
2
22
Oε1
H29
Glu319
H32
H31
F
Cluster 5
Hε2
6
AC6
Glu496
C1
O3
25
C4
O8
27
9
O6
M16
C5
10
M17
O18
1
2
20
22
Oε1
Glu319
H29
H32
H31

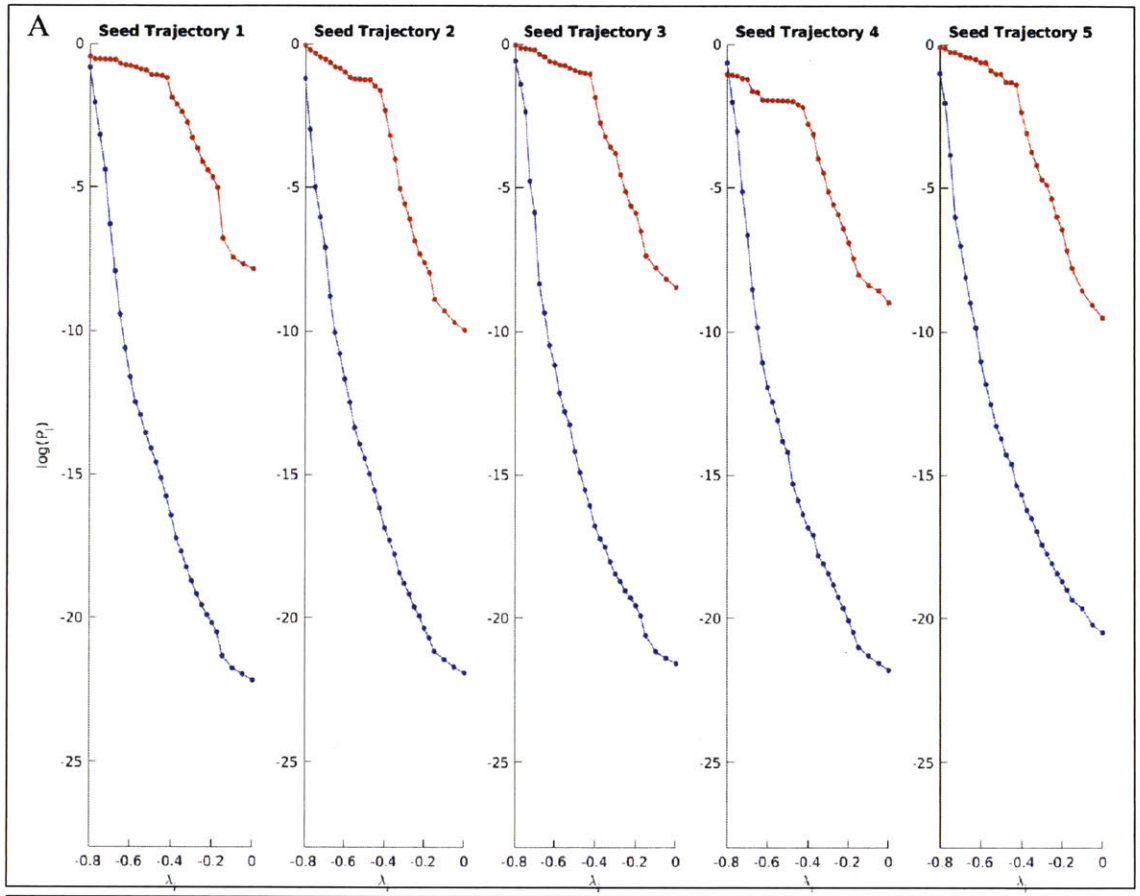
Table 4: Top 10 LASSO selected features at 0 fs time point and coefficients β_j used to define reactive region A' in constrained TIS simulations. Note that classification was performed on the fly through the TIS Markov chain and thus features were not normalized by Z-scores, so non-standardized coefficients β_j are reported. The bias β_0 used was -18.603.

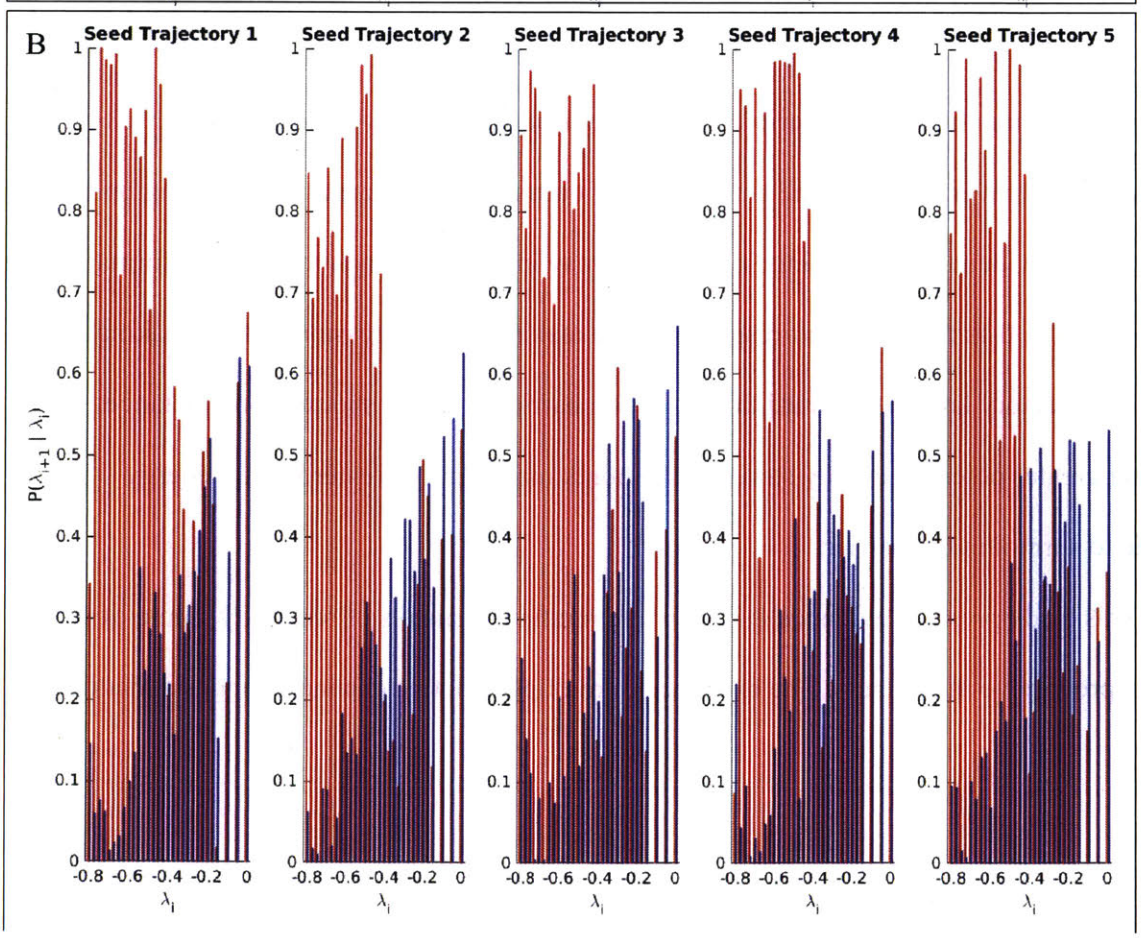
j	Feature	β_j
1	Distance GLU`319/Oε1,AC6/C5	2.1944
2	Distance MG6/M16,AC6/O3	-12.093
3	Distance AC6/C1,AC6/C4	13.447
4	Distance AC6/C4,AC6/O6	20.561
5	Angle NDP/C4N,NDP/N1N,NDP/C1NQ	-2.8234
6	Distance AC6/O8,GLU`496/Hε2	-3.4298
7	Distance AC6/C5,AC6/C4	-8.8403
8	Distance AC6/O8,MG6/M17	8.8193
9	Dihedral AC6/C5,AC6/C4,AC6/C7,AC6/C9	-3.7307
10	Distance MG6/H28,AC6/O6	-0.5615

Table 5: Computed rate constants, probability factors and flux factors for each seed studied. Values of upper and lower bounds represent 95% confidence intervals computed using three independent sets of simulations.

Seed	Experiment	Mean P	Mean Flux (1/fs)	Mean Rate Constant (1/s)	Mean Test/Control Fold Increase
1	Control	6.7×10^{-23}	1.0×10^{-03}	6.7×10^{-11}	
1	Test	1.4×10^{-08}	4.2×10^{02}	$5.8 \times 10^{+09}$	$8.7 \times 10^{+19}$
2	Control	1.2×10^{-22}	9.0×10^{-04}	1.1×10^{-10}	
2	Test	1.1×10^{-10}	1.2×10^{02}	$1.4 \times 10^{+07}$	$1.3 \times 10^{+17}$
3	Control	2.7×10^{-22}	1.0×10^{-03}	2.7×10^{-10}	
3	Test	3.5×10^{-09}	$9.6 \times 10^{+01}$	$3.4 \times 10^{+08}$	$1.2 \times 10^{+18}$
4	Control	1.6×10^{-22}	7.0×10^{-04}	1.1×10^{-10}	
4	Test	1.0×10^{-09}	$8.7 \times 10^{+01}$	$8.7 \times 10^{+07}$	$7.8 \times 10^{+17}$
5	Control	3.2×10^{-21}	1.3×10^{-03}	4.2×10^{-09}	
5	Test	3.0×10^{-10}	$2.7 \times 10^{+02}$	$8.2 \times 10^{+07}$	$2.0 \times 10^{+16}$

Figure 8: (A) Cumulative log(P) for increasing interface placement for each of the 5 seeds trajectories tested. Red lines indicate trajectories sampled with the reactant basin constrained to only include the region where the 10 feature classifier evaluated to true. Blue lines indicate unconstrained control simulations. (B) Individual values of $P(\lambda_{i+1}|\ \lambda_i)$ for each λ_i ensemble computed. Error bars correspond to two standard errors of the mean across three independent Markov chains at each λ_i ensemble. Red bars indicate test simulations, while blue bars indicate unconstrained control simulations.


A
Seed Trajectory 1
Seed Trajectory 2
Seed Trajectory 3
Seed Trajectory 4
Seed Trajectory 5
log(P_i)
λ_i


B
Seed Trajectory 1
Seed Trajectory 2
Seed Trajectory 3
Seed Trajectory 4
Seed Trajectory 5
P(λ_i+1 | λ_i)
λ_i

Discussion

In this work, we find that features evident in the enzyme–substrate complex before it departs the reactant well are highly predictive of reactivity through the identification of relatively subtle conformational effects. These structural characteristics include internal substrate conformation, interactions of substrate with its environment, and details of the electronic environment of the two magnesium ions that coordinate the substrate. A consensus set of 30 features serve as determinants of reactivity that operate across the pre-launch window, although the detailed roles of some descriptors change across the window.

Interestingly, velocities are not needed to reliably distinguish reactive from non-reactive trajectories. This does not mean that velocities cannot also be useful or important, but only that conformations alone are sufficient. In fact, in preliminary work leading up to this study, we saw that velocities alone, without direct conformational measures, were also sufficient to distinguish reactive from almost-reactive trajectories. This points to the redundancy of the information, and that different descriptions can be equally useful in understanding and predicting reactivity. Interestingly, a more thorough description might be necessary to truly understand reactivity than to predict it. Moreover, although the analysis in the current work appears static, relying on conformations evident at fixed points in time, this may implicitly contain dynamic information. For example, the 0-fs time point corresponds to the maximum compression of the breaking bond before the trajectory launches toward the activation barrier, and so a shorter bond distance, indicating greater potential energy stored in the bond, may signify greater kinetic energy available to surmount the barrier.

This study presents evidence that there are multiple channels of reactivity, some of which are more productive than others. The existence of multiple reactive channels suggests that there

are identifiably different sub-pathways of reaction. Results further suggest that within each channel there could be more ways of not reacting than reacting, consistent with the notion that there are many conditions that must be met in order to produce a reactive trajectory, and failing to achieve any of multiple combinations of those features can be detrimental to reactivity.

This transition interface sampling study highlights the important role that early active-site conformational effects play in driving chemical catalysis, an idea that underlies existing theories of the importance of early conformational effects such electrostatic preorganization (Warshel 1998; Kamerlin et al. 2010) and enzyme-stabilized "near-attack conformations" in certain catalytic systems (Lau and Bruice 1998; Hur and Bruice 2003). That we are able to use machine-learning methods to identify early conformations predictive of reactivity lends additional support to the preorganization and near-attack conformation hypotheses of enzymatic activity, although further research would be necessary to determine whether electrostatic preorganization or stabilization of near-attack conformations is a primary driver of catalysis in the KARI isomerization reaction studied.

A key distinguishing feature between this work and prior studies of near-attack conformations is that we have defined reactivity at time points relative to the temporal progress of the prospective catalytic event rather than purely configurational states (Sadiq and Coveney 2014; Hur and Bruice 2003; Lau and Bruice 1998). Although the sampling constraints during the TIS simulations were enforced at specific time points relative to the progress of the prospective catalytic event, e.g. the "last trough" that we have defined as the 0 time point in the reaction, future work is need to test how critical the time point is on the effectiveness of the constraint in leading to more reactive trajectories. Initial results (unpublished) for set of constrained TIS simulations in which a classifier was learned that was predictive of reactivity across the entire

pre-launch window, suggests that reactive trajectories spend significantly more time in the reactive sub-region of the reactant well than almost-reactive trajectories. This result implies that constraints broadly applied across multiple early time points may be just as effective, if not more effective at enhancing reactivity than constraints applied at one specific time point.

In this work we also show that that path-sampling simulation techniques such as TIS combined with QM/MM simulations, although computationally expensive, can be used to generate large valuable data sets that allow the question of reactivity to be phrased as a binary classification problem well suited for machine learning techniques. We believe this represents both an exciting and promising application of machine learning, but also a productive strategy for elucidating subtle yet meaningful drivers of catalysis in enzymatic systems. While this work utilized features selected through human intuition and a linear classification model (LASSO), the application of unsupervised learning techniques such as auto-encoders to identify perhaps better features combined with non-linear classification models represents an opportunity to understand further the early events that lead to enzymatic catalysis. Finally, although this work utilized TIS to generate only two types of data sets, reactive and almost-reactive, TIS can also be used to generate many more types of data (for example, to generate sets of trajectories that reach progressively higher points along the barrier). Applying the machine learning to trajectory outcomes representing more than two states of reactivity can potentially yield new insights as to precisely when and how reactive and non-reactive trajectories diverge. The TIS probability factor calculation is well suited to this type of analysis.

Although this study identified features indicative of reactivity, an understanding of how those structural and potentially electronic effects cooperate to facilitate the reaction is not

obvious from structures alone. It is possible that more detailed quantum chemical analysis, perhaps with a focus on orbital behaviors, will lend more insight.

The features identified are more than indicators that reaction will likely occur; they are also control levers that alone can guide and enhance reactivity. Our studies demonstrate that enforcing the indicators of reactivity leads to dramatic rate enhancements, largely through increasing the probability of trajectories reaching the product state. This increased probability is exerted across a broad region of the reaction path, past where the constraint is applied, rather than speeding passage through one particularly slow region. That is, by analogy, the constraints act to have an effect more like entering the highway commuter lane to avoid miles of stop-and-go traffic, rather than like taking a shortcut in traffic to avoid one slow intersection.

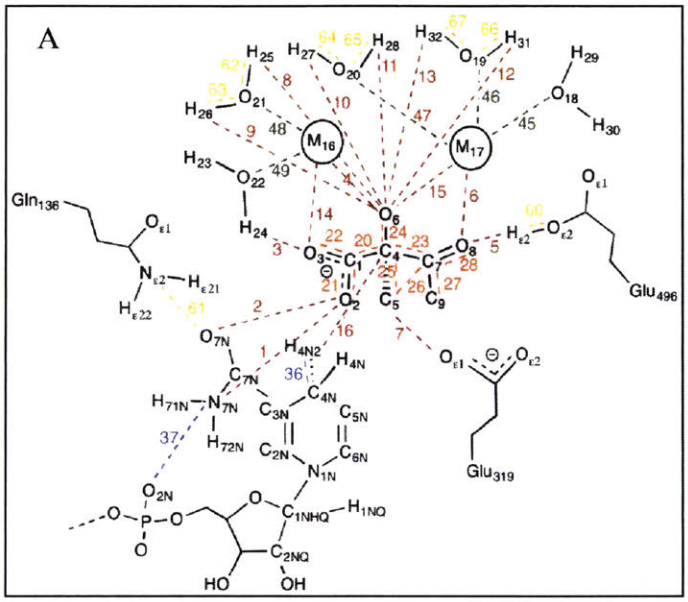
Given the enormous enhancement in reactivity that can be attributed to the selective visitation of early conformations predicted to be reactive, it becomes interesting to contemplate whether mutations can be identified whose predominant effect is to reshape the reactant well so that the more reactive portions (such as those identified by feature constraints here) are more highly populated. If mutations can be found that have this effect with minimal effects elsewhere on the reactive energy surface, they may similarly show useful, measurable rate enhancements. Indeed, in other ways, several recent studies have attempted to leverage insights from path-sampling simulations in order to design enzyme variants (Zoi et al. 2016; Harijan et al. 2017), which represents a promising and novel framework for biocatalyst design.

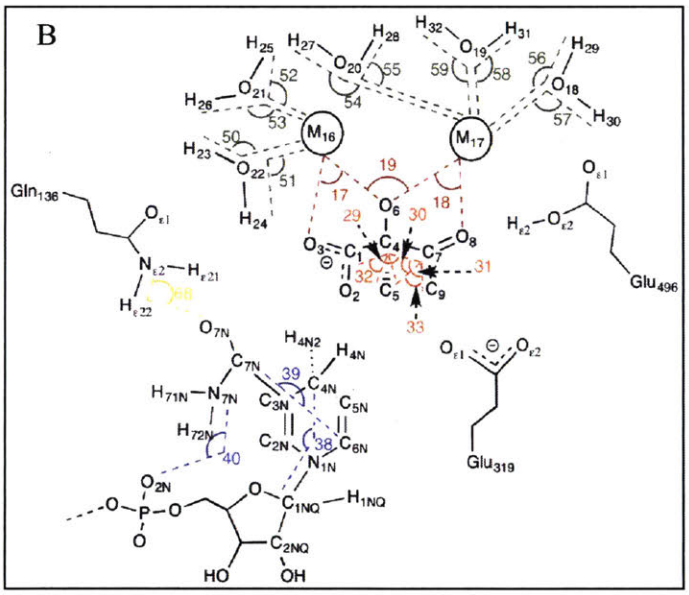
Acknowledgements

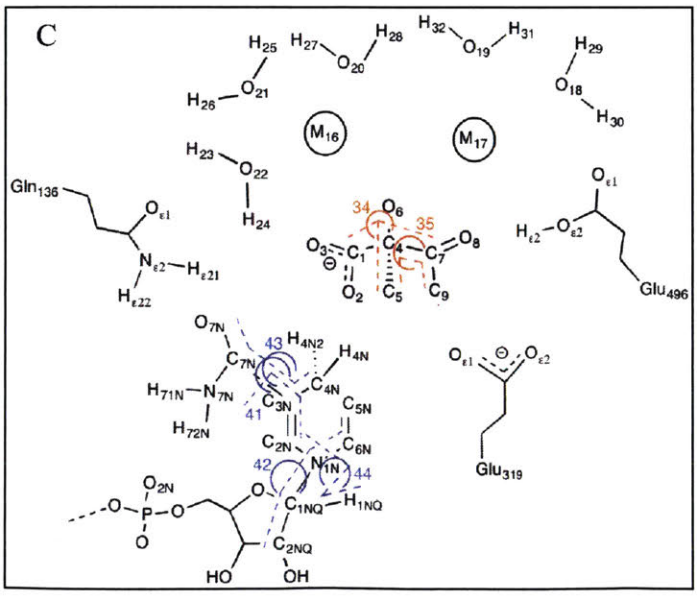
We thank Nathaniel Silver for performing the initial QM fitting and KARI structure preparation, as well as Ishan Patel for providing the starting MATLAB code base for performing WHAM and TIS calculations. Valuable conversations with Catherine Gibson, Mark Nelson, Daniel O'Keefe, Nathaniel Silver, and members of our research group are gratefully acknowledged. This work was supported by awards from the NDSEG Fellowship program (to BMB and JWW) and National Institute of General Medical Sciences of the US National Institutes of Health (R01 GM082209 and R01 GM065418 to BT).

Supplementary Figures

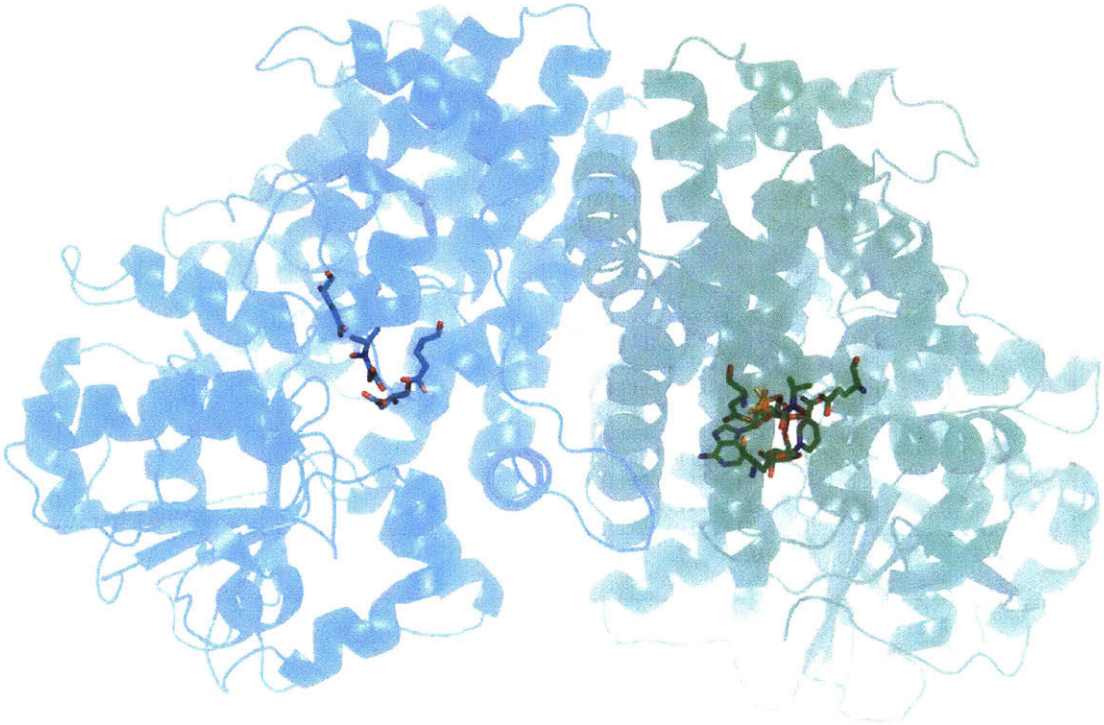
Supplementary Figure 1: Structural representation of (A) distances computed, (B) angles computed, and (C) dihedrals computed at each time point. Numbering of features corresponds to that of Table 1. Coloring of features corresponds to the feature type with red indicating substrate-environment interactions, orange indicating intra-substrate conformations, blue indicating intra-cofactor conformations, green indicating water-metal interactions and gold indicating other environment interactions.


A
M16
M17
Gln136
Glu496
Glu319

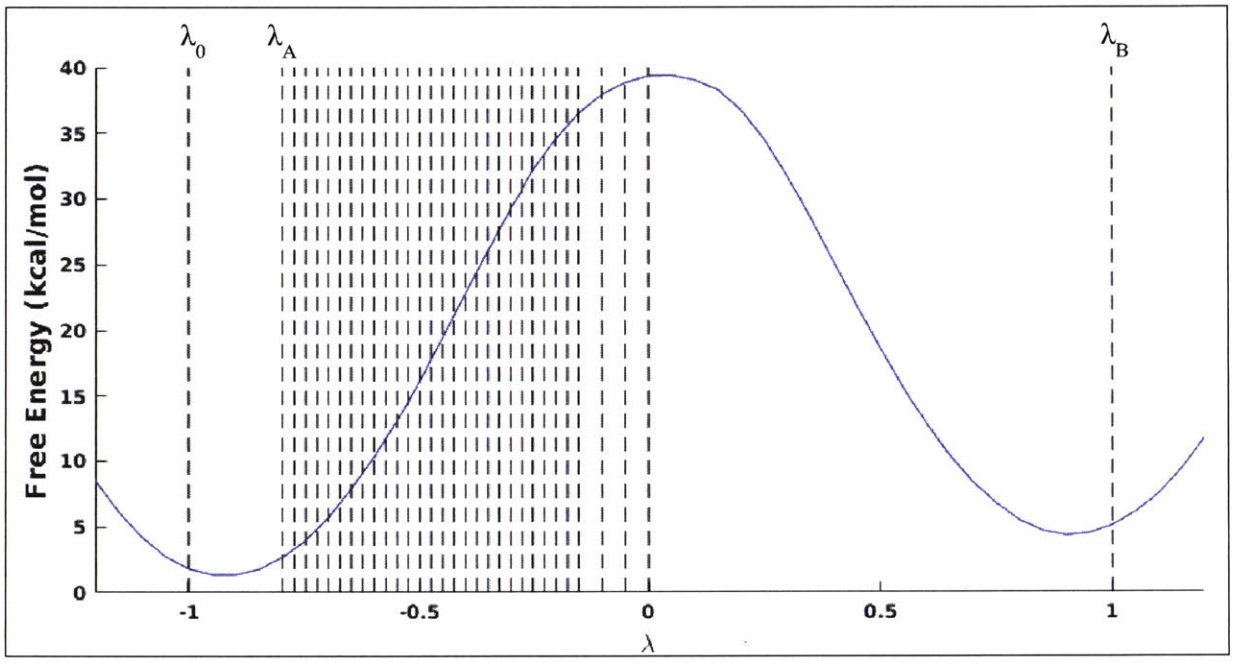

B
M16
M17
Gln136
Glu496
Glu319


C
M16
M17
Gln136
Glu496
Glu319

Supplementary Figure 2: Illustration of both KARI homodimer subunits (PDB ID: 1YVE), with active site residues Asp 315, Glu 319, Glu 496, bound transition state analog N-hydroxy-N-isopropyloxamate and NADPH cofactor shown as sticks to indicate active-site separation and to support the choice of using a single subunit in simulations.



Supplementary Figure 3: Placement of interfaces used in TIS probability factor calculations superimposed onto the potential of mean force surface used to generate initial seed trajectories. Key interfaces λ_0=-1, λ_A = -0.8 and λ_B = 1 are labeled.



Chapter 4 : General Conclusions

This thesis has presented a set of novel methods and applications for the computer-aided design and understanding of enzymatic catalysis. This work has applied rigorous theoretical and simulation methods to a set of challenging problems in biocatalyst modeling and design. The two enzymes studied in this thesis both represent industrially useful systems with potential commercial applications in sustainable chemical and biofuel production.

Chapter 2 presented a combined biophysical modeling and experimental study of substrate selectivity in a key enzymatic step (the thiolase-catalyzed condensation of two different acyl-CoA substrates) of an industrially useful *de novo* metabolic platform pathway, the 3-hydroxyacid pathway (Martin et al. 2013). The objective of the study presented in Chapter 2 was to identify thiolase mutations capable of enhancing C6 product production *in vivo* relative to C4 product formation. Although numerous previous studies have characterized thiolase enzymes (Slater et al. 1998; Fage, Meinke, and Keatinge-Clay 2015; Kim, Clomburg, and Gonzalez 2015) for metabolic engineering and other applications, this work is one of few attempts to rationally design thiolase specificity. The two-step bi-bi ping pong mechanism of the thiolases (Modis and Wierenga 1999) studied presented a number of modeling challenges in our effort to identify mutations capable of enhancing the relative C6 / C4 output of the 3-hydroxyacid pathway, and we found that we were able to increase this ratio primarily engineering a decrease in the C4 product output.

Future work is needed to experimentally validate whether the mutations discussed that decreased C4 product production also increased C6 production as predicted by the biophysical calculations. A significant challenge in the thiolase engineering study presented in Chapter 2 was

the fact that the immediate C6 product resulting from the thiolase catalyzed condensation of butyryl-CoA and acetyl-CoA could not be assayed directly *in vivo* or *in vitro*. Because the thermodynamics of thiolase binding heavily favor the thiolysis reaction (the reverse of the condensation reaction under study), in both the *in vivo* and *in vitro* studies, the thiolase enzyme had to be coupled with a downstream reductase to force the reaction equilibrium in the condensation direction. Accordingly, all experimental measurements of thiolase activity in the forward condensation direction were potentially biased by the specificity of downstream enzymes. By providing the cofactor required by this reductase in excess concentrations, *in vitro* studies could ensure that the overall forward rate of the reductase was not rate-limiting in the assays, but relative rate of production of the product produced the condensation of two acetyl-CoA substrates compared to the product of butyryl-CoA serving as the priming acyl-CoA substrate followed by the acetyl-CoA serving as the extending acyl-CoA substrate was impossible to measure directly. Experimentally, the assay readout for the thiolase forward condensation rate was the rate of cofactor oxidation by the reductase, which meant that only limited information could be inferred about the relative rate of production of C6 and C4 products *in vitro*. The *in vivo* picture was even more complicated, as the final experimental readout was the affected by a significantly greater number of downstream enzymes such as the thioesterases and polymerases, for which none of the C6 / C4 specificities were known at the beginning of the study. Although the results of the thiolase study were promising in that the relative C6 / C4 *in vivo* was increased, the results show that significantly more work is required to eliminate confounding factors presented by downstream enzymes and accurately measure the level of increase that can actually be attributed to the thiolase mutations. A combination of further experimental work and kinetic modeling studies may prove fruitful for such future endeavors.

The effect of multiple active site thiolase mutations beyond the point mutations tested in Chapter 2 may also lead to thiolase variants with even greater C6 / C4 activity. Ultimately however, metabolic engineering approaches for chemical and fuel production represent an attractive synthesis route less reliant on non-renewable feedstocks and studies like that presented in Chapter 2 represent a step toward making such approaches, particularly using the versatile 3-hydroxyacid pathway and other pathways which utilize thiolase enzymes to facilitate carbon-carbon bond formation, more feasible commercially.

In contrast to Chapter 2, Chapter 3 presented a purely theoretical and simulation study, and applied methods from machine learning, in particular LASSO, to a large ensemble of reactive and non-reactive molecular dynamics trajectories generated using transition interface sampling in order to elucidate catalytic drivers in another industrially important model enzyme system, ketol-acid reductoisomerase (KARI). Although machine learning has been applied with with great success to a number of problems in structural biology such as prediction of protein structure, protein folding pathways, protein-ligand binding affinities and drug design (Wallach, Dzamba, and Heifets 2015; Radivojac et al. 2013; Wu et al. 2017; Ramsundar and Pande 2016), few previous studies have applied machine learning approaches to ensembles of reactive trajectories harvested using path sampling methods such as transition path sampling (Zhang et al. 2017; Antoniou and Schwartz 2011), an application we and others believe represents a promising approach to elucidate the subtle mechanisms by which enzymes can achieve such enormous rate enhancements over uncatalyzed reactions as well as improve existing frameworks for biocatalyst design. Chapter 3 explored the central idea that underlies electrostatic preorganization (Kamerlin et al. 2010; Warshel 1998) and near attack conformation theories of reactivity (Hur and Bruice 2003; Sadiq and Coveney 2014) – that certain subsets of phase space are inherently more

reactive than others, and that successful reactive trajectories selectively visit these inherently reactive regions of phase space.

In the first part of Chapter 3, using a small number of molecular features, we used a logistic classifier to identify conformational states that are highly predictive of reactivity, which represent examples of such inherently reactive regions, at specific time points relative to the progress of the prospective catalytic event. The specific features learned, although not intended to be a complete and exhaustive set, provided mechanistic insight into the rate-limiting isomerization catalyzed by KARI. The specific features learned by LASSO and described by the logistic classifier in particular highlight the importance of the compression of the substrate breaking bond, the orientation of active site water molecules relative to the substrate and active site metal ions and the subtle positioning of key side chain residues. Our results lend evidence to the near-attack conformation theory of enzyme catalysis in the KARI system, and underscore the importance of the dynamics and timing at which these reactive states are visited.

In the second part of Chapter 3, we then presented a novel theoretical framework based on transition interface sapling for evaluating the contribution to the overall catalytic rate of the conformational states found to be highly predictive of reactivity. We showed that ensembles of trajectories sampled in such a manner as to selectively visit the conformations predicted to be characteristic of reactivity exhibit rate constants many orders of magnitude greater than trajectories not required to visit these reactive conformations. The results show the enormous extent to which early conformational effects can define reactivity in the model system studied.

Future studies in the same vein as Chapter 3, utilizing different model systems, better quantum mechanical and solvation models and more sophisticated machine learning, can likely provide greater insight into the mechanisms governing enzyme catalysis. Although the results of

Chapter 3 are promising, their generality is somewhat limited by relatively short sampling times and limiting assumptions of the specific semi-empirical quantum mechanics / molecular mechanics simulation approach utilized. As computing power continues to grow, data storage becomes cheaper, and force fields continue to improve, combined machine learning and transition path sampling studies such as that presented in Chapter 3 will likely continue to yield greater insights into mechanism of reactive catalysis. We and others (van Erp et al. 2016; Moqadam et al. 2017; Zoi et al. 2016) believe that path sampling studies have great continued potential to test specific hypotheses about reactivity, analyze subtly different reaction channels, and be used as part of design programs to identify catalytic variants which bias reactions in a favorable manner.

References

Agnew, Daniel E., and Brian F. Pfleger. 2013. "Synthetic Biology Strategies for Synthesizing Polyhydroxyalkanoates from Unrelated Carbon Sources." *Chemical Engineering Science* 103: 58–67. doi:10.1016/j.ces.2012.12.023.

Aldor, I. S., S.-W. Kim, K. L. J. Prather, and J. D. Keasling. 2002. "Metabolic Engineering of a Novel Propionate-Independent Pathway for the Production of Poly(3-Hydroxybutyrate-Co-3-Hydroxyvalerate) in Recombinant Salmonella Enterica Serovar Typhimurium." *Applied and Environmental Microbiology* 68 (8): 3848–3854. doi:10.1128/AEM.68.8.3848-3854.2002.

Aleksandrov, Alexey, and Martin Field. 2011. "Efficient Solvent Boundary Potential for Hybrid Potential Simulations." *Physical Chemistry Chemical Physics : PCCP* 13 (22): 10503–9. doi:10.1039/c0cp02828b.

Antoniou, Dimitri, and Steven D. Schwartz. 2011. "Towards Identification of the Reaction Coordinate Directly from the Transition State Ensemble Using the Kernel PCA Method." *The Journal of Physical Chemistry. B* 115 (10): 2465–69. doi:10.1021/jp111682x.

Baker, David. 2010. "An Exciting but Challenging Road Ahead for Computational Enzyme Design." *Protein Science : A Publication of the Protein Society* 19 (10): 1817–9. doi:10.1002/pro.481.

Basner, Jodi E, and Steven D Schwartz. 2005. "How Enzyme Dynamics Helps Catalyze a Reaction in Atomic Detail: A Transition Path Sampling Study." *Journal of the American Chemical Society* 127 (40): 13822–31. doi:10.1021/ja043320h.

Bastian, Sabine, Xiang Liu, Joseph T Meyerowitz, Christopher D Snow, Mike MY Chen, and Frances H Arnold. 2011. "Engineered Ketol-Acid Reductoisomerase and Alcohol Dehydrogenase Enable

Anaerobic 2-Methylpropan-1-Ol Production at Theoretical Yield in Escherichia Coli." *Metabolic Engineering* 13 (3): 345–352.

Bayly, Christopher I., Piotr Cieplak, Wendy Cornell, and Peter A. Kollman. 1993. "A Well-Behaved Electrostatic Potential Based Method Using Charge Restraints for Deriving Atomic Charges: The RESP Model." *The Journal of Physical Chemistry* 97 (40): 10269–10280. doi:10.1021/j100142a004.

Berman, Helen M., John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. 2000. "The Protein Data Bank." *Nucleic Acids Research* 28 (1): 235–42.

Bernstein, F.C., T.F. Koetzle, G. J. Williams, E.E. Jr. Meyer, M.D. Brice, J. R. Rodgers, O Kennard, T. Shimanouchi, and M Tasumi. 1977. "The Protein Data Bank: A Computer-Based Archival File for Macromolecular Structures." *Journal of Molecular Biology* 112: 535.

Biou, V, R Dumas, C Cohen-Addad, R Douce, D Job, and E Pebay-Peyroula. 1997. "The Crystal Structure of Plant Acetohydroxy Acid Isomeroreductase Complexed with NADPH, Two Magnesium Ions and a Herbicidal Transition State Analog Determined at 1.65 A Resolution." *The EMBO Journal* 16 (12): 3405–15. doi:10.1093/emboj/16.12.3405.

Bolhuis, P. G., and C. Dellago. 2015. "Practical and Conceptual Path Sampling Issues." *The European Physical Journal Special Topics* 224 (12): 2409–27. doi:10.1140/epjst/e2015-02419-6.

Bolhuis, Peter G., David Chandler, Christoph Dellago, and Phillip L. Geissler. 2002. "Transition Path Sampling : Throwing Ropes Over Rough Mountain Passes, in the Dark." *Annual Review of Physical Chemistry* 53 (1): 291–318. doi:10.1146/annurev.physchem.53.082301.113146.

Bolhuis, Phillip L., Peter G.Chandler, DavidDellago, ChristophGeissler. n.d. "TRANSITION PATH SAMPLING: Throwing Ropes Over Rough Mountain Passes, in the Dark." *Annual Review of Physical Chemistry. 2002* 53 (1): 291. 28p. 1 Diagram.

Bond-Watts, Brooks B, Robert J Bellerose, and Michelle C Y Chang. 2011. "Enzyme Mechanism as a Kinetic Control Element for Designing Synthetic Biofuel Pathways." *Nature Chemical Biology* 7 (4): 222–7. doi:10.1038/nchembio.537.

Borrero, Ernesto E., Marcus Weinwurm, and Christoph Dellago. 2011. "Optimizing Transition Interface Sampling Simulations." *The Journal of Chemical Physics* 134 (24): 244118. doi:10.1063/1.3601919.

Brandl, H, R a Gross, R W Lenz, and R C Fuller. 1988. "Pseudomonas Oleovorans as a Source for Novel Poly(Beta-Hydroxyalkanoates)." *Applied and Environmental Microbiology* 54 (8): 1977–1982.

Brooks, Bernard R., Robert E. Bruccoleri, Barry D. Olafson, David J. States, S. Swaminathan, and Martin Karplus. 1983. "CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations." *Journal of Computational Chemistry* 4 (2):187–217. https://doi.org/10.1002/jcc.540040211.

Brooks, B R, C L Brooks, A D Mackerell, L Nilsson, R J Petrella, B Roux, Y Won, et al. 2009. "CHARMM: The Biomolecular Simulation Program." *Journal of Computational Chemistry* 30 (10): 1545–614. doi:10.1002/jcc.21287.

Bruice, Thomas C. 2002. "A View at the Millennium: The Efficiency of Enzymatic Catalysis." *Accounts of Chemical Research* 35 (3): 139–48. doi:10.1021/ar0001665.

Bruice, Thomas C, and Felice C Lightstone. 1999. "Ground State and Transition State Contributions to the Rates of Intramolecular and Enzymatic Reactions." *Accounts of Chemical Research* 32 (2): 127–136.

Burton, S.G., Cowan, D.A., Woodley, J.M. 2002. "The Search for the Ideal Biocatalyst." *Nature Biotechnology* 20 (6): 36–45. doi:10.1517/14656566.3.6.681.

Cahn, Jackson KB, Sabine Brinkmann-Chen, Thomas Spatzal, Jared A Wiig, Andrew R Buller, Oliver Einsle, Yilin Hu, Markus W Ribbe, and Frances H Arnold. 2015. "Cofactor Specificity Motifs and the Induced Fit Mechanism in Class I Ketol-Acid Reductoisomerases." *Biochemical Journal* 468 (3): 475–484.

Cameron, D Ewen, Caleb J Bashor, and James J Collins. 2014. "A Brief History of Synthetic Biology." *Nature Reviews Microbiology* 12 (5): 381–390.

Carter, E. A., Giovanni Ciccotti, James T. Hynes, and Raymond Kapral. 1989. "Constrained Reaction Coordinate Dynamics for the Simulation of Rare Events." *Chemical Physics Letters* 156 (5): 472–77. doi:10.1016/S0009-2614(89)87314-2.

Chandler, David. 1978. "Statistical Mechanics of Isomerization Dynamics in Liquids and the Transition State Approximation" 68 (6): 2959.

Chen, Chang-Ting, and James C Liao. 2016. "Frontiers in Microbial 1-Butanol and Isobutanol Production." *FEMS Microbiology Letters* 363 (5): fnw020.

Cheong, Seokjung, James M. Clomburg, and Ramon Gonzalez. 2016. "Energy- and Carbon-Efficient Synthesis of Functionalized Small Molecules in Bacteria Using Non-Decarboxylative Claisen Condensation Reactions." *Nature Biotechnology* In press (April): 1–8. doi:10.1038/nbt.3505.

Chunduru, S. K., G. T. Mrachko, and K. C. Calvo. 1989. "Mechanism of Ketol Acid Reductoisomerase--Steady-State Analysis and Metal Ion Requirement." *Biochemistry* 28 (2): 486–93.

Church, George M, Michael B Elowitz, Christina D Smolke, Christopher A Voigt, and Ron Weiss. 2014. "Realizing the Potential of Synthetic Biology." *Nature Reviews. Molecular Cell Biology* 15 (4): 289.

Clomburg, James M., Matthew D. Blankschien, Jacob E. Vick, Alexander Chou, Seohyoung Kim, and Ramon Gonzalez. 2015. "Integrated Engineering of β-Oxidation Reversal and $ømega$-Oxidation Pathways for the Synthesis of Medium Chain $ømega$-Functionalized Carboxylic Acids." *Metabolic Engineering* 28: 202–212. doi:10.1016/j.ymben.2015.01.007.

Clomburg, James M., Jacob E. Vick, Matthew D. Blankschien, María Rodríguez-Moyá, and Ramon Gonzalez. 2012. "A Synthetic Biology Approach to Engineer a Functional Reversal of the β-Oxidation Cycle." *ACS Synthetic Biology* 1 (11): 541–554. doi:10.1021/sb3000782.

Crehuet, Ramon, and Martin J Field. 2007. "A Transition Path Sampling Study of the Reaction Catalyzed by the Enzyme Chorismate Mutase." *The Journal of Physical Chemistry. B* 111 (20): 5708–18. doi:10.1021/jp067629u.

Dametto, Mariangela, Dimitri Antoniou, and Steven D. Schwartz. 2012. "Barrier Crossing in Dihydrofolate Reductasedoes Not Involve a Rate-Promoting Vibration." *Molecular Physics* 110 (9–10): 531–36. doi:10.1080/00268976.2012.655337.

Dellago, Christoph, Peter G. Bolhuis, Félix S. Csajka, and David Chandler. 1998. "Transition Path Sampling and the Calculation of Rate Constants." *The Journal of Chemical Physics* 108 (5): 1964. doi:10.1063/1.475562.

Dellago, Christoph, Peter G. Bolhuis, and Phillip L. Geissler. 2002. "Transition Path Sampling." In *Advances in Chemical Physics*, edited by I. Prigogine and Stuart A. Rice, 1–78. John Wiley & Sons, Inc. doi:10.1002/0471231509.ch1.

Dhamankar, Himanshu, Yekaterina Tarasova, Collin H. Martin, and Kristala L J Prather. 2014. "Engineering E. Coli for the Biosynthesis of 3-Hydroxy-γ-Butyrolactone (3HBL) and 3,4-Dihydroxybutyric Acid (3,4-DHBA) as Value-Added Chemicals from Glucose as a Sole Carbon Source." *Metabolic Engineering* 25: 72–81. doi:10.1016/j.ymben.2014.06.004.

Dewar, Michael J. S., Eve G. Zoebisch, Eamonn F. Healy, and James J. P. Stewart. 1985. "Development and Use of Quantum Mechanical Molecular Models. 76. AM1: A New General Purpose Quantum Mechanical Molecular Model." *Journal of the American Chemical Society* 107 (13):3902–9. https://doi.org/10.1021/ja00299a024.

Dumas, R., D. Job, J. Y. Ortholand, G. Emeric, A. Greiner, and R. Douce. 1992. "Isolation and Kinetic Properties of Acetohydroxy Acid Isomeroreductase from Spinach (Spinacia Oleracea) Chloroplasts Overexpressed in Escherichia Coli." *The Biochemical Journal* 288 (Pt 3) (December): 865–74.

Dumas, Renaud, Valérie Biou, Frédéric Halgand, Roland Douce, and Ronald G. Duggleby. 2001. "Enzymology, Structure, and Dynamics of Acetohydroxy Acid Isomeroreductase." *Accounts of Chemical Research* 34 (5): 399–408. doi:10.1021/ar000082w.

Dunbrack, R. L., and F. E. Cohen. 1997. "Bayesian Statistical Analysis of Protein Side-Chain Rotamer Preferences." *Protein Science* 6 (8): 1661–81. doi:10.1002/pro.5560060807.

Dunn, Briana J, Katharine R Watts, Thomas Robbins, David E Cane, and Chaitan Khosla. 2014. "Comparative Analysis of the Substrate Specificity of Trans- versus Cis-Acyltransferases of Assembly Line Polyketide Synthases." *Biochemistry* 53 (23): 3796–806. doi:10.1021/bi5004316.

Eastman, P., N. Gronbech-Jensen, and S. Doniach. 2001. "Simulation of Protein Folding by Reaction Path Annealing." *Journal of Chemical Physics* 114 (8): 3823–41. doi:10.1063/1.1342162.

Elber, R., and M. Karplus. 1987. "A Method for Determining Reaction Paths in Large Molecules - Application." *Chemical Physics Letters* 139 (5): 375–80. doi:10.1016/0009-2614(87)80576-6.

Erickson, Brent, Nelson, and Paul Winters. 2012. "Perspective on Opportunities in Industrial Biotechnology in Renewable Chemicals." *Biotechnology Journal* 7 (2): 176–85. doi:10.1002/biot.201100069.

Erp, Titus S. van, and Peter G. Bolhuis. 2005. "Elaborating Transition Interface Sampling Methods." *Journal of Computational Physics* 205 (1): 157–181. doi:10.1016/j.jcp.2004.11.003.

Erp, Titus S. van, Mahmoud Moqadam, Enrico Riccardi, and Anders Lervik. 2016. "Analyzing Complex Reaction Mechanisms Using Path Sampling." *Journal of Chemical Theory and Computation* 12 (11): 5398–5410. doi:10.1021/acs.jctc.6b00642.

Erp, Titus S. van, Daniele Moroni, and Peter G Bolhuis. 2003. "A Novel Path Sampling Method for the Calculation of Rate Constants" 118: 7762–74.

Eyring, Henry, and Stearn. 1939. "The Application of the Theory of Absolute Reacton Rates to Proteins." 24 (2): 253–70.

Fage, Christopher D., Jessica L. Meinke, and Adrian T. Keatinge-Clay. 2015. "Coenzyme A-Free Activity, Crystal Structure, and Rational Engineering of a Promiscuous Beta-Ketoacyl Thiolase from Ralstonia Eutropha." *Journal of Molecular Catalysis B: Enzymatic* 121: 113–121. doi:10.1016/j.molcatb.2015.08.007.

Fisher, Amanda K., Benjamin G. Freedman, David R. Bevan, and Ryan S. Senger. 2014. "A Review of Metabolic and Enzymatic Engineering Strategies for Designing and Optimizing Performance

of Microbial Cell Factories." *Computational and Structural Biotechnology Journal* 11 (18): 91–99. doi:10.1016/j.csbj.2014.08.010.

Frisch, M. J., G. W. Trucks, H. B. Schlegel, M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, et al. 2003. *Gaussian 03: Revision B.05*. Pittsburgh, PA: Gaussian Inc. https://www.scienceopen.com/document?vid=2d5bdd72-8930-4f27-97f7-25a88aefdaa3.

Gao, J, Patricia Amara, Cristobal Alhambra, and Martin Field. 1998. "A Generalized Hybrid Orbital (GHO) Method for the Treatment of Boundary Atoms in Combined QM/MM Calculations - The Journal of Physical Chemistry A (ACS Publications)." *J. Phys. Chem. A* 102: 4714–21.

Garcia-Viloca, Mireia, Jiali Gao, Martin Karplus, and Donald G. Truhlar. 2004. "How Enzymes Work: Analysis by Modern Rate Theory and Computer Simulations." *Science* 303 (5655): 186–95. doi:10.1126/science.1088172.

Gawehn, Erik, Jan A. Hiss, and Gisbert Schneider. 2016. "Deep Learning in Drug Discovery." *Molecular Informatics* 35 (1): 3–14. doi:10.1002/minf.201501008.

Glowacki, David R, Jeremy N Harvey, and Adrian J Mulholland. 2012. "Taking Ockham's Razor to Enzyme Dynamics and Catalysis." *Nature Chemistry* 4 (3): 169–76. doi:10.1038/nchem.1244.

Guo, Weihua, Jiayuan Sheng, and Xueyang Feng. 2017. "Mini-Review: In Vitro Metabolic Engineering for Biomanufacturing of High-Value Products." *Computational and Structural Biotechnology Journal* 15 (January): 161–67. doi:10.1016/j.csbj.2017.01.006.

Haapalainen, Antti M., Gitte Meriläinen, and Rik K. Wierenga. 2006. "The Thiolase Superfamily: Condensing Enzymes with Diverse Reaction Specificities." *Trends in Biochemical Sciences* 31 (1): 64–71.

Harijan, Rajesh K., Ioanna Zoi, Dimitri Antoniou, Steven D. Schwartz, and Vern L. Schramm. 2017. "Catalytic-Site Design for Inverse Heavy-Enzyme Isotope Effects in Human Purine Nucleoside

Phosphorylase." *Proceedings of the National Academy of Sciences* 114 (25): 6456–61. doi:10.1073/pnas.1704786114.

Hilvert, Donald. 2013. "Design of Protein Catalysts." *Annual Review of Biochemistry* 82 (January): 447–70. doi:10.1146/annurev-biochem-072611-101825.

Holland, J. Todd, Jason C. Harper, Patricia L. Dolan, Monica M. Manginell, Dulce C. Arango, Julia A. Rawlings, Christopher A. Apblett, and Susan M. Brozik. 2012. "Rational Redesign of Glucose Oxidase for Improved Catalytic Function and Stability." *PLOS ONE* 7 (6): e37924. doi:10.1371/journal.pone.0037924.

Houk, K. N., and Fang Liu. 2017. "Holy Grails for Computational Organic Chemistry and Biochemistry." *Accounts of Chemical Research* 50 (3): 539–43. doi:10.1021/acs.accounts.6b00532.

Huang, Jing, and Alexander D. MacKerell. 2013. "CHARMM36 All-Atom Additive Protein Force Field: Validation Based on Comparison to NMR Data." *Journal of Computational Chemistry* 34 (25): 2135–45. doi:10.1002/jcc.23354.

Hummer, Gerhard. 2010. "Catching a Protein in the Act." *Proceedings of the National Academy of Sciences of the United States of America* 107 (6): 2381–2. doi:10.1073/pnas.0914486107.

Hur, Sun, and Thomas C. Bruice. 2003. "The near Attack Conformation Approach to the Study of the Chorismate to Prephenate Reaction." *Proceedings of the National Academy of Sciences* 100 (21): 12015–20. doi:10.1073/pnas.1534873100.

Jemli, Sonia, Dorra Ayadi-Zouari, Hajer Ben Hlima, and Samir Bejar. 2016. "Biocatalysts: Application and Engineering for Industrial Purposes." *Critical Reviews in Biotechnology* 36 (2): 246–258.

Kamerlin, Shina C L, Pankaz K Sharma, Zhen T Chu, and Arieh Warshel. 2010. "Ketosteroid Isomerase Provides Further Support for the Idea That Enzymes Work by Electrostatic Preorganization." *Proceedings of the National Academy of Sciences of the United States of America* 107 (9): 4075–80. doi:10.1073/pnas.0914579107.

Kamerlin, Shina C. L., and Arieh Warshel. 2010. "At the Dawn of the 21st Century: Is Dynamics the Missing Link for Understanding Enzyme Catalysis?" *Proteins* 78 (6): 1339–75. doi:10.1002/prot.22654.

Keasling, Jay D. 2009. "Manufacturing Molecules Through Metabolic Engineering." *Science* 323 (January): 1355–1359. doi:10.1126/science.1226338.

Khersonsky, Olga, Daniela Röthlisberger, Andrew M. Wollacott, Paul Murphy, Orly Dym, Shira Albeck, Gert Kiss, K.N. Houk, David Baker, and Dan S. Tawfik. 2011. "Optimization of the In-Silico-Designed Kemp Eliminase KE70 by Computational Design and Directed Evolution." *Journal of Molecular Biology* 407 (3): 391–412.

Kim, Eun Jung, Hyeoncheol Francis Son, Sangwoo Kim, Jae Woo Ahn, and Kyung Jin Kim. 2014. "Crystal Structure and Biochemical Characterization of Beta-Keto Thiolase B from Polyhydroxyalkanoate-Producing Bacterium Ralstonia Eutropha H16." *Biochemical and Biophysical Research Communications* 444 (3): 365–369. doi:10.1016/j.bbrc.2014.01.055.

Kim, Seohyoung, James M. Clomburg, and Ramon Gonzalez. 2015. "Synthesis of Medium-Chain Length (C6-C10) Fuels and Chemicals via β-Oxidation Reversal in Escherichia Coli." *Journal of Industrial Microbiology and Biotechnology* 42 (3): 465–475. doi:10.1007/s10295-015-1589-6.

Kiss, Gert, Nihan Çelebi-Ölçüm, Rocco Moretti, David Baker, and K. N. Houk. 2013. "Computational Enzyme Design." *Angewandte Chemie International Edition* 52 (22): 5700–5725. doi:10.1002/anie.201204077.

Kiss, Gert, Daniela Röthlisberger, David Baker, and K N Houk. 2010. "Evaluation and Ranking of Enzyme Designs." *Protein Science : A Publication of the Protein Society* 19 (9): 1760–73. doi:10.1002/pro.462.

Kursula, Petri, Juha Ojala, Anne-Marie Lambeir, and Rik K. Wierenga. 2002. "The Catalytic Cycle of Biosynthetic Thiolase: A Conformational Journey of an Acetyl Group through Four Binding Modes and Two Oxyanion Holes‡." *Biochemistry* 41 (52): 15543–15556.

Laio, Alessandro, and Michele Parrinello. 2002. "Escaping Free-Energy Minima." *Proceedings of the National Academy of Sciences* 99 (20): 12562–66. doi:10.1073/pnas.202427399.

Laprevote, O., L. Serani, B. C. Das, F. Halgand, E. Forest, and R Dumas. 1999. "Stepwise Building of a 115-KDa Macromolecular Edifice Monitored by Electrospray Mass Spectrometry. The Case of Acetohydroxy Acid Isomeroreductase" 256. https://www.ncbi.nlm.nih.gov/pubmed/9914514.

Lau, Edmond Y, and Thomas C Bruice. 1998. "Importance of Correlated Motions in Forming Highly Reactive near Attack Conformations in Catechol O-Methyltransferase." *Journal of the American Chemical Society* 120 (48): 12387–12394.

Laube, B., S. Winkler, B. Ladstetter, T. Scheller, and L.R. Schwarz. 2000. "Establishment of a Novel in Vitro System for Studying the Interaction of Xenobiotic Metabolism of Liver and Intestinal Microflora." *Archives of Toxicology* 74 (7): 379–387. doi:10.1007/s002040000137.

Lee, Jeong Wook, Dokyun Na, Jong Myoung Park, Joungmin Lee, Sol Choi, and Sang Yup Lee. 2012. "Systems Metabolic Engineering of Microorganisms for Natural and Non-Natural Chemicals." *Nature Chemical Biology* 8 (6): 536–46. doi:10.1038/nchembio.970.

Li, Xiaosong, and Michael J. Frisch. 2006. "Energy-Represented Direct Inversion in the Iterative Subspace within a Hybrid Geometry Optimization Method." *Journal of Chemical Theory and Computation* 2 (3): 835–39. doi:10.1021/ct050275a.

Linder, Mats, Adam Johannes Johansson, Tjelvar S G Olsson, John Liebeschuetz, and Tore Brinck. 2012. "Computational Design of a Diels-Alderase from a Thermophilic Esterase: The Importance of Dynamics." *Journal of Computer-Aided Molecular Design* 26 (9): 1079–95. doi:10.1007/s10822-012-9601-y.

Lippow, Shaun M., and Bruce Tidor. 2007. "Progress in Computational Protein Design." *Current Opinion in Biotechnology* 18 (4): 305–11. doi:10.1016/j.copbio.2007.04.009.

Lippow, Shaun M, K Dane Wittrup, and Bruce Tidor. 2007. "Computational Design of Antibody-Affinity Improvement beyond in Vivo Maturation." *Nature Biotechnology* 25 (10): 1171–6. doi:10.1038/nbt1336.

Machado, Hidevaldo B., Yasumasa Dekishima, Hao Luo, Ethan I. Lan, and James C. Liao. 2012. "A Selection Platform for Carbon Chain Elongation Using the CoA-Dependent Pathway to Produce Linear Higher Alcohols." *Metabolic Engineering* 14 (5): 504–511. doi:10.1016/j.ymben.2012.07.002.

Mann, Miriam S, and Tina Lütke-Eversloh. 2012. "Thiolase Engineering for Enhanced Butanol Production in Clostridium Acetobutylicum." *Biotechnology and Bioengineering* xxx (xxx): 1–11. doi:10.1002/bit.24758.

Martin, Collin H, Himanshu Dhamankar, Hsien-Chung Tseng, Micah J Sheppard, Christopher R Reisch, and Kristala L J Prather. 2013. "A Platform Pathway for Production of 3-Hydroxyacids Provides a Biosynthetic Route to 3-Hydroxy-γ-Butyrolactone." *Nature Communications* 4: 1414. doi:10.1038/ncomms2418.

McMahon, Matthew D., and Kristala L J Prather. 2014. "Functional Screening and in Vitro Analysis Reveal Thioesterases with Enhanced Substrate Specificity Profiles That Improve Short-Chain

Fatty Acid Production in Escherichia Coli." *Applied and Environmental Microbiology* 80 (3): 1042–1050. doi:10.1128/AEM.03303-13.

Meriläinen, Gitte, Visa Poikela, Petri Kursula, and Rik K Wierenga. 2009. "The Thiolase Reaction Mechanism: The Importance of Asn316 and His348 for Stabilizing the Enolate Intermediate of the Claisen Condensation." *Biochemistry* 48 (46): 11011–25. doi:10.1021/bi901069h.

Meriläinen, Gitte, Werner Schmitz, Rik K Wierenga, and Petri Kursula. 2008. "The Sulfur Atoms of the Substrate CoA and the Catalytic Cysteine Are Required for a Productive Mode of Substrate Binding in Bacterial Biosynthetic Thiolase, a Thioester-Dependent Enzyme." *The FEBS Journal* 275 (24): 6136–48. doi:10.1111/j.1742-4658.2008.06737.x.

Modis, Y, and R K Wierenga. 2000. "Crystallographic Analysis of the Reaction Pathway of Zoogloea Ramigera Biosynthetic Thiolase." *Journal of Molecular Biology* 297 (5): 1171–82. doi:10.1006/jmbi.2000.3638.

Modis, Yorgo, and Rik K Wierenga. 1999. "A Biosynthetic Thiolase in Complex with a Reaction Intermediate: The Crystal Structure Provides New Insights into the Catalytic Mechanism." *Structure* 7 (10): 1279–1290. doi:10.1016/S0969-2126(00)80061-1.

Moqadam, Mahmoud, Enrico Riccardi, Thuat T. Trinh, Anders Lervik, and Titus S. van Erp. 2017. "Rare Event Simulations Reveal Subtle Key Steps in Aqueous Silicate Condensation." *Physical Chemistry Chemical Physics* 19 (20): 13361–71. doi:10.1039/C7CP01268C.

Morissette, Sherry L, Stephen Soukasene, Douglas Levinson, Michael J Cima, and Orn Almarsson. 2003. "Elucidation of Crystal Form Diversity of the HIV Protease Inhibitor Ritonavir by High-Throughput Crystallization." *Proceedings of the National Academy of Sciences of the United States of America* 100 (5): 2180–4. doi:10.1073/pnas.0437744100.

Moroni, Daniele, Titus S. van Erp, and Peter G. Bolhuis. 2004. "Investigating Rare Events by Transition Interface Sampling." *Physica A: Statistical Mechanics and Its Applications* 340 (1–3): 395–401. doi:10.1016/j.physa.2004.04.033.

Nieves-Quinones, Yexenia, and Daniel A. Singleton. 2016. "Dynamics and the Regiochemistry of Nitration of Toluene." *Journal of the American Chemical Society* 138 (46): 15167–76. doi:10.1021/jacs.6b07328.

Pauling, Linus. 1946. "Molecular Architecture and Biological Reactions." *Chemical & Engineering News Archive* 24 (10): 1375–77. doi:10.1021/cen-v024n010.p1375.

Peng, Chunyang, Philippe Y. Ayala, H. Bernhard Schlegel, and Michael J. Frisch. 1996. "Using Redundant Internal Coordinates to Optimize Equilibrium Geometries and Transition States." *Journal of Computational Chemistry* 17 (1): 49–56. doi:10.1002/(SICI)1096-987X(19960115)17:1<49::AID-JCC5>3.0.CO;2-0.

Peng, Chunyang, and H. Bernhard Schlegel. 1993. "Combining Synchronous Transit and Quasi-Newton Methods to Find Transition States." *Israel Journal of Chemistry* 33 (4): 449–54. doi:10.1002/ijch.199300051.

Pierce, Niles A., and Erik Winfree. 2002. "Protein Design Is NP-Hard." *Protein Engineering* 15 (10): 779–82.

Poree, Carl, and Franziska Schoenebeck. 2017. "A Holy Grail in Chemistry: Computational Catalyst Design: Feasible or Fiction?" *Accounts of Chemical Research* 50 (3): 605–8. doi:10.1021/acs.accounts.6b00606.

Privett, Heidi K, Gert Kiss, Toni M Lee, Rebecca Blomberg, Roberto A Chica, Leonard M Thomas, Donald Hilvert, Kendall N Houk, and Stephen L Mayo. 2012. "Iterative Approach to

Computational Enzyme Design." *Proceedings of the National Academy of Sciences of the United States of America* 109 (10): 3790–5. doi:10.1073/pnas.1118082108.

Proust-De Martin, Flavien, Renaud Dumas, and Martin J. Field. 2000. "A Hybrid-Potential Free-Energy Study of the Isomerization Step of the Acetohydroxy Acid Isomeroreductase Reaction." *Journal of the American Chemical Society* 122 (32): 7688–7697. doi:10.1021/ja000414s.

Quaytman, Sara L, and Steven D Schwartz. 2007. "Reaction Coordinate of an Enzymatic Reaction Revealed by Transition Path Sampling." *Proceedings of the National Academy of Sciences of the United States of America* 104 (30): 12253–8. doi:10.1073/pnas.0704304104.

Quaytman, Sara L., and Steven D. Schwartz. 2009. "Comparison Studies of the Human Heart and Bacillus Stearothermophilus Lactate Dehydrogreanse by Transition Path Sampling." *The Journal of Physical Chemistry A* 113 (10): 1892–97. doi:10.1021/jp804874p.

Radivojac, Predrag, Wyatt T. Clark, Tal Ronnen Oron, Alexandra M. Schnoes, Tobias Wittkop, Artem Sokolov, Kiley Graim, et al. 2013. "A Large-Scale Evaluation of Computational Protein Function Prediction." *Nature Methods* 10 (3): 221–27. doi:10.1038/nmeth.2340.

Ramsundar, Bharath, and Vijay S. Pande. 2016. "Learning Protein Dynamics with Metastable Switching Systems." *ArXiv:1610.01642 [Cs, Stat]*, October. http://arxiv.org/abs/1610.01642.

Reinecke, Frank, and Alexander Steinbüchel. 2008. "Ralstonia Eutropha Strain H16 as Model Organism for PHA Metabolism and for Biotechnological Production of Technically Interesting Biopolymers." *Journal of Molecular Microbiology and Biotechnology* 16 (1–2): 91–108. doi:10.1159/000142897.

Richter, Florian, Rebecca Blomberg, Sagar D Khare, Gert Kiss, Alexandre P Kuzin, Adam J T Smith, Jasmine Gallaher, et al. 2012. "Computational Design of Catalytic Dyads and Oxyanion Holes

for Ester Hydrolysis." *Journal of the American Chemical Society* 134 (39): 16197–206. doi:10.1021/ja3037367.

Rose, Peter W., Andreas Prlić, Ali Altunkaya, Chunxiao Bi, Anthony R. Bradley, Cole H. Christie, Luigi Di Costanzo, et al. 2017. "The RCSB Protein Data Bank: Integrative View of Protein, Gene and 3D Structural Information." *Nucleic Acids Research* 45 (Database issue): D271–81. doi:10.1093/nar/gkw1000.

Röthlisberger, Daniela, Olga Khersonsky, Andrew M Wollacott, Lin Jiang, Jason DeChancie, Jamie Betker, Jasmine L Gallaher, et al. 2008. "Kemp Elimination Catalysts by Computational Enzyme Design." *Nature* 453 (7192): 190–5. doi:10.1038/nature06879.

Ruscio, Jory Z, Jonathan E Kohn, K Aurelia Ball, and Teresa Head-Gordon. 2009. "The Influence of Protein Dynamics on the Success of Computational Enzyme Design." *Journal of the American Chemical Society* 131 (39): 14111–5. doi:10.1021/ja905396s.

Sadiq, S Kashif, and Peter V Coveney. 2014. "Computing the Role of near Attack Conformations in an Enzyme-Catalyzed Nucleophilic Bimolecular Reaction." *Journal of Chemical Theory and Computation* 11 (1): 316–324.

Saen-Oon, Suwipa, Sara Quaytman-Machleder, Vern L Schramm, and Steven D Schwartz. 2008. "Atomic Detail of Chemical Transformation at the Transition State of an Enzymatic Reaction." *Proceedings of the National Academy of Sciences of the United States of America* 105 (43): 16543–8. doi:10.1073/pnas.0808413105.

Saen-Oon, Suwipa, Vern L Schramm, and Steven D Schwartz. 2008. "Transition Path Sampling Study of the Reaction Catalyzed by Purine Nucleoside Phosphorylase." *Zeitschrift Fur Physikalische Chemie (Frankfurt Am Main, Germany)* 222 (8–9): 1359–1374. doi:10.1524/zpch.2008.5395.

Schomburg, I, A Chang, S Placzek, and C Söhngen. 2013. "BRENDA in 2013: Integrated Reactions, Kinetic Data, Enzyme Function Data, Improved Disease Classification: New Options and Contents in BRENDA." *Nucleic Acids Research* 41: 764–772.

Shankar Kumar, John M. Rosenberg, Bouzida Djamal, Robert H. Swendsen, and Peter A. Kollman. 1992. "The Weighted Histogram Analysis Method for Free-Energy Calculations on Biomolecules." 13 (8). http://onlinelibrary.wiley.com/doi/10.1002/jcc.540130812/abstract.

Shen, Yang, Michael K Gilson, and Bruce Tidor. 2012. "Charge Optimization Theory for Induced-Fit Ligands." *Journal of Chemical Theory and Computation* 8 (11): 4580–4592. doi:10.1021/ct200931c.

Sheppard, Micah J, Aditya M Kunjapur, Spencer J Wenck, and Kristala L J Prather. 2014. "Retro-Biosynthetic Screening of a Modular Pathway Design Achieves Selective Route for Microbial Synthesis of 4-Methyl-Pentanol." *Nature Communications* 5: 5031. doi:10.1038/ncomms6031.

Shurki, A., M. Štrajbl, J. Villà, and A. Warshel. 2002. "How Much Do Enzymes Really Gain by Restraining Their Reacting Fragments?" *Journal of the American Chemical Society* 124 (15): 4097–4107. doi:10.1021/ja012230z.

Siegel, Justin B., Alexandre Zanghellini, Helena M. Lovick, Gert Kiss, Abigail R. Lambert, Jennifer L. St.Clair, Jasmine L. Gallaher, et al. 2010. "Computational Design of an Enzyme Catalyst for a Stereoselective Bimolecular Diels-Alder Reaction." *Science* 329 (5989): 309–13. doi:10.1126/science.1190239.

Silver, Nathan. 2011. "Ensemble Methods in Computational Protein and Ligand Design: Applications to the Fcr Immunoglobulin, HIV-1 Protease, and Ketol (Doctoral Dissertation)." Massachusetts Institute of Technology.

Slater, Steven, Kathryn L Houmiel, Minhtien Tran, A Timothy, Nancy B Taylor, Stephen R Padgette, J Kenneth, Timothy A Mitsky, and Kenneth J Gruys. 1998. "Multiple β -Ketothiolases Mediate Poly (β -Hydroxyalkanoate) Copolymer Synthesis in Ralstonia Eutropha." *Journal of Bacteriology* 180 (8): 1979–1987.

Stewart, J.P. 2004. "Optimization of Parameters for Semiempirical Methods IV: Extension of MNDO, AM1 and PM3 to More Main Group Elements" 10: 155–164.

Štrajbl, Marek, Avital Shurki, Mitsunori Kato, and Arieh Warshel. 2003. "Apparent NAC Effect in Chorismate Mutase Reflects Electrostatic Transition State Stabilization." *Journal of the American Chemical Society* 125 (34): 10228–37. doi:10.1021/ja0356481.

Stubbe, JoAnne, and Jiamin Tian. 2003. "Polyhydroxyalkanoate (PHA) Hemeostasis: The Role of PHA Synthase." *Natural Product Reports* 20 (5): 445–57.

Swendsen, D.W., and Peter G Bolhuis. 2014. "A Replica Exchange Transition Interface Sampling Method with Multiple Interface Sets for Investigating Networks of Rare Events: The Journal of Chemical Physics: Vol 141, No 4" 141 (4): 044101.

Tadrowski, Sonya, Marcelo M Pedroso, Volker Sieber, James A Larrabee, Luke W Guddat, and Gerhard Schenk. 2016. "Metal Ions Play an Essential Catalytic Role in the Mechanism of Ketol–Acid Reductoisomerase." *Chemistry-A European Journal* 22 (22): 7427–7436.

Thompson, S, F Mayerl, O P Peoples, S Masamune, a J Sinskey, and C T Walsh. 1989. "Mechanistic Studies on Beta-Ketoacyl Thiolase from Zoogloea Ramigera: Identification of the Active-Site Nucleophile as Cys89, Its Mutation to Ser89, and Kinetic and Thermodynamic Characterization of Wild-Type and Mutant Enzymes." *Biochemistry* 28 (14): 5735–42.

Thorpe, Colin. 1986. "A Method for the Preparation of 3-Ketoacyl-CoA Derivatives." *Analytical Biochemistry* 155 (2): 391–394. doi:10.1016/0003-2697(86)90452-5.

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 58: 267–88.

Torrie, Glenn M., and John P. Valleau. 1974. "Monte Carlo Free Energy Estimates Using Non-Boltzmann Sampling: Application to the Sub-Critical Lennard-Jones Fluid." *Chemical Physics Letters* 28 (4): 578–81. doi:10.1016/0009-2614(74)80109-0.

Tseng, H.-C., and K. L. J. Prather. 2012. "Controlled Biosynthesis of Odd-Chain Fuels and Chemicals via Engineered Modular Metabolic Pathways." *Proceedings of the National Academy of Sciences* 109 (44): 17925–17930. doi:10.1073/pnas.1209002109.

Tseng, Hsien-Chung, Catey L Harwell, Collin H Martin, and Kristala L J Prather. 2010. "Biosynthesis of Chiral 3-Hydroxyvalerate from Single Propionate-Unrelated Carbon Sources in Metabolically Engineered E. Coli." *Microbial Cell Factories* 9 (1): 96. doi:10.1186/1475-2859-9-96.

Tseng, Hsien-Chung, Collin H Martin, David R Nielsen, and Kristala L Jones Prather. 2009. "Metabolic Engineering of Escherichia Coli for Enhanced Production of (R)- and (S)-3-Hydroxybutyrate." *Applied and Environmental Microbiology* 75 (10): 3137–45. doi:10.1128/AEM.02667-08.

Tyagi, Rajiv, Yu-Ting Lee, Luke W Guddat, and Ronald G Duggleby. 2005. "Probing the Mechanism of the Bifunctional Enzyme Ketol-Acid Reductoisomerase by Site-Directed Mutagenesis of the Active Site." *The FEBS Journal* 272 (2): 593–602. doi:10.1111/j.1742-4658.2004.04506.x.

Vagelos, Roy P., Alberts, A.W. 1960. "Chemical Synthesis of Beta-Ketoocanoyl Coenzyme A." *Analytical Biochemistry* 1: 8–16.

Van Erp, Titus S. 2012. "Dynamical Rare Event Simulation Techniques for Equilibrium and Nonequilibrium Systems." In , 27–60. John Wiley & Sons, Inc. doi:10.1002/9781118309513.ch2.

Villà, Jordi, and Arieh Warshel. 2001. "Energetics and Dynamics of Enzymatic Reactions." *The Journal of Physical Chemistry B* 105 (33): 7887–7907. doi:10.1021/jp011048h.

Voter, A. F. 1997. "Hyperdynamics: Accelerated Molecular Dynamics of Infrequent Events." *Physical Review Letters* 78 (20): 3908–11. doi:10.1103/PhysRevLett.78.3908.

Wallach, Izhar, Michael Dzamba, and Abraham Heifets. 2015. "AtomNet: A Deep Convolutional Neural Network for Bioactivity Prediction in Structure-Based Drug Discovery." *ArXiv:1510.02855 [Cs, q-Bio, Stat]*, October. http://arxiv.org/abs/1510.02855.

Wang, Bao-Lei, Ronald G Duggleby, Zheng-Ming Li, Jian-Guo Wang, Yong-Hong Li, Su-Hua Wang, and Hai-Bin Song. 2005. "Synthesis, Crystal Structure and Herbicidal Activity of Mimics of Intermediates of the KARI Reaction." *Pest Management Science* 61 (4): 407–12. doi:10.1002/ps.972.

Warshel, A. 1998. "Electrostatic Origin of the Catalytic Power of Enzymes and the Role of Preorganized Active Sites." *Journal of Biological Chemistry* 273 (42): 27035–27038. doi:10.1074/jbc.273.42.27035.

Weis, James W. (James Woodward). 2017. "Artificial Intelligence and Protein Engineering : Information Theoretical Approaches to Modeling Enzymatic Catalysis." https://dspace.mit.edu/handle/1721.1/108969#files-area.

Wen, Fei, Nikhil U Nair, and Huimin Zhao. 2009. "Protein Engineering in Designing Tailored Enzymes and Microorganisms for Biofuels Production." *Current Opinion in Biotechnology*, Protein technologies / Systems and synthetic biology, 20 (4): 412–19. doi:10.1016/j.copbio.2009.07.001.

Wu, Zhenqin, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. 2017. "MoleculeNet: A Benchmark for Molecular Machine Learning." *ArXiv:1703.00564 [Physics, Stat]*, March. http://arxiv.org/abs/1703.00564.

Zhang, Jun, Zhen Zhang, Yi Isaac Yang, Sirui Liu, Lijiang Yang, and Yi Qin Gao. 2017. "Rich Dynamics Underlying Solution Reactions Revealed by Sampling and Data Mining of Reactive Trajectories." *ACS Central Science* 3 (5): 407–14. doi:10.1021/acscentsci.7b00037.

Zheng, Ya-Jun, and Thomas C. Bruice. 1997. "A Theoretical Examination of the Factors Controlling the Catalytic Efficiency of a Transmethylation Enzyme: Catechol O-Methyltransferase." *Journal of the American Chemical Society* 119 (35): 8137–45. doi:10.1021/ja971019d.

Zoi, Ioanna, Dimitri Antoniou, and Steven D. Schwartz. 2017. "Incorporating Fast Protein Dynamics into Enzyme Design: A Proposed Mutant Aromatic Amine Dehydrogenase." *The Journal of Physical Chemistry B* 121 (30): 7290–98. doi:10.1021/acs.jpcb.7b05319.

Zoi, Ioanna, Matthew W. Motley, Dimitri Antoniou, Vern L. Schramm, and Steven D. Schwartz. 2015. "Enzyme Homologues Have Distinct Reaction Paths through Their Transition States." *The Journal of Physical Chemistry B* 119 (9): 3662–68. doi:10.1021/jp511983h.

Zoi, Ioanna, Javier Suarez, Dimitri Antoniou, Scott A. Cameron, Vern L. Schramm, and Steven D. Schwartz. 2016. "Modulating Enzyme Catalysis through Mutations Designed to Alter Rapid Protein Dynamics." *Journal of the American Chemical Society* 138 (10): 3403–9. doi:10.1021/jacs.5b12551.