Computational approaches to understand the atomistic drivers of enzyme catalysis

by

Natasha Seelam

B.S. Chemical and Biomolecular Engineering
Johns Hopkins University, 2013

Submitted to the Department of Chemical Engineering
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Chemical Engineering
at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2021

Signature of Author: _____
Department of Chemical Engineering
January 20, 2021

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

Accepted by: _____
Patrick S. Doyle
Robert T Haslam (1911) Professor
Chairman, Committee for Graduate Students

Computational approaches to understand the atomistic drivers of enzyme catalysis

by
Natasha Seelam

Submitted to the Department of Chemical Engineering on January 20, 2021, in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Chemical Engineering

Enzymes readily perform chemical reactions several orders of magnitude faster than their uncatalyzed versions in ambient conditions with high specificity, making them attractive design targets for industrial purposes. Traditionally, enzyme reactivity has been contextualized through transition-state theory (TST), in which catalytic strategies are described by their ability to minimize the activation energy to cross the reaction barrier through a combination of ground-state destabilization (GSD) and transition-state stabilization (TSS). While excellent progress has been made to rationally design enzymes, the complexity of the design space and the highly optimized nature of enzymes make general application of these approaches difficult. This thesis presents a set of computational methods and applications in order to investigate the larger perspective of enzyme-assisted kinetic processes.

For the first part of the thesis, we analyzed the energetics and dynamics of proficient catalyst orotidine 5′-monophosphate decarboxylase (OMPDC), an enzyme that catalyzes decarboxylation nearly 17 orders of magnitude more proficiently than the uncatalyzed reaction in aqueous solvent. Potential-of-mean-force (PMF) calculations on wild type (WT) and two catalytically hindered mutants, S127A and V155D (representing TSS and GSD, respectively), characterized the energy barriers associated with decarboxylation as a function of two parameters: the distance between the breaking C–C bond and a proton-transfer coordinate from the nearby side chain of K72, a conserved lysine in the active site. Coupling PMF analyses with transition path sampling (TPS) approaches revealed two distinct decarboxylation strategies: a simultaneous, K72-assisted pathway and a stepwise, relatively K72-independent pathway. Both PMF and TPS rate calculations reasonably reproduced the empirical differences in relative rates between WT and mutant systems, suggesting these approaches can enable *in silico* inquiry into both pathway and mechanism identification in enzyme kinetics.

For the second study, we investigated the electronic determinants of reactivity, using the enzyme ketol-acid reductoisomerase (KARI). KARI catalyzes first a methyl isomerization and then reduction with an active site comprised of several polar residues, two magnesium divalent cations, and NADPH. This study focused on isomerization, which is rate limiting, with two objectives: characterization of chemical mechanism in successful catalytic events (“reactive”) versus failed attempts to cross the barrier (“non-reactive”), and the interplay between atomic positions, electronic descriptors, and reactivity. Natural bonding orbital (NBO) analyses provided detailed electronic description of the dynamics through the reaction and revealed that successful

catalytic events crossed the reaction barrier through a 3-center-2-electron (3C) bond, concurrent to isomerization of hydroxyl/carbonyls on the substrate. Interestingly, the non-reactive ensemble adopted a similar electronic pathway as the reactive ensemble, but its members were generally unable to form and sustain the 3C bond. Supervised machine learning classifiers then identified small subsets of geometric and electronic descriptors, "features", that predicted reactivity; our results indicated that fewer electronic features were able to predict reactivity as effectively as a larger set of geometric features. Of these electronic features, the models selected diverse descriptors representing several facets of the chemical mechanism (charge, breaking–bond order, atomic orbital hybridization states, etc.). We then inquired how geometric features reported on electronic features with classifiers that leveraged pairs of geometric features to predict the relative magnitude of each electronic feature. Our findings indicated that the geometric, pair-feature models predicted electronic structure with comparable performance as cumulative geometric models, suggesting small subsets of features were capable of reporting on electronic descriptors, and that different subsets could be leveraged to describe various aspects of a chemical mechanism.

Lastly, we revisited OMPDC in order to learn the key geometric features that distinguished between the simultaneous and stepwise pathways of decarboxylation, aggregating and labeling pathways drawn from WT and mutant systems ensembles. We leveraged classifiers that predicted between reactive pathways by selecting small subsets of structural features from 620 geometric features comprised of atoms from the active site. The classifiers performed comparably, with greater than 80% testing accuracy and AUC, between times starting from in the reactant basin to 30 fs into crossing the reaction barrier. Remarkably, model-selected features reported on chemically meaningful interactions despite no explicit prior knowledge of the mechanism in training. To illustrate this, we focused analyses on two particular features shown to be predictive while in the reactant basin, prior to crossing the barrier: a potential hydrogen-bond between D75*, an aspartate in the active site, and the 2'-hydroxyl of OMP, and electrostatic repulsion through the proximity of a different aspartate, D70, to the leaving group carboxylate of OMP. Analysis between the simultaneous and stepwise ensembles demonstrated that the simultaneous ensemble adopted shorter distances for both features, generally suggesting stronger interactions. Both features were additionally shown to be associated with the ability to distort the planarity of the orotidyl ring, where shorter distances for either feature were correlated with larger degrees of distortion. Taken together, this suggested the simultaneous ensemble was more effective at distorting the ground state structure prior to crossing the reaction barrier.

Thesis Supervisor: Bruce Tidor

Title: Professor of Biological Engineering and Computer Science

Acknowledgements

I am fortunate to have been surrounded by many wonderful people who have supported, mentored, encouraged, and given me the strength to be where I am now. While this list is non–exhaustive, I wish to express my intense gratitude to the many who shaped my future.

I want to begin by relaying my gratitude to my thesis advisor, Professor Bruce Tidor. There are many, many things I could say to thank you – but I would like to focus on the ones that made the most pronounced impact on my life. First, I have greatly developed as a researcher because you have taught me how to pursue the right questions. Secondly, I have deeply appreciated your passion for details, and that intellectual rigor has left a lasting mark on everything I do. Lastly, you instilled in me courage and confidence because you believed in me. I remember how daunted I felt at the beginning of the PhD. Every year, though, I would solve a problem(s!) that I had initially worried would be impossible until it became virtually routine. The impossible is easy, because you had faith in me.

I am so appreciative of the wonderful thesis committee who has shaped this work: Professor Catherine L. Drennan, Professor William H. Green, and Professor Kristala L. Jones Prather. Every committee meeting has left me brimming with inspiration and ideas. Professor Drennan's keen insight into enzymes has consistently guided me in the right directions, Professor Green's excellent chemoinformatic and kinetic advice has shaped my intuition around reactivity, and Professor Prather's ability to tie my work to experimental intuition kept the work grounded. I owe Professor Prather a huge debt of gratitude for keeping me connected to ChemE, and all her sage advice.

I am fortunate to have had several other faculty mentors, including Professor Troy Van Voorhis, who welcomed me into his research group so I could learn quantum methods from the pioneers in the field; my understanding vastly expanded from interactions within the group. I also would like to extend my deep gratitude to Professor Anne McCants (and Bill McCants!), who has been a true role model for me, and whose compassion and strength has kept me afloat.

The Tidor lab has been a space where graduate students could freely explore ideas – even across groups! I am so grateful to have been in the energetic CSAIL environment. I would like to thank Ishan Patel (and Mary Orczykowski by proxy), whose love of software design taught me best practices, David Flowers for being an excellent friend to joust philosophies, Erika DeBenedictis for her incredible love of enzymes, protein design, and keen scientific advice, and Eli Karvelis, whose attention to detail and sharpness are astounding. A craftsman is enhanced by their tools – so a small thank you also to my lab desktop, Sputnik, with its egregious amounts of RAM to enable me to do virtually any calculation I want to prototype without concern. I also must underscore the brilliance of the wonderful students I've mentored: Heather Sweeney, Zi-Ning Choo, Anne Y. Kim, Emma Bernstein, Allison Tam, Diana Gong, and Corban Swain. I have cherished my time with every single one of you, and have learned so much from your beautiful perspectives. Lastly, to my office mates from the Mueller lab: thank you for the company, Dishita Turakhia, Maroula Bacharidou, Isabel Qamar, and Mustafa Doga Dogan.

The Van Voorhis group and theoretical chemists at MIT have been a source of strength and community in my time at MIT – I am so grateful to have been part of the Zoo! Lexie McIsaac, Henry Tran, Hong-Zhou Ye, Nathan Ricke, Changhae Andrew Kim, Jacq Tan, and Tami Goldzak have taught me so much electronic structure theory. I am eternally grateful for your support!

Scheduling at MIT is a miraculous feat, and I owe a huge thank you to the wonderful admins of my professors: Nira Manokharan, Gwen Wilcox, Barbara Balkwill (and Lilly!), and Read Schusky – thank you for being so helpful, and so approachable. Your impact in keeping the labs happy, afloat, and organized cannot be stated enough!

I care deeply about helping others – I was fortunate enough to meet those who shared my sentiments. Together with Garrett Dowdy and Anasuya Mandal-Sresht, I am proud of the work we accomplished with REFS-X, and am delighted by its growth over the years.

I have also had the pleasure of being a Burton Conner (B1) GRA where I had the single greatest group of undergrads over the years. I am so happy to have had the chance to watch you grow; and being a part of your lives has been one of the most rewarding experiences of my time at MIT that I am eternally grateful for. To my fellow GRAs/AD – Leilani Gilpin, Erik and Monique Eisenach, Kat Howell, and Cameron Cler particularly for their wine and whines, I am so thankful to have had great camaraderie with people who acutely understand the experience!

A large part of my resilience is owed to my friends – those who loved me and were my tireless champions. From ChemE, thank you to Yamini Krishnan, Kat Tarasova, Nathan Yee, Alex Bourque, Sakul Rantanalert, Raman Ganti, Hursh Sureka, Krishna Srinivas, Winston Chern (not ChemE but might as well be), Lionel Lam, Eric Miller, Mark Keibler, and Mark Goldman. Kara Rodby and Maddie Dery: you are my proteges and I am proud of the things you have done; you are both balm for the soul. To my incredible UCSB clique: (Professor!) Alexandra Bayles, John Abel, and Corinne Carpenter: thank you for bringing sunshine, and intense technical rigor into my life. To Mordhaus and pals: Maciej Murakowski, Emily Wilson, Natalie Woodard, Miranda Dobbs, Joe Giuliano, Kep Peterson, I am grateful for always cheering me up (in addition to amusing adventures such as 'Meatstrosity', 'Fried Ice', and 'Snakesgiving'). To Nicole Kogan and Hope Watson: you are my pride and joy, and your successes make me so proud. From across the pond in the other Cambridge, Greg Lever (and Maggie and baby Ophelia), thank you for your incredible technical insight – you inspire me! To John Santa Maria Jr, Mark Kalinich, Audra Amasino, Dori Katas, Mariola Szenk, Krzysztof Franaszek, Brian Chow, and Alexandria Cogdill: your strength and toughness kept me going. Nicole Faut: your tenacity is a legacy. Serena Booth (and by proxy, James): I am so grateful for your spirit and all the coffee and snack breaks in CSAIL; I am thankful I met someone who shared my philosophies and expanded my research curiosity.

Khoi Nguyen, Deena Rennerfeldt, and Adrienne Rothschilds: you have made me feel so worthy and I am grateful every day to have you in my life. You continue to motivate and inspire me throughout our careers, and I am looking forward to our futures together. Andrew Doyle, my k-pop king, you understand me to my core, and I have learned so much from your incredible ability to teach and mentor. Nicholas Delateur and David C Miller: you have been my biochemistry heroes since day one, thank you! Myungsun (Sunny) Kang: I am grateful to have found a soul sister in grad school who speaks algorithms and stat mech fluently; your hard work and brilliance never fails to give me courage. Emma Lagan: you capture the spirit of adventure in research and I am so grateful to have had such a scholarly friend since grade school. To Janusz and Renata Murakowski: thank you for welcoming me into your family.

To Aaron Newfield, Erika Takahashi, and Mama B (Barbara Snook, my second mom): our friendship now extends beyond a decade of love, and you have saved my life on many an occasion. Thank you for all that you do, I am eternally indebted to you. My three wonderful pets: Walter, Fang, and Curie – thank you for (mostly) unconditional snuggles and love.

The journey of my PhD required the selfless sacrifices of two generations to make me who I am – to my grandparents Srinivasan Amudhanar and Vijaya Srinivasan, my uncles and aunts Sridhar Amudhanar and Dagmar Richert Amudhanar, Shailu and Nagi Rampa (Nishi and Babu), and Anand Suchindrum and Sheela Amudhanar (and Neeraj), and to my parents Pushpa Seelam and Suresh Seelam: your selfless sacrifices allowed me to be who I am today, and want for nothing. All my successes in this world are from my family. To mom in particular, I have never needed a role model because I still want to be just like you when I "grow up". To my plucky younger sister, Nikita Seelam, your spirit has kept me positive and cheerful in the darkest of times. I am fortunate to have a sibling who is so empathetic, intuitive, and brilliant.

And lastly and most importantly, to my future family and fiancé – Dariusz Murakowski: I knew from the first Chandler paper we talked about together that you were the one. You are my raison d'etre.

Contents

3. Identifying the electronic determinants of reactivity in enzyme catalysis

List of Figures

Chapter 1: Introduction

Chapter 2: Catalytic Strategies of OMPDC elucidated by path sampling methods

Chapter 3: Identifying the electronic determinants of reactivity in enzyme catalysis

Chapter 4: Dynamic drivers of catalytic strategy in OMPDC and mutants

List of Tables

Chapter 1: Introduction

Natasha Seelam[1,2]

1. Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge MA

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

1.1 Overview

Of the diverse functions of proteins, enzymes perform the role of biological catalysts, facilitating reactions on molecules that are easily millions of times faster than their uncatalyzed version at physiological pH and temperature, and with high specificity [1, 32]. To date, enzymes are known to catalyze at least 5000 reactions, many critical to life as the product molecules of these reactions would be imperceptible at uncatalyzed rates [2, 3]. Harnessing the exquisite specificity of enzymes has long been a desirable goal for industrial purposes [30–33]. Problematically, it remains a major challenge to re–engineer proteins for custom function or repurpose them for novel substrates [34].

Several experimental and computational strategies exist toward the design of custom enzymes, including but not exclusive to directed evolution, computer–aided rational design, and a newly emerging branch employing machine learning. While such approaches have had remarkable successes, the complexity of design space and the highly optimized nature of enzymes for their native function thwart ubiquitous application of these approaches [26]. Often, engineered enzymes (or antibodies with catalytic function) are many orders of magnitude less efficient than their natural counterparts, and may still require rounds of directed evolution or random mutagenesis to improve their efficacy [29, 35–39], This suggests the need for further inquiry into the facets of enzyme chemistry to develop a holistic view on how to engineer an enzyme toward a desired goal (either *de novo* or repurposed), and to understand what structural facets of the enzyme govern successful catalysis. Three key directions toward that end include: (1) modeling the complex chemistries performed by enzymes in rich detail while preserving the underlying dynamics, (2) identifying the salient structural components most indicative of reactivity, and lastly (3) understanding the refined interplay between atomic positions within the active site and its influence on electronic structure

that encourage reactivity. The focus of this thesis is to delve into the catalytic nuances of enzyme reactivity, and to provide a computational framework in order to generate representative simulation data of enzyme–facilitated reactivity, analyze the catalytic strategies, and interpret the way enzymes achieve incredible chemistries.

1.2 Enzyme kinetics

1.2.1 Transition–state theory and origins

Inquiry into the origins of enzymes' remarkable catalytic prowess has led to numerous theories as to how enzymes facilitate these great enhancements. The prevailing theory of enzyme catalysis has been traditionally couched in the language of transition–state theory (TST), arising from foundational work beginning in the 1940s [4–7]. Broadly, transition–state theory provides a framework in which to study chemical reactions: namely, in this framework, 'activated–complexes' (i.e. the transition state) exist in quasi–equilibrium with the reactant–state molecules, and they have the ability to convert to products [4–6, 40, 41]. While the quasi–equilibrium assumption differs from the classical interpretation of equilibrium, the same thermodynamic treatment can be used to express the formation of product [5, 42].

The definition of the transition–state structure relies on the concept of a potential energy surface (PES), in which the progress of a reaction is defined as a function of atomic positions and momenta. The 'reaction coordinate' is a hypothetical variable(s) often used to describe this reaction progress, and the transition state is often cast in the context of it. Identifying the transition–state structure then corresponds to identifying the saddle points on the PES of the reaction [40, 41]. Due to the nature of this high energy, unstable structure, the transition state is extraordinarily short–lived, with a proposed lifetime of barely 10^{-13} s, on the scale of a bond vibration [47].

Using the framework of TST, Pauling proposed that these biological catalysts exhibit tight–binding to transition–state structures, thereby reducing the activation energy and improving the rate of reactivity without altering the equilibria of the unbound reactant and product molecules [7, 12] (Figure 1.1).

1.2.2 Michaelis–Menten kinetics in enzyme–substrate catalysis

Figure 1.1 shows the description of enzyme kinetics in the Michaels–Menten formalization [86]. Free enzyme (E) and substrate (S) exist in equilibrium with the formation of the enzyme–substrate complex (ES), with this equilibrium association constant denoted as $K_1 = \frac{k_1}{k_{-1}}$. The ES complex also has the ability to become product (P) with rate k_{cat} (measured in units of inverse time for 'unimolecular' reactions, of which ES is considered to be). When considering catalysis, the prowess of the enzyme influences k_{cat}. The rate, k_{cat}, incorporates additional terms that quantify the equilibrium of the ES complex and the (activated) transition state ($ES^{\ddagger}$), as well as the formation of EP, or the enzyme–product bound state, that then separates into enzyme and product molecules (E + P).

$$\mathrm{E} + \mathrm{S} \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} \mathrm{ES} \underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} \mathrm{ES}^{\ddagger} \xrightarrow{k_3} \mathrm{E+P}$$

Figure 1.1: Kinetics of enzyme–substrate chemistry, as described by Michaelis–Menten kinetics [86]. The free (unbound) enzyme is "E", the unbound substrate is "S", the unbound product is "P", the bound enzyme–substrate complex is "ES", and the activated (transition state) complex is "$ES^{\ddagger}$". The left–hand portion denotes the equilibrium kinetics at which the enzyme and substrate association leading to the enzyme substrate complex. The right–hand portion of the equation denotes the enzyme–substrate complex becoming unbound enzyme and product, for brevity simplified into E+P.

Figure 1.2 illustrates these kinetics in the context of a free–energy diagram (i.e., potential energy surface). The corresponding schematic illustrates a hypothetical reaction coordinate that marks the progress of the reaction. The left–hand portion of the equation in Figure 1.1 is represented in the 'binding' stage of Figure 1.2, in which the ES complex is formed. Catalysis is the subsequent activation of the ES complex, marked by the transition state (TS, or $ES^{\ddagger}$) with a concomitant increase in energy, followed by the formation of the enzyme–product complex (EP). This transition–state complex is energetically unfavorable and it quickly dissociates to free enzyme and product, for favorable reaction attempts [7, 12, 86].

The region of the protein that surrounds the substrate molecule for binding and subsequent reactivity is characterized as the 'active site'. The active site often has side–chain amino acids that are capable of exerting several forces (hydrogen bonds, van der Waals, and electrostatics) that provide an appropriate environment for catalysis, and that reduce the activation energy of the reaction. Comparisons of the uncatalyzed versus catalyzed rate of enzymatic enhancements have estimated that this reduction in activation energy (E_a) ranges from 11 to 38 kcal/mol from the uncatalyzed reaction [8–11].

Rigorous treatment of transition–state theory also suggests the inclusion of a transmission coefficient [4]. The transmission coefficient accounts for the fact that not every vibration may lead to successful barrier crossing [4]. While considerable work exists to characterize transmission coefficients in enzymes, it is not always practical to explicitly consider such effects [133–135].

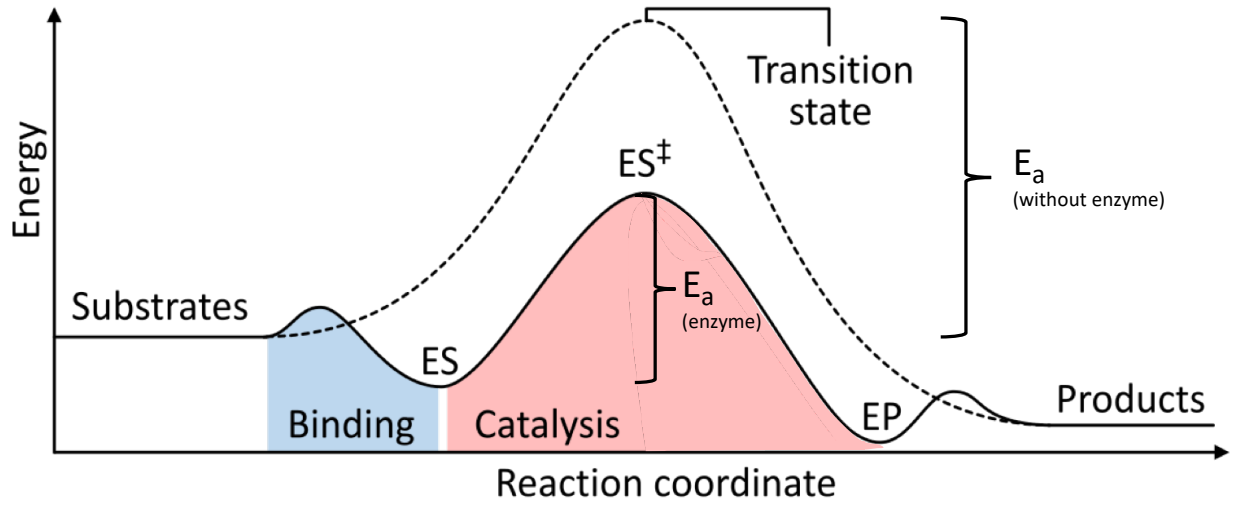


Figure 1.2 [Figure adapted from reference 141]: Schematic of the canonical interpretation of enzyme kinetics. The binding step incorporates free enzyme (E) and substrate (S) into the enzyme–substrate (ES) complex. This complex becomes activated into $ES^{\ddagger}$, the transition–state. The $ES^{\ddagger}$ exists in quasi–equilibrium with ES, but also has the ability to become EP, or the enzyme–product complex. Subsequently, the enzyme–product state rapidly dissociates to form free enzyme and product.

1.2.3 Catalytic strategies employed by enzymes

Transition–state theory has played a pronounced role in the investigation of enzyme catalysis, and has led to the rational design of inhibitor drug molecules that mimic hypothesized transition states of their respective enzymes [47]. Figure 3 represents a schematic of a hypothetical reaction from the context of enzyme–substrate catalysis, contrasting both the uncatalyzed and catalyzed version of the reaction. Two prevalent catalytic strategies put forward in the context of transition–state theory are ground–state destabilization (GSD) and transition–state stabilization (TSS), both of which both aim to reduce the activation energy of the reaction [24, 55]. It should be noted that enzymes may also provide for an alternative reaction pathway compared to what may occur in solvent [66].

As the name suggests, GSD increases the free energy of the enzyme and/or substrate in the enzyme–substrate complex relative to the same molecules in the unbound substrate and enzyme, often through using some of the binding energy to induce a distortion. On the other hand, TSS decreases the free energy of the bound transition–state complex. Conventional mechanistic proposals that support GSD–based hypotheses are often centered around electronic strain, bond–distortion, desolvation–effects, and conformational restriction after binding the substrate [43]. Similarly, central tenets of TSS include favorable interactions promoted by electrostatic interactions, such as hydrogen bonding, solvent environment, and occasionally promoting efficient and rapid proton abstractions or additions [48].

While transition–state theory has provided an organized methodology toward studying reactivity in enzymes, increasingly, studies have shown that there are other contributions that influence enzyme reactivity [13–17]; thus, these necessitate extending upon the classic transition–state theory characterization of enzymes. Several hypotheses have been put forth to help fill out the complete picture of enzyme kinetics, including enzyme pre–organization, and the role of near–attack conformations (NACs) which are conformations of the ground state that lie on the transition path of the reaction [13–17].

The hypothesis of electrostatic preorganization describes the enzyme providing a (typically) polar environment that encourages catalysis [54, 55]. Traditionally, this region is defined to include generally the first and occasionally second coordination sphere residues (typically corresponding to the amino acids within the active site) [52, 53]. A core tenet of this hypothesis is that the enzyme positions residues to sample NACs more effectively; this rearrangement of the environment can be as effective in enabling catalysis as lowering the reaction barrier [53, 138–140].

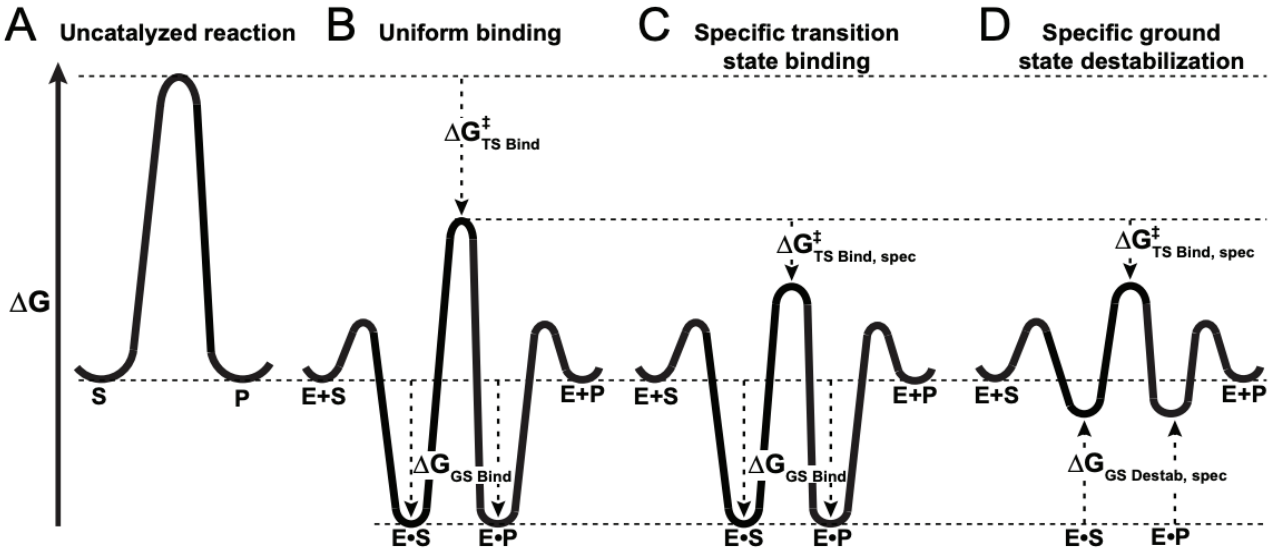


Figure 1.3 [Figure used from reference 46] Hypothetical 1D reaction profile. Free enzyme is represented as "E", unbound substrate as "S", unbound product as "P", enzyme–substrate complex as "ES", and the bound enzyme–product complex as "EP". Note, in all cases, these examples are simplified; enzymes can drastically change the reaction path compared to the solvated reaction. (A) The uncatalyzed reaction has some barrier indicated by the top-most dashed line. The activation energy would be the difference in energy from the TS, appearing at the topmost dashed black line, to the substrate "S". (B) This hypothetical uniform binder enzyme is not considered a true catalyst, as it does not change the activation energy of the enzyme. The decrease in energy after binding substrate is the same as the decrease in energy in attaining the enzyme-assisted TS complex. Thus, the activation energy is unchanged. (C) This hypothetical enzyme stabilizes the transition state compared to Fig 3B, as the difference between TS and ES is smaller than in Fig 3A and 3B. (D) This enzyme additionally includes ground-state destabilizing interactions compared to Fig 3A–3C, as the ground state is higher in energy; the difference in energy from TS and ES is smaller than 3A–3C, also suggesting this is enzyme would be a more effective catalyst.

A parallel question emerging from these catalytic strategies is what role atomic motions and dynamics play in the context of enzyme catalysis [17, 50, 51]. Reactivity (bond breaking/formation) often occurs on time scales that span mere femtoseconds up to picoseconds – on the order of larger–scale atomic motions [29]. In contrast, binding events are often on the scale of microseconds to milliseconds, suggesting these events could be decoupled from the catalytic act [29, 49]. A compelling hypothesis put forth by Schwartz and Schramm is the idea that the enzyme active site "increases the probability of rare dynamic interactions that permit rapid barrier crossing" [29]. Stated otherwise, after the relatively slow events of substrate collision and binding (and any conformational changes that may accompany them), the active site of an enzyme may favor or even encourage rate–promoting vibrations that help the system cross the barrier and facilitate catalysis [17, 29, 50]. Consensus among enzymologists suggests that enzymes likely employ numerous strategies, not limited to just GSD, TSS, preorganization, or NACs, to attain their catalytic performance [17–24, 52–56].

1.3 Biophysical modeling and molecular simulation

1.3.1 Quantum–mechanical/molecular–mechanical methods

As reactivity is a rapid event, careful experimental work has been able to supplement and support mechanistic proposals supporting these hypotheses [56–58]. However, many of these methods can be quite challenging to employ in experimental settings due to the short–lived nature of many of these states; hence computational approaches are an attractive strategy to investigate and quantify refined atomistic details about how enzymes facilitate their chemistries, and to inspire subsequent experiments that attempt to alter their functions. When considering theoretical studies, enzymes are large, high–dimensional, and complex molecules to model. With the advent of

increased processing power and specialized algorithms, biophysical modeling at the atomic level of detail is possible for many of these proteins [45]. Molecular mechanics employs the use of force fields to describe the physics of atomic interactions from a classical perspective; such approaches are versatile in describing dynamics and binding events [86–94]. However, force fields alone do not describe electronic structure phenomena, such as bond formation and destruction. Chemical reactions require quantum mechanical descriptions to characterize the transient changes in electronic structure. Quantum mechanical/molecular mechanical (QM/MM) methods were developed to efficiently address this discrepancy, in which a system is simultaneously modeled with both levels of theory, focusing the more expensive QM model only on the reactive portion [95–98]. For protein systems, a region within the active site where the reaction occurs (typically catalytic/conserved residues and the substrate) is often characterized at the quantum level of theory, while the remainder of the environment is quantified with an appropriate molecular–mechanics forcefield [45].

1.3.2 Enhanced sampling approaches to identify energetic landscapes

Most QM/MM simulations are run at ambient conditions; given a Boltzmann energy distribution centered around these conditions, sampling high–energy configurations, such as transition states, can be extraordinarily rare. To illustrate this point, consider orotidine 5′-monophosphate decarboxylase (OMPDC); its reaction barrier is nearly 17 kcal/mol [66]. The Boltzmann–associated likelihood of generating a configuration with this energy at 300 K would be $e^{-\frac{Ea}{RT}} \cong 4 \times 10^{-13}$ – a virtually impossible event!

To tackle this problem, clever physical methods have been developed that efficiently sample configurations in these high–energy regions. These methods encompass techniques such as umbrella sampling, blue–moon sampling, metadynamics, quantum–mechanical band methods, and empirical–valence bond theory [38, 59–62, 99–101]. A unifying element of these techniques is to apply either a biasing potential or higher temperatures to make rare configurations more accessible [59–62]. Methods like Weighted Histogram Analysis Method (WHAM) can then be used to estimate the unbiased potential at the desired temperature. Although these techniques provide accurate potential energy landscapes for reactive paths, they alter the dynamics of the simulation [63]. Moreover, the performance of these techniques in estimating barrier heights and reactivity is sensitive to the definition of the reaction coordinate(s) [63–65].

1.3.3 Transition path sampling methods for catalysis

Path algorithms, led by Transition Path Sampling (TPS), can explore rugged energy landscapes and reactivity without distorting dynamics. These methods harness Markov chain Monte Carlo (MCMC) techniques to sample ensembles of 'transitions', and they are formulated to be agnostic to reaction coordinates [63–65]. The only requirement for path sampling methods is the definition of an order parameter that appropriately identifies structures in the starting (reactant) basin and the ending (product) basin [63, 64] (Figure 4). Equipped with this parameter, one only needs an initial 'seed' path that connects between the starting and ending basins to generate ensembles of the transition of interest [63–65].

Generating ensembles with path sampling methodologies requires the use of several types of moves: this work specifically focuses on shifting and shooting moves, as indicated in Figure 4 [63–65]. Shifting moves preserve the majority of the trajectory, but alter a relatively few time steps

at the start or end of the trajectory [63–65]. By contrast, shooting moves can generate entirely *de novo* pathways. The Monte Carlo move chooses a time slice of a given trajectory, and creates a small perturbation to the velocity of all the atomic coordinates at this selected point. This perturbation is integrated forward and backward in time by the prescribed physics of the QM/MM simulation. A new candidate trajectory is accepted if the move resulted in a trajectory that successfully begins and ends in the appropriate starting and ending basins [63–65].

Prior studies have employed path–sampling to study several enzymatic systems due to the methods' versatility and rich data generation process [67–76]. Additionally, a rigorous statistical mechanical formulation has been developed to analyze the trajectories computed by path sampling methodologies and compute rate constants from the generated ensembles [77, 78]. Historically, path–sampling simulation studies have often focused on the mechanistic details of enzymatic reactions as opposed to full rate computations [128–132].

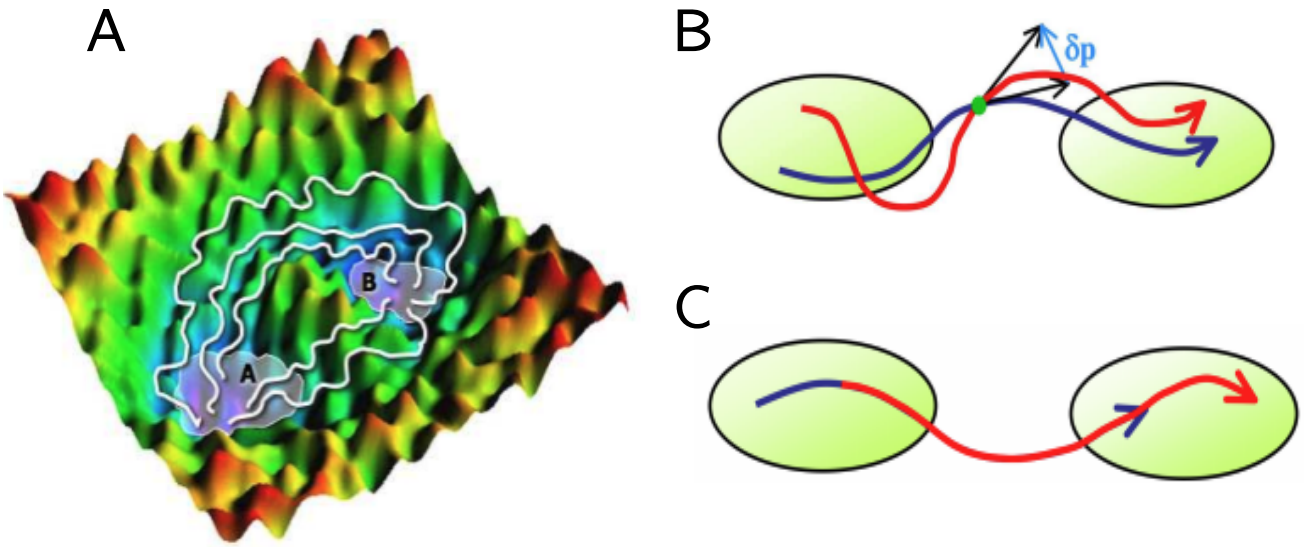


Figure 1.4 [Figure used from reference 79] (A) Hypothetical potential energy surface (PES) and reactive paths drawn across the landscape. Path sampling strategies require the order parameter to provide a clear delineation between the starting basin "A", and the product basin "B". In the following diagram, the advantage of path sampling strategies is that the Markov chain Monte Carlo (MCMC) sampling procedures allow for new paths to be constructed that can navigate across rugged landscapes that conventional sampling may not traverse without an assisting biasing potential. The ensemble of paths that connect A with B can then be analyzed for insights. In the case of reactivity, the starting basin is the reactant, and the ending basin is the product. The paths that connect between these states represent the catalytic trajectories. (B) A schematic of a shooting move; a shooting move within the MCMC path sampling ensemble uses a slice chosen from the prior trajectory within the ensemble (guaranteed to connect between the starting and ending basins), and applies a momentum or velocity perturbation to all the atoms in the system. Molecular mechanics models can propagate the resulting forces due to this perturbation forward and backward in time to assess whether the new trial trajectory also connects between the desired basins. If it does, the trajectory is accepted into the ensemble and a new move is computed using this as the starting trajectory. If it is rejected, the original seed trajectory is added again to the ensemble. (C) A schematic representing a shifting move; the TPS shifting move temporally "shifts" a path so that the actual path connecting between the basins remains the same, the beginning and ending are changed by lengthening or shortening.

1.3.4 Quantum calculation and natural bonding orbitals (NBO)

The path sampling methodology preserves the underlying dynamics of crossing a reactive barrier. For this reason, the atomic positions of the enzyme reaction can be probed in richer detail to quantify the electronic transitions as the reaction proceeds. To analyze electronic structure, pure *ab initio* methods can be employed on a system of interest to construct a wavefunction (or proxy of one) that characterizes the electron density across the atoms. However, full quantum characterization of biological systems remains a difficult feat. The complexity and size of large biomolecules (for example, the fully solvated protein system in Chapter 2 with orotidine 5′-monophosphate and explicit water is nearly 54,000 atoms!) makes it difficult to overcome the computational requirements that often scale as order 3 or greater with respect to the system size (i.e. $O(N^3)$ where N is the number of electrons in the system) [38, 99–103]. However, analyses may be focused on the critical region of reactivity (the substrate and the active site) to quantify the manner in which the enzyme facilitates reactivity [98, 104, 105]. While quantum techniques provide descriptions of electron density, further refinement of the orbital structure, as localized between atoms, can afford interpretable chemical insight. To that end, techniques such as Natural Bonding Orbital (NBO) theory allow the elucidation of mechanism through the lens of an organic chemist; namely, NBO theory quantifies bonding orbitals and lone–pair orbitals from a Lewis–like perspective [80–82].

NBOs aim to characterize the electron density between “centers”, typically one or two (occasionally three or more) atoms via an orthonormal set of “localized maximum occupancy orbitals” [80–82, 106–108]. This technique makes no pre–supposed hypotheses toward either the form, or location of the bonding orbitals. Instead, it searches across all possible ways of drawing bonds and lone pairs that describe the highest percentage of total electron density in leading

"Lewis–type" NBOs [106–108]. This description is considered the "Natural Lewis Structure" (NLS), which attributes the percentage of the computed wavefunction toward a "Lewis–like" interpretation, and the remaining orbitals are labeled as "non–Lewis" type NBOs that represent the residual effects not well–captured by Lewis theory (such as resonance or other types of delocalization).

The NBO approach works through a series of steps; first, the non–orthogonalized functions corresponding to the atomic wavefunctions calculated via some high level of theory (such as Density Functional Theory "DFT" or Hartree–Fock "HF") are converted into individual, "atom–centered", orthogonal natural atomic orbitals (NAOs) [81, 109]. These NAOs represent a localized, 1–center orbital that are ascribed to a given atom. The formulation of the NAOs allows for two physical attributes to be qualitatively described: (a) the spatial diffuseness of the orbital (i.e. the delocalized versus contracted nature) will depend on the molecular environment and (b) the valence NAOs properly incorporate nodal features from steric confinement. NAOs are strictly orthogonal, and are used in Natural Population Analysis (NPA) that appropriates the number of electrons associated to each atom in a more basis–insensitive way compared to the Mulliken Population Analysis [81, 109].

NAOs are then subsequently transformed into natural hybrid orbitals (NHOs), which represent atom–centered hybrids that are a linear combination of NAOs. The NHOs are eventually used in the two–centered natural bond orbital or NBO [110]. One advantage of NBOs is that they are uniquely associated to the wavefunction, with orbitals that are generally non–degenerate. Departures from classic Lewis–like descriptions also have meaningful interpretations, typically regarded as delocalization or resonance effects that a single Lewis structure cannot easily articulate. Common orbitals found in the NBO formulation include core shell, lone pair ("non–

bonded"), bonding, antibonding, and Rydberg orbitals. More recently, the formulation adopted a definition to encompass 3–center–2–electron (3C) descriptors [111, 112]. For biological systems, the NBO analysis has been performed on localized regions of enzymes in order to explore the chemical mechanism of catalysis [113], and to link key atomistic features and the way they influence the electronic environment of a reaction.

While quantum calculations provide refined interpretation of the electronic structure underlying a chemical reaction, understanding how a protein facilitates reactivity requires also capturing what geometric changes of the active site drive catalysis and are amenable to successfully crossing a reactive barrier. To highlight those geometric facets of the active site that give rise to reactivity, machine learning models can highlight key features that are linked to catalysis, which can then lead to structure and dynamics–based mechanistic hypotheses.

1.4 Machine learning methods in protein catalysis

Protein models are complex, diverse–atom, high–dimensional systems which are not necessarily intuitive, even to the keen enzymologist. Most classes of machine learning models are adept at parsing complex, multivariate relationships between key features within data, and can be applied to biology. Recently, several pioneering papers have employed various architectures of machine learning models to investigate relationships in structural biology in the fields of protein–protein interaction networks, optimized directed evolution, protein–ligand binding, protein structure and/or protein folding prediction [114–121].

Given the multitude of machine learning models, the choice of which algorithm depends on the problem posed. No single architecture or algorithm suffices for all tasks ubiquitously; thus, the formulation of the problem will drive the types of models deployed. The goal of understanding

the drivers of enzyme reactivity with machine learning requires two criteria be met: models should (i) identify whether or not a system crosses the barrier (reactive versus "non–reactive"), and (ii) select features most important toward predicting reactivity that are readily interpretable and/or easily accessible.

Given these criteria, neural networks are not necessarily appropriate for the enzyme reactivity problem. While they have excellent predictive performances [115–121], interpretability of the key features that influence predictive power in neural networks is still an open problem in the field [122–124]. Moreover, such models are often data hungry and require extensive training data in order to learn appropriate representations, given the number of parameters in the model [115–121]. Simpler machine learning models typically apply some type of linear transformation on the input features in order to predict a property or class of interest, and are often the first baseline predictors employed before more expressive models are used [117]. The work within chapters 3 and 4 employs the use of a simpler machine learning model, logistic regression, chosen for its ability to classify while directly underscoring the relevant features involved.

1.5 Thesis scope and organization

This thesis aims to synthesize the above progress in enzymology and computational techniques to investigate high–dimensional data in order to understand how enzymes facilitate their magnificent chemistries. Toward that end, the following three chapters describe key studies to explore these methods.

In Chapter 2, we studied the catalytic proficiency of orotidine 5′-monophosphate decarboxylase (OMPDC), an enzyme found in many biological systems that facilitates the decarboxylation of its substrate, orotidine 5′-monophosphate (OMP). This enzyme is known for

its incredible catalytic proficiency, enhancing reactivity by nearly 17 orders of magnitude from the solvated, uncatalyzed reaction [125]. Elegant experimental methods have dissected several key residues that facilitate reactivity in the enzyme, and our work focuses on 2 empirically verified mutants, S127A and V155D [66]. These variants, respectively, address the catalytic strategies of transition–state stabilization and ground–state destabilization. This investigation employed the use of potentials of mean force (PMFs) and path sampling strategies to provide both an energetic and a dynamic perspective on the decarboxylation of OMPDC. The PMF and path sampling trajectories showed reasonable agreement with the ranked, relative orders of the empirically identified rates, suggesting that both approaches were capable of reflecting systematic changes in catalysis from the local environment of the enzyme. The path sampling trajectories across WT, S127A, V155D revealed that there were at least two distinct reactive pathways to facilitate decarboxylation. WT and S127A frequently decarboxylated via a "simultaneous" mechanism, where the position of K72 was observed to be "coordinated" with the decarboxylation event. The V155D mutant displayed two types of mechanisms: one that was similar to WT and S127A, and another that explicitly broke the carbon–carbon bond before the proton transfers to the substrate and finishes the reaction ("stepwise").

We applied quantum–chemical techniques to explore the electronic description of the drivers of reactivity in Chapter 3. This chapter investigated the catalysis performed by ketol–acid reductoisomerase (KARI), an enzyme that facilitates a methyl–transfer isomerization with the assistance of two magnesium ion cofactors and NADPH. This enzyme has been the target of several studies, as the reaction it catalyzes plays a critical role in the synthesis of biofuels [127, 137, 138]. Prior work in the group identified a subset of 30 geometric features, constructed from distances, angles, and torsions of the active site of KARI that predicted reactivity and influenced

the likelihood of crossing the barrier when sampled in the reactant basin [136]. Chapter 3 analyzed the successful, methyl transferring reactive simulations and the simulations that failed to cross the barrier from the work of Bonk et al. [136] and extended the chemical significance by revealing the mechanism of KARI via Natural Bonding Orbital (NBO) analyses to represent electronic structure in a Lewis–like representation. Our calculations revealed that methyl transfer occurred simultaneously with carbonyl formation and lone–pair formation with the adjacent oxygens (O6, O8) of the substrate, and that the reaction formed a three–center–two–electron bond, formed from the participating carbons (C4, C5, C7) of the substrate. In order to investigate the relationship between the geometric features of the active site and underlying electronic structure, we performed classification tasks to predict reactivity with both geometric and electronic features. Subsequent analyses identified a subset of 10 geometric features or a set of 6 electronic features, either of which were sufficient to predict reactive and non–reactive simulations with ROC AUC > 0.9, suggesting that while both types of features were predictive, fewer electronic features were required to predict reactivity with the same performance. We identified a possible catalytic strategy in which one of the geometric features involving a catalytically conserved glutamate, E319, and the distance to the transferring methyl (E319/OE1–C5) influenced the torsional orientation of the methyl prior to transferring such energy of the breaking bond orbital (C4–C5) increased, suggesting ground–state destabilization. This torsional orientation was also related to an electronic descriptor reporting on the electron density of the breaking bond (C4–C5 bond index), demonstrating that eclipsed orientations weakened the breaking bond prior to reacting. Lastly, we showed that pair–feature models of the 10 geometric features in predicting each of the 6 electronic features performed comparably to a full cumulative model, suggesting small subsets of geometric features were enough to predict the underlying electronic structure.

Chapter 4 revisited the OMPDC system and harnessed the framework of Chapter 3 in order to investigate the decarboxylation mechanism across the WT and mutant systems. From the results of chapter 2, energetic and dynamic characterization of the decarboxylation of OMPDC revealed two distinct pathways: a 'simultaneous' mechanism where the decarboxylation is coordinated with the positioning of catalytically important lysine, K72, and a 'stepwise' mechanism in which the decarboxylation is relatively independent of K72's position. The reactive pathways of the WT, S127A, and V155D mutants were combined to create two aggregate ensembles for each pathway, and supervised machine learning models are trained to classify between the two mechanisms for several time points starting from the reactant basin to 30 fs after the systems were committed to crossing the reaction barrier. Pair–feature models were able to predict reactivity with (ROC) AUC > 0.8 across all time points tested in this work. Moreover, model–selected features underscored several mechanistic hypotheses, of which we showed that two features, the distance between residue D75 in proximity to the 2'–hydroxyl of the OMP substrate ribose and the distance between D70's carboxylate oxygen and OMP's carboxylate oxygen, influenced the ability to distort the planarity of the orotidyl prior to reacting. Taken together, this highlights the ability for machine learning models to recognize chemically meaningful features in enzyme catalysis, without explicit knowledge of the mechanism.

1.6 References

1. X. Zhang, K.N. Houk. Why Enzymes are Proficient Catalysts: Beyond the Pauling Paradigm. *Acc. Chem. Res.* 38, 5, 379–385, 2005.
2. Schomburg, A. Chang, S. Placzek, C. Söhngen, M. Rother, M. Lang, C. Munaretto, S. Ulas, M. Stelzer, A. Grote, M. Scheer, D. Schomburg. BRENDA in 2013: integrated reactions, kinetic data, enzyme function data, improved disease classification: new options and contents in BRENDA. *Nucleic Acids Res*. 41:D1 D764–D772, 2013.
3. J.M. Berg, J.L. Tymoczko, L. Stryer, L. Stryer. Biochemistry. *New York: W.H. Freeman.* 2002.
4. H.Eyring. The Activated Complex in Chemical Reactions. *J. Chem. Phys*. 3 (2): 107–115, 1935.
5. K.J Laidler, M.C. King. The development of Transition–State Theory. *J. Phys. Chem.* 87 (15): 2657–2664, 1983
6. M. G. Evans and M. Polanyi. Some applications of the transition state method to the calculation of reaction velocities, especially in solution. *Trans. Faraday Soc. 31*, 875, 1935.
7. L. Pauling. Chemical achievement and hope for the future. *Am. Scientist.* 36, 51–58, 1948.
8. B.G. Miller, R. Wolfenden. Catalytic proficiency: the unusual case of OMP decarboxylase. *Annu. Rev. Biochem*. 71: 847–885, 2002.
9. M.J. Snider, R. Wolfenden. The rate of spontaneous decarboxylation of amino acids. *J. Am. Chem. Soc.*, 122, 11507–11508, 2000.
10. C. Lad, H. Williams, R. Wolfenden. The rate of hydrolysis of phosphomonoester dianions and the exceptional catalytic proficiencies of protein and inositol phosphatases. *Proc. Natl. Acad. Sci. USA*, 100, 5607–5610, 2003.

11. A.C. Reyes, A.P. Koudelka, T.L. Amyes, J.P. Richard. Enzyme Architecture: Optimization of Transition–State Stabilization from Cation–Phosphodianion Pair. *J. Am. Chem. Soc.* 137, 16, 5312–5315, 2015.

12. L., Pauling. Molecular Architecture and Biological Reactions. *Chem. Eng. News*. 24 (10): 1375–77, 1946.

13. M. Garcia–Viloca, J. Gao, M. Karplus, D.G. Truhlar. How enzymes work: analysis by modern rate theory and computer simulations. *Science*. 303:186–195, 2004.

14. S.J. Benkovic, S.A. Hammes–Schiffer. Perspective on Enzyme Catalysis. *Science*. 301(5637), 1196–11202, 2003.

15. K. Zinovjev, I. Tunon. Quantifying the limits of transition state theory in enzymatic catalysis. *Proc. Natl. Acad. Sci. USA* 114:12390–12395, 2017.

16. D.D. Boehr, R. Nussinov, P.E. Wright. The role of dynamic conformational ensembles in biomolecular recognition. *Nat Chem Biol.* 5:789–796, 2009.

17. S. Hur, T.C. Bruice. The near attack conformation approach to the study of chorismite to prephenate reaction. *Proc. Natl. Acad. Sci. USA,* 100(21) 12015–12020, 2003.

18. J.R. Knowles. To build an enzyme. *Philos. Trans. R. Soc.,* B 332, 115−121, 1991

19. J.R. Knowles. Enzyme catalysis: Not different, just better. *Nature* 350, 121−124, 1991.

20. P. Carter, J.A. Wells. Functional interaction among catalytic residues in subtilisin BPN′. *Proteins* 7, 335−342, 1990.

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

22. G.G. Hammes, S.J. Benkovic, S. Hammes–Schiffer. Flexibility, diversity, and cooperativity: Pillars of enzyme catalysis. *Biochemistry* 50, 10422−10430, 2011.

23. A, R. Fersht. Structure and Mechanism in Protein Science, 2nd ed., *W. H. Freeman and Co., New York*, 1999.

24. W.P. Jencks. Catalysis in Chemistry and Enzymology, 2nd ed., *Dover, New York,* 1987.

25. J.P. Richard. Protein Flexibility and Stiffness Enable Efficient Enzyme Catalysis. *J. Am. Chem. Soc.* 141, 3320–3331, 2019.

26. B. Kuhlman, D. Baker. Native protein sequences are close to optimal for their structures. *Proc. Natl. Acad. Sci. USA*, 97(19): 10383–103888, 2000.

27. M. Roca, A. Vardi–Kilshtain, A. Warshel. Toward accurate screening in computer aided enzyme design. *Biochemistry* 48:3046–305, 2009.

28. Warshel, P.K. Sarma, M. Kato, Y. Xiang, H. Liu, M.H.M. Olsson. Electrostatic basis for enzyme catalysis. *Chem. Rev.* 106:3210–3235, 2006.

29. S.D. Schwartz, V.L. Schramm. Enzymatic transition states and dynamic motion in barrier crossing. *Nat Chem Biol.* 5(8), 551–558, 2009.

30. H. Nam, N.E. Lewis, J. A. Lerman, D. Lee, R.L. Chang, D. Kim, B.O. Palsson. Network Context and Selection in the Evolution to Enzyme Specificity. *Science*, 337(6098): 1101–1104, 2012.

31. S. Li, X. Yang, S. Yang, M. Zhu, X. Wang. Technology prospecting on enzymes: application, marketing and engineering. *Comput Struct Biotechnol J.* 2:1–11, 2012.

32. J.M. Choi, S.S. Han, H.S. Kim. Industrial applications of enzyme biocatalysis: current status and future aspect. *Biotechnol Adv.* 33:1443–1454, 2015.

33. C. Li, R. Zhang, J. Wang, L.M. Wilson, Y.Yan. Protein Engineering for improving and diversifying natural product biosynthesis. *Trends in Biotechnol*. 38(7):729–744, 2019.

34. S. Jemli, D. Ayadi–Zouari, H.B. Hlima, S.Bejar. Bioacatalysts: application and engineering for industrial purposes. *Crit Rev Biotechnol.* 36(2):246–58, 2016.

35. A.M. David, A.T. Plowright, E. Valeur. Directing Evolution: the next revolution in drug discovery?. *Nat. Rev. Drug Discovery.* 16(10):681–698, 2017

36. F,H. Arnold. The nature of chemical innovation: new enzymes by evolution. *Q. Rev. Biophys.* 48(4) 404–410, 2015.

37. T.C. Bruice, K. Kahn. Computational Enzymology. *Curr Opin Chem Biol*. 4(5) 540–544, 2000.

38. M.P. Frushicheva, M.J. Mills, P. Schopf, M.J. Singh, R.B. Prasad, A. Warshel. Computer aided enzyme design and catalytic concepts. *Curr Opin Chem Biol*. 21:56–62, 2014.

39. L. Regan, D. Cabellero, M.R. Hinrichsen, A. Virrueta, D.M. Williams, C.S. O'Hern. Protein Design: Past, Present, and Future. *Biopolymers*. 104(4): 334–350, 2015.

40. R. Marcelin. Contribution a l'étude de la cinétique physico–chimique. *Annales de Physique.* 9(3):120–231, 1915.

41. J. Bigeleisen. The effect of isotopic substitution on the rates of chemical reaction. *J. Phys. Chem.* 56(7):823–858, 1952.

42. J.L. Steinfeld, J.S. Francisco, W.L Hase. Chemical Kinetics and Dynamics (2nd ed.). *Prentice–Hall.* pp. 289–293, 1999.

43. V.E. Anderson. Ground State Destabilization. *John Wiley and Sons.* 2001. https://doi.org/10.1038/npg.els.0000625

44. J.G. Belasco, J.R. Knowles. Direct observation of substrate distortion by triosephosphate isomerase using Fourier transform infrared spectroscopy. *Biochemistry* 19: 472–477, 1980.

45. M.W. van der Kamp, A.J. Mulholland, Combined Quantum Mechanics/Molecular Mechanics (QM/MM) Methods in Computational Enzymology. *Biochemistry*, 52, 2708–2728, 2013

46. L.D. Andrews, T.D. Fenn, D. Herschlag. Ground–State Destabilization by Anionic Nucleophiles Contributes to the Activity of Phosphoryl Transfer Enzymes. *PLOS Biology,* 11:7, 1–18, 2013.

47. V.L. Schramm, B.A. Horenstein, P.C. Kline. Transition State Analysis and Inhibitor Designs for Enzymatic Reactions. *The Journal of Biological Chemistry*, 269, 28, 18259–18262, 1994.

48. Schramm, VL. Enzymatic Transition States and Transition–State Analog Design. *Annu. Rev. Biochem*. 693–720, 1998.

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

50. G. Bhabha, J. Lee, D.C. Ekiert, J. Gam, I.A. Wilson, H.J. Dyson, S.J. Benkovic, P.E. Wright. A Dynamic Knockout Reveals That Conformational Fluctuations Influence the Chemical Step of Enzyme Catalysis. *Science*. (2), 332, 6026, 234–238, 2011.

51. A.J. Adamcyzk, J. Cao, S.C.L. Kamerlin, A. Warshel, Catalysis by dihydrofolate reductase and other enzymes arises from electrostatic preorganization, not conformational motions. *Proc. Natl. Acad. Sci. USA108*, 34, 14115–14120, 2011.

52. Morgenstern, M. Jaszai, M.E. Eberhart, A.N Alexandrova. Quantified electrostatic preognization in enzymes using the geometry of electron charge density. *Chem. Sci.*, 8, 5010–5018, 2017.

53. J. Fuller III, T.R. Wilson, M.E. Eberhart, A.N. Alexandrova. Charge Density in Enzyme Active Site as a Descriptor of Electrostatic Preorganization. *J. Chem. Inf. Model.* 59, 5, 2367–2373, 2019.

54. Warshel. Electrostatic Origin of the Catalytic Power of Enzymes and the Role of Preorganized Active Sites. *J. Biol. Chem.* 273, 27035−27038, 1998.

55. Warshel, A.; Sharma, P. K.; Kato, M.; Xiang, Y.; Liu, H.; Olsson, M. H. M. Electrostatic Basis for Enzyme Catalysis. *Chem. Rev.* 2006, 106, 3210−3235.

56. W. Childs, S.G. Boxer. Solvation Response along the Reaction Coordinate in the Active Site of Ketosteroid Isomerase. *J. Am. Chem. Soc.*, 132, 6474−6480, 2010.

57. S.D. Fried, S. Bagchi, S.G. Boxer. Extreme Electric Fields Power Catalysis in the Active Site of Ketosteroid Isomerase. *Science*, 346, 1510−1514, 2014.

58. Y. Wu, S.G. Boxer. A Critical Test of the Electrostatic Contribution to Catalysis with Noncanonical Amino Acids in Ketosteroid Isomerase. *J. Am. Chem. Soc.*, 138, 11890−11895, 2016.

59. E. Vanden–Eijnden. Some Recent Techniques for Free Energy Calculations. *J. Comput. Chem.*, 30, 1737–1747, 2009.

60. C.D. Christ, A.E. Mark, W.F. van Gunsteren. Basic Ingredients of Free Energy Calculations: A Review. *J. Comput. Chem.*, 31, 1569–1582, 2010.

61. D. Trzesniak, A.P.E. Kunz, W.F. van Gunsteren. A comparison of methods to compute the potential of mean force. *Chem Phys Chem*, 8, 162–169, 2007.

62. P. Kollman. Free–energy calculations – applications to chemical and biochemical phenomena. *Chem. Rev.,* 93, 2395–2417, 1993.

63. D.W. Swendsen, and Peter G Bolhuis. "A Replica Exchange Transition Interface Sampling Method with Multiple Interface Sets for Investigating Networks of Rare Events. *J. Chem. Phys.* 141(4): 044101, 2014.

64. P.G. Bolhuis, D. Chandler, C. Dellago, P.L. Geissler. Transition Path Sampling: Throwing Ropes Over Rough Mountain Passes, in the Dark. *Annu. Rev. Phys. Chem.* 53: 291–318, 2002.

65. P.G. Bolhuis, C. Dellago. Practical and Conceptual Path Sampling Issues. *Eur Phys J Spec Top*, 224 (12): 2409–27, 2015.

66. V. Iiams, B.J. Desai, A.A Fedorov, E.V. Fedorov, S.C Almo, J.A. Gerlt. Mechanism of the orotidine 5′–monophosphate decarboxylase–catalyzed reaction: Importance of residues in the orotate binding site. *Biochemistry*. 50(39): 8497–8507, 2011.

67. J.E. Basner, S.D. Schwartz. How Enzyme Dynamics Helps Catalyze a Reaction in

68. Atomic Detail: A Transition Path Sampling Study. *J Am Chem Soc* 127 (40): 13822–31, 2005.

69. G. Hummer. Catching a Protein in the Act. *Proc. Natl. Acad. Sci. USA* 107 (6): 2381–2, 2010.

70. S.L. Quaytman, S.D. Schwartz. Reaction Coordinate of an Enzymatic Reaction Revealed by Transition Path Sampling. *Proc. Natl. Acad. Sci. USA* 104 (30): 12253–8, 2007.

71. H.B. Mayes, B.C. Knott, M.F. Crowlery, L.J. Broadbelt, J. Stahlberg, G.T. Beckham. Whos on Base? Rebealing the catalytic mechanism of inverting family 6 glycoside hydrolases. *Chem Sci.* 7(9): 5955–5968, 2016.

72. S.L. Quaytman, S.D. Schwartz. Comparison Studies of the Human Heart and Bacillus Stearothermophilus Lactate Dehydrogreanse by Transition Path Sampling. *J. Phys. Chem. A*, 113 (10): 1892–1897, 2009.

73. R. Crehuet, M.J. Field. A Transition Path Sampling Study of the Reaction Catalyzed by the Enzyme Chorismate Mutase. *J. Phys. Chem. B.,* 111 (20): 5708–5718, 2007..

74. S. Saen–Oon, S. Quaytman–Machleder, V.L. Schramm, S.D Schwartz. Atomic Detail of Chemical Transformation at the Transition State of an Enzymatic Reaction. *Proc. Natl. Acad. Sci. USA*, 105 (43):16543–16548, 2008.

75. T. Burgin, H. B. Mayes. Mehcanism of oligosaccharide synthesis via a mutant GH29 fucosidase. *React. Chem. Eng.* 4, 402–409, 2018.

76. S. Saen–Oon, S. Quaytman–Machleder, V.L. Schramm, S.D Schwartz. 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, 2008..

77. C. Dellago, P.G. Bolhuis, F.S Csajka, D. Chandler. Transition path sampling and the calculation of rate constants. *J Phys Chem*, 108, 1964–1977, 1998.

78. T.S. van Erp, M. Moqadam, E. Riccardi. A. Lervik. Analyzing Complex Reaction Mechanisms Using Path Sampling. *J. Chem. Theory Comput.*, 12 (11): 5398–5410. 2016.

79. M. Ferrario, G. Ciccotti, K. Binder. Computer Simulations in Condensed Matter Systems: From Materials to Chemical Biology Volume 1. *Springer, Berlin, Heidelberg.* 703, 350–391, 2007.

80. J.P. Foster, F. Weinhold. Natural Hybrid Orbitals. *J. Am. Chem. Soc.* 1980, 102, 7211−7218.

81. A.E. Reed, F. Weinhold. Natural Bond Orbital Analysis of Near–Hartree−Fock Water Dimer. *J. Chem. Phys.* 78, 4066−4073, 1983.

82. F. Weinhold. Natural Bond Orbital Analysis: A Critical Overview of its Relationship to Alternative Bonding Perspectives. *J. Comp. Chem.* 33, 2363–2379, 2012.

83. P.–O. Löwdin. Quantum Theory of Many–Particle Systems. I. Physical Interpretations by Means of Density Matrices, Natural Spin–Orbitals, and Convergence Problems in the Method of Configurational Interaction. *Phys. Rev.* 97, 1474 (1955);

84. E. Davidson. Reduced Density Matrices in Quantum Chemistry. *Academic Press*, Vol 1, 1976.

85. F. Weinhold, "Natural Bond Orbital Methods," in *Encyclopedia of Computational Chemistry*, P. v.R. Schleyer, N. L. Allinger, T. Clark, J. Gasteiger, P. A. Kollman, H. F. Schaefer III, P. R. Schreiner (Eds.), (John Wiley & Sons, Chichester, UK, 1998), Vol. 3, pp. 1792–1811.

86. L. Michaelis, M.L. Menten; Kinetik der Invertinwirkung. *Biochem*. Zeitung, 49, pp. 333–369, 1913.

87. E.P. Barros, J.M. Schiffer, A. Vorobieva, J. Dou, D. Baker, R.E. Amaro. Improving the Efficiency of ligand–binding protein design with molecular dynamics simulations. *J. Chem. Theory Comput*., 15, 10, 5703–5715, 2019.

88. A.C. Pan, D. Jacobson, K. Yatsenko, D. Sritharan, T.M. Weinreich, D.E. Shaw. Atomic–level characterization of protein–protein association. *Proc. Natl. Acad. Sci. USA,* 116 (10) 4244–4249 March 5, 2019

89. N. Barbera, M.A.A Ayee, B.S. Akpa, I. Levitan. Molecular Dynamics Simulations of Kir2.2 Interactions with an Ensemble of Cholesterol Molecules. *Biophys. J.* 115, 7, 1264–1280, 2018.

90. A. Spitaleri, S. Decherchi, A. Cavalli, W. Rocchia. Fast Dynamic Docking Guided by Adaptive Electrostatic Bias: The MD–Binding Approach. *J. Chem. Theory Comput.*, 14, 3, 1727–1736, 2018.

91. A. Spitaleri, W. Rocchia. Molecular dynamics–based approaches describing protein binding," in Biomolecular simulations in structure–based drug discovery. *Wiley VCH*: *Weinheim, Germany,* 29–39, 2019.

92. A. Frenkel, B. Smit. Chapter 4 – Molecular Dynamics Simulations, Understanding Molecular Simulation (Second Edition). *Academic Press*, 63–107, ISBN 9780122673511, 2002.

93. J. Wang, W. Wang, P.A. Kollman, D.A. Case. Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graph. Model.* 25, 2006, 247260.

94. J. Wang, R.M. Wolf, J.W. Caldwell, P.A. Kollman, D.A. Case. Development and testing of a general AMBER force field. *J. Comp. Chem.*, 25, 1157–1174, 2004.

95. A. Warshel, M. Levitt. Theoretical studies of enzymic reactions: dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *J. Mol. Bio.* 103 (2): 227–49, May 1976.

96. E. Brunk, U. Rothlisberger. Mixed Quantum Mechanical/Molecular Mechanical Molecular Dynamics Simulations of Biological Systems in Ground and Electronically Excited States. *Chem. Rev.* 115 (12): 6217–63, 2015.

97. H. Lin, D.G. Truhlar. QM/MM: what have we learned, where are we, and where do we go from here?. *Theor. Chem. Acc.* 117 (2): 185–199, Feb. 2007.

98. H.M Senn, W. Thiel. QM/MM methods for biomolecular systems. *Angew. Chem.* Int. Ed., 48: 1198–1229, 2009.

99. A. Warshel, R.M. Weiss. An empirical valence bond approach for comparing reactions in solutions and in enzymes. *J. Am. Chem. Soc.,* 102 (20): 6218–6226, Sept. 1980.

100. H. Jónsson, G. Mills, K. W. Jacobsen. Nudged Elastic Band Method for Finding Minimum Energy Paths of Transitions, in Classical and Quantum Dynamics in Condensed Phase Simulations. Ed. B. J. Berne, G. Ciccotti and D. F. Coker, 385. *World Scientific*, 1998

101. M.J. Field, P.A. Bash, M. Karplus. A combined quantum–mechanical and molecular mechanical potential for molecular–dynamics simulations. *J Comput Chem.* 11(6):700–33, 1990.

102. P. Hohenberg; W. Kohn. Inhomogeneous electron gas. *Phys. Rev*. 136 (3B): B864–B871, 1964.

103. F Jensen. Introduction to Computational Chemistry. *John Wiley & Sons, Inc.* ISBN: 0470011874, 2006.

104. K.E. Ranaghan, A.J. Mulholland. Investigations of enzyme–catalysed reactions with combined quantum mechanics/molecular mechanics (QM/MM) methods. *Int Rev Phys Chem.,* 29(1):65–133, 2010.

105. G. Jindal, A. Warshel. Exploring the Dependence of QM/MM Calculations of Enzyme Catalysis on the Size of the QM Region. *J. Phys. Chem. B*, 120, 37, 9913–9921, 2016.

106. C. R. Landis, F. Weinhold, "The NBO View of Chemical Bonding", in, G. Frenking and S. Shaik (eds.), The Chemical Bond: Fundamental Aspects of Chemical Bonding (*Wiley*), pp. 91–120, 2014.

107. F. Weinhold, C. R. Landis, E. D. Glendening. What is NBO Analysis and How is it Useful? *Int. Rev. Phys. Chem.,* 35, 399–440, 2016.

108. F. Weinhold, C. R. Landis. Natural Bond Orbitals and Extensions of Localized Bonding Concepts. *Chem. Educ. Res. Pract.* 2, 91–104, 2001.

109. E. Reed, R. B. Weinstock, F. Weinhold. Natural Population Analysis. *J. Chem. Phys.* 83, 735–746 (1985).

110. J. P. Foster, F. Weinhold. Natural Hybrid Orbitals. *J. Am. Chem. Soc.* 102, 7211–7218 (1980).

111. E. D. Glendening, C. R. Landis, F. Weinhold. NBO 7.0: New Vistas in Localized and Delocalized Chemical Bonding Theory. *J. Comput. Chem.* 40, 2234–2241, 2019.

112. NBO 7.0. E. D. Glendening, J, K. Badenhoop, A. E. Reed, J. E. Carpenter, J. A. Bohmann, C. M. Morales, P. Karafiloglou, C. R. Landis, and F. Weinhold, Theoretical Chemistry Institute, University of Wisconsin, Madison (2018).

113. I.S. Patel, Large scale simulation and analysis to understand enzymatic chemical mechanisms. Doctoral Dissertation, Massachusetts Institute of Technology, 2015.

114. M. AlQuraishi. End–to–end differentiable learning of protein structure. *Cell Systems.* 8, 4, p292–301, 2019.

115. F.Yang, K. Fan, D. Song, H. Lin. Graph–based prediction of Protein–protein interactions with attributed signed graph embedding. *BMC Bioinformatics,* 21, 323, 2020.

116. K.K. Yang, Z. Wu, F.H. Arnold. Machine–learning–guided directed evolution for protein engineering. *Nat. Methods*, 16(8):687–694, 2019

117. Z. Wu, S.B.J. Kan, R.D. Lewis, B.J. Wittmann, F.H, Arnold. Machine learning–assisted directed protein evolution with combinatorial libraries. *Proc. Natl. Acad. Sci. USA,* 116, 18, 8852–8858, 2019.

118. A.W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Zidek, A.W.R. Nelson, A. Bridgland, H. Penedones, S. Petersen, K. Simonyan, S. Crossnan, P. Kohli, D.T. Jones, D. Silver, K. Kavukcuoglu, D. Hassabis. Improved protein structure prediction using potentials from deep learning. *Nature*, 577, 706–710, 2020.

119. F. Noé, G. De Fabritiis, C. Clementi. Machine learning for protein folding and dynamics. *Curr. Opin. Struct. Biol*. 60, 77–84, 2020.

120. J.M. Cunningham, G. Koytiger, P.K. Sorger, M. AlQuraishi. Biophysical prediction of protein–peptide interactions and signaling networks using machine learning. *Nat Methods* 17, 175–183, 2020.

121. E.C. Alley, G. Khimulya, S. Biswas, M. AlQuraishi, G.M. Church. Unified rational protein engineering with sequence–based deep representation learning. *Nat Methods*, 16, 1315–1322, 2019.

122. Z.C. Lipton. The Mythos of Model Interpretability. *2016 ICML Workshop on Human Interpretability in Machine Learning* (WHI 2016), New York, NY

123. L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter and L. Kagal, Explaining Explanations: An Overview of Interpretability of Machine Learning, *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA),* Turin, Italy, 2018, pp. 80–89

124. F. Doshi–Velez, B. Kim. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv:1702.08608*

125. A. Radzicka, R. Wolfenden R. A proficient enzyme. *Science*. 267, 90–93, 1995.

126. M.M Cox, D.L Nelson. Chapter 6.2: How enzymes work. Lehninger Principles of Biochemistry (6th ed.). New York, N.Y.: *W.H. Freeman*. p. 195, 2013.

127. S. Bastian, X. Liu, J.T. Meyerowitz, C.D. Snow, M.M.Y Chen, F.H. Arnold. Engineered ketol–acid reductoisomerase and alcohol dehydrogenase enable anaerobic 2–methylpropan–1–ol production at theoretical yield in Escherichia coli. *Metab. Eng.*, 13, 3, p345–352, 2011.

128. I. Zoi, D. Antoniou, S.D. Schwartz. Incorporating Fast Protein Dynamics into Enzyme Design: A Proposed Mutant Aromatic Amine Dehydrogenase. *J. Phys. Chem. B*., 121 (30): 7290–7298, 2017.

129. I, Zoi, M,W. Motley, D, Antoniou, V,L. Schramm, S,D. Schwartz. Enzyme Homologues Have Distinct Reaction Paths through Their Transition States. *J. Phys. Chem. B*. 119 (9): 3662–3668, 2015.

130. I. Zoi, J. Suarez, D. Antoniou, S.A. Cameron, V.L. Schramm, S.D. Schwartz. Modulating Enzyme Catalysis through Mutations Designed to Alter Rapid Protein Dynamics. *J. Am. Chem. Soc*. 138 (10): 3403–3409, 2016.

131. R.K Harijan, I. Zoi, D. Antoniou, S.D. Schwartz, V.L. Schramm. Catalytic–Site Design for Inverse Heavy–Enzyme Isotope Effects in Human Purine Nucleoside Phosphorylase. *Proc. Natl. Acad. Sci. USA,* 114 (25): 6456–6461, 2017.

132. M. Dametto, D. Antoniou, S.D. Schwartz. Barrier Crossing in Dihydrofolate Reductase does Not Involve a Rate–Promoting Vibration. *Mol. Phys.* 110 (9–10): 531–536, 2012.

133. N. Boekelheide, R. Salomón–Ferrer, T.F. Miller III. Dynamics and dissipation in enzyme catalysis. *Proc. Natl. Acad. Sci. USA,* 108(39):16159–16163, 2011.

134. J. Pu, J. Gao, D.G Truhlar. Multidimensional Tunneling, Recrossing, and the Transmission Coefficient for Enzymatic Reactions. *Chem. Rev.*, 106, 8, 3140–3169, 2006.

135. D.G Truhlar. Transition state theory for enzyme kinetics. *Arch Biochem Biophys.* Sep 15; 582: 10–17, 2015.

136. B.M. Bonk, J. Weis, B. Tidor. Machine Learning Identifies the Chemical Characteristics That Promote Enzyme Catalysis. *J. Am. Chem. Soc.* 141, 4108–4118, 2019.

137. Y. Kung, W. Runguphan, J.D. Keasling. From Fields to Fuels: Recent Advances in the Microbial Production of Biofuels. *ACS Synth. Bio.* 1, 11, 498–513, 2012.

138. G. Jindal, A. Warshel. Misunderstanding the Preorganization Concept can lead to Confusions about the Origin of Enzyme Catalysis. *Proteins*. 85, 12, 2157-2161, 2017.

139. P.T.R. Rajagopalan, S.J. Benkovic. Preorganization and protein dynamics in enzyme catalysis. *Chem. Rec*. 2, 24-36, 2002.

140. F.M. Menger, F. Nome. Interaction vs preorganization in enzyme catalysis. A despite calls for resolution. *ACS Chem. Biol.* 14, 1386-1392, 2019.

141. "Enzyme." T. Shafee, Wikipedia, Wikimedia Foundation, Dec 2020, https://en.wikipedia.org/wiki/Enzyme.

Chapter 2: Catalytic strategies of OMPDC elucidated by path sampling methods

Natasha Seelam[1,2], Bruce Tidor[2,3,4]

1. Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge MA

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

3. Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge MA

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

2.1 Abstract

Orotodine 5′-monophosphate decarboxylase (OMPDC) is notable for its extreme catalytic proficiency and has been the subject of numerous experimental and theoretical studies. This enzyme facilitates conversion of its substrate, orotidine 5′-monophosphate (OMP), to uridine 5′-monophosphate (UMP), a precursor in pyrimidine biosynthesis, with a rate enhancement of nearly 17 orders of magnitude over the uncatalyzed reaction in aqueous solvent. Empirical studies have implicated direct decarboxylation, yet there are several catalytic strategies that remain compatible with this proposal. While prior theoretical work often focused on studying the enzyme in the context of explicit transition–state stabilization (TSS) or ground–state destabilization (GSD), we examine both energetics and dynamics. In the current work, we characterize the potentials of mean force for the enzyme catalyzed reaction with umbrella sampling using QM/MM simulations for wild type (WT) and two catalytically hindered mutants, S127A and V155D, empirically identified to address different aspects of the catalysis. Our results suggest that two reaction mechanisms are compatible with direct decarboxylation: one in which a catalytically conserved lysine, K72, stabilizes the carbanion formed through direct decarboxylation; and another in which decarboxylation could be independent of K72's position. We performed path–sampling calculations that simulated the reaction dynamics to further study contributions to catalytic proficiency. The simulations showed reduced catalytic proficiency for both mutants, consistent with the rate depreciation established from experiments. Interestingly, dynamic paths collected from WT and each mutant demonstrated that each system sampled different reaction paths: WT and S127A preferentially adopted a "simultaneous" path, in which K72's position plays a role in the decarboxylation, whereas V155D decarboxylated with both the former pathway, and a second, K72–independent, "stepwise" path. These results suggest the role of dynamics in adopting paths toward reactivity that are altered when the chemical environment is changed.

2.2 Introduction

Orotidine 5′-monophosphate decarboxylase (OMPDC) is an enzyme that catalyzes the precursor material in *de novo* synthesis of pyrimidines [1]. This pathway is present in eukaryotic organisms such as plants, fungi and even protists as well as prokaryotic organisms like bacteria [15, 17, 20–22]. Strikingly, this enzyme is noted as one of the most proficient catalysts in existence, producing a rate enhancement of over 17 orders of magnitude compared to aqueous solvent, without assistance of co–factors or metal ions to facilitate the chemistry [1, 2]. The enzyme facilitates the decarboxylation of the ring–structure connected to the ribophosphate group of the reactant–state orotidine 5′-monophosphate (OMP) to make uridine 5′-monophosphate (UMP) (Figure 2.1).

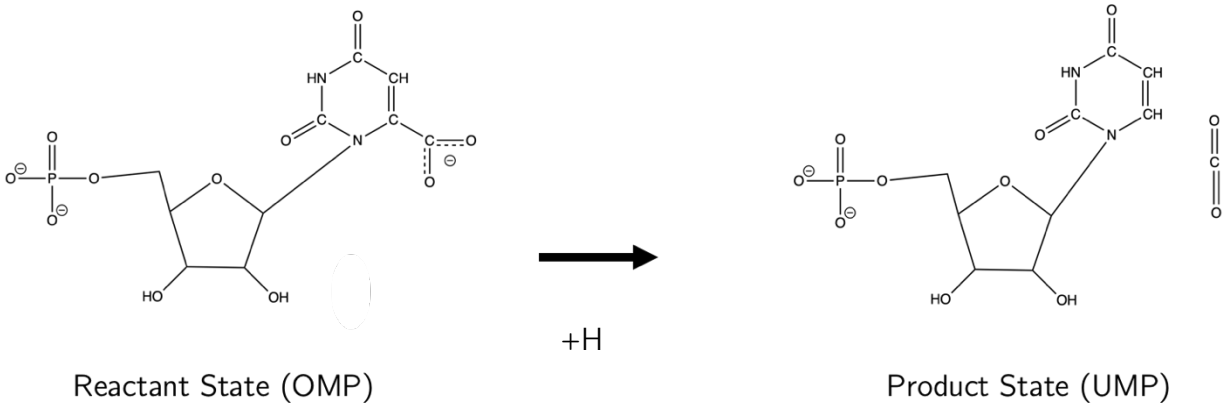


Figure 2.1: Reaction catalyzed by OMPDC. This enzyme is able to facilitate the decarboxylation of the pyrimidine precursor OMP into UMP. Cleavage removes a carboxylate group from the base and replaces it with a proton for two products molecules: UMP and carbon dioxide.

Since the discovery of OMPDC, several mechanisms have been posited to describe the decarboxylation reaction [3–7]. While mechanistic proposals have included the possibility of a

nucleophilic attack on C5 [5], or a zwitterionic structure formed via protonation of O2 [4], experimental and theoretical characterization has led to the prevailing hypothesis that the enzyme stabilizes a carbanion formed through decarboxylation [3, 6–11]. Experiments have also demonstrated that the decarboxylation likely happens before protonation [11], and the rate–determining step is considered to be dependent on the decarboxylation, or breaking of the C–C bond on the nitrogenous base of the nucleotide precursor substrate [53]. Evidence also seems to suggest that the enzyme favors catalytic strategies that can stabilize the carbanion formed upon decarboxylation to help facilitate the rate enhancement [13–20]. Indeed, evidence suggests that methyl–orotate, a truncated version of the substrate methylated at the N1 location (omitting the ribophosphate), reacted up to three orders of magnitude faster in solvents less polar than water [12]. Another experiment demonstrated that a truncated substrate representing the orotate ring fluorinated at position of C5 (replacing a proton) enhanced the rate up to 400–fold, suggesting that the electronegativity of a proximal atom can help alleviate the strain of decarboxylation by potentially delocalizing negative charge from a formed carbanion [13].

OMPDC is well characterized across yeast and bacterial species, with a highly conserved, charged active–site network, and substrate positioning that seems to be conserved across organisms [15, 17, 20–22], despite divergence in the sequence itself. This study specifically focused on the *Methanothermobacter thermoautotrophicus* variant (MtOMPDC) due to the existence of an excellent crystal structure with a resolution of 1.4 Å and empirically well–characterized mutants [23, 11].

MtOMPDC exists as an obligate homodimer, in which two active–site pockets are formed using the interface of the dimer; the opposing dimer provides a catalytically important residue, thus is necessary for the formation of the appropriate active site. The active–site possesses highly

charged residues that assist in reactivity and create a tetrad of lysines and aspartates: K42, D70, K72, D75* (the asterisk indicating the second monomer) [7, 20, 23] (Figure 2.2). This particular motif is seen in several other species, including yeast and *E. coli* [20, 24].

The aforementioned residues correspond to K59, D91, K93, D96* respectively in *Saccharomyces cerevisiae* yeast OMPDC (ScOMPDC). Experimental validation of this yeast analog has demonstrated that individual alanine substitution of each of these residues, except K59 (analogous to K42 in MtOMPDC), has detrimental effects on catalytic function [20, 28]. This charged network appears to also be important to MtOMPDC, as the enzyme exhibits a hundred–fold reduction in catalytic performance with a D70N mutation [27].

Considerable theoretical discussion has also explored the role of these specific catalytic residues in the context of the reaction, resulting in two leading hypotheses framed in the context of ground–state destabilization and transition–state stabilization. The specific functional significance of the catalytic tetrad residues in the active site in close proximity to the substrate carboxylate is considered unclear; initial hypotheses posed by Wu et al. suggested evidence for ground–state stabilization (GSD) [15], compatible with the theory originally proposed by Jencks that an enzyme first may employ luring attractive forces to later compensate for a destabilizing effect [30]. Wu et al. claimed that the role of the active–site lysines (namely K72) would 'lure' the substrate into the binding pocket, whereupon the aspartyl side chains would provide considerable electrostatic stress to the Michaelis–Menten complex to reduce the $\Delta\Delta G$ between the ground state and transition state [15, 23]. Contrary to the GSD hypothesis, Warshel et al. proposed that the origin of rate enhancement for OMPDC arises from transition–state stabilization (TSS), claiming that K72 provides a stabilizing role toward a possible carbanionic transition–structure [14]. In support of the TSS–based hypotheses, an additional residue (namely S127) is thought to delocalize

the transition state to allow for a carbene–like intermediate through a transient hydrogen–bond with a carbonyl present on the ring structure [6, 27]. Experimental mutation of S127 to alanine or proline does point to some role for this residue, as these mutants showed marked reduction in the catalytic rate of the decarboxylation [27].

In addition to the charged tetrad, several other residues are implicated to play an important role toward the catalysis of OMPDC. The residues binding the phosphodianion moiety (namely R203 and Q185) provide a stabilizing role for the substrate to bind into the active site. For MtOMPDC, substitution of either of these residues dramatically affects catalysis by changing K_m, suggesting these residues affect the binding affinity of the substrate to the enzyme [32]. Moreover, prior work identified a hydrophobic pocket adjacent to the highly charged catalytic tetrad, indicating the importance of the charged environment in substrate destabilization (i.e. GSD), as mutation of any of these participating pocket side–chains to hydrophilic and/or polar residues resulted in a 10^1 to 10^4–fold reduction in catalysis [27].

Evidence converges to suggest that the enzyme actively employs multiple attributes to facilitate catalysis. Mechanisms compatible with experimental observations underscore the following two facets: (1) the ability to either directly stabilize a carbanion or create a carbene–like resonance structure during the reaction via transition–state stabilization (TSS) and (2) the capacity to destabilize the carboxylate group via ground–state/substrate destabilization (GSD).

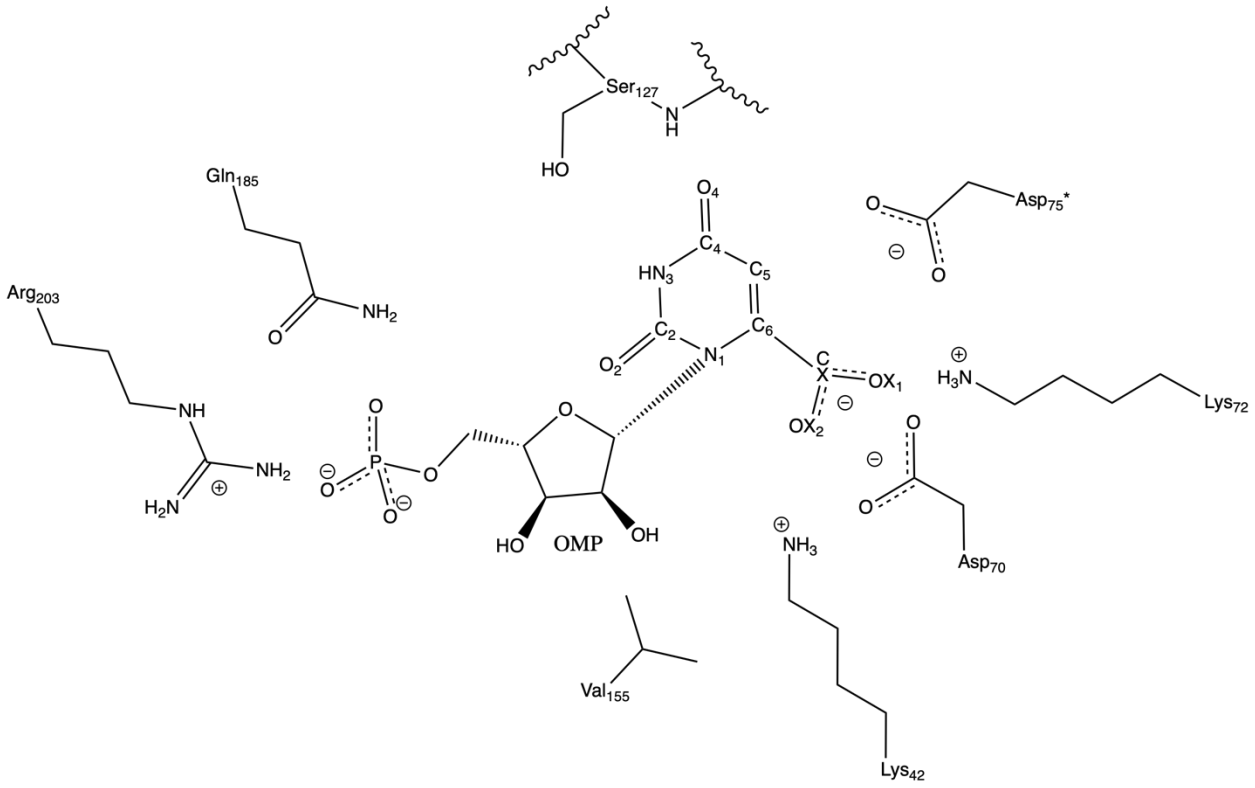


Figure 2.2: Active–site of MtOMPDC highlighting the substrate and several active-site residues: the catalytic tetrad K42, D70, K72, and D75* (asterisk denoting it comes from the other monomer), S127, V155, R203 and Q185. The decarboxylation reaction removes the carboxylate group at C6 of the orotidyl ring. Subsequently, C6 is protonated, and the carboxylate group of CX leaves as an uncharged CO_2. For a walleyed stereoview, see Figure 4.S17.

Iiams et al. provided extensive characterization of residues implicated in catalytic strategies of TSS and GSD by creating mutants of MtOMPDC demonstrating dramatic reductions in k_{cat} when contrasted with the wild type (WT) system [27]. In particular, that study focused on comparing WT with several mutants including V155D and S127A; both mutants exhibited large reduction in reactivity and link changes to the local environment to each of the two aforementioned catalytic strategies. The mutant V155D represents a polar substitution to the hydrophobic pocket near the catalytic, charged tetrad of the active site, and is thought to hinder GSD [27]. S127A

substitutes the hydroxyl side group of a conserved serine that participates in hydrogen bonding the amide of the substrate ring. While it does not directly discount delocalization via the amide proton and carbonyl O4 of the substrate, it has been shown to decrease the rate by 2 orders of magnitude, potentially by disrupting the hydrogen bond formed by the side–chain hydroxyl and carbonyl O2 from the ring, disrupting effective TSS [27].

The objectives of the current work are two–fold: first, we aimed to characterize and contrast the free–energy surfaces of the WT with the two catalytically hindered mutants, and secondly, we aimed to elucidate reaction paths that the WT and mutants may take to facilitate reactivity. These aims were accomplished through the use of potential of mean force (PMF) calculations coupled with transition path sampling (TPS) methodologies to provide a free–energy perspective of the reaction, and to explore how the WT and mutants navigated across the reaction barrier of facilitated decarboxylation.

Our work extends prior work by Vardi–Kilshtain et al. that produced the free–energy landscape of the decarboxylation of WT MtOMPDC [18]. Their work employed the use of umbrella–sampling and QM/MM methodologies to characterize the energetics of the reaction as a function of three reaction coordinates, namely: the decarboxylation coordinate of the breaking bond between the ring carbon and the carboxylate carbon; a proton–transfer coordinate monitoring the distance of a proton on K72 as it formed a bond with the ring carbon versus the distance of the originally attached amide group of K72; and a hybridization coordinate measuring the linearity of the carboxylate group. In their findings, they reported that the minimum free–energy path implicated the lysine, K72, providing a directly stabilizing effect – the position of the nearest proton of K72 was coordinated with the decarboxylation event and concurrently approached the ring carbon while the decarboxylation occurred. While well–characterized, their sampling

approach was restricted to a narrowly defined region of the decarboxylation/proton–transfer coordinate phase space, hence omitting the possibility of exploring if any other mechanistic paths may be energetically plausible [18]. Our work extends theirs and explicitly samples a larger grid of decarboxylation and proton–transfer coordinates and additionally includes the empirically characterized mutants.

The introduction of empirically characterized mutants that probe different aspects of the catalysis allows for a holistic view of how decarboxylation proceeds in the OMPDC chemical environment. Our calculations applied umbrella sampling methods to decarboxylation and proton–transfer coordinates of the WT and catalytically hindered mutants, capturing absolute and relative reactivity reasonably well, as measured by barrier heights (PMF) and rate (TPS). The resulting PMFs for WT and S127A mutant also suggested two energetically comparable pathways toward decarboxylation: one in which the decarboxylation occurs concurrently to the K72 proton moving closer to the ring, and another independent of the K72 proton's general location. In contrast, the V155D mutant exhibited a higher barrier toward the concurrent route, favoring the K72–independent approach.

While umbrella sampling methods provide an excellent view of the energetics of a landscape, subsequent methods are required to identify reaction paths. Examples of such approaches include nudged–band methods, which have been employed in the use of identifying reaction paths across energetic barriers [18]. While such methods provide rigorous structures of the transition–state saddle point, they often omit the dynamic contributions that occur throughout the reaction, and are also dependent on the reaction coordinates chosen to describe the reaction. For protein systems, there has been much debate on the role of dynamics in the ability of an enzyme to facilitate reactivity [48, 49]. In order to be able to explore the role of dynamics, in addition to

potential reactive paths, transition–path sampling methods offer an amenable way to explore these reactive pathways without the constraints and assumptions directly imposed by explicit transition–state theory (TST) methodologies. While compatible with TST, the theory does not explicitly require a formal definition of the transition–state structure or reaction coordinate, and it is possible to directly compute a rate that takes account of dynamic effects [39, 40]. TPS, in the context of chemical reactions, only requires an order parameter, denoted λ, capable of describing the reactant and product basins, with basin boundaries being independent of one another [39, 40]. For this reason, TPS methods were suitable for our inquiry into the dynamic paths that the WT and mutant OMPDCs may employ to facilitate decarboxylation. The resulting TPS calculations not only recapitulated the PMF results, in particular suggesting that the likelihood of crossing the barrier diminished for hindered mutants compared to WT, but additionally demonstrated that the WT and mutants preferentially crossed the barrier in different ways. Simulations of successful bond breakage showed the WT and S127A preferentially decarboxylating in a "simultaneous" pathway which leveraged the position of a nearby side chain, K72, while decarboxylating, whereas the V155D mutant had two distinct pathways: one that resembled the WT and S127A pathway and a separate approach independent of K72's position dubbed "stepwise".

2.3 Methods

2.3.1 Structure Preparation

The crystal structure of native *M. thermobacterautotrophicus* OMPDC with bound transition–state inhibitor 6'–hydroxyuridine–5'–phosphate (BMP) (accession code 3LTP) from the Protein Data Bank was used for the preparation of WT and mutant structures [23]. The crystal structure was solved with data collected to 1.4 Å resolution and a pH of 7.1. Missing hydrogen–atom positions were generated with CHARMM's internal HBUILD module using parameters from CHARMM36's all–atom forcefield. The structure contained two histidine residues per monomer; both were protonated in the δ–position in order to optimize hydrogen–bonding to nearby side chains. OMPDC comprises two independent active sites from the assembly of two monomers. In both active sites of the crystal structure, BMP is complexed with the enzyme. BMP was converted to OMP in each active–site by assigning analog atom positions to the OMP parameters (indicated below in 2.3.2) and allowing CHARMM's native internal–coordinate building procedures to replace the missing carboxylate. requiring it to be co–planar with the aromatic ring based on geometry optimization with GAUSSIAN03 (see 'Constructing force field parameters'). Starting coordinates for point mutants of the protein were obtained by ablating the side–chain of the canonical residue in the WT, and re–building the mutated side chain from internal coordinates using the remaining peptide backbone atoms. No significant changes in the pucker of the ribose were exhibited between the OMP structure and the BMP structure.

2.3.2 Constructing force–field parameters for OMP

An initial structure and molecular mechanics parameters for the substrate orotidine–5′-monophosphate (OMP) were constructed based on the structure of BMP (present in the original

structure) and existing parameters for the related thymine derivative Gamma 2–carboxy phenyl GA CDCA amide (2XBD) [45]. The full OMP structure was optimized by quantum mechanical optimization at the RHF/6–31G* level and partial atomic charges were assigned with RESP [54, 55]. The carboxylate group (CX, OX1, OX2) was found to be co-planar with the orotidyl ring after optimization.

2.3.3 Equilibration of OMPDC and substrate

All molecular mechanics and dynamics calculations were carried out with the CHARMM program package [46, 47]. Energy minimization using adopted basis Newton–Raphson (ABNR) [46] was performed to reduce poor contacts for wild–type and mutant structures, progressively relaxing hydrogen atoms, side chains and substrate, and then finally the backbone. All crystallographic water molecules present in the original crystal structure were preserved in the active–site preparation, and the entire system was solvated with a pre–equilibrated rhombic dodecahedron box (75 Å x 100 Å x 75 Å) of TIP3P water molecules [37]. All water molecules were treated with SHAKE [56]. Twenty–four potassium counterions were randomly placed in the box to restore charge neutrality. The neutral cell of protein and water complex was heated to 298 K over the course of 100 ps using a 1–fs timestep, subjected to NPT molecular dynamics at one atmosphere pressure with Leapfrog integration. A subsequent equilibration run of 100 ps without heating was performed [57]. Periodic boundary conditions with Ewald summation was applied for all molecular dynamics (MD) simulations. Electrostatic interactions were computed with particle–mesh Ewald summation using a real space cutoff of 14 Å and 1 Å grid spacing.

2.3.4 Umbrella Sampling and PMF Construction

Umbrella sampling was performed with CHARMM's SQUANTUM QM/MM implementation at the level of AM1 in order to both allow for electronic descriptions of bond–breaking and formation of the decarboxylation and proton–transfer coordinates, and to balance accuracy and speed [46, 47, 58]. The quantum mechanical region included the side groups, all from the active site primarily comprised of chain A, starting from C_β, of K42, D70, K72, D75*, and the orotidyl group of the OMP substrate from the N1 amide of the ring, where boundary atoms across the $C_\alpha - C_\beta$ bond were treated with the Generalized Hybrid Orbital (GHO) method [34].

Explicit treatment of umbrella sampling potentials and metrics were treated with the RXNCOR module of CHARMM. A pair of coordinates was chosen to drive the chemical reaction, corresponding to the observed chemistry: (1) the decarboxylation coordinate corresponding to breaking the C6–CX bond (and equal to that bond length), and (2) a coordinate for the proton transfer from the ammonium group of K72 to the C6 carbon of the orotate ring (and equal to the breaking bond length minus the making bond length). In other words, the definition of the decarboxylation coordinate was cast as:

$$\lambda_1 = distance(C_6 - CX)$$

and the proton transfer coordinate as

$$\lambda_1 = distance(NZ_{K72} - HZ1_{K72}) - distance(C_6 - HZ1_{K72})$$

Henceforth, these coordinates are represented as an ordered pair (λ_1, λ_2) (Figure 2.3). Umbrella sampling simulations were performed across the two–dimensional space of these reaction coordinates (at 0.1Å spacing for the decarboxylation coordinate and 0.2Å spacing for the proton–transfer coordinate) with harmonic restraint of 80 kcal/Å^2 with the use of CHARMM's RXNCOR module. To maintain the integrity of the proton–transfer coordinate and prevent transfer of a

different amide proton (HZ2 and HZ3), the protons that formed hydrogen bonds with proximal aspartates near K72 were restrained to K72's amide nitrogen, NZ, with an additional harmonic potential of 200 kcal/Å^2. Each simulation began with the same template equilibrated structure, and was perturbed iteratively in increments of 200 fs until the desired umbrella's center was attained for both reaction coordinates. Once the center was attained, each simulation sampled 50,000 timesteps of production dynamics. To construct the free energy profile as a function of these two reaction coordinates, the use of the Weighted Histogram Analysis Method (WHAM) was employed [31].

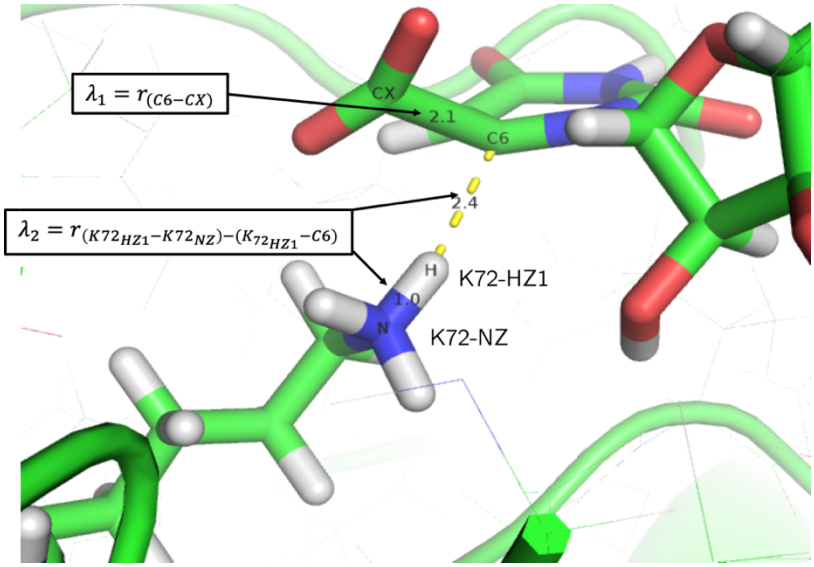


Figure 2.3: Schematic of reaction coordinates used in potential of mean force calculations to characterize the decarboxylation landscape. As indicated, the decarboxylation reaction coordinate (λ_1) was the C6–CX distance, and the proton–transfer reaction coordinate (λ_2) combined the distance between proton HZ1 of K72 and the amide nitrogen NZ of K72 versus the distance between the same K72–HZ1 and the orotidyl ring C6. In the illustrated example, λ_1 = 2.1 Å and λ_2 = 1.0 Å – 2.4 Å = –1.4 Å.

2.3.5 TPS Procedure

Transition path sampling methods enabled inquiry into the reactive pathways of WT and mutants used to catalyze decarboxylation. In particular, it allowed an exploration of the routes by which WT and mutants approached the reaction barrier independent of equilibrium assumptions. A path sampling formulation allows the generation of transition ensembles that reflect the native probability distribution of reactivity, without the need of defining an actual reaction coordinate. The use of this method allowed for analysis for subsequent, successive transitions from a reactant–bound state to a product–bound state and enabled the calculation of rate constants (k_{cat}) for WT and mutants. For TPS, an order parameter was required to define the reactant and product basins. The order parameter was defined as the breaking bond decarboxylation coordinate, C6–CX, where the reactant basin definition was the interval [1 Å, 1.7 Å] and the product basin definition was [2.5 Å, 5 Å]. All path sampling dynamics were performed with the same quantum region as indicated in umbrella sampling, also at the AM1 level of theory. All harmonic restraints that were placed on K72's protons HZ2 and HZ3 were removed.

TPS calculates rate by combining two quantities: the frequency factor, $\dot{v}(t)$, and the probability factor, P, which are detailed below [59]. Briefly, the frequency factor accounts for the rate by which trajectories enter the product basin, across all successful catalytic events. This contrasts the P–calculation, which explicitly probes the ratio of trajectories that exactly end, within a given time–frame, within a window's limits. The actual probability factor is the integrated probability for all values of the order parameter corresponding to the product basin, and it represents the likelihood of a trajectory starting in the reactant basin ending in the product basin within a fixed amount of time. The product of the frequency factor and the probability factor yields the rate:

$$k_{cat}^{TPS} = (\dot{v} \text{ in linear regime}) \times (\text{area under } P \text{ curve})$$

2.3.6 Seed Trajectory

An initial path connecting the reactant–bound state with the product–bound state was obtained by extracting coordinate and velocity frames from the umbrella simulations near the hypothesized transition–state (TS) region and subsequently used to create starting structures to perform forward and backward integration [REF]. Each frame was seeded with random momenta drawn from a Boltzmann distribution centered at 298 K. The aforementioned initial seed trajectory regions were identified for WT, S127A, and V155D using all umbrella sampling frames from (2.2 to 2.5 Å, –1.8 to –1 Å), (2 to 2.5 Å, –2 to –1 Å), and (2.2 to 2.7 Å, –2 to –1 Å), using the decarboxylation, proton–transfer coordinate dimensions respectively. One seed trajectory was selected for each protein model for subsequent TPS calculations. Seed trajectories were a total of 2000 timesteps, integrated for 1000 timesteps forward and backward from the frame chosen, utilizing Leapfrog NPT dynamics at a timestep of 1 fs.

2.3.7 Frequency factor: $\dot{\nu}$ Calculation

The $\dot{\nu}$–calculation path ensemble used a path length of 601 fs, as this ensured that over 90% of the trajectories were able to successfully reach the product basin by the end of the timespan, hence longer than the molecular transit time. To obtain the path ensemble for each protein, we attempted 2000 shooting and shifting moves with equal probability, where 50% of shooting moves were allowed as half–shoots equally likely in both directions [39, 40, 59]. Perturbed path velocities were provided with the use of Langevin integrator, capitalizing on the Langevin perturbation. The maximum shift allowed was 50 fs, and shooting locations were restrained to the last 50 fs of the prior accepted move. To compute the $\dot{\nu}$ term, 200 independently computed ensembles were

constructed, and a first–order polynomial fit was applied to the linear regime of each curve using *polyfit* in Python's NumPy package [60].

2.3.8 Probability factor: P Calculation

Computing the P–ensemble required partitioning the order parameter into windows with overlap to compute the relative probability of progressively traversing the reaction barrier. Windows were constructed in 0.1 Å intervals of the order parameter, starting from 1.6 Å to 3.45 Å, allowing a 0.05 Å overlap with a prior and subsequent window. A boundary window to describe the reactant and product basins was set from 1 Å to 1.6 Å and 3.45 Å to 5 Å. For each window, 80% of all valid moves were shooting moves, of which we chose 20% as half–shoots equally likely in either direction, and the remaining 20% of all valid moves were shifting moves. The very last window from 3.45 to 5 Å was run with a shooting/shifting ratio at 50%/50%, as we observed this generated the greatest diversity in trajectories, which was critical for downstream analyses of reactive paths. As indicated in the v–calculation, Langevin dynamics provided the new momenta assigned to a frame to explore a novel path. Each window sampled 2000 moves each with an accepted trajectory required to end at the edges of a window after 651 fs. Mutants were run with 10 replicate windows for a total of 420 total independent simulations for each mutant, and the WT was run with 5 replicates. To calculate the P–factor, windows were re–weighted using the overlap–distribution of the preceding window such that reweighted windows could be aggregated to compute the overall probability distribution. The normalized P–factor was computed by integrating over all values in the product basin.

2.3.9 Visitation probability of sampled TPS paths for decarboxylation and proton transfer

To compare dynamical TPS ensembles to the PMF energetic surface, we track how the reaction coordinates used in the PMF evolved over the course of the dynamic simulations. For each path in the last window of the TPS ensemble (corresponding to the formation of product), we extracted the decarboxylation coordinate (C6–CX distance), the distance of each proton of K72's amide group to the amide nitrogen (NZ–HZ1, NZ–HZ2, NZ–HZ3), and the distance of each proton of K72's amide group to the orotidyl ring carbon (C6–HZ1, C6–HZ2, C6–HZ3). In order to compare with the reaction coordinates of the PMF, two order parameters were calculated for each path. The decarboxylation coordinate was defined as such:

$$\lambda_1 = distance(C_6 - CX)$$

As the dynamic paths employ no biasing potentials or restraints, the proton transfer coordinate was defined to account for the closest of K72 protons to C6 as such:

$$\lambda_2 = \max_{i\in\{1,2,3\}} \left[distance\left(NZ_{K72} - HZ_{i_{K72}}\right) - distance\left(C6 - HZ_{i_{K72}}\right)\right]$$

We define the visitation probability as the probability that a trajectory would visit some value of the pair of coordinates (λ_1, λ_2). Using the same grid as the PMF, we counted the number of trajectories that entered a given (λ_1, λ_2) bin at least once.

2.4 Results and Discussion

2.4.1 Energetic landscapes of WT and mutants suggests two possible mechanisms for decarboxylation

Before studying dynamic reactive pathways across wild–type and mutant OMPDC systems, the energetic landscapes of WT and mutant were characterized using QM/MM simulations and umbrella sampling to compute a PMF. While previous studies describe the catalysis as decarboxylation dependent, the inclusion in our quantum mechanical region of the catalytic residues K72, D70, K42, and D75* in addition to the orotidyl ring of the substrate permitted protonation of the substrate ring; therefore, the energetics were modeled as a function of both the decarboxylation coordinate and the proton–transfer coordinate. First, to account for bond–breaking reaction, the decarboxylation reaction coordinate (λ_1) included the ring carbon (C6) and the carboxylate carbon (CX). Second, to account for proton transfer, a combined reaction coordinate (λ_2) was employed, comprising the difference between the breaking bond of the lysine's proton (K72) and the forming bond between the ring carbon and the lysine proton: explicitly, the was the distance from K72–HZ1 to K72–NZ, and the distance from K72–HZ1 to OMP–C6. HZ1 was chosen because this proton did not coordinate the adjacent carboxylate groups of the aspartates post–equilibration. The other two protons of K72's amide group were physically restrained to prevent transfer, thus ensuring the proton–transfer coordinate indeed reflected that of the transferring proton. For the proton–transfer coordinate, negative values correspond to the state with the proton still attached to the lysine, whereas positive values indicate that the proton partially or fully transferred to the ring carbon C6, with a value of approximately 1 Å corresponding to a successful protonation onto the ring. Umbrella simulations for window pairs of the decarboxylation and proton–transfer coordinate, (λ_1, λ_2), allowed us to obtain a smooth 2–

dimensional histogram of the free energy (PMF) of WT and mutant systems as a function of the two reaction coordinates (Figure 2.4).

The PMFs of the WT and mutants exhibited similar reactant basins and two product basins compatible with the common definition of reactivity that, strictly speaking, only includes decarboxylation (Figure 2.5). In all PMFs, the reactant basin appeared in the lower left around (λ_1, λ_2) = (1.5 Å, –3.1 Å). Moreover, in all PMFs, there existed a 'lower–right' basin with λ_1 around 3.3 Å to 3.5 Å and λ_2 between –2.5 Å to –1.7 Å suggesting a local minimum. This basin corresponded to a decarboxylation event in which the decarboxylation coordinate was sufficiently extended to be considered broken, yet the system had not yet transferred a proton to create the UMP molecule. Literature reported that the proton transfer to the ring is faster than the decarboxylation event, suggesting the decarboxylation is rate limiting [27, 53]. In all PMFs, two valid product definitions in the context of the simulations included an 'upper–right' basin, corresponding to the final, expected observable product state – a protonated ring substrate in addition to the formation of carbon dioxide due to decarboxylation, and a 'lower–right' basin, corresponding to an intermediate where decarboxylation occurred but the ring was unprotonated. The energetics of the upper–right (protonated) product basin were comparable among WT and mutants, but the lower–right (unprotonated) basin was 4 to 10 kcal/mol higher in energy for mutants S127A and V155D respectively. For WT and both mutants, the upper–right basin was lower in energy than the lower–right basin.

Between the reactant and product states, particularly for WT and S127A systems, we observed a diffuse transition zone that indicated two potential reaction path hypotheses compatible with decarboxylation. Coupled with the existence of the two basins, this suggests the following two scenarios: (1) a reaction path may be free to react independent of the proton transfer

coordinate, entering the lower–right basin by first successfully decarboxylating, and then inevitably protonating in a stepwise manner; or (2) the carboxyl group may transfer in a simultaneous manner, where K72 may play a role in stabilizing a potential transition–state (TS) and enabling reactivity. For the WT enzyme, the transition state was (2.1 Å, –1.2 Å), in terms of the decarboxylation and proton–transfer coordinates, respectively, with estimated barrier height of 21 kcal/mol above the reactant state. Notably, the path from reactant to the lower–right basin was comparable energetically. Similarly, the S127A mutant exhibited a familiar landscape with increased activation barrier, in which the expected transition–state structure was (2.2 Å, –1.4 Å) at about 24 kcal/mol above the bound substrate. In contrast, the V155D mutant exhibited the most dramatic difference in PMF compared to the WT and S127A mutants. Inspection of its PMF indicated that the second scenario (i.e. the stepwise pathway) appeared most energetically favorable. The expected transition state of this mutant was near (2.4 Å, –1.6 Å), further along in the decarboxylation coordinate and protonation coordinate than either WT or S127A. The V155D mutant's corresponding transition structure was also energetically higher, at 27 kcal/mol above the bound substrate.

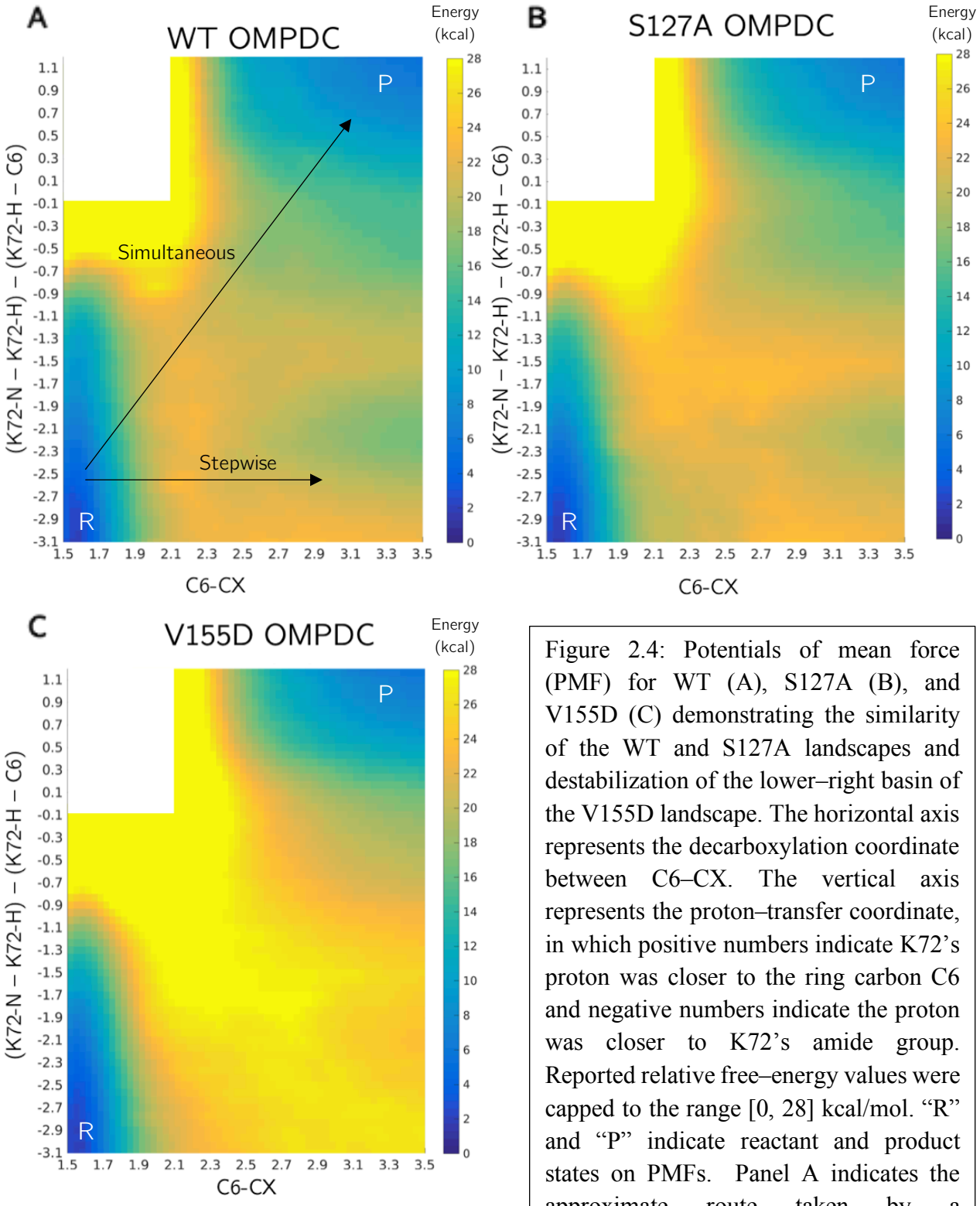


Figure 2.4: Potentials of mean force (PMF) for WT (A), S127A (B), and V155D (C) demonstrating the similarity of the WT and S127A landscapes and destabilization of the lower–right basin of the V155D landscape. The horizontal axis represents the decarboxylation coordinate between C6–CX. The vertical axis represents the proton–transfer coordinate, in which positive numbers indicate K72's proton was closer to the ring carbon C6 and negative numbers indicate the proton was closer to K72's amide group. Reported relative free–energy values were capped to the range [0, 28] kcal/mol. "R" and "P" indicate reactant and product states on PMFs. Panel A indicates the approximate route taken by a simultaneous or stepwise mechanism.

2.4.2 Predicted energetics of the transition state match experimental values in relative order but not absolute magnitude

The experimental rate constants were identified by Iiams et al. (2011) [27], who measured the k_{cat} for the WT and mutants. This enabled the use of the Eyring approach to estimate expected energetics of barrier–height (activation energy), given the empirical rate constants [61–63]. The experimentally derived activation energies were then compared to the predicted barrier height, provided by the transition–state structure estimates above (Table 1). While the energetics were 4–5 kcal/mol higher than the predicted empirical values for WT and prior umbrella sampling by Vardi–Kilshtain [18], the barriers were consistently ordered from WT to mutants. The activation energy was systematically higher for each catalytically impaired mutant compared to WT, and the relative separation of the energetics is in good agreement to the empirical literature.

Protein variant	Experimental rate (s^{-1})	Experimental barrier height (kcal/mol)	Predicted barrier height (kcal/mol)	Expt. Relative k_{cat}	Predicted Relative k_{cat}
WT	4.0 ± 0.2	16.5	21	1	1
S127A	$3.5 \times 10^{-2} \pm 0.002$	19.4	24	8.8×10^{-3}	6.4×10^{-3}
V155D	$5 \times 10^{-4} \pm .0002$	21.9	27	1.3×10^{-4}	4.0×10^{-5}

Table 1: Comparison of the transition–state barrier height from experiments (Iiams et al. 2011 [27]) and from the potentials of mean force calculations of WT and mutant systems of MtOMPDC. Eyring rates were applied to convert k_{cat} measurements from Iiams et al. to approximate the barrier–height of WT and its mutants, assuming the transmission coefficients were the same for WT and mutants. The predicted barrier heights from PMF calculations indicate that the calculations were systematically 4–5 kcal/mol higher than experiment, but with the correct relative spacing among mutants to preserve the fold–reduction in rate. Relative rates were scaled with the WT as reference.

2.4.3 Reaction rates obtained from transition path sampling match experimental rates in relative order

Transition path sampling (TPS) methods, with their explicit treatment of reaction dynamics, were then subsequently employed to further explore the routes that WT and mutants take in order to traverse the reaction barrier. Although the PMF results suggested the existence of a protonation–independent pathway, TPS methods overcome the limitations of umbrella sampling methods, and allow elucidation of dynamic pathways.

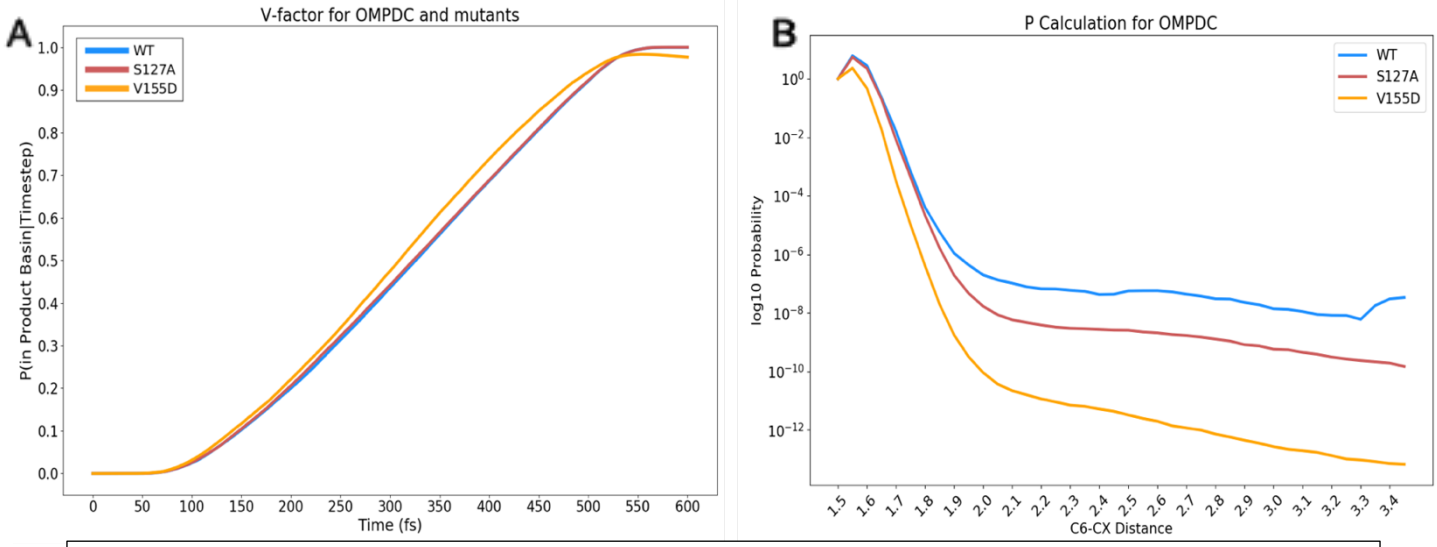


Figure 2.5: Components of the TPS rate calculation, where (A) represents the $\dot{\nu}$ calculation or frequency factor and (B) represents the P calculation or probability factor. (A) The $\dot{\nu}$ measures the rate at which trajectories enter the defined product basin (decarboxylation reaction coordinate, $\lambda_1 \geq$ 2.5 Å) over time, as indicated by the slope of the line in the plot provided. The WT and S127A mutant possessed the same slope value (2.3 x 10^{12} s^{-1}), as indicated by the linear regime between 100 fs and 525 fs. The V155D mutant possessed a slope that was nearly 5% larger (2.4 x 10^{12} s^{-1}) for the same indicated regime. Before the linear regime, there was a lag of about 100 fs, which suggests the minimum time required for the order parameter to reach the product basin is at least 100 fs. (B) P–calculation represents the probability of a trajectory reaching a certain window of the order parameter within a fixed time (651 fs). The WT and S127A mutant show similar likelihoods of reaching an order parameter stretched to 1.8 Å, but then differ beyond that: the WT is consistently more likely to reach the subsequent windows (i.e. larger values of order parameter) than the S127A mutant. The V155D mutant is consistently hampered and less likely to reach extended values of the order parameter than the WT and S127A mutant, both below and beyond the 2.1 Å distance.

The frequency factor for the WT and S127A were both 5% smaller than for V155D. The frequency factor is related to the average transition time, which in Eyring theory is treated as a function of the breaking bond's (C6–CX) vibrational frequency (Figure 2.5A). The small difference between WT and mutant frequency factors has been observed in other enzymes [51, 52]. For the P–factor, the WT was more effective at reaching the product state compared to the mutants. Notably, the likelihood of extending the CX–C6 bond remained comparable for S127A and WT until 1.8 Å, at which point the WT became more likely to reach further windows toward the product side (Figure 2.5B). The V155D mutant was compromised even at much smaller values of the breaking bond, being less likely to reach every window from the reactant basin definition.

The computed rates from TPS were faster than both the expected rates computed by umbrella sampling and the experimental benchmarks (Table 2). While the predicted rates were larger when compared to experiment, there was a systematic ordering between the WT and mutants with regard to rate. Encouragingly, the relative rates of the TPS calculations match the empirical results well, with the S127A nearly two orders of magnitude slower and the V155D mutant nearly 4 orders slower than the WT.

System	$\dot{v}(t)$ (s^{-1})	P	Computed k_{cat} (s^{-1})	Expt. k_{cat} (s^{-1})	Predicted Relative k_{cat}	Expt. Relative k_{cat}
WT	2.3×10^{12}	2.2×10^{-8}	5.1×10^{4}	$4.0 \pm .02$	1	1
S127A	2.3×10^{12}	1.6×10^{-9}	3.7×10^{3}	$(3.5 \pm 0.2) \times 10^{-2}$	7.3×10^{-2}	8.8×10^{-3}
V155D	2.4×10^{12}	2.0×10^{-12}	4.8×10^{0}	$(5 \pm 2) \times 10^{-4}$	8.4×10^{-5}	1.3×10^{-4}

Table 2: Reaction rates calculated from TPS match systematically with relative rates computed from empirical characterization by Iiams et al. [27]. The frequency factor, $\dot{v}(t)$, indicates the slope of the v–calculation line for WT and mutant systems. The P factor indicates the normalized likelihood of reaching product basin given that a path has exited the reactant basin. The predicted k_{cat} is the product of the frequency factor and the P factor, contrasted with the experimentally reported k_{cat} from Iiams et al. [27]. Relative rates are scaled to the WT for both computed and experimental k_{cat} values.

2.4.4 Analysis of the visitation probability from productive trajectories of WT and mutant OMPDC suggests that the V155D mutant is more likely to decarboxylate independently of K72

To describe the dynamical route by which each system successfully decarboxylates (i.e. break the C6–CX bond), the ensemble of trajectories from the last window of TPS simulations was analyzed, as these represent full reactive trajectories. For each simulation in the WT and each mutant ensemble, a decarboxylation coordinate λ_1 and a proton transfer coordinate λ_2 for each proton of K72 was computed, where the K72 proton with the closest proximity to C6 was used to define the proton–transfer coordinate for any given timepoint. Using the same grid spacing for the proton–transfer and decarboxylation coordinates as the PMF, we computed the number of trajectories that visited a grid point at least once, and then scaled by the total number of trajectories, to estimate the distribution of decarboxylation paths the ensemble took.

This analysis revealed that the WT and S127A mutant exhibited similar paths; most trajectories sampled a path wherein the proton transfer coordinate increased linearly while the decarboxylation coordinate stretched (Figure 2.6). Curiously, for the WT only about 50% of the trajectories sampled the PMF–derived transition state at $(\lambda_1, \lambda_2) = (2.1\ \text{Å}, -1.2\ \text{Å})$, but over 60% of the S127A mutant trajectories sampled its PMF–derived transition state at (2.2 Å, –1.4 Å) (Figure 2.S1). For both the WT and S127A mutant, no trajectories within the ensemble were able to cross into the product state (in which the decarboxylation coordinate stretched to 2.5 Å) with the proton transfer coordinate below –2.7 Å (Figure 2.S1). Fewer than 5% of trajectories for either WT or S217A were able to enter the corresponding lower–right basin of the PMF with $\lambda_1 \geq 3.3$ Å and $\lambda_2 \in (-2.5\ \text{Å}, -1.7\ \text{Å})$ (Figure 2.S1). Contrarily, the V155D mutant exhibited a different landscape than either WT or the S127A mutant; unlike either WT or mutant, the visitation

probability of the proton–transfer coordinate concurrently changing with the decarboxylation coordinate was much lower. Curiously, about 40% of trajectories sampled the transition state of (2.4 Å, –1.6 Å). Additionally, some trajectories did cross into the product state with the proton–transfer coordinate below –2.7 Å, and at least 20% of trajectories in the V155D ensemble explicitly visited the lower basin of the PMF. These results indicate that the V155D mutant takes different paths to decarboxylate than the WT and S127A OMPDC.

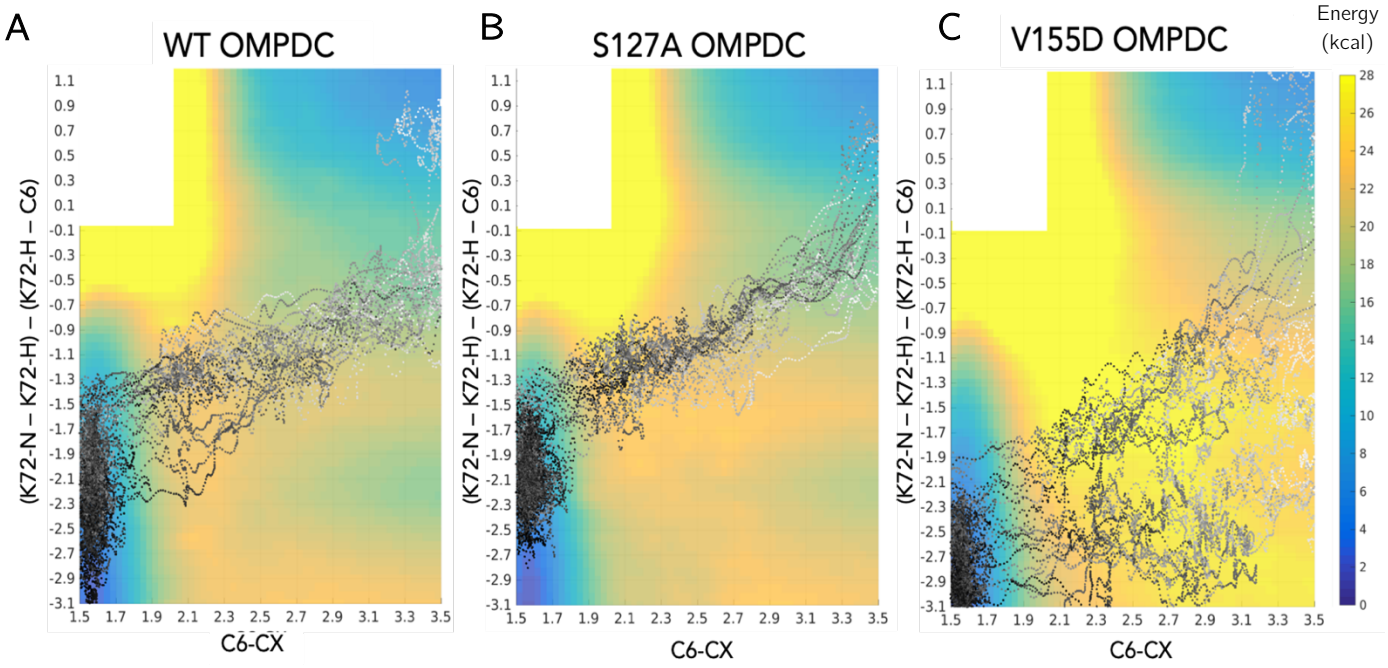


Figure 2.6: Time points corresponding to the decarboxylation and proton–transfer coordinate as a function of time overlaid on the PMFs for WT (A), S127A (B), and V155D (C). In order to denote time in the trajectories, the color saturates from darker to lighter, with black indicating the start of the trajectory and white its end. The above illustrates a randomly selected subset of the trajectories within each ensemble.

2.4.5 Analysis of proton transfer coordinate as a function of the decarboxylation coordinate

For each trajectory, we stratified the C6–CX order parameter in 0.1–Å increments and monitored the proton–transfer coordinate distribution for all time steps within a trajectory

considered within the stratum (Figure 2.7). This stratification allowed inquiry into how the proton of K72 may approach C6 as a function of the decarboxylation coordinate. Notably, for any stratum of the order parameter C6–CX, the distribution between all pairs of the WT, S127A, and V155D mutants had a p–value less than 1×10^{-5}, suggesting the ensembles were indeed distinct distributions. Notably, for C6–CX values within the reactant basin (C6–CX < 1.7 Å), V155D proton–transfer coordinate was shifted left to –3 Å, nearly 1 Å further left than the WT or S127A ensemble, each centered between (–2.3 Å to –2.1 Å). As the reaction proceeded and C6–CX stretched, all distributions shifted to the right, but the WT and S127A distributions remained relatively tight compared to that of V155D. Consistent with a concurrent transfer mechanism, the proton–transfer coordinate distributions for the WT and S127A ensembles were unimodal and shifted to the right as decarboxylation occurred. By contrast, the V155D proton–transfer distribution exhibited a local peak around –2.9 Å for multiple strata of the decarboxylation coordinate; while some trajectories transfer the proton, this particular peak did not disappear until the decarboxylation coordinate exceeds approximately 3.0 Å. This analysis showed how the proton–transfer coordinate depended on the decarboxylation coordinate in a manner that differs between systems. Specifically, the V155D ensemble's trajectories more frequently decarboxylated in paths such that the proton–transfer coordinate did not change concurrently with the decarboxylation coordinate, as compared to the WT and S127A mutant.

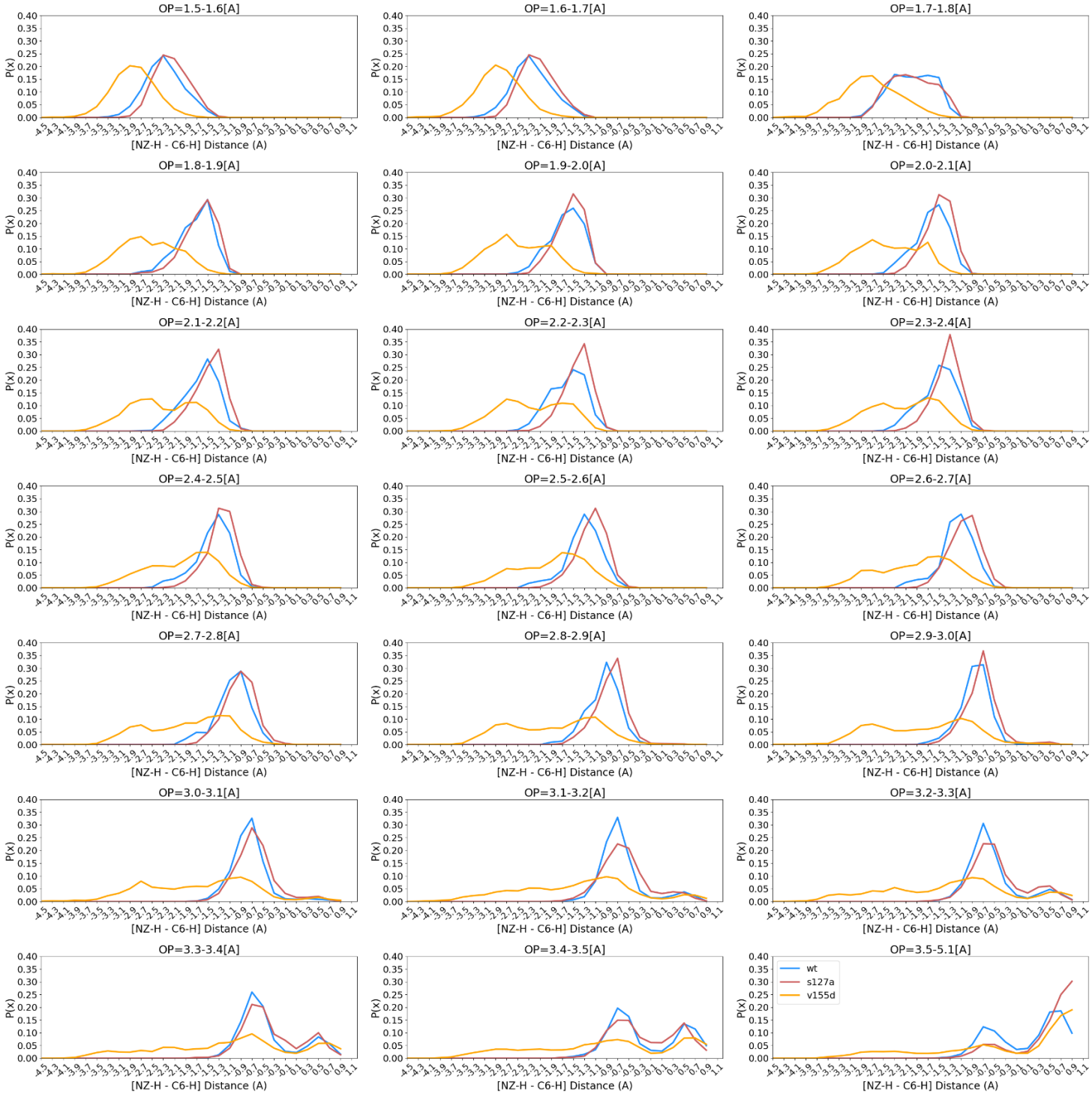


Figure 2.7: Dependence of the proton–transfer coordinate distribution on decarboxylation coordinate, for WT, S127A, and V155D, among TPS–sampled trajectories that achieve successful decarboxylation (i.e. for which the decarboxylation coordinate stretches from at least 1.7 Å to up to 5 Å.).

2.5 Conclusion and Future Directions

Considerable theoretical and experimental inquiry has explored the origin of catalytic proficiency of the OMPDC enzyme. Consensus in the field suggests decarboxylation as rate–limiting, but provides compatible evidence for several mechanistic strategies couched in the language of Transition–State Theory. In this study, we analyzed energetic and dynamic contributions to OMPDC catalysis for wild type and two catalytically impaired variants, S127A and V155D.

Comprehensive analysis of the PMF profiles and dynamical trajectories describing the decarboxylation of WT and mutant enzymes indicated the possibility of two possible catalytic strategies: (1) decarboxylation that occurs concurrently with a proton from K72 approaching C6 and (2) decarboxylation independent of K72's proton position. The energetic landscape of the PMFs are in reasonable agreement with both theory and experiment [18, 27] but suggest the V155D mutant may prefer to decarboxylate independently of the position of K72. We subsequently applied path–sampling strategies to investigate how dynamics may play a role in the catalysis of both the WT and mutant reactions. Notably, the concurrent route was sampled in both WT and each mutant system, but the WT and S127A mutant showed strong preference for a K72–assisted decarboxylation, whereas the V155D showed a lesser preference and frequently exhibited proton–independent decarboxylation.

The catalytic strategies presented are compatible with the prevailing belief that decarboxylation occurs prior to proton transfer. While all ensembles possessed some fraction of trajectories that participated in K72–assisted decarboxylation, no trajectories transferred the proton before the decarboxylation coordinate reached 2.5 Å, suggesting that bond–breakage occurred prior to proton abstraction by C6. The existence of the lower–right hand basin, in which

decarboxylation is assured but K72 proton transfer has not occurred, may correspond to a state in which nearby solvent molecules provide the necessary stabilization and proton required for catalysis. One potential extension of the work is to include quantum–mechanically described solvent molecules in the active site to permit the formation or destruction of a bond facilitated by water. Another potential source of further investigation is to quantify the pK_a of K72 for WT and variant systems, which has been reported to be around 8.0 for OMPDC homologs [6].

In the present work, a single seed trajectory obtained from the resulting PMFs for WT and mutant systems was employed to launch the ensembles. In follow–up work, we would initialize ensembles from multiple different seeds to ascertain the ergodicity of the sampling procedure.

Taken together, the methodologies employed and results obtained in this work illustrate the importance of the chemical environment in altering dynamics and reaction mechanism, particularly in explaining the catalytic proficiency of OMPDC and its hindered mutants. Such insights could be employed for enzyme–catalyzed reactions observed in other systems.

2.6 References

1. A. Radzicka, R. Wolfenden. A proficient enzyme. *Science*. 267, 90–93, 1995.
2. C.A. Lewis Jr, R. Wolfenden. Uroporphyrinogen decarboxylation as a benchmark for the catalytic proficiency of enzymes. *Proc. Natl. Acad. Sci USA* 105(45): 17328–17333, 2008.
3. T.L. Amyes, B.M. Wood, K. Chan, J.A. Gerlt, J.P. Richard. Formation and stability of a vinyl carbanion at the active–site of orotidine 5'–monosphophate decarboxylase: pKa of the C–6 proton of the enzyme–bound UMP. *J. Am. Chem. Soc*. 130, 1574–1575, 2008.
4. P. Beak, B. Siegel. Mechanism of decarboxylation of 1,3–dimethylorotic acid. A model for orotidine 5'–phosphate decarboxylase. *J. Am. Chem. Soc.* 98 (12): 3601–6, 1976.
5. R.B. Silverman, M.P. Groziak. Model chemistry for a covalent mechanism of action of orotidine 5'–phosphate decarboxylase. *J. Am. Chem. Soc.* 104 (23): 6434–6439, 1982.
6. K.K. Chan, B.M. Wood, A.A. Fedorov, E.V. Fedorov, H.J. Imker, T.L. Amyes, J.P. Richard, S.C. Almo, J.A. Gerlt. Mechanism of the orotidine 5'– monophosphate decarboxylase–catalyzed reaction: evidence for substrate destabilization. *Biochemistry* 48, 5518–553, 2009.
7. K.N. Houk, J.K Lee, D.J. Tantillo, S. Bahmanyar, B.N Hietbrink. Crystal structures of orotidine monophosphate decarboxylase: does the structure reveal the mechanism of nature's most proficient enzyme? *Chem BioChem* 2, 113–118, 2001.
8. M.A. Rishavy, W.W. Cleland. Determination of the mechanism of orotidine 5'–monophosphate decarboxylase by isotope effects. *Biochemistry*. 39 (16), 4569–4574, 2000.
9. K. Toth, T.L. Amyes, B.M Wood, K. Chan, J.A. Gerlt, J.P. Richard. Product deuterium isotope effects for orotidine 5'–monophosphate decarboxylase: effect of changing substrate and enzyme structure on the partitioning of the vinyl carbanion reaction intermediate. *J. Am. Chem. Soc.* 132, 7018–7024, 2010.

10. J.L. Van Vleet, L.A. Reinhardt, B.G. Miller, A. Sievers, W.W. Cleland. Carbon isotope effect study on orotidine 5'–monophosphate decarboxylase: support for an anionic intermediate. *Biochemistry*. 47, 798–803, 2008.
11. K. Toth, T.L. Amyes, B.M. Wood, K. Chan, J.A. Gerlt, J.P. Richard. Product deuterium isotope effect for orotidine 5'–monophosphate decarboxylase: evidence for the existence of a short–lived carbanion intermediate. *J. Am. Chem. Soc.* 129, 12946–12947, 2007.
12. C.A. Lewis Jr, R. Wolfenden. Orotic acid decarboxylation in water and nonpolar solvents: a potential role for desolvation in the action of OMP decarboxylase. *Biochemistry*. 48 (36), 8738–8745, 2009.
13. B. Goryanova, K. Spong, T.L. Amyes, J.P. Richard, *Biochemistry*. 52, 537–546. 2013.
14. A. Warshel, M. Strajbl, J. Villa, J. Florian. Remarkable rate enhancement of orotidine 5‘–monophosphate decarboxylase is due to transition–state stabilization rather than to ground–state destabilization. *Biochemistry*. 39: 14728–14738, 2000.
15. N. Wu, Y. Mo, J. Gao, E.F Pai. Electrostatic stress in catalysis: Structure and mechanism of the enzyme orotidine monophosphate decarboxylase. *Proc. Natl. Acad. Sci USA* 97, 2017–2022, 2000
16. J. Gao. Catalysis by enzyme conformational change as illustrated by orotidine 5′–monophosphate decarboxylase. *Curr. Opin. Struct. Biol.* 13, 184–192, 2003
17. B.G. Miller, R. Wolfenden. Catalytic proficiency: the unusual case of OMP decarboxylase. *Annu. Rev. Biochem.* 71: 847–885, 2002.
18. A. Vardi–Kilshtain, D. Doron, D. Major. Quantum and classical simulations of orotidine monophosphate decarboxylase: support for a direct decarboxylation mechanism. *Biochemistry*. 52 (25), 4382–4390, 2013.

19. H. Hu, A. Boone, W. Yang. Mechanism of OMP decarboxylation in orotidine 5′–monophosphate decarboxylase. *J. Am. Chem. Soc.* 130(44), 14493–503, 2008.

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

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

22. N. Wu, W. Gillon, E.F. Pai. Mapping the Active–Site Ligand Interactions of Orotidine 5′-monophosphate Decarboxylase by Crystallography. *Biochemistry*. 41 (12), 4002–4011, 2002.

23. A.A Federov, E.V. Federov, B.M Wood, J.A. Gerlt, S.C Almo. Conformational changes in orotidine 5′-monophosphate decarboxylase: "remote" residues that stabilize the active conformation. *Biochemistry*. 49; 3514–3516, 2010.

24. P. Harris, J.N. Poulsen, K.F. Jensen, S. Larsen. Structural basis for the Catalytic Mechanism of a Proficient Enzyme: Orotidine 5′-monophosphate Decarboxylase. *Biochemistry*. 39, 4217–4224, 2000.

25. T.W. Traut, B.R. Temple. The chemistry of the reaction determines the invariant amino acids during the evolution and divergence of orotidine 5′–monophosphate decarboxylase. *J. Biol. Chem.* 275: 28675–81, 2000.

26. T.L. Amyes, S.A. Ming, L.M Goldman, B.M Wood, B.J. Desai, J.A. Gerlt, J.P. Richard. Orotidine 5′–monophosphate decarboxylase: transition–state stabilization from remote protein–phosphodianion interactions. *Biochemistry*. 51(23), 4630–4632, 2012.

27. V. Iiams, B.J. Desai, A.A Fedorov, E.V. Fedorov, S.C Almo, J.A. Gerlt. Mechanism of the orotidine 5′–monophosphate decarboxylase–catalyzed reaction: Importance of residues in the orotate binding site. *Biochemistry*. 50(39): 8497–8507, 2011.

28. J. Yuan, A.M. Cardenas, H.F. Gilbert, T. Palzkill. Determination of the amino acid sequence requirements for catalysis by the highly proficient orotodine 5′-monophosphate decarboxylase. *Protein Sci*. 20; 1891–1906, 2011.

29. S.A. Barnett, T. L. Amyes, B.M. Wood, J.A. Gerlt, J. P. Richard. Dissecting the total transition state stabilization provided by amino acid side chains at orotidine 5′–monophosphate decarboxylase: A two–part substrate approach. *Biochemistry*, *47*(30), 7785–7787, 2008.

30. W.P. Jencks. Advances in Enzymology and Related Areas of Molecular Biology; *J Wiley & Sons, Inc: New York*; Vol. 43. 1975

31. S. Kumar, J.M. Rosenberg, D. Bouzida, R.H. Swendsen, P.A. Kollman. The weighted histogram analysis method for free–energy calculations on biomolecules. *J. Comput. Chem*. 13, 1011−1021, 1992.

32. B.J. Desai, M. Wood, A.A. Federov, E.V. Federov, B. Goryanova, T.L. Amyes, J.P. Richard, S.C. Almo, J.A. Gerlt. Conformational changes in orotidine 5′-monophosphate decarboxylase: A structure–based explanation for how the 5'–phosphate group activates the enzyme. *Biochemistry*. 51, 43, 8665–8678, 2012.

33. H.M Senn, W. Thiel. QM/MM methods for biomolecular systems. *Angew. Chem. Int. Ed.,* 48: 1198–1229, 2009.

34. J. Gao, P. Amara, C. Alhambra, M.J. Field. A generalized hybrid orbital (GHO) Method for the treatment of boundary atoms in combined QM/MM calculations. *J. Phys. Chem. A*, 102 (24), 4714–4721, 1998.

35. T. Steinbrecher, M. Elstner. QM and QM/MM simulations of proteins. *Methods Mol. Bio.* 924, 91–124, 2012.

36. G.M Torrie, J.P Valleau. Nonphysical sampling distributions in Monte Carlo free–energy estimation: umbrella sampling. *J. Comput. Phys.,* 23, 187–199, 1977.

37. A.D MacKerell, D. Bashford, M. Bellott, R.L. Dunbrack, J.D. Evanseck, M.J Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph–McCarthy, L. Kuchnir, K. Kuczera, F.T. Lau, C. Mattos, S. Michnick, T. Ngo, D.T. Nguyen, B. Prodhom, W.E. Reiher, B. Roux, M. Schlenkrich, J.C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiórkiewicz–Kuczera, D. Yin, M. Karplus. All–atom empirical potential for molecular modeling and dynamics studies of proteins. *J Phys Chem B*. 102(18):3586–616, 1998

38. M. Souaille, B. Roux. Extension to the weighted histogram analysis method: combining umbrella sampling with free energy calculations. *Comput. Phys. Commun.;* 135, 40–57, 2001.

39. P.G. Bolhuis, D. Chandler, C. Dellago, P.L. Geissler. Transition Path Sampling: Throwing Ropes Over Rough Mountain Passes, in the Dark. *Annu. Rev. Phys. Chem*. 53: 291–318, 2002.

40. T.S. Van Erp, D. Moroni, P.G. Bolhuis. A novel path sampling method for the calculation of rate constants. *J. Chem. Phys.* 118 (17): 7762–7774, 2003.

41. J.T. Berryman, T. Schilling. Sampling rare events in nonequilibrium and nonstationary systems. *J. Chem. Phys.* 133 (24): 244101, 2010.

42. D.T. Major, J.L. Gao. A combined quantum mechanical and molecular mechanical study of the reaction mechanism and alpha–amino acidity in alanine racemase. *J. Am. Chem. Soc.* 128, 16345–16357, 2006.

43. D. Baker, An exciting but challenging road ahead for computational enzyme design. *Protein Sci.* 19(10): 1817–1819, 2010.

44. J.Z. Ruscio, J.E. Kohn, K.A. Ball, T. Head–Gordon, The influence of protein dynamics on the success of computational enzyme design. *J. Am. Chem. Soc.,* 131 (39):14111–14115, 2009.

45. J. Huang, A.D. MacKerell,. CHARMM36 all–atom additive protein force field: Validation based on comparison to NMR data. *J. Comput. Chem.*, 34(25): 2135–2145, 2013.

46. B.R. Brooks, R.E. Bruccoleri, B.D. Olafson, D.J. States, S. Swaminathan, M. Karplus,. CHARMM: A program for macromolecular energy, minimization and dynamics calculations. *J. Comput. Chem.* 4(2):187–217, 1983.

47. B.R. Brooks, C.L. Brooks, III, A.D. MacKerell, Jr., L. Nilsson, R.J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A.R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R.W. Pastor, C.B. Post, J.Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D.M. York, and M. Karplus. CHARMM: The Biomolecular Simulation Package. *J. Comput. Chem*, 30(10): 1545–1614, 2009.

48. S. Hur, T.C. Bruice. The near attack conformation approach to the study of chorismite to prephenate reaction. *Proc. Natl. Acad. Sci USA.*, 100(21) 12015–12020, 2003.

49. S.J. Benkovic, S.A. Hammes–Schiffer. Perspective on Enzyme Catalysis. *Science*. 301(5637), 1196–11202, 2003.

50. T. L Amyes, J.P. Richard, J.J Tait. Activation of Orotidine 5‘–Monophosphate Decarboxylase by Phosphite Dianion: The Whole Substrate is the Sum of Two Parts. *J. Am. Chem. Soc.* 127, 15708–15709, 2005.

51. N.W. Silver, Ensemble methods in computational protein and ligand design: Applications to the Fc–γ immunoglobulin, HIV– 1 protease, and ketol–acid reductoisomerase system. Doctoral Dissertation, Massachusetts Institute of Technology, 2012.

52. I.S. Patel, Large scale simulation and analysis to understand enzymatic chemical mechanisms. Doctoral Dissertation, Massachusetts Institute of Technology, 2015.

53. B.M. Wood, K.K. Chan, T.L Amyes, J.P Richard, J.A. Gerlt. Mechanism of the Orotidine 5´-monophosphate Decarboxylase Catalyzed Reaction: Effect of Solvent Viscosity on Kinetic Constants. *Biochemistry*. 48(24), 5510–5517, 2009.

54. C.I Bayly, P. Cieplak, W.D. Cornell, P.A. Kollman. A Well–Behaved Electrostatic Potential Based Method Using Charge Restraints for Deriving Atomic Charges: The RESP Model. *J Phys Chem*. 97, 10269–10280, 1993.

55. Gaussian 03, Revision C.02, Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, Jr., J. A.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.;

Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al–Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; and Pople, J. A.; *Gaussian, Inc.,* Wallingford CT, 2004.

56. J.P. Ryckaert, G. Ciccotti; H.J.C. Berendsen. Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n–Alkanes. *J. Comp. Phys.* 23 (3): 327–341, 1977.
57. W. F. Van Gunsteren, H. J. C. Berendsen. A Leap–frog Algorithm for Stochastic Dynamics. *Molecular Simulation*. 1:3, 173–185, 1988.
58. A. Warshel, M. Levitt. Theoretical studies of enzymic reactions: dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *J. Mol. Bio.* 103 (2): 227–49, May 1976.
59. C. Dellago, P.G. Bolhuis, F.S Csajka, D. Chandler. Transition path sampling and the calculation of rate constants. *J Phys Chem*, 108, 1964–1977, 1998.
60. C.R. Harris, K. Jarrod Millman, S.J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N.J. Smith, R. Kern, M. Picus , S. Hoyer, M.H. van Kerkwijk, M. Brett, A. Haldane, J. Fernandez del Rio, M. Wiebe, P. Peterson, P. Gerard–Marchant, K/ Sheppard, T. Reddy , W. Weckesser, H. Abbasi, C. Gohlke, T.E. Oliphant. Array programming with NumPy. *Nature* 585, 357–362, 2020.
61. H. Eyring. The Activated Complex in Chemical Reactions. *J. Chem. Phys*. 3 (2): 107–115, 1935.

62. K.J Laidler, M.C. King. The development of Transition–State Theory. *J. Phys. Chem.* 87 (15): 2657–2664, 1983

63. M. G. Evans and M. Polanyi. Some applications of the transition state method to the calculation of reaction velocities, especially in solution. *Trans. Faraday Soc.* 31, 875, 1935.

2.7 Supplementary Information

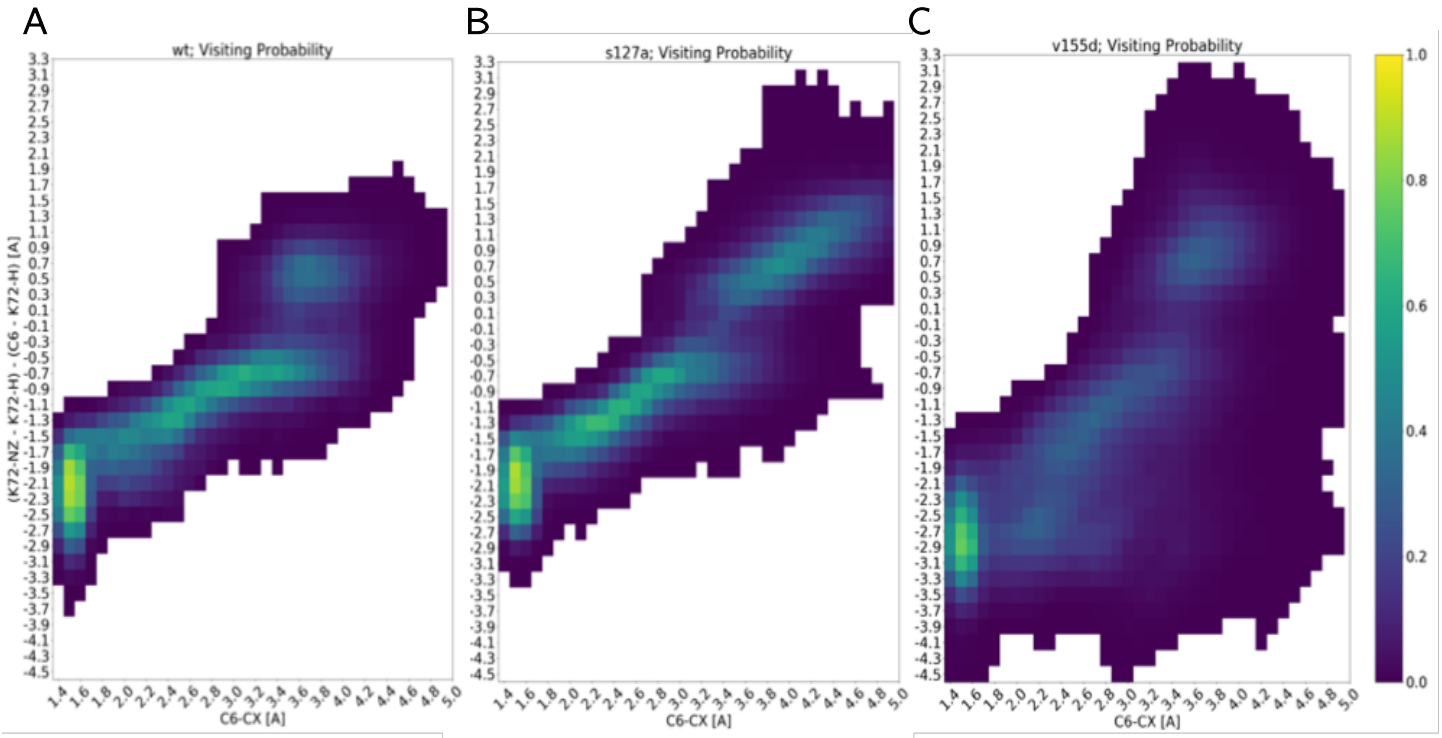


Figure 2.S1: The visitation probability of the path ensembles of the WT and mutant OMPDC as a function of the decarboxylation coordinate and the proton–transfer coordinate. Each panel indicates the proportion of trajectories that sampled a decarboxylation/proton transfer coordinate at least once during the duration of the trajectory. (A) Most trajectories of the WT ensemble appeared to decarboxylate concurrent to the K72 proton approaching the C6 carbon. Moreover, less than 5% of trajectories were capable of sampling the lower–right hand basin corresponding to proton–independent decarboxylation. (B) The S127A ensemble appeared to follow a similar path to the WT ensemble, with a majority of trajectories decarboxylating as the K72 proton approached the C6 carbon. (C) The V155D ensemble appeared most different of the three ensembles. A concurrent route, like the WT and S127A mutant, did appear, but at a lower probability than the other two ensembles. Additionally, the V155D mutant sampled the lower–right basin of the PMFs more often than either the WT or S127A mutant.

Chapter 3: Identifying the electronic determinants of reactivity in enzyme catalysis

Natasha Seelam[1,2], Brian M. Bonk[2,3], Bruce Tidor[2,3,4]

1. Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge MA

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

3. Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge MA

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

Author Contributions:
NS performed the NBO analyses and pairwise model classification for geometric and electronic calculations. NS performed the exploratory analyses and discussion of feature interpretability. BMB performed the CHARMM calculations on the KARI system. NS and BT designed the exploration around explainability of NBO features and project.

3.1 Abstract

Path sampling rate calculations provide tremendously detailed statistical mechanical descriptions of kinetic processes. Using these powerful datasets to understand the structural and dynamical processes that drive catalysis remains a challenge due to their high dimensionality. Here, we use combined quantum and molecular mechanical models together with machine learning tools to explore the interplay of atomic and electronic interactions responsible for the catalytic activity of the enzyme ketol–acid reductoisomerase (KARI), an enzyme that facilitates branched–chain amino acid synthesis via rate limiting alkyl migration followed by reduction. Previous work identified relatively small sets of structural features describing the enzyme–substrate complex in the reaction basin that were capable of predicting whether or not a particular trajectory would succeed in crossing the barrier and becoming a product complex [1]. In the current work, we extend the chemical significance of these findings by identifying the electronic determinants of reactivity and studying the relationships among atomic geometric features, electronic structure, and reactivity.

To characterize the electronic descriptors most relevant to the reaction, Natural Bonding Orbital (NBO) calculations were used to probe simulations that successfully (reactive) and unsuccessfully (non–reactive) facilitate methyl transfer within the substrate. Analyses of the Wiberg bond index, a proxy for bond order, indicated that methyl transfer occurred concurrent to isomerization between the substrate carbonyls/hydroxyls. Further inspection of the atomic orbital hybridization states across various bonding orbitals throughout the course of the reaction revealed that the KARI transition state possessed a three–center–two–electron bond formed with the three adjacent carbon atoms (C_4, C_5, C_7), primarily formed through a network of 2pz orbitals orthogonal to the plane constructed by the oxygens (O_6, O_8) and carbons of the substrate (C_4, C_7).

Supervised machine learning methods were used to identify a reduced subset of geometric features, and electronic features that predicted reactivity. A subset of 10 geometric or set of 6 electronic features were capable of over 90% AUC predictive performance, suggesting electronic descriptors were more indicative of reactivity as fewer were required for similar predictive power. For both the geometric and electronic–feature models, the cumulative models outperformed any pair–feature model, suggesting that multiple descriptors were synergistic and predicted reactivity more effectively than isolated pairs.

Analysis of the dynamic trajectories revealed that the torsional orientation of the methyl group prior to reacting influenced the ability to climb the reaction barrier by destabilizing the breaking $C_4 - C_5$ σ–bond orbital energy. This orientation was also related to the first model–selected geometric feature (the distance between the transferring methyl C_5 and a neighboring, conserved residue's carboxylate oxygen E319/OE_1), and an electronic feature ($C_4 - C_5$ bond index) in this work. Stratifying the geometric feature on the methyl torsional orientation demonstrated that farther proximity between the methyl and E319 permitted the eclipsed orientation of the methyl group that encouraged reactivity in the reactive ensemble, and that the non–reactive ensemble was in closer contact on average than the reactive simulations and less likely to adopt this orientation. Similar stratification for the electronic feature also demonstrated that the breaking $C_4 - C_5$ bond index was weaker in the reactive ensemble for eclipsed orientations, and that the non–reactive ensemble on average had a slightly stronger $C_4 - C_5$ bond index overall.

The cumulative electronic feature model possessed a variety of features that reported on two facets of the mechanism: the breaking $C_4 - C_5$ σ–bond and the diminishing $C_7 - O_8$ π–bond. To investigate how the geometric features may report on these different components, we

constructed classifiers that predicted the relative magnitude for each of the 6 model–selected electronic features, given the combined average of the reactive and non–reactive ensembles. Our results showed that geometric pair-feature models were capable of predicting the relative magnitude of the electronic features as well as the cumulative geometric feature model leveraging all 10 unique features. This suggested that small subsets of geometric features were capable of reporting on an electronic descriptor, and that different subsets could be leveraged to predict various aspects of the chemical reaction.

3.2 Introduction

Enzymes readily facilitate difficult reactions with high selectivity at ambient temperatures and pressures, making them desirable targets for industrial repurposing. However, preparing enzymes for industrial purposes often requires tailoring their catalytic prowess toward the specific target of interest [2]. Despite excellent progress from directed evolution and rational design strategies, facile enzyme engineering remains a challenging feat due to the incomplete perspective of the determinants of enzyme reactivity, and vast design space [2–7, 11].

Approaches to enzyme design currently involve directed evolution [8–10] and computational design that spans from tight–binding approaches based on Transition–State Theory [4, 11], to machine learning strategies based on inferring function from aspects of a sequence [6, 7]. While such methods have demonstrable success, they have yet to become generalizable tools. Moreover, it is not always intuitive why certain refinements on rational design models, or changes accrued from evolutionary processes influence reactivity in a way that may be desirable [11–13]. For this reason, it is useful for *in silico* models' features to be able to explain the way they report on reactivity and report on how such changes in the local environment may influence the electronic characteristics of a reaction of interest.

Prior work by Bonk et al. identified classifiers, trained on geometric features (features that indicate structural interactions such as distances, angles, and torsions of active–site atoms) chosen for their ability to describe the local environment of the active site, and potentially report on key drivers of reactivity [1], as they improved the computed rate of reactivity. This work indicated a subset of 30 consensus features that were part of different classifiers that were trained at several time intervals prior to the simulations' approach to the reaction barrier. These geometric features were subsequently shown to not only predict reactivity, but also improve the likelihood of crossing

a reactive barrier. Moreover, clustering on these geometric features showed that the same features could be employed in different ways, suggesting the existence of multiple reaction channels representing slight variations toward the overall mechanism.

The current study extends on the work done by Bonk et al. [1] to analyze the results, as reflected through the machine learning models computed, and to understand how the models may distinguish between the reactive and nearly reactive (non–reactive) simulations. Moreover, this work aims to compare determinants of reactivity, discovered by machine learning methods, to the chemical mechanism constructed from prior experimental and theoretical studies, and further elucidated by our theoretical analyses.

3.2.1 Catalytic strategies of ketol–acid reductoisomerase (KARI)

The system studied in this work involves the enzyme ketol–acid reductoisomerase (KARI), an enzyme crucial to the creation of branched–chain amino acid synthesis in plants, archaea, algae, and fungi [14, 25, 26]. There are two linked reactions carried out by KARI: (1) an isomerization resulting in the migration of an alkyl group and interconversion between carbonyl and hydroxyl moieties and (2) an NADPH–assisted reduction [14–16]. Native KARI includes two divalent cationic magnesium ions and one molecule of NADPH as required co–factors, and the substrate either acetolactate (AL; methyl R–group) or acetohydroxybutyrate (AHB; ethyl–R group) [15–18]. Ordinarily substrate binds and is isomerized and then reduced without unbinding in between [14, 15, 17–19]. The reduction reaction has been shown to work independently of isomerization by binding other 2–ketoacids that are unable to isomerize but are capable of being reduced further [20]. Further experiments have demonstrated that the alkyl migration step is specifically Mg^{2+} dependent. Reduction is possible in the presence of other divalent cation, as Mn^{2+} and Co^{2+} have

been shown to salvage reductase activity on compounds that do not require isomerization [16, 17, 20]. The isomerization step is considered rate–limiting, and the product does not release after isomerization until a subsequent reduction step occurs, creating the final product dihydroxyisovalerate (DHIV) or dihydroxymethylvalerate (DHMV) (respectively if the alkyl is a methyl or ethyl substrate) [19, 21].

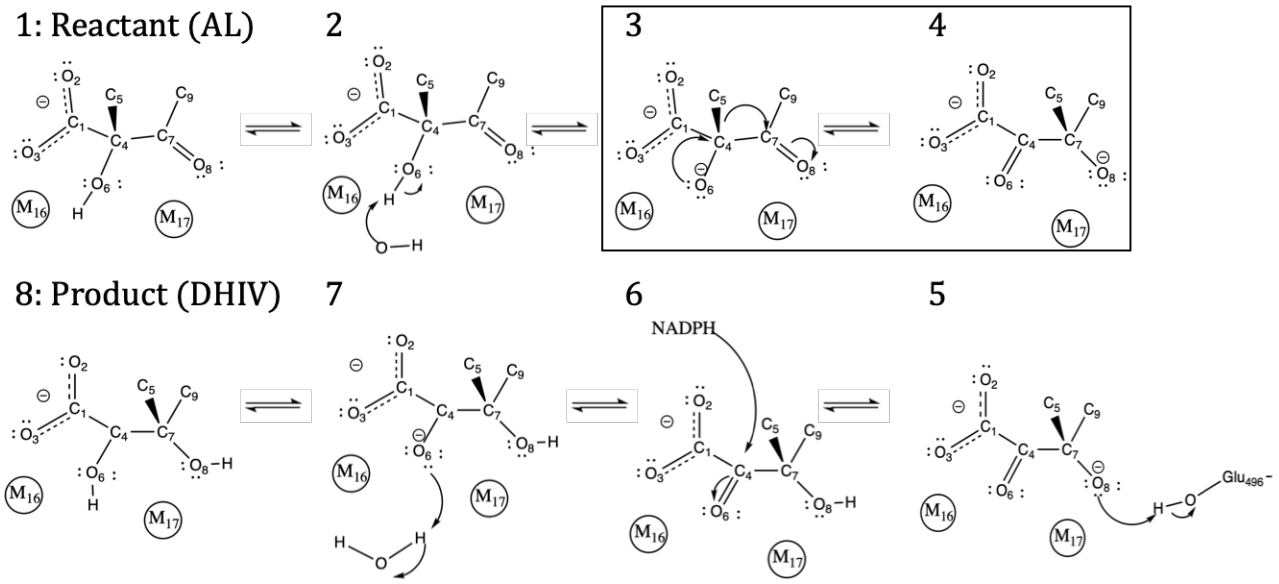


Figure 3.1: Hypothesized mechanism of the two reactions of KARI, of which this work focuses on modeling the rate–limiting isomerization, indicated in the box for step 3 and 4. Magnesium atoms are labeled M16 and M17, as are substrate atoms.

Mechanistic studies both experimental and theoretical have posited that a proton abstraction, initiated from a hydroxyl group or active–site residue, occurs quickly from the substrate's O_6 (Figure 3.1) [18, 21]. Subsequently, the alkyl group (C_5) migrates from carbon C_4 to C_7 employing the Mg^{2+} cations in the process [18, 21]. Potent inhibitors, IpOHA and Hoe704, possess similar oxalyl motifs that are thought to mimic the isomerization transition–state by bridging Mg^{2+} (specifically M17) in the process of the reaction (Figure 3.2B) [22, 28]. Notably,

the hypothesized transition state of the reaction is thought to possess a three–center–2–electron (3C) bond bridging between carbon atoms C_4, C_5, and C_7.

There is considerable structural diversity in the KARI family: a short form (Class I) and long form (Class II) enzyme isoform [26]. Despite these differences, active sites within KARI family remain highly conserved for a binding pocket for NADPH and charged active–site residues, namely E496, E319, and D315 (*S. oleracea* numbering). These active–site residues are thought to play a key role, together with water molecules, in the hexa–coordination of the divalent cationic Mg found in the active site [15, 16, 20]. Studies have validated that these residues are critical, and even a charge–conservative mutation of aspartyl to glutamyl (or vice–versa) moieties results in reduced substrate and/or co–factor binding and/or loss of activity [20].

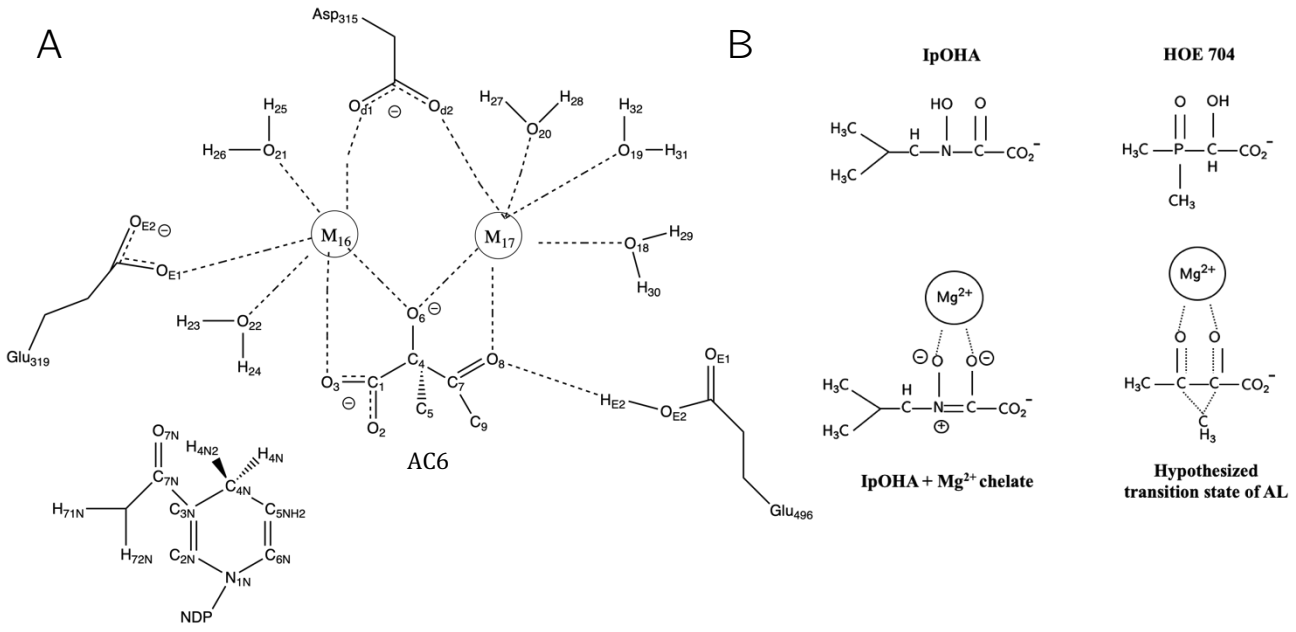


Figure 2: (A) Quantum mechanical region of the QM/MM simulations including the co–factor NADPH (indicated by acronym NDP); side chains E496, E319, and D315; two Mg^{2+} cations; 5 water molecules; and the substrate, AL. Mg^{2+} cations are hexacoordinated through the water molecules and polar side chains. (B) Known transition–state inhibitors of KARI and the hypothetical transition state [21–23].

3.2.2 QM/MM ensembles generate simulations that successfully catalyze methyl transfer or rebound to the reactant state

Prior work by Bonk et al. generated ensembles of molecular dynamic trajectories that traversed from enzyme–bound substrate, over the barrier, to enzyme–bound product (of the methyl migration reaction), as well as trajectories that return to the reactant basin after an attempt to cross the barrier [1, 23]. In this work, we analyze these simulations and augment them with additional quantum mechanical calculations to identify the underlying electronic determinants of reactivity. In order to do this, a common timeline was established by identifying the last bond compression between the C_4 and C_5 atoms, and marking this as time $t = 0$ fs (Figure 3.3), for the 2000 sampled reactive and non–reactive trajectories. By establishing a common timeline, direct comparisons could be made by how the reactive and non–reactive ensemble diverge electronically over time. For timepoints prior to $t = 0$ fs, trajectories were still in the reactant basin, and at times $t > 0$ fs, they began to ascend the barrier.

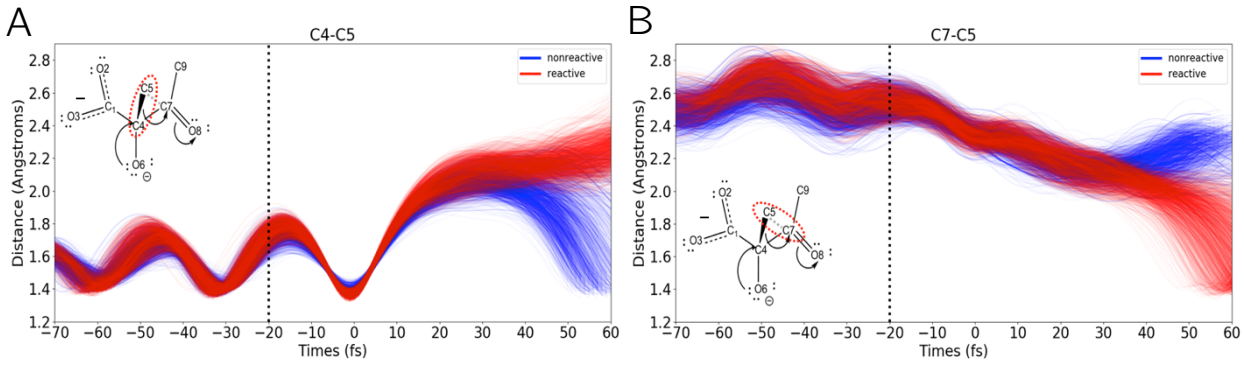


Figure 3.3: Distances between the breaking bond, $C_4 - C_5$, and the forming bond $C_7 - C_5$ for the reactive (red) and non–reactive (blue). The time point for subsequent analyses (-20 fs) is indicated by the dashed vertical black lines on both plots. (A) The reactive and non–reactive ensembles exhibited similar oscillatory behavior for times $t \leq 0$ fs while in the reactant basin. As the reaction proceeded, the reactive ensemble was able to successfully transfer the methyl group and continued to maintain separation beyond 2.0 Å, whereas the non–reactive ensemble returned to baseline values. (B) The distance between the product–side carbon, C_7, and the methyl carbon, C_5 for the reactive and non–reactive ensembles was on average 2.4 Å away while in the reactant state. As the reaction proceeded, only the reactive ensemble was able to achieve bonding between these atoms, below 1.8 Å. The non–reactive ensemble did not approach any closer than 2.0 Å.

3.2.3 Feature selection for enzyme catalysis and machine learning

We selected chemically meaningful features capable of distinguishing reactive from non–reactive trajectories while still in the reactant well. We chose both geometric features describing the relative configurations and electronic features describing the relationships between orbitals and bonding patterns. Feature selection was pursued with two priorities in mind; we wanted to allow for synergistic sets of features to the extent that they enhanced predictive performance due to their ability to represent complex, or coordinated chemistry. We also desired redundant features (multiple features indicative of the same underlying chemistry) to increase the signal highlighting important chemistry. To meet these priorities, we chose a sequence–of–pairs approach.

The geometric features identified were chosen from the 30 consensus features of prior work [1]. These features included a collection of distances, angles, and dihedrals that appeared in classifiers, trained at several time points prior to reactivity, that could distinguish between a reactive and non–reactive ensemble. These features were part of a greater collection initially chosen due to their relationship with mechanistic hypotheses proposed in the literature.

The features chosen for the electronic analysis included: charge, bond index, and atomic orbital hybridization state, derived from Natural Bonding Orbital (NBO) computations to provide a Lewis–like interpretation of the electronic wavefunction [29, 30]. Charge was computed through natural population analysis (NPA), which uses the sum total of the occupancies of Natural Atomic Orbitals (NAOs), centered on the atom of interest, prescribed by NBO formulation [29–31]. The Wiberg NAO bond–index computes the atomic coefficient overlap between two atoms specified; this corresponds to half the electrons shared between two atoms—a proxy for bond order. Covalent bonds are considered greater than one, whereas ionic bonds are historically less than one unless negligible [32, 33]. Lastly, atomic orbital hybridization state was derived from the coefficients,

centered on each atom, considered to be participating in a given molecular orbital (whether bonding or lone–pair). This definition of charge and bond index are robust to basis–set dependencies [29–33].

3.3 Methods

3.3.1 Electronic Structure and Natural Bonding Orbital (NBO) Calculations

A reactive region encompassing the quantum–mechanical component of the reactive and non–reactive simulations was extracted, including 32 atoms of the substrate, and divalent cationic complex of magnesium ions and corresponding bound water molecules. This subset of the reaction required no imputations of electronic structure corresponding to the boundary atoms, as these are all complete fragments. To derive NBO descriptors out of NBO 5.0, implemented within Q–Chem, the following descriptors were used: "NBOSUM" – an NBO summary table, "CMO" – bonding character of canonical molecular orbitals, "BNDIDX" – the Natural Atomic Orbital (NAO)–Wiberg bond index as a proxy for bond order, "3CHB" – three–center four–electron bond hybrid molecular orbital (MO) searches, and "3CBOND" – three–center two–electron bond hybrid molecular orbital searches.

3.3.2 Machine Learning

Scikit–learn's *LogisticRegression* module was employed to perform the classification task [34]. Classifiers employing all pairwise subsets of only geometric features or only electronic features were constructed using this python module. The 30 geometric features were chosen from the top populated features of Bonk et al. [1], and 19 electronic features were chosen to capture electronic descriptors of the mechanism described with Natural Population Analysis charge, Natural Atomic Orbital (NAO) Wiberg Bond Index for bond order, and atomic orbital composition based on the coefficients of bonding orbitals described by NBO on the atoms where the reaction is localized: C4, C5, C7, O6, O8, M16, and M17 (Supplemental Table 1).

The dataset pooled all five reactive clusters from the prior paper [1] in equal proportions; in total, there were 2000 simulations in the reactive ensemble and 2000 simulations for the non–reactive ensemble. To perform the classification task, trajectories were aligned according to the last bond compression defined as time 0, and subsequently the timepoint of –20 fs was chosen to extract for the training and testing sets. To verify the performance of the models, 10 randomized training/testing splits were created, assigning 70% of the data as train, and 30% toward testing, and the average performance of accuracy and AUC on the testing sets was reported.

The testing performance (accuracy) for each pair was rank ordered, and subsequently, the best performing pair was selected until a combined model (5 pairs for geometric, 3 pairs for electronic) testing performance improved no further. Performance metrics included were accuracy and ROC AUC.

Classifiers with explicitly only geometric, or only electronic features predicted reactivity (from the reactive ensemble, versus the non–reactive ensemble). Labeling of reactivity is inherent to whether the simulation successfully catalyzed the methyl transfer and ended in the product basin, versus failed to react and returned to the reactant basin. From the subset of geometric features capable of predicting reactivity, these features were subsequently tasked to predict the expectation of a large/small value of an electronic descriptor, of the electronic features that predicted reactivity. To label the electronic descriptor, the reactive and non–reactive ensembles were combined, and the average value across the total ensemble was calculated. A “large” electronic feature is labeled if a simulation exhibited an electronic descriptor larger than the average of the distribution and subsequently vice versa for a “small” electronic feature.

3.3.3 Geometric feature analysis with torsional order parameter

We defined a torsion order parameter to measure the degree of stagger/eclipse orientation between the C_5 and C_4 by averaging the minimum, absolute angle between each proton of C_5 to C_4's constituent bonded partners (C_1, O_6, and C_7). This order parameter describes the degree to which the C_5 methyl was purely eclipsed (0 degrees) or purely staggered (60 degrees) while in the reactant state when referenced to C_4.

3.3.4 Generation of QM/MM reactive and non–reactive ensembles

Quantum mechanical/molecular mechanical (QM/MM) simulations were performed on KARI facilitating the methyl transfer of the substrate using CHARMM version 41 with the SQUANTUM semi–empirical methodology. The simulations employed the use of AM1 to describe the active site including the substrate, active–site–coordinated waters (5 total), both magnesium cations, the R–groups of D315, E319, E496, and the nicotinamide group of NADPH for a total of 77 atoms. Boundary–condition atoms were treated using the General Hybridized Orbital (GHO) method to handle the QM/MM region across covalent bonds. To account for the rate limiting step, O6 of the substrate was deprotonated to emulate the hypothesized, transient fast–step proton–abstraction, as indicated by prior studies [21]. E496 is modeled at neutral charge, acquiring this proton.

In generating the reactive and non–reactive ensembles, reactivity was defined through the use of a combined order parameter:

$$\lambda = distance(C_4 - C_5) - distance(C_7 - C_5)$$

The combined parameter measures the breaking bond ($C_4 - C_5$) versus the forming bond ($C_7 - C_5$) wherein the reactant interface was $\lambda_R = -1$ Å and the canonical product is defined as $\lambda_R = 1$ Å. Reactive ensemble trajectories were expected to reach the product interface, but non–reactive simulations had a modified product definition of $\lambda_{NR} = -0.2$ Å, and were advanced until they returned to the reactant basin. Detailed information of this particular QM/MM system setup can be found in prior work [1].

3.3.5 Structure Preparation

The crystal structure of *S. oleracea* (spinach) KARI with accession code 1YVE from the Protein Data Bank (PDB) was acquired. The native homodimer of KARI enzyme contains two active sites, significantly separated between both monomers. To facilitate computational efficiency, only chain A of the KARI monomer was used. To represent the enzyme–substrate (ES) bound complex for QM/MM calculations, *in vacuo* ground–state electronic calculations were performed with two magnesium cations, five coordinated water molecules, and the side–chains of 3 active–site residues (E496, E319, D315), and substrate at the level of *RHF/3–21G** theory using GAUSSIAN03. Detailed information of structure preparation can be found in prior work [1, 23].

3.4 Results and Discussion

3.4.1 Electronic description of KARI methyl transfer reaction for reactive and non–reactive ensemble

To construct an electronic level description of the chemical mechanism of KARI facilitated methyl transfer, quantum mechanical calculations were carried out on the reaction region of both reactive and non–reactive simulations with the NBO method employed to interpret the resulting wavefunction as orbital populations. Simulations were time–aligned such that $t = 0$ fs corresponded to the last compression of the breaking bond, $C_4 - C_5$, before attempting to cross the barrier.

The electronic "flow" (the "arrows" of the reaction, per se), was quantified through the reaction by tracking the Wiberg NAO bond–index between core atoms of the reaction: C_4, C_5, C_7, O_8 and O_6 (summarized in Figure 3.4). The simulations for both reactive and non–reactive ensembles indicated that the methyl migration – namely the breaking of the $C_4 - C_5$ bond (Figure 3.5A) and formation of the $C_7 - C_5$ bond (Figure 3.5B) – occurred concurrently with the formation of a double bond between $C_4 - O_6$ (Figure 3.5C) and single bond between $C_7 - O_8$ (Figure 3.5D). Shortly after $t = 0$ fs (i.e. after the last compression of the $C_4 - C_5$ bond), in both non–reactive and reactive ensembles, as all trajectories exhibited a simultaneous reduction in bond index between $C_4 - C_5$ from an average of 1.0 to roughly 0.4, concomitant with an increase in bond index between $C_7 - C_5$ from 0 to 0.25. At time $t = +20$ fs, the reactive and non–reactive ensembles diverged, as the bond index between $C_4 - C_5$ steadily decreased (heading toward 0) while $C_7 - C_5$ increased (toward 1.0) for the reactive ensemble only. In the non–reactive ensemble, both of these bond indices returned to the values adopted before $t = 0$ fs.

Concurrently with the changing methyl–migration bond indices, the oxygens attached to both reactant–side carbon (C_4) and product–side carbon (C_7) exhibited changes in bond indices.

The bond index for the carbonyl between $C_4 - O_6$ was nearly 0.9–1.0 at times prior to time $t = 0$ fs for both reactive and non–reactive ensembles. As the reaction proceeded for times after $t = 0$ fs, the bond index between $C_4 - O_6$ increased to approximately 1.0–1.15 until time $t = +20$ fs, where the non–reactive ensemble returned to their baseline value of 0.9, and the reactive ensemble continued on upward, reaching a value of 1.1–1.5 during the simulated time. Likewise, the bond between $C_7 - O_8$ started with a baseline of 1.5–1.65 for times prior to time $t = 0$ fs. As the reaction proceeded, both the reactive and non–reactive ensemble fell to ~1.3 at around time $t = +20$ fs, after which the non–reactive ensemble subsequently rebounded toward the initial value, and the reactive ensemble continued toward 1.0.

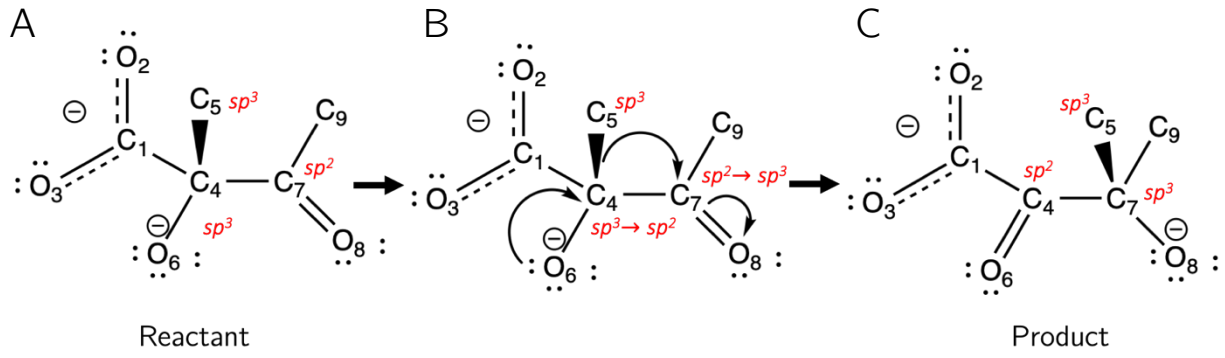


Figure 3.4: Lewis–representation of the KARI methyl transfer reaction. (A) The reactant state in a Lewis–like depiction (omitting side chains, magnesium ions and NADPH) indicates that the reactant–side carbon, C_4, initially starts as an sp^3 hybrid, in addition to the transferring methyl carbon C_5. The product–side carbon, C_7, is an sp^2 hybrid. (B) As the reaction proceeded, NBO calculations suggested a concurrent change in bond–order across the methyl migration coordinate and the increase/decrease of bond index between $C_4 - O_6$ and $C_7 - O_8$ respectively. As the reaction proceeded, a change in hybridization also occurred to account for the appropriate amount of carbon valences. (C) The Lewis–like product state has a 'mirrored' hybridization state to the reactant, where the C_4 carbon adopts an sp^2 hybridization state whereas the C_7 carbon becomes sp^3.

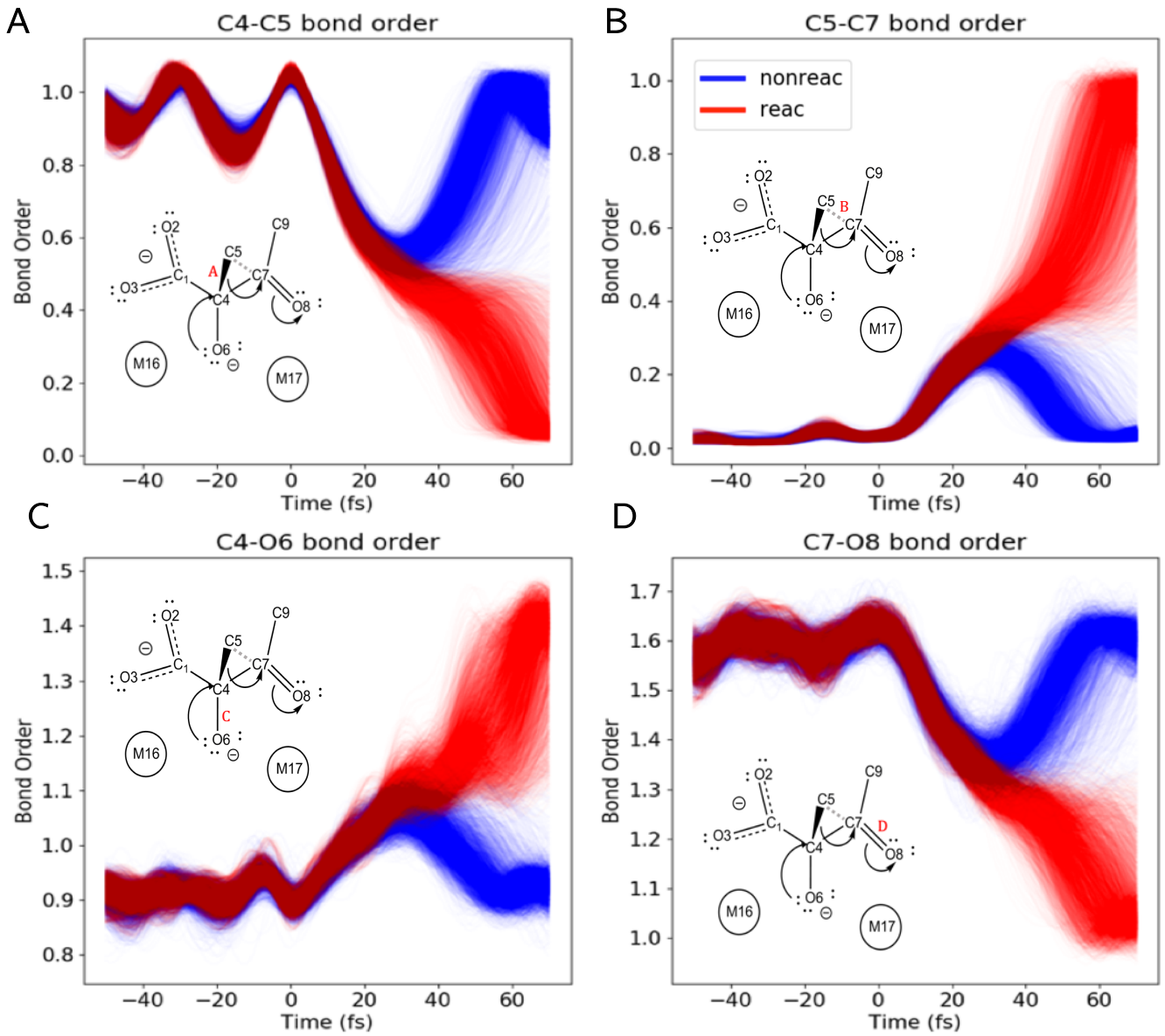


Figure 5: NAO–Wiberg bond indices changed concurrently for the four changing bonds across the methyl transfer reaction of the KARI substrate. Each panel shows a bond index over time for simulations in the reactive (red) and non–reactive (blue) ensembles. (A) The bond index for $C_4 - C_5$ oscillated around the expected value of 1.0 as a function of the harmonic stretching and compression for the reactant state prior to time $t = 0$ fs. The bond index for $C_4 - C_5$ diminished until nonexistent for the reactive ensemble but rebounded to baseline values for the non–reactive ensemble, indicating that the latter simulations do not successfully sever the bond between $C_4 - C_5$. (B) The bond index for $C_7 - C_5$ was marginal prior to times $t = 0$ fs for both reactive and non–reactive ensembles, suggesting no bond was formed prior to the reaction. Only the reactive ensemble successfully enabled bond–formation for $C_7 - C_5$, whereas the non–reactive ensemble rebounded to baseline value. (C) For $C_4 - O_6$, the bond index near 1.0 suggested a roughly single bond between these two atoms at times prior to $t = 0$ fs. As the reaction proceeded, an increase in bond index was sustained by the reactive ensemble only, but less than the expected 2.0 for double bond formation. (D) The bond index for $C_7 - O_8$ started marginally lower than the expected 2.0 for a double bond, but reduced to nearly 1.0 for the reactive ensemble only, suggesting a single–bond formation between these two atoms that did not occur for the non–reactive ensemble.

To provide an orbital level view of the reaction, the electronic NBO calculations reported on orbital hybridization character through the composition of atomic orbital coefficients comprising key bonding orbitals between the reactant, transient, and product states, including the σ–bond $C_4 - C_5$, the σ–bond $C_7 - C_5$, π–bond between $C_4 - O_6$, the π–bond $C_7 - O_8$, the lone pair orbital (indicated "LP") of O_6, the lone pair orbital of O_8, and the transient methyl–migration bond considered the three–center–two–electron (3C) bond encompassing $C_4 - C_5 - C_7$ (Figure 3.6). Geometrically, the 3C atoms constitute a plane roughly normal to the plane formed by O6, C4, C7, and O8. Monitoring the atomic coefficients throughout the reaction, the most notable change to these orbitals resided in the composition of $2s$ and $2p_z$ atomic orbitals used to construct natural bonding orbitals. To measure hybridization state, the $2s$ and $2p_z$ composition was used to indicate whether a bond was mostly sp^3 (comprised of 25% $2s$ character and 75% $2p_z$ character), sp^2 (comprised of 33.3% $2s$ character and 66.7% $2p_z$ character), or sp (comprised of 50% $2s$ character and 50% $2p_z$ character). For any sp–hybrid orbital, the $2p$ character may be distributed across the $2p_x$, $2p_y$, or $2p_z$ orbitals.

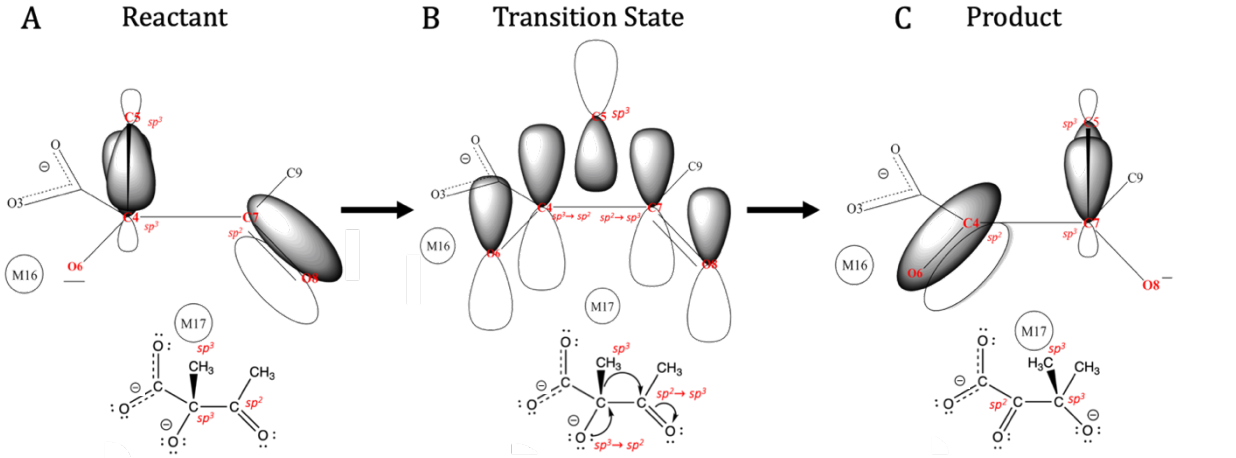


Figure 3.6: Corresponding orbital diagram of the NBO–computed hybridizations for the KARI methyl transfer reaction. (A) Reactant state orbitals of interest include the $C_4 - C_5$ σ–bond, the $C_7 - O_8$ π bond, and the lone pair of O_6. (B) In the transition state, the transient orbitals used mainly $2p_z$ orbitals in the 3C bond between C_4, C_5 and C_7, and the lone pairs on O_8 and O_6. (C) Simulations that attained the successful product state exhibited a σ–bond between $C_7 - C_5$, a π–bond between $C_4 - O_6$, and a lone pair on O_8.

The C_4 atom participated in bonds that change character over the course of the reaction. The natural bonding orbital describing the $C_4 - C_5$ σ–bond exhibited, on average, 25% $2s$ character (Figure 3.7, Panels A and C) and 75% $2p_z$ character (Figure 3.7, Panels B and D) specifically for the C_4 atom before time $t = 0$ fs for both the reactive and non–reactive ensemble, suggesting sp^3 behavior in the reactant state, which matches the expected reactant hybridization. As the reaction proceeded, both the reactive and non–reactive ensembles decreased in the amount of $2s$ character that C_4 exhibited in the $C_4 - C_5$ σ–bond, diminishing in value until nearly 0%, and they both increased in this atom's amount of $2p_z$ character in that bond until nearly 100%. Trajectories in the reactive ensemble, around time $t = +30$ fs, all successfully formed a transient 3C bond with the carbon atoms of the substrate, (C_4, C_5 and C_7) for at least 20 fs, whereas only a small fraction of the non–reactive simulations could sustain this 3C orbital. As time proceeded, the non–reactive ensemble returned to the substrate and established $2s$ character of the $C_4 - C_5$ σ–bond to baseline values, whereas the reactive ensemble never regained any substantial $2s$ character. By about time $t = +60$ fs, the atom C_4 participated in a π–bond $C_4 - O_6$ with nearly 100% $2p_z$ character, consistent with the expectation that a product–state π–bond would be comprised of overlapping $2p$–atomic orbitals. Contrary to the reactive ensemble, the non–reactive ensemble was unable to form any substantial π–bond character between C_4 and O_6.

The next atom, C_5, participated in three natural bonding orbitals including the $C_4 - C_5$ σ–bond, the 3C bond involving the central three substrate carbons, and the eventual formed $C_7 - C_5$ σ–bond. Atomic coefficient analysis of the natural bonding orbitals over time (Figure 3.8) for both the reactive and non–reactive ensemble showed that C_5 was sp^3 hybridized, and the $C_4 - C_5$ σ–bond averaged 25% $2s$ character and 75% $2p_z$ character while in the reactant basin (for times $t \leq 0$ fs). As the reaction proceeded, a distinct decrease in $2s$ character to about 10−15% and

simultaneous increase in $2p_z$ character by a comparable amount was reflected in both ensembles. Notably, this change in $2s$ character contrasts that of C_4, where the observed $2s$ character was far lower at the comparable time points than for C_5. For times $t = +30$ fs onward, all reactive trajectories formed a 3C bond with predominantly $2p_z$ character, whereas only a small subset of the non–reactive simulations was able to transiently form a 3C bond. Throughout the 3C bond existence, trajectories maintained some $2s$ character for the C_5 atom. At times $t = +60$ fs onward, the 3C bond was dissolved and reactive simulations successfully established the $C_7 - C_5$ σ–bond, with roughly 25% $2s$ character and 75% $2p_z$ character, suggesting C_5 adopted an sp^3 hybridization state once more. Likewise, non–reactive simulations returned to the reactant state and sp^3 hybridization. Notably for the nonreactive ensemble, the few trajectories capable of forming the 3C bond possessed less $2s$ character in the transient time they existed.

The final carbon atom involved in the methyl migration was C_7, which initially participates in a double bond comprised of a $C_7 - O_8$ σ–bond and $C_7 - O_8$ π–bond. As seen in Figure 3.9 for early times, the π–bond was observed to have no $2s$ character for either reactive or non–reactive trajectories, but steadily increased in relative $2p_z$ character as the reaction approached time $t = 0$ fs. At time $t = +20$ fs, the C_7 atom appeared to no longer participate in the π–bond between $C_7 - O_8$ for either the reactive or non–reactive ensembles. However, almost immediately after, all the members of the reactive ensemble were able to establish the 3C bond between the three carbon atoms of the substrate in contrast to the few non–reactive trajectories that were able to do so. The formation of the 3C bond also coincided with an increase in $2s$ character for the C_7 atom for the reactive ensemble. Time $t = +40$ fs marked the first instance of the $C_7 - C_5$ σ–bond with almost 25% $2s$ character for all trajectories in the reactive ensemble. In contrast, the non–reactive ensemble was unable to form this σ–bond between $C_7 - C_5$.

Addressing the carbonyl involving atom O_8, Figure 3.10 shows that this atom initially participated in the $C_7 - O_8$ π–bond comprised of mostly $2p_z$ character that increased in $2p_z$ character as the reaction proceeded for both the reactive and non–reactive ensembles. At time $t = +20$ fs, atom O_8 demonstrated additional a lone–pair formation, claiming the $2p_z$ orbital originally used in the $C_7 - O_8$ π–bond. Between times $t = +20$ fs and $t = +40$ fs, the new lone pair persisted for the non–reactive ensemble but was then reformed back into the $C_7 - O_8$ π–bond. The reactive ensemble sustained the lone–pair orbital for the duration of the reaction, where at times $t = +45$ fs and beyond, rehybridization with other $2p$ atomic orbitals occurred.

The last atom participating in the reaction is atom O_6 on the reactant–side carbon C_4 (Figure 3.11). Atom O_6 was modeled as negatively charged, labeled as "LP 3" in the QM calculations. As the reaction proceeded, this LP began to exhibit high $2p_z$ character for both reactive and non–reactive ensembles and diminishing $2s$ character, as if in preparation for folding the lone pair into the incipient $C_4 - O_6$ π–bond. During time $t = +20$ fs to time $t = +40$ fs, the lone pair on O_6 maintained almost 100% $2p_z$ character which subsequently evolved into the $C_4 - O_6$ π–bond for only the reactive ensemble around time $t = +60$ fs. The non–reactive simulations, while they first shifted in a similar manner to the reactive, were not capable of establishing the π–bond with C_4, and regained their $2s$ character by time $t = +40$ fs while ultimately decreasing in $2p_z$ character.

Concurrent to the change in orbitals was a slight shortening of the distance between specifically O_8 and M_{17} (Figure 3.12). From $t = -20$ fs, both the reactive and non–reactive ensembles were centered at 2.27 Å ± 0.06 Å and 2.24 Å ± 0.06 Å. Roughly corresponding to the time of the first (reactive) simulations' 3C bond formation ($t = 20$ fs), these distributions shifted to 2.11 Å ± 0.04 Å and 2.15 Å ± 0.06 Å respectively, suggesting a slight shortening of the O_8 and

M_{17} coordination. This shortened distance persisted for at least 20 fs after the 3C bond was formed ($t = 40$ fs).

Incidentally, the opposite trend held for the distance between O_6 and M_{17} (Figure 3.13). While in the reactant basin ($t = -20$ fs), the reactive ensemble and non–reactive ensemble were centered at 2.07 Å ± 0.05 Å and 2.05 Å ± 0.04 Å respectively. At time $t = 20$ fs, these distributions extended to 2.16 Å ± 0.05 Å and 2.14 Å ± 0.05 Å, suggesting an extension of the O_6 and M_{17} distance. By $t = 40$ fs, these distributions were roughly centered at the same distance as the latter time point, at 2.18 Å ± 0.05 Å and 2.14 Å ± 0.05 Å respectively.

Multiple studies have postulated the methyl transfer occurs concurrently with isomerization of the carbonyl/hydroxyls (O_8 and O_6 respectively while in the reactant basin), and that the KARI transition state adopts a 3C bond between the central carbons of the substrate (C_4, C_5, C_7) while tightening the interaction between O_8 and O_6 and the bridging magnesium ion (M_{17}) [15, 19, 21–23, 28]. This work supports that hypothesis, as the changing bond index of the bonds $C_4 - C_5$, $C_7 - C_5$, $C_4 - O_6$, and $C_7 - O_8$ are concurrent and agree with the anticipated changes in bond order expected of a methyl transfer and carbonyl/hydroxyl isomerization. The orbital analyses of this work indicated that the reactive ensemble formed an electronic structure compatible with the NBO definition of 3C bond, as did a small minority of the non–reactive ensemble. Interestingly, our simulations reflected a tightening in coordination of O_8 with M_{17} but the same shortening of distance does not occur with O_6 and M_{17}. The substrate oxygen–magnesium distributions, for both reactive and non–reactive simulations, over time are supported by prior studies expectations of hexa–coordinated Mg atoms [35, 36]. The possible shortening of the $O_8 - M_{17}$ distance may correspond to the formation of a lone–pair orbital that interacts more effectively

with M_{17}. Subsequently, the lengthening of $O_6 - M_{17}$ may correspond to the loss of the extra O_6 lone–pair orbital that goes on to form the $C_4 - O_6$ π–bond in the product state.

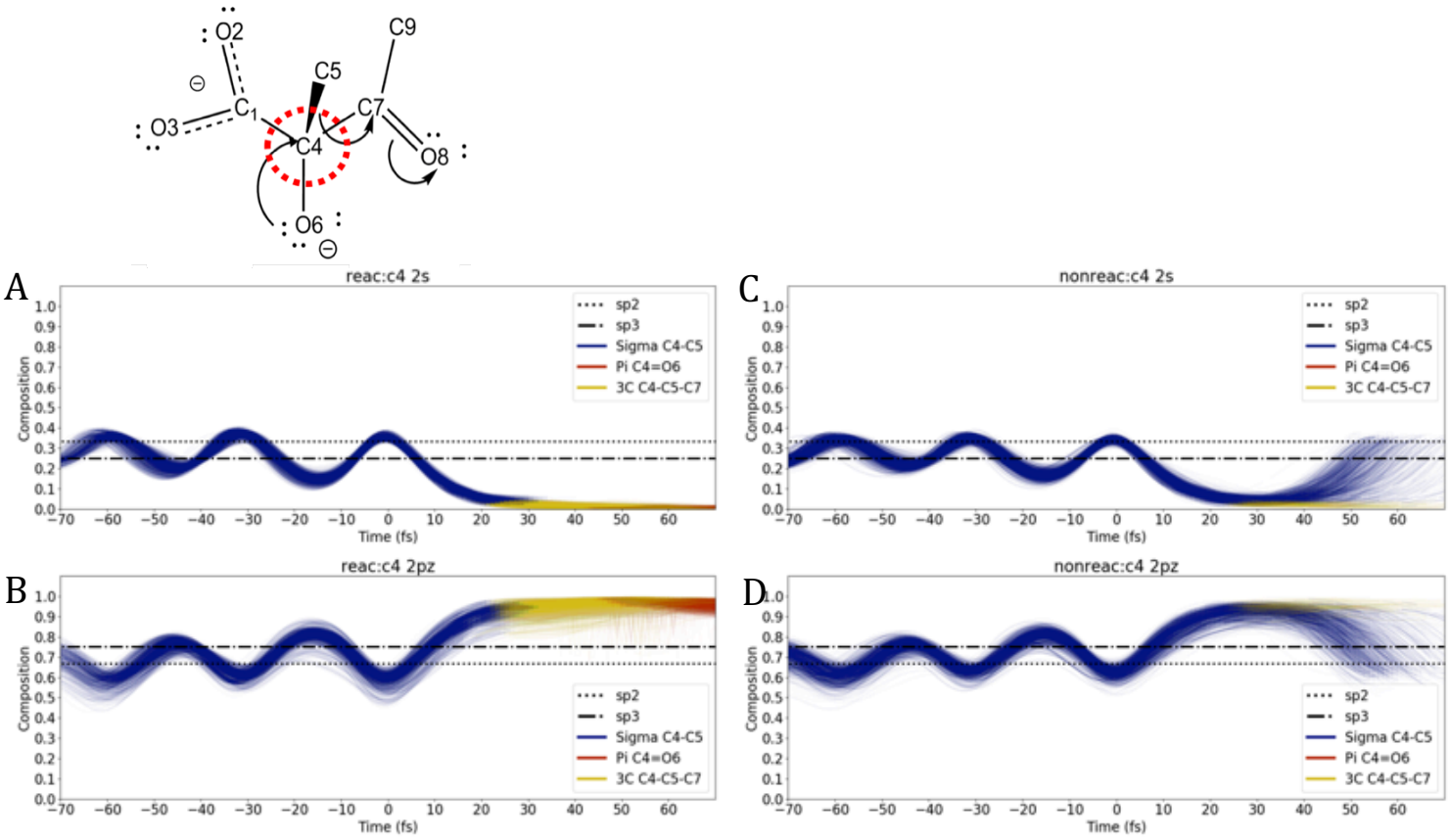


Figure 3.7: C_4 atomic orbital composition across trajectories, indicated by changing colors, for reactive (panels A and B) and non–reactive (panels C and D) ensembles. (A) The $2s$ character of the C_4 atom as it participated in several different natural bonding orbitals through the course of the reaction diminishes toward the product state for the reactive ensemble. Subsequent bonding orbitals that occurred through the course of the reaction required less $2s$ character for the reactive ensemble. Taken with panel B, the subsequent reduction of $2s$ character from 25% to 0% suggests that the C_4 atom transitions from sp^3 hybridization to sp^2 hybridization. (B) Successful reactions developed more $2p_z$ character in bonding orbitals than reactant. The reactive ensemble was able to facilitate not only the formation of the 3C bond with mostly $2p_z$ character, but also eventually form the $C_4 - O_6$ π–bond with the $2p_z$ atomic orbital. (C) The non–reactive ensemble exhibited diminishing $2s$ character comparable to the reactive ensemble until time $t = +20$ fs, but was unable to entirely eliminate it. (D) The non–reactive ensemble exhibited high $2p_z$ character until time $t = +30$ fs, when the $2p_z$ character began to return to its baseline value. Few non–reactive ensembles were able to establish the existence of the 3C bond, but even then were unable to generate the $C_7 - C_5$ σ–bond. In all panels, dashed horizontal black lines show canonical orbital proportions for static sp^2 hybridization and sp^3 hybridization.

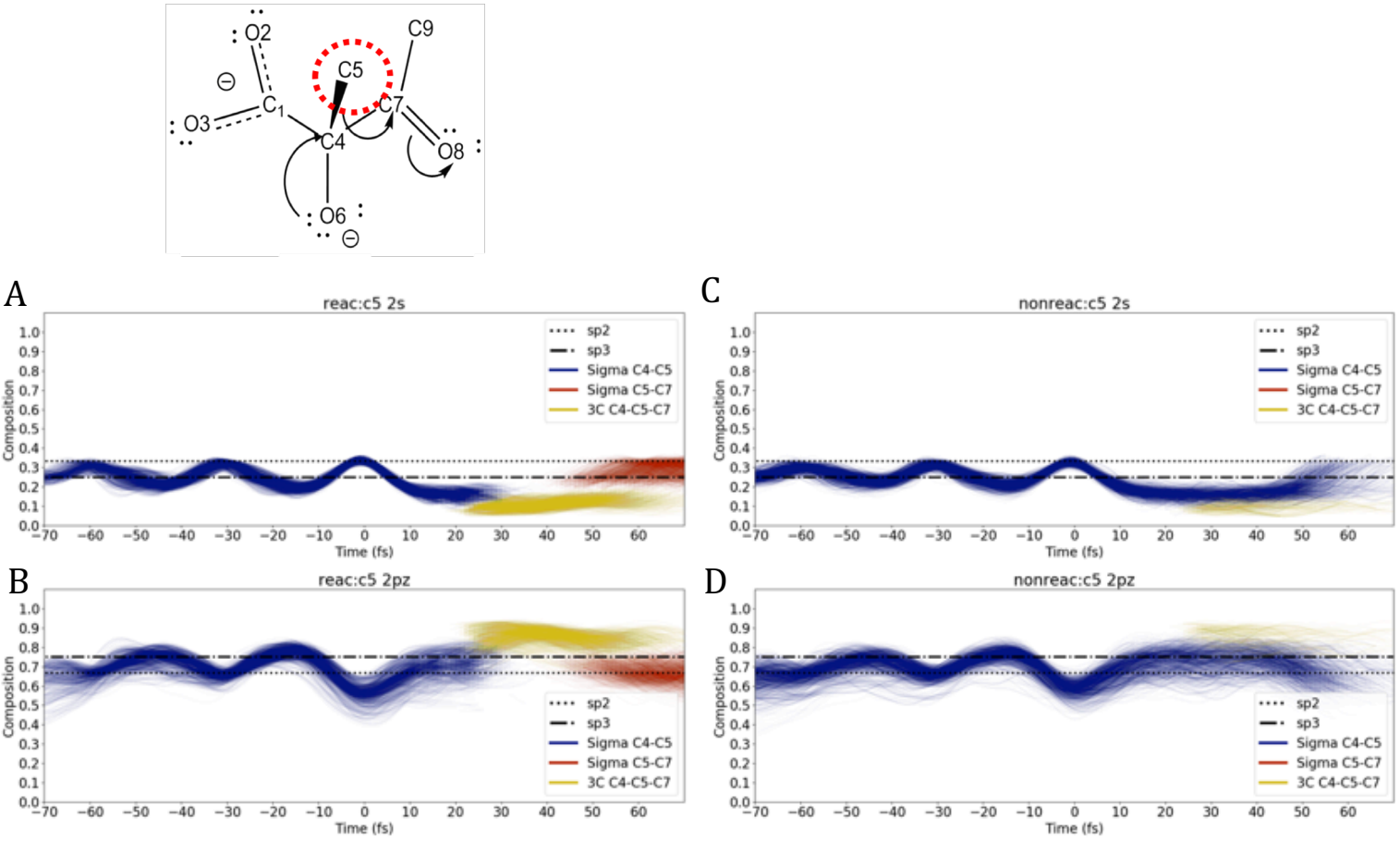


Figure 8: C_5 atomic orbital composition across trajectories, indicated by changing colors, for reactive (panels A and B) and non–reactive (panels C and D) ensembles. (A) The $2s$ character for the C_5 atom of the reactive ensemble exhibited similar oscillatory behavior as C_4 at times prior to $t = 0$ fs. As the reaction proceeded, the $2s$ character diminished by $10-15\%$, but did not entirely disappear as in the case of C_4. As the reactive ensemble established the $C_7 - C_5$ σ–bond, the $2s$ character returned to baseline sp^3 values. (B) The reactive ensemble showed increased $2p_z$ character through the participation of the 3C bond but possessed equivalent values of the baseline between the reactant–state σ–bond between $C_4 - C_5$ and the product–state σ–bond between $C_7 - C_5$. (C) The non–reactive ensemble observed a persistent decrease in the $2s$ character of the $C_4 - C_5$ σ–bond over the times including $t = +20$ fs to $t = +40$ fs but returned to baseline values of about 25% as the reaction failed to proceed. (D) The non–reactive ensemble showed increased $2p_z$ character throughout the reaction, but returned to a baseline value of about 75% as the reaction attempted to proceed. The few trajectories that were capable of forming the 3C bond exhibited slightly higher $2p_z$ character than those that could not. Taken with panel (C), no significant change in hybridization state occurred for C_5, suggesting this carbon remained sp^3–hybridized throughout the reaction. In all panels, dashed horizontal black lines show canonical orbital proportions for static sp^2 and sp^3 hybridization.

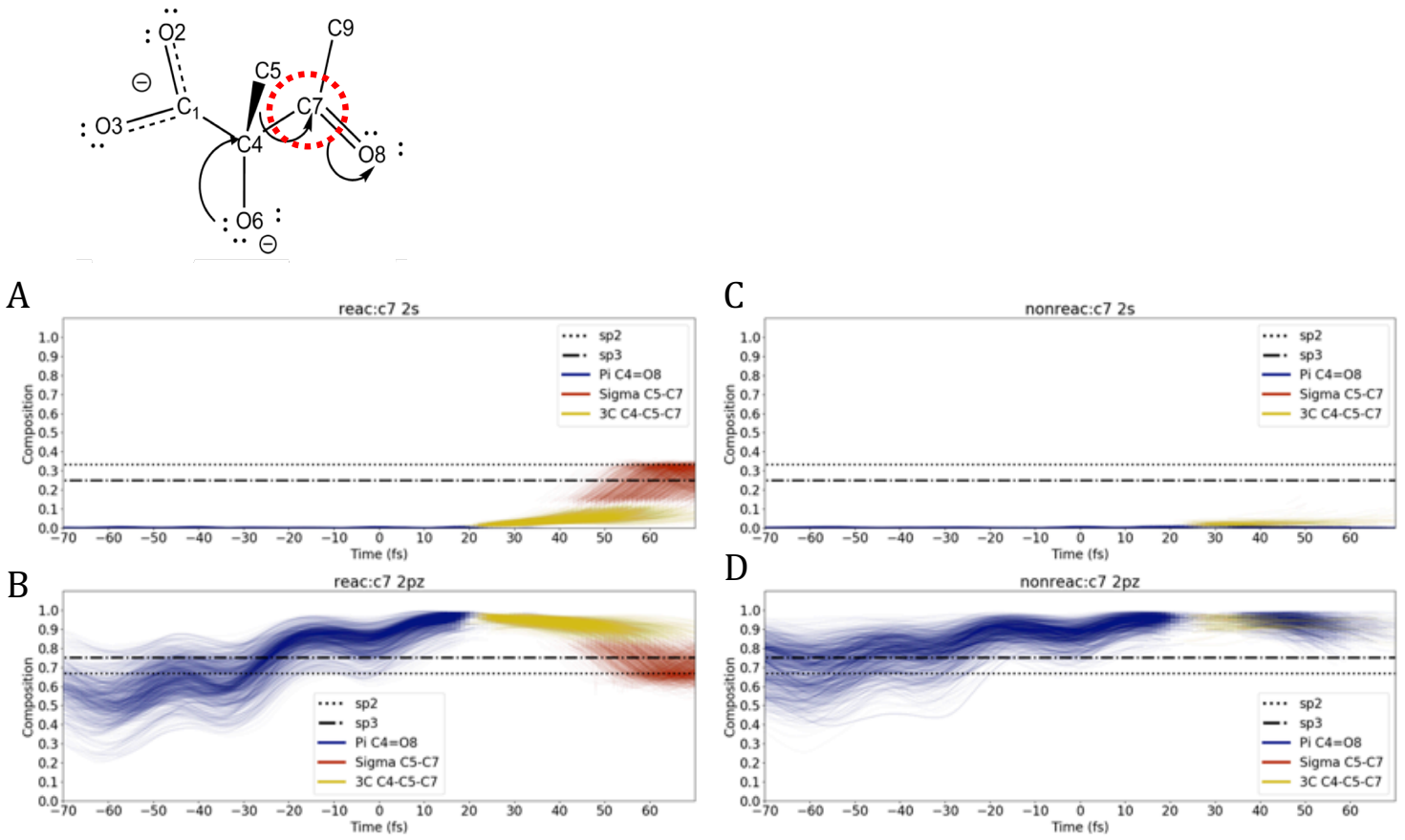


Figure 9: C_7 atomic orbital composition across trajectories, indicated by changing colors, for reactive (panels A and B) and non–reactive (panels C and D) ensembles. (A) Atom C_7 participated in a π–bond in the reactant state, indicated by a lack of $2s$ character in the π–bond between $C_7 - O_8$. As the reaction proceeded toward 3C bond formation, $2s$ character grew until reaching nearly 25% as the $C_7 - C_5$. σ–bond formed. (B) The $C_7 - O_8$ π–bond grew in $2p_z$ character until formation of the 3C bond. As the $C_7 - C_5$ σ–bond formed, the $2p_z$ character reduced to about 75%. Taken with panel (A), this suggested that C_7 underwent a hybridization change from sp^2, with 1 free 2p orbital (specifically $2p_z$), to sp^3 hybridized. (C) The non–reactive ensemble was not capable of facilitating a truly successful methyl migration, and showed no developing $2s$ character in the $C_7 - O_8$ π–bond. For the few trajectories capable of forming the 3C bond, these contained negligible $2s$ character in the 3C bond, and also returned to 0% as the reaction proceeded to transition back to the π–bond. (D) Conversely, while there was little $2s$ character, the $C_7 - O_8$ π bond increased in $2p_z$ character as the reaction proceeded. Between $t = +20$ fs and $t = +40$ fs, the π–bond was not sustained, but the non–reactive ensemble reestablished this orbital. In all panels, dashed horizontal black lines show canonical orbital proportions for static sp^2 and sp^3 hybridization.

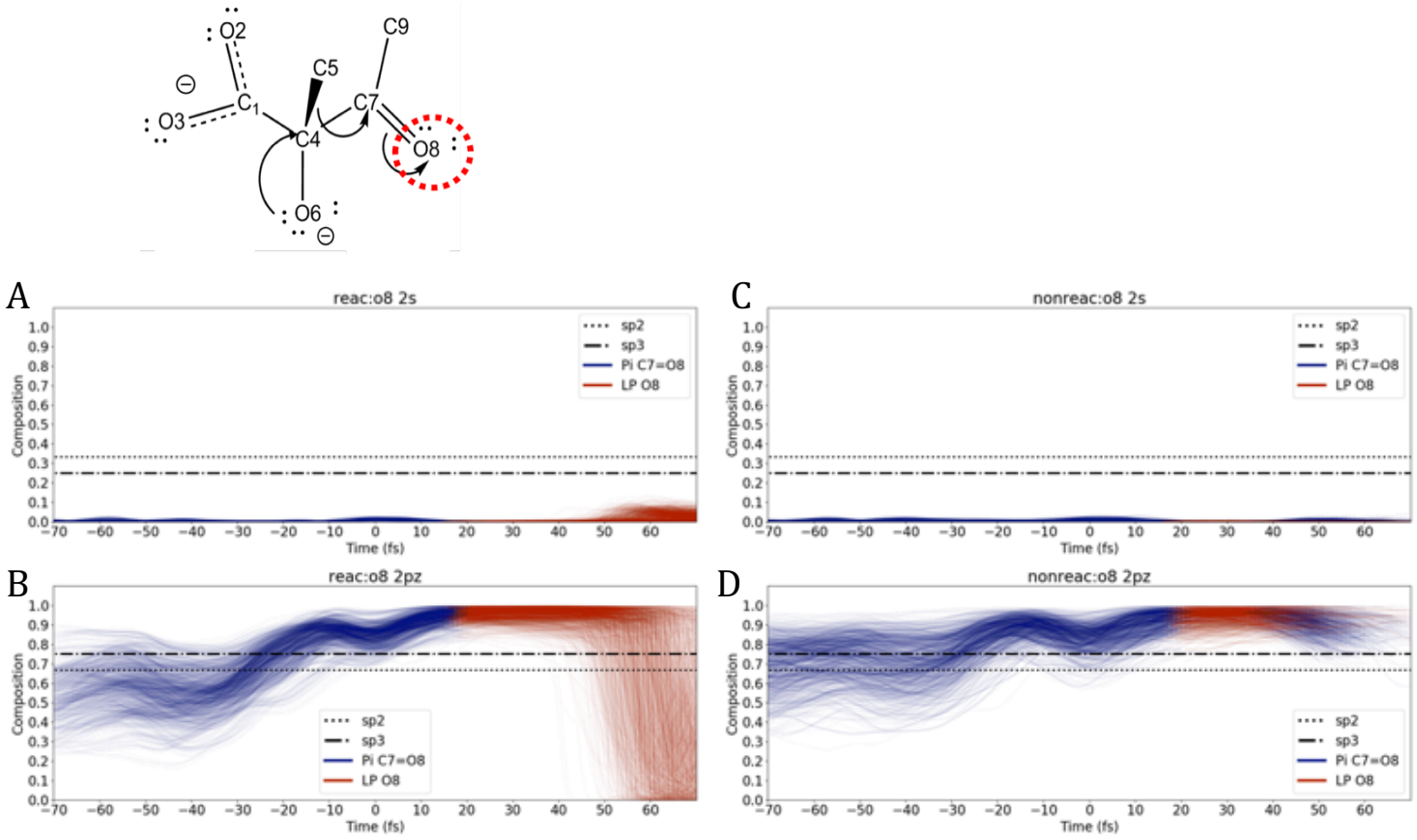
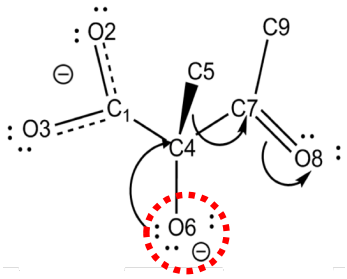


Figure 10: O_8 atomic orbital composition across trajectories, indicated by changing colors, for reactive (panels A and B) and non–reactive (panels C and D) ensembles. (A) While in the reactant basin, the reactive ensemble exhibited negligible $2s$ character in the $C_7 - O_8$ π bond. Near the time the reactive ensemble successfully established the $C_7 - C_5$ σ–bond (around $t = +50$ fs), the $2s$ character of the lone pair on O_8 marginally increased. (B) The $C_7 - O_8$ π–bond grew in $2p_z$ character until $t = +20$ fs where the $2p_z$ orbital was classified as part of a new lone pair orbital on O_8. The new lone pair orbital was maintained at nearly 100% $2p_Z$ character until $t \geq 50$ fs, where a rapid rehybridization occurred. (C) Similar to the reactive ensemble, marginal $2s$ character existed in either the $C_7 - O_8$ π–bond or the lone pair orbital of O_8. (D) The non–reactive ensemble approached the reactive barrier comparably to the reactive ensemble with nearly 100% $2p_z$ character. Like the reactive ensemble, all non–reactive trajectories sustained the formation of the lone pair of O_8 between $t = +20$ fs to $t = +40$ fs. Subsequently however, the non–reactive ensemble reestablished the π–bond between $C_7 - O_8$. In all panels, dashed horizontal black lines show canonical orbital proportions for static sp^2 and sp^3 hybridization.



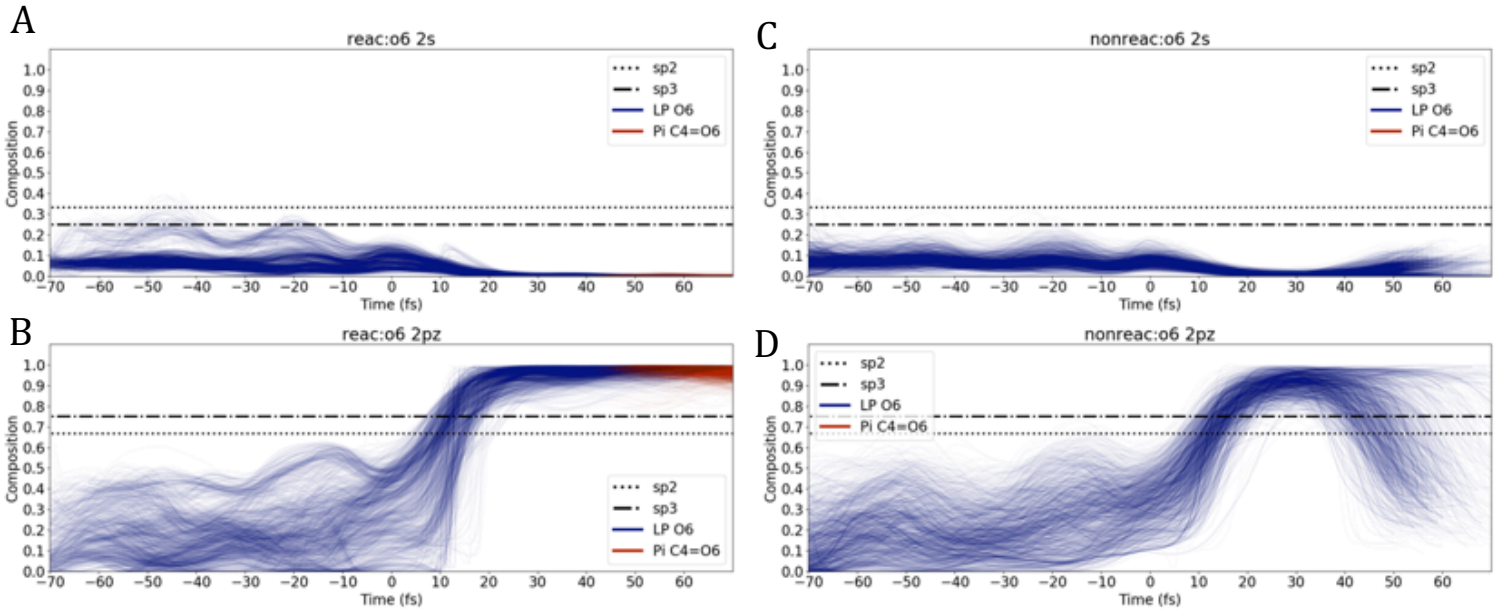


Figure 11: O_6 atomic orbital composition across trajectories, indicated by through changing colors, for reactive (panels A and B) and non–reactive (panels C and D) ensembles. (A) The lone pair orbital corresponding to the negative charge on O_6 for the reactive ensemble exhibited marginal $2s$ character. By time $t = +30$ fs, negligible $2s$ character existed in either the lone pair orbital, or the subsequent $C_4 - O_6$ π–bond. (B) The lone pair orbital had mixed $2p_z$ character that steadily increased in composition until $t = +10$ fs when the composition was nearly 100% for the reactive ensemble. Throughout the remainder of the reaction, the $2p_z$ character remained at 100% until transitioning into the $C_4 - O_6$ π bond. (C) The non–reactive ensemble had similar $2s$ character to the reactive ensemble. From $t = +20$ fs to $t = +40$ fs, corresponding to the start of the 3C bond formation and the lone pair formation on O_8, there was no $2s$ character, but it marginally increased near the end of the failed reaction. (D) The non–reactive ensemble followed a similar increase in $2p_z$ composition as the reactive ensemble, but rebounded back to the original value after time $t = +30$ fs. In all panels, dashed horizontal black lines show canonical orbital proportions for static sp^2 and sp^3 hybridization.

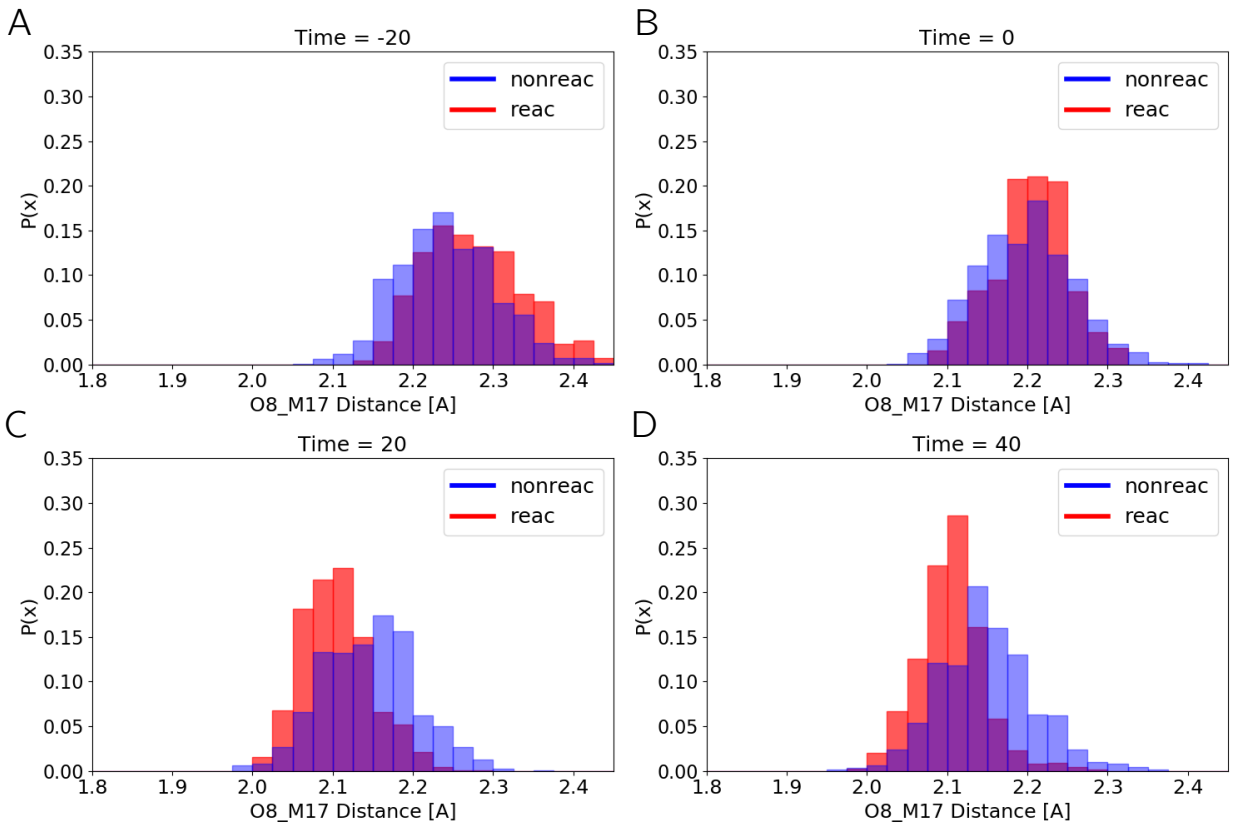


Figure 12: Histograms of reactive (red) and non–reactive (blue) simulations for the $O_8 - M_{17}$ distance (in Å). The distance between $O_8 - M_{17}$ shortened as the simulations grew closer to crossing the reaction barrier. (A) For time $t = -20$ fs, the reactive distribution was centered at (mean ± std. dev) 2.27 Å ± 0.06 Å whereas the non–reactive distribution was centered at 2.24 Å ± 0.06 Å. (B) For time $t = 0$ fs, the reactive distribution shifted left slightly to center at 2.20 Å ± 0.05 Å, whereas the non–reactive distribution remained at 2.20 Å ± 0.06 Å. (C) For time $t = 20$ fs, the reactive distribution shifted closer 2.11 Å ± 0.04 Å, whereas the non–reactive distribution shifted slightly left to 2.15 Å ± 0.06 Å. (D) For time $t = 40$ fs, the reactive distribution remained at 2.10 Å ± 0.04 Å, whereas the non–reactive distribution consistently stayed at 2.15 Å ± 0.06 Å.

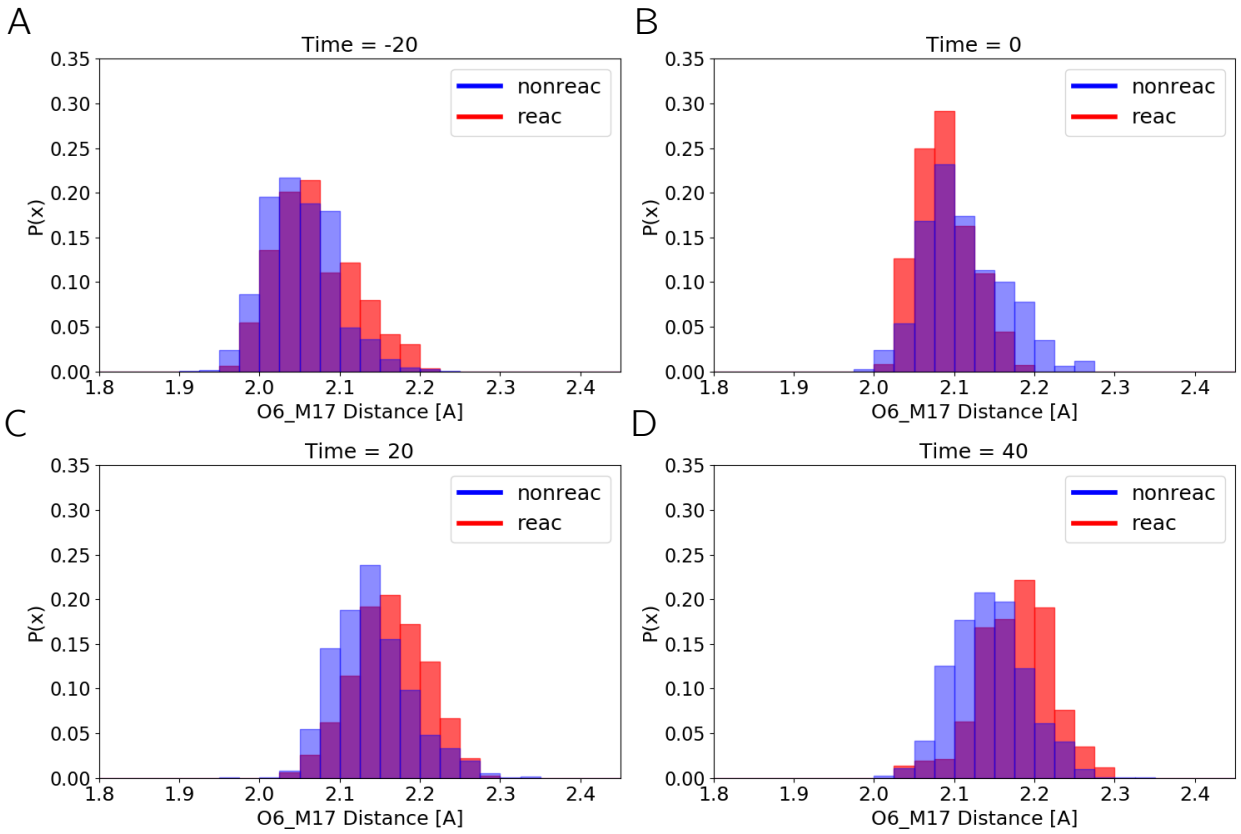


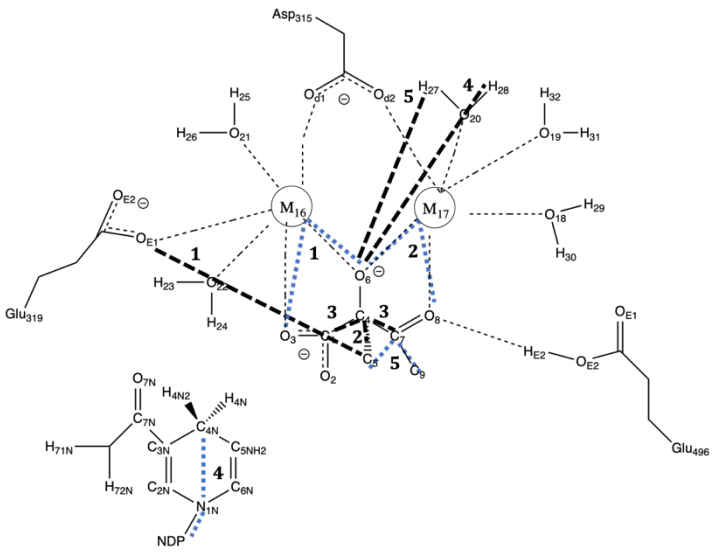
Figure 13: Histograms of reactive (red) and non–reactive (blue) simulations for the $O_6 - M_{17}$ distance (in Å). The $O_6 - M_{17}$ distance was shorter in the reactant basin, prior to attempting to cross the barrier, than they were during the course of the reaction. (A) For time $t = -20$ fs, the reactive distribution was centered at (mean ± std. dev) 2.07 Å ± 0.05 Å, whereas the non–reactive distribution was centered at 2.05 Å ± 0.04 Å. (B) For time $t = 0$ fs, the reactive distribution shifted right slightly to center at 2.09 Å ± 0.03 Å, whereas the non–reactive distribution remained at 2.11 Å ± 0.05 Å. (C) For time $t = 20$ fs, the reactive distribution shifted further to 2.16 Å ± 0.05 Å, whereas the non–reactive distribution shifted slightly right to 2.14 Å ± 0.05 Å. (D) For time $t = 40$ fs, the reactive distribution was centered at 2.18 Å ± 0.05 Å, whereas the non–reactive distribution consistently stayed at 2.14 Å ± 0.05 Å.

3.4.2 Geometric feature classifiers predict reactivity with a subset of the 30 consensus features

We performed feature selection calculations to identify predictive features from 30 relevant features we previously identified (See Methods) [1]. Specifically, we ranked all two–feature classifiers and selected the top five with unique features. We also studied the ten–feature model containing all five pairs of features which had 89% accuracy and 94% AUC (results shown in Table 1). Across the two–feature models, pair accuracy was in the range of 66%–79% with an AUC of 70%–80%. The logistic regression model values are also shown in Table 1. Most coefficients were similar in scale, and all were identical in sign between the corresponding two–feature model and the ten–feature model. The consistent sign suggests that each feature may have a similar role in predictive reactivity in a pair model or combined model.

Model	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}	Acc	AUC
Set 1	5.32	5.55	–0.28	–	–	–	–	–	–	–	–	0.79	0.88
Set 2	2.21	–	–	9.35	–0.14	–	–	–	–	–	–	0.74	0.81
Set 3	21.93	–	–	–	–	–7.2	–7.34	–	–	–	–	0.67	0.73
Set 4	12.27	–	–	–	–	–	–	–1.18	–0.05	–	–	0.68	0.71
Set 5	12.92	–	–	–	–	–	–	–	–	–1.02	–0.08	0.66	0.70
All	79.45	5.37	–0.37	10.27	–0.19	–17.5	–5.82	–1.19	–0.06	–1.18	–0.06	0.89	0.94

Table 1: Average performance (accuracy and AUC) and coefficients of geometric feature models across 10 randomized training/testing splits, for 5 two–feature models (Set 1–5) and the combined ten–feature model (all sets). Standard error was 0.02 for performance. Dashes indicate unused features in a model for the pair–feature models. Feature coefficients are numbered in Figure 14, with β_0 representing the bias term.



Feature Set	Identity	Type
Set 1	GLU319 OE1– C5	Distance
	O6–M16–O3	Angle
Set 2	C5–C4	Distance
	M17–O6–M16	Angle
Set 3	C4–C7	Distance
	C1–C4	Distance
Set 4	H28–O6	Distance
	NDP C4N–N1N–C1NQ	Angle
Set 5	H27–O6	Distance
	C5–C7–C9	Angle

Figure 3.14: Schematic indicating the geometric feature pairs, collected from the $t = -20$ fs time point, where numbering indicates in which pair the feature was identified. Out of 10 features, 6 were distance features (indicated in black) and the remaining were angles (indicated in blue).

3.4.3 Electronic feature classifiers predict reactivity

We used quantum mechanical methods to compute properties describing the electronic structure of the substrate and its environment, and then trained classifiers to identify which of these electronic features were most predictive of reactivity. Specifically, we computed 171 pairwise classifiers from the 19 electronic features identified as potentially important in the mechanism extracted from time–aligned simulations at time $t = -20$ fs. We chose the top performing unique two–feature models for subsequent analysis (Figure 3.15).

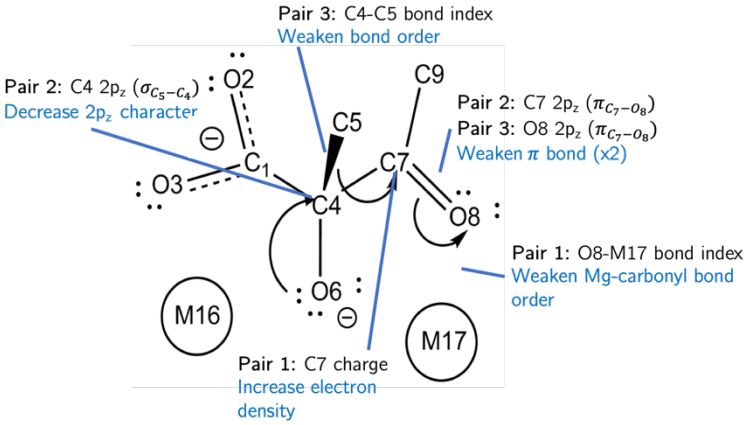
Each two–feature model, taken alone, had an accuracy in the range 80%–85% with AUC 88%–92%, with standard error of performance under 1% across training–testing splits (Table 2).

By design, the first pair of features had highest performance metrics that subsequently fell monotonically with each subsequent pair. A collective set of the top three pairs, corresponding to 6 electronic features, was sufficient to produce a classifier with 91% accuracy and 97% AUC, outperforming significantly the best individual pair. Both the top performing pair features, and the cumulative 6–feature classifier of the electronic features, outperformed the geometric top performing pair and cumulative 10–feature classifier respectively, suggesting that the electronic structure is more diagnostic of reactivity in the sense that fewer electronic descriptors than geometric descriptors are required.

Unlike the geometric classifiers, several of the individual pairs of features exhibited varied magnitude or sign when used in a combined model. Most of the weight of the combined classifier seemed placed on features corresponding to the charge on C_7 and the bond index between $O_8 - Mg_{17}$. The coefficients of $C_4 - C_5$ bond index and the $2p_Z$ character of O_8 of the π–bond between $C_7 - O_8$ exhibited similar magnitude either in an individual pair classifier or as a combined model. The other features, while varied in magnitude, retained the same sign with exception of the $2p_Z$ character of C_4 of the σ–bond between $C_4 - C_5$, which switched sign from positive in the pair–feature model to negative in the combined model.

Model	β_0	β_1	β_2	β_3	β_4	β_5	β_6	Acc	AUC
Set 1	41.71	–55.78	–77.62	–	–	–	–	0.85	0.92
Set 2	–9.17	–	–	49.05	–34.49	–	–	0.82	0.88
Set 3	49.36	–	–	–	–	–19.15	–38.01	0.80	0.88
All	133.11	–104.25	–164.24	–1.53	–7.34	–14.80	–38.00	0.91	0.97

Table 2: Average performance (accuracy and AUC) and coefficients of electronic feature models across 10 randomized training/testing splits, for three two–feature models (Set 1–3) and the combined 6–feature model (All). Standard error was 0.01 for performance. Feature coefficients are numbered as in Figure 15, with β_0 representing the bias term.



Feature Set	Identity	Type
Set 1	C7	Charge
	O8–M17	Bond Index
Set 2	2pz C4; σ C4–C5	Atomic Orbital Hybridization
	2pz C7; π C7–O8	Atomic Orbital Hybridization
Set 3	2pz O8; π C7–O8	Atomic Orbital Hybridization
	C4–C5	Bond index

Figure 15: Schematic of electronic features that were most predictive of reactivity. Out of 6 features, three were atomic orbital hybridization states, explicitly of $2p_z$ character, two were bond–indices, and one was partial atomic charge. Text in blue indicates how the electronic feature changes as the reaction progresses at the $t = -20$ fs timepoint.

3.4.4 Geometric feature OE1–C5 influences torsional orientation of methyl that weakens electronic feature C4–C5 bond order

Our dynamic trajectories revealed that the reactive ensemble, compared to the non–reactive ensemble, had a larger fraction of trajectories that started to cross the reaction barrier with an eclipsed conformation of the methyl, C_5, with respect to the reactant carbon, C_4. (Supplementary Figure 3.S1, See Methods for torsion angle calculation). Further inspection of both ensembles, stratified by the torsion angle of C_5, indicated that trajectories crossing the reaction barrier starting from the eclipsed conformation versus the staggered conformation, particularly for the reactive ensemble, had higher $C_4 - C_5$ σ–bond orbital energies and lower $C_4 - C_5$ σ*–antibond orbital energies, corresponding to a destabilization of the breaking–bond orbital and a stabilization of the anti–bonding orbital respectively, prior to 3C bond formation (Supplementary Figure 3.S2).

Moreover, the $C_5 - C_7$ σ–bond and 3C bond orbital energies were lower for the reactive populations that left the reactant basin in the eclipsed orientation compared to the staggered orientation (Supplementary Figure 3.S3). There was an anti–correlated relationship between the torsional orientation with respect to C_4 and the torsional orientation with respect to C_7 which was time invariant (Supplementary Figure 3.S4); said otherwise, an eclipsed orientation with respect to the reactant transferred in the staggered orientation to the product. This suggested that the torsional orientation of the methyl prior to crossing the reaction barrier assisted catalysis by potentially creating a conformation that destabilized the ground state and stabilized the transition–state.

The torsional freedom and orientation of the methyl group can be influenced by the nearby environment, namely, interactions with nearby side–chains in close proximity. Correspondingly, residue E319's carboxylate oxygens were considered in close enough proximity to create a CH–O type bond with the methyl protons [37–44]. The first model–selected geometric feature in this work, and the most commonly selected feature in prior work by Bonk et al., E319/$OE_1 - C_5$, signaled a CH–O type bond between the transferring methyl's protons and one of E319's carboxylate oxygens (OE_1), indicated by the strong correlation between the E319/$OE_1 - C_5$ and the E319/$OE_1 - H$ distance where "H" represented the closest methyl proton to OE_1 [1] (Supplementary Figure 3.S5). Prior work studying methyl transfer reactions in enzymes have indicated that CH–O type bonds can restrict rotational mobility, which influences catalysis [43–45]. In the case of KARI, the E319/$OE_1 - H$ distance for the reactive ensemble was on average 0.3 Å further away than the non–reactive ensemble (the mean ± std.dev for reactive was 2.3 ± 0.2 Å versus non–reactive 2.0 ± 0.2 Å; Figure 3.16A). The corresponding p–value between these distributions, via the Student's t–test, was $p < 1 \times 10^{-10}$, suggesting the difference between the

reactive and non–reactive ensembles were significant. Stratifying the ensembles by the C_5 torsion revealed that, for the reactive ensemble, the eclipsed oriented methyl groups were more likely to have elongated $OE_1 - H$ distances than increasingly staggered orientations, suggesting that the larger distance between $OE_1 - H$ encouraged the eclipsed conformation. In contrast, there was no notable trend with the non–reactive ensemble, but the median population of the E319/$OE_1 - H$ distance was shorter than the reactive ensemble for all comparable populations.

We observed that the torsion angle also influenced one of the electronic features, the $C_4 - C_5$ bond index. The reactive ensemble was observed to have a smaller bond–order (the mean ± std.dev for reactive was 0.87 ± 0.03 versus non–reactive 0.90 ± 0.03, $p < 1 \times 10^{-10}$), corresponding to a weakened electronic overlap between the $C_4 - C_5$ σ–bond compared to the non–reactive ensemble (Figure 3.16C). Stratified on torsion angle, the eclipsed orientation in the reactive ensemble showed nearly a 6% decrease in bond index when compared to the staggered orientation (Figure 3.16D). Similar to the $OE_1 - H$ analysis, no clear trend was identified for the non–reactive ensemble but likewise, bond index was higher for each comparable population between non–reactive and reactive. A comparable analysis performed on the $C_4 - C_5$ σ–bond orbital energy to verify the diminished strength of this bond corroborated that the reactive ensemble was nearly 16 kcal higher in energy (Supplemental Figure 3.S6A; reactive mean ± std. dev = –454 ± 21 kcal vs non–reactive = –470 ± 21 kcal, with $p < 1 \times 10^{-10}$ via t–test), and that the reactive ensemble's eclipsed orientation was higher in energy than the staggered orientation.

Taken together, these results suggest a potential catalytic strategy whereby an eclipsed torsion angle assists reactivity by weakening the $C_4 - C_5$ σ–bond, and such an orientation is promoted when the CH–O bond from E319 is weakest (as indicated by a longer $OE_1 - C_5$ distance). While the torsion angle may play some role in reactivity, we note that the majority of

both ensembles, reactive and non–reactive, were more likely to be staggered than eclipsed (Supplementary Figure 3.S1), suggesting it is not a necessary requirement for catalysis. It is very likely there are other strategies that allow for the reactive ensemble's staggered trajectories to cross the barrier, further underscored by Bonk et al.'s claim that clustering the same features indicate 5 disparate reactive pathways [1]. As seen in Figure 3.16A and Figure 3.16C, the reactive and non–reactive ensembles share considerable overlap despite being statistically different distributions. It is possible the trajectories that lie in the overlap region deviate significantly with respect to other features, which is potentially an indicator of why cumulative classifiers are capable of predicting reactivity better than any given pair of features.

3.4.5 Pairwise geometric feature models predict electronic features as well as the cumulative geometric model

The model–selected set of electronic features that predicted reactivity reported on different aspects of the mechanism; from the NBO analyses, the features indicated information on the breaking–bond ($C_4 - C_5$ bond index, C_4 2pz orbital character in the $C_4 - C_5$ σ–bond) and the diminishing $C_7 - O_8$ π–bond/eventual lone pair formation on O_8 (C_7 charge, C_7 2pz and O_8 2pz character in the $C_7 - O_8$ π–bond, and $O_8 - M_{17}$ bond index). Given that predictive power increased for the cumulative electronic model as opposed to the pairwise models, a diverse set of descriptors that reported on different details of the mechanism was important for reactivity. A pair of geometric features (or even a single geometric feature) may not be capable of faithfully capturing all components of the chemical mechanism alone; diversity in geometric features can ensure implicit electronic structure was described adequately.

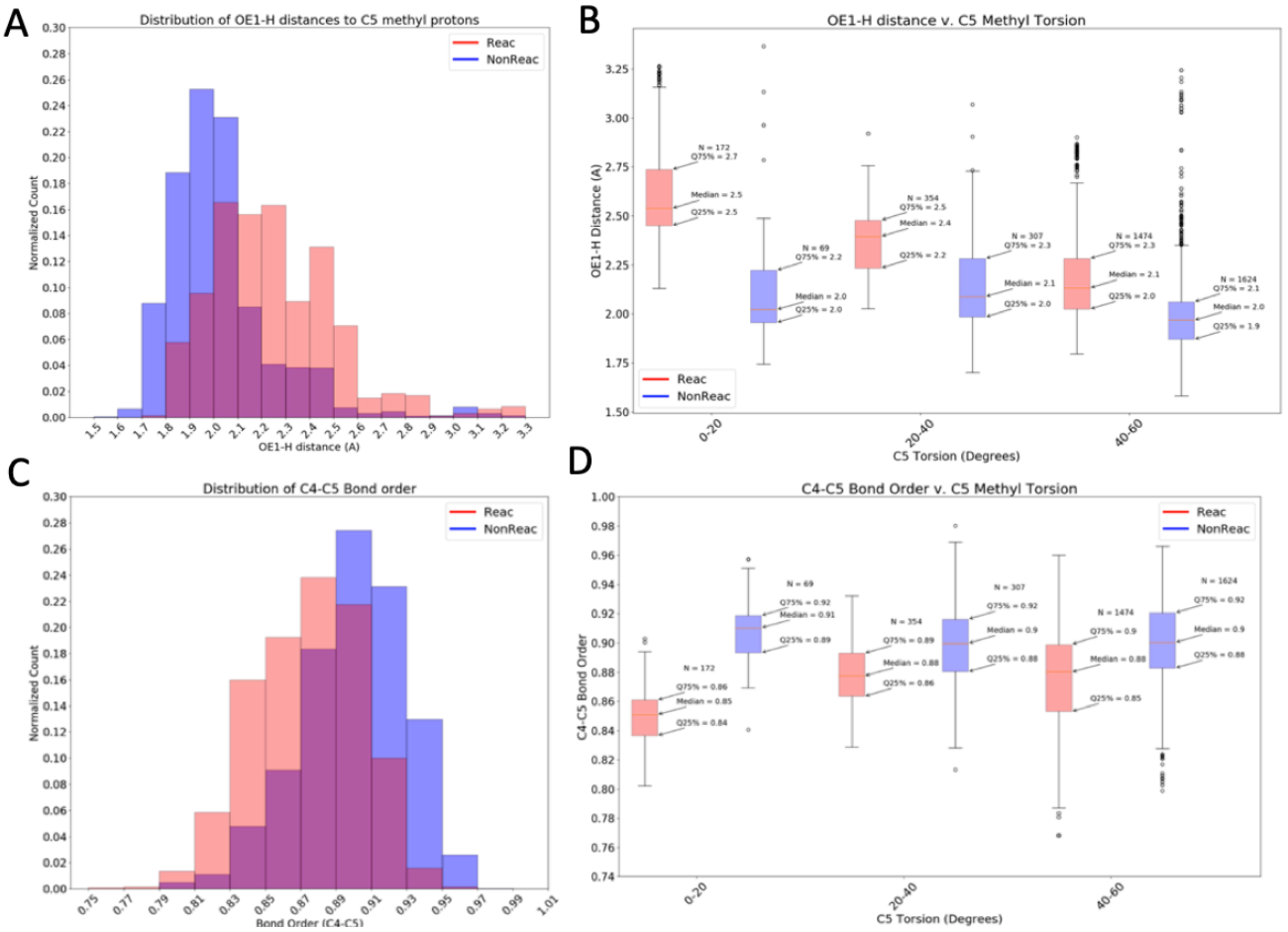


Figure 16: Relationship of $OE_1 - C_5$ distance to reactivity. $OE_1 - C_5$ is the distance between the methyl carbon (C_5) and Glu319's carboxylate oxygen (OE_1) used to coordinate Mg. This distance was strongly correlated to the distances of the methyl protons to Glu319's oxygen. (A) Methyl–proton to Glu319–O distances ($OE_1 - H$) were significantly shorter in the non–reactive simulations (mean ± std. dev: 2.0 Å ± 0.2 Å) than the reactive simulations (2.3 ± 0.2 Å). (B) Distribution of methyl–proton to carboxylate oxygen distance as a function of methyl torsional conformation for reactive and non–reactive simulations. The data was separated into three bins: "eclipsed" between 0–20 degrees; "staggered" between 40–60 degrees; and intermediate between 20–40 degrees. Reactive simulations showed an association between an eclipsed state of the methyl prior to reacting ($t = -20$ fs) and a farther $OE_1 - H$ distance. The median $OE_1 - H$ distance for the eclipsed orientation was 2.5 Å for reactive simulations, contrasted with 2.1 Å for the staggered orientation. The non–reactive distribution did not exhibit a trend between the methyl–proton–$H - OE_1$ distance and methyl torsion, as all proton–distances were relatively short across any torsional conformation. (C) Distribution of the reactive and non–reactive bond–order between $C_4 - C_5$. Reactive simulations adopted lower overall bond–order (mean ± std. dev: 0.87 ± 0.03) than non–reactive simulations on average (mean ± std. dev: 0.90 ± 0.03). (D) Association of bond order as a function of methyl torsional orientation. Reactive simulations that adopted torsional conformations of 0–20 degrees (i.e. eclipsed rather than staggered) had a median $C_4 - C_5$ bond order centered at 0.85 compared to the 40–60 (i.e. staggered) orientations at 0.88. This difference of 0.03 bond order is notable, as it is identical to the difference in the distributions' overall average, and equal to the standard deviation of both distributions.

In order to investigate the relationship between the geometric and the electronic features, we trained geometric feature models (with the 10 geometric features that were shown to predict reactivity) on a different task where instead of predicting reactivity, models classified whether a given electronic feature was *larger* or *smaller* than the combined (reactive and non–reactive) ensemble average. Similar to prior analyses, we selected the top performing geometric pairs for each of the six electronic features, and compared it to a cumulative model trained with all 10 geometric features predicting the electronic feature (Table 3 and 4).

The performance of the geometric pair–feature models in predicting the electronic features spanned from 68% to 87% with an average (across 10 training/testing splits) of 75% for accuracy, and 70% to 95% with an average of 80% for AUC and a standard error of 1% or less, unless otherwise stated (Table 3). Comparatively, the performance of a 10–feature model, employing the use of all geometric features in distinguishing an electronic feature, spanned from 69% to 89% with an average of 77% for accuracy and 77% to 96% for AUC, with an average of 84%. Notably, only 2 out of the 6 electronic features exhibited an increase of over 2% accuracy prediction when a classifier was trained with all 10 features as opposed to 2 features, which was within twice the standard error. When considering each pairwise model for every electronic feature, we observed 7 out of the 10 geometric features were represented. Some geometric features were popular among models: this included the E319/$OE_1 - C_5$ distance, $C_4 - C_5$ distance, and $C_4 - C_7$ distance, which appeared in three models for the former two, and 2 for the latter.

The marginal increase in predictive performance, given additional geometric features beyond the original pair, and the diversity of which geometric feature pairs were most effective in predicting the different electronic classification tasks suggested two things: (i) small subsets of geometric features were enough to describe the electronic feature effectively and (ii) different

subsets of geometric features were capable of reporting on various parts of the chemical mechanism, which may account for the boost in predictive performance for the cumulative classifier in predicting reactivity. The popularity of some features may suggest that there are some geometric features that influence many electronic features, hence other descriptors were required to account for the other facets of the electronic structure.

Electronic Feature	Geometric Features	Coefficient	Bias	2ft:Mean Accuracy	2ft:Mean AUC	10ft: Mean Accuracy	10 ft: Mean AUC
C7 charge	C5-C4	-2.41	-23.47	0.79	0.87	0.84	0.91
	C4-C7	4.42					
O8-M17 bond order	H27-O6	8.00	-13.74	0.68	0.70	0.69	0.77
	M17-O6-M16	8.73					
C4 2pz; σ C4-C5	GLN319 OE1-C5	-4.31	-32.36	0.73	0.81	0.77	0.85
	C5-C4	2.64					
C7 2pz; π C7=O8	GLN319 OE1-C5	-4.09	6.57	0.74	0.76	0.73	0.77
	C5-C7-C9	5.08					
O8 2pz; π C7=O8	GLN319 OE1-C5	-3.69	5.93	0.71	0.73	0.72	0.78
	NDP C4N-N1N-C1NQ	3.27					
C4-C5 bond order	C5-C4	-3.96	31.80	0.87	0.95	0.89	0.96
	C4-C7	2.5					

Table 3: Coefficients and model performance for classifiers that predict each electronic feature from geometric features. "2ft" represents the model with the best performance among all pairwise geometric classifiers; it; for each electronic feature, the corresponding geometric features and coefficients are shown. The bias corresponds to the 2ft model. "10ft" represents the model that uses all ten geometric features.

Ele model	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
C7	-31.4	-1.32	-25.66	42.44	-12.40	0.86	1.34	0.09	0.07	0.04	0.06
O8-M17	-48.7	0.59	8.62	-11.85	4.28	0.83	0.67	-0.01	0.21	0.02	0.07
C4 2pz	-2.3	-3.80	25.25	-7.03	-16.17	-1.05	-0.54	-0.02	0.04	-0.00	0.06
C7 2pz	-16.1	-4.23	1.89	9.32	-3.49	-0.08	0.15	0.08	0.04	0.02	0.03
O8 2pz	-28.9	-4.32	4.53	11.08	-1.49	0.18	0.54	0.18	-0.03	0.01	0.05
C4-C5	6.5	0.41	-51.70	25.18	23.24	-0.68	-0.18	0.01	0.05	0.02	0.02

Table 4: Coefficients for the 10–feature geometric models that predict each of the six electronic features. Features for the geometric classifier are numbered as in the geometric schematic in Figure 14.

3.5 Conclusion and Future Directions

In the current work, we explored the electronic drivers of reactivity and how structural components of the active site of KARI influence the reaction through their effect on electronic properties. To our knowledge, this work represents the first dynamical electronic description of the KARI methyl transfer reaction supporting the existence of a three–center–two–electron bond during concurrent transfer and isomerization.

NBO analyses of the reactive and non–reactive ensembles revealed methyl transfer and isomerization substrate carbonyls/hydroxyl groups occurred concurrently. Further inspection of the atomic orbital compositions throughout the course of the reaction showed that the reaction proceeded through a transition–state that had a three–center–two–electron bond via overlapping $2p_z$ atomic orbitals of the central carbons on the substrate, and orthogonal to the plane formed by the atoms O_6, C_4, C_7 and O_8. Interestingly, the non–reactive ensemble adopted an electronic structure similar to the reactive ensemble, and failure to react was more of a consequence of the extent. However, further inquiry is needed to determine how the electronic mechanism changes as a function of the methyl–transfer coordinate; the non–reactive ensemble analyzed in this work allowed the transferring methyl group to stretch considerably (up to 2.1 Å in some cases) before failing to proceed further up the reaction barrier. Establishing different order parameter criteria in generating the non–reactive ensemble (wherein a threshold of a lower order parameter could result in a shorter stretch of the methyl transfer coordinate before returning to the reactant basin) and repeating this analysis could reveal other electronic departures from successful catalysis beyond the observed mechanism that could provide useful features for enzyme design. Delocalized systems are difficult to characterize with NBO, as there can be many compatible Lewis–like interpretations (i.e. resonance structures); while the mechanism characterized in this work

corroborates prior empirical observations and hypotheses, higher levels of theory and direct inspection of the density matrix components could reveal more subtle interactions that NBO approach may simplify [21–23, 46–51].

Inspection of the dynamic trajectories revealed the torsional orientation of the transferring methyl group influenced whether the system crossed the barrier with higher $C_4 - C_5$ σ–bond orbital energies, such that an eclipsed orientation promoted reactivity. The first selected model feature, the $OE_1 - C_5$ distance (E319's glutamyl side chain oxygen and the substrate's methyl group) highlighted a CH–O type interaction with the glutamyl side–chain oxygen and the methyl protons where close proximity affected the methyl's ability to adopt certain conformations, as eclipsed orientations possessed larger distances compared to staggered orientations. Comparison of the reactive and non–reactive ensembles revealed the non–reactive ensemble had closer contact with E319, which taken with the earlier findings, may have discouraged reactivity by restricting the orientation of the methyl group. A similar analysis indicated this torsional orientation also influenced one of the electronic features, the $C_4 - C_5$ bond index. The reactive ensemble distribution on average had lower bond index than the non–reactive ensemble, suggesting a weakened electronic overlap between the $C_4 - C_5$ σ–bond. Similarly, the eclipsed orientation corresponded to lower $C_4 - C_5$ bond index than the staggered orientation for the reactive ensemble. While the torsional angle may assist in some part with successful barrier crossing, the conformation alone may not be sufficient to cross the activation barrier of KARI, as seen by the large percentage of trajectories in both reactive and non–reactive ensembles that still crossed the barrier in the staggered orientation [20, 27]. Several catalytic strategies may need to be leveraged, perhaps simultaneously, in order to make substantial progress up the reaction barrier.

Logistic regression classifiers of electric–only and geometric–only models demonstrated that fewer electronic features (6) could predict reactivity as effectively as a larger set of geometric features (10). The cumulative set of electronic features reported on at least two distinct components of the mechanism: the breaking of the $C_4 - C_5$ σ–bond and $C_7 - O_8$ π–bond. Given the cumulative model outperformed any pairwise classifier, it implied that a larger set of descriptors capable of capturing more attributes of the mechanism pinpoint the drivers of reactivity. In subsequent classification tasks where models, using the subset of geometric features, predicted whether an electronic feature was larger or smaller than the combined–ensemble average, we observed pairwise classifiers were nearly as effective as the entire cumulative model in predictive performance, and that 7 out of 10 of the subset of geometric features were incorporated into at least one electronic model. Taken together, this implied that the cumulative classifier may be employing subsets of features to describe different components of reactivity which may explain the increased predictive performance.

A key extension of this work would identify the correlated relationships between geometric features and their influence on an electronic state. A perturbation in isolation to only one (geometric) feature would not necessarily be physically feasible (e.g. stretching the distance of two atoms in the active–site would make at least one of them closer to the other atoms in the system); for this reason, identifying the strongest relationships between the geometric features and studying their effect energetically could provide insight into what conformational sets promote reactivity. This could potentially be performed by investigating the informatic overlap (such as via mutual information, or conditional/joint probabilities) to provide a robust view on which geometric features are most uniquely representative of different desirable electronic properties, and which are truly redundant.

3.6 References

1. B.M. Bonk, J. Weis, B. Tidor. Machine Learning Identifies the Chemical Characteristics That Promote Enzyme Catalysis. *J. Am. Chem. Soc*. 141, 4108–4118, 2019.
2. A.D. St–Jacques, O. Gagnon, R.A. Chica. Computational enzyme design: Successes, challenges and future directions. *Modern BioCatalysis: Advances toward synthetic biological systems. Roy. Soc. Chem*, 32, 88–111, 2018.
3. S.C.L. Kamerlin, A. Warshel. At the Dawn of the 21st century: Is Dynamics the Missing Link for Understanding Enzyme Catalysis?. *Proteins*. 78, 6, 1339–1375, 2010.
4. P. Huang, S.E. Boyken, D. Baker. The coming age of de novo protein design. *Nature*, 537, 32, 320–327, 2016.
5. J.Z. Ruscio, J.E. Kohn, K.A. Ball, T. Head–Gordon. The influence of protein dynamics on the success of computational enzyme design. *J. Am. Chem. Soc.,* 131 (39), 14111, 2009.
6. E.C. Alley, G. Khimulya, S. Biswas, M. AlQuaraishi, G.M. Church. Unified rational protein engineering with sequence–based deep representation learning. *Nat. Methods*, 16, 1315–1322, 2019.
7. K.K. Yang, Z. Wu, F.H. Arnold. Machine–learning–guided directed evolution for protein engineering. *Nat. Methods*, 16(8):687–694, 2019.
8. S.C. Hammer, A.M. Knight, F.H. Arnold. Design and evolution of enzymes for non–natural chemistry. *Current Opinion in Green and Sustainable Chemistry*, 7, 23–30, 2017.
9. G. Qu, A. Li, C.G. Acevedo–Rocha, Z. Sun, M.T. Reetz. The crucial role of methodology development in directed evolution of selective enzymes. *Angew. Chem.,* 59, 32, 13204–13221, 2019.

10. C. Zeymer, D. Hilvert. Directed Evolution of Protein Catalysts. *Annu. Rev. Biochem*., 87, 131–157, 2018.
11. G. Kiss, N. Celebi–Olcum, R. Moretti. D. Baker, K.N. Houk. Computational Enzyme Design. *Angew. Chem. Intl. Ed.*, 52, 5700–5725, 2013.
12. D. Verma, G. Grigoryan, C. Bailey–Kellog. Structure–based design of combinatorial mutagenesis libraries. *Pro. Sci.,* 24, 895–908, 2015.
13. E. Wrenbeck, L.R. Azouz, T.A. Whitehead. Single mutation fitness landscapes for an enzyme on multiple substrates reveal specificity is globally encoded. *Nat. Comm*., 8, 15695, 1–10, 2017
14. J.A. McCourt, R.G. Duggleby. Acetohydroxy synthase and its role in the biosynthetic pathway for branched–chain amino acids. *Amino Acids*., 31: 173–210, 2006.
15. V. Biou, R. Dumas, C. Cohen–Addad, R. Douce, D. Job, E. Pebay–Peyroula. The crystal structure of plant acetohydroxy acid isomeroreductase complexed with NADPH, two magnesium ions and a herbicidal transition state analog determined at 1.65 Å resolution. *EMBO J,* 16(12): 3405–3415, 1997.
16. R. Dumas, M.C. Butikofer, D. Job, R. Douce. Evidence for two catalytically different magnesium– binding sites in acetohydroxy acid isomeroreductase by site–directed mutagenesis. *Biochemistry*, 34, 6026–6036, 1995.
17. R. Dumas, D. Job, J. Ortholand, G. Emeric, and A. Greiner. Isolation and kinetic properties of acetohydroxy acid isomeroreductase from spinach *(Spinacia oleracea)* chloroplasts overexpressed in *Escherichiacoli. Biochem. J.,* 288, 1992.
18. S. K. Chunduru, G. T. Mrachko, K. C. Calvo. Mechanism of ketol acid reductoisomerase steady-state analysis and metal ion requirement. *Biochemistry*, 28(2):486-93, 1989.

19. R. Dumas, V. Biou, F. Halgand, R. Douce, R. G. Duggleby. Enzymology, structure, and dynamics of acetohydroxy acid isomeroreductase. *Acc. Chem. Res.,* 34(5):399–408, 2001.

20. R. Tyagi, Y. Lee, L.W. Guddat, R.G. Duggleby. Probing the mechanism of the bifunctional enzyme ketol–acid reductoisomerase by site–directed mutagenesis of the active site. *FEBS Journal*, 272, 593–602, 2005.

21. F. Proust–De Martin, R. Dumas, M.J. Field. A Hybrid–Potential Free–Energy Study of the Isomerization Step of the Acetohydroxy Acid Isomeroreductase Reaction. *J. Am. Chem. Soc.,* 122, 7688–7697, 2000.

22. A. Aulabaugh, J.V. Schloss. Oxalyl Hydroxamates as Reaction–Intermediate Analogues for Ketol–Acid Reductoisomerase. *Biochem*. 29, 2824–2830, 1990.

23. N.W. Silver, Ensemble methods in computational protein and ligand design: Applications to the Fc–γ immunoglobulin, HIV– 1 protease, and ketol–acid reductoisomerase system. Doctoral Dissertation, Massachusetts Institute of Technology, 2012.

24. V.S. Krishna, S. Zheng, E.M. Rekha, L.W. Guddat, D. Sriram. Discovery and evaluation of novel Mycobacterium tuberculosis ketol–acid reductoisomerase inhibitors as therapeutic drug leads. *Journal of Computer–Aided Molecular Design*, 33, 357–366, 2019.

25. E.W.W. Leung, L.W. Guddat. Conformational changes in plant ketol–acid reductoisomerase upon Mg2+ and NADPH binding as revealed by two crystal structures. *J. Mol. Bio*, 389, 167–182, 2009.

26. R. Tyagi, S. Duquerroy, J. Navaza, L.W. Guddat, R.G. Duggleby. The crystal structure of a bacterial Class II ketol–acid reductoisomerase: Domain conservation and evolution. *Protein Sci.*, 14, 3089–3100, 2005.

27. S. Tadrowski, M.M. Pedroso, V. Sieber, J.A. Larrabee, L.W. Guddat, G. Schenk. Metal ions play an essential catalytic role in the mechanism of ketol–acid reductoisomerase. *Chem. Eur. J.,* 22, 7427–7436, 2016.

28. A. Schulz, P. Sponemann, H. Kocher, F. Wengenmayer. The herbicidally active experimental compound Hoe 704 is a potent inhibitor of the enzyme 2–acetolactate reductoisomerase. *FEBS Lett,* 238, 375–378, 1988.

29. J.P. Foster, F. Weinhold. Natural Hybrid Orbitals. *J. Am. Chem. Soc.*, 1980, 102, 7211−7218.

30. A.E. Reed, F. Weinhold. Natural Bond Orbital Analysis of Near–Hartree−Fock Water Dimer. *J. Chem. Phys*., 78, 4066−4073, 1983.

31. F. Martin, H. Zipse, Charge distribution in the water molecule– a comparison of methods. *J. Comp. Chem.*, *26*, 97 – 105, 2005.

32. K.B. Wiberg. Application of the Pople–Santry–Segal CNDO Method to the Cyclopropylcarbinyl and Cyclobutyl Cation and to Bicyclobutane. *Tetrahedron*, 24, 1083−1096, 1968.

33. I. Mayer. Bond order and valence indices: a personal account. *J. Comput. Chem.,* 28, 204−221, 2007.

34. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay. Scikit–learn: Machine learning in python. *JLMR*, 12, 2825–2830, 2011.

35. G. Kuppuraj, M. Dudev, C. Lim. Factors governing metal–ligand distances and coordination geometries of metal complexes. *J. Phys. Chem B*, 113, 9, 2952–2960

36. A. Moomaw, M.E. Maguire. The unique nature of Mg2+ channels. *Physiology*, 23, 275–285, 2008.

37. Y.L. Gu, T. Kar, S. Scheiner. Fundamental properties of the CH center dot center dot center dot O interaction: Is it a true hydrogen bond?. *J Am Chem Soc*, 121, 9411–9422, 1999.

38. Desiraju, G. R. The C–H center dot center dot center dot O hydrogen bond: Structural implications and supramolecular design. *Accounts Chem Res*, 29, 441– 449, 1996.

39. Z.S. Derewenda, L. Lee, U. Derewenda. The Occurrence of CH···O Hydrogen–Bonds in Proteins, *J Mol Biol*, 252, 248–262, 1995.

40. A.M. Vibhute, U. Deva–Priyakumar, A. Ravi, K.M. Sureshan. Model Molecules to Classify C–H···O Hydrogen–Bonds. *Chem. Commun.,* 54, 4629–4632, 2018.

41. S. Scheiner. Relative strengths of NH–O and CH–O H–bonds in Polypeptide chain segments. *J Phys Chem B*, 109, 16132–16141, 2005.

42. J.J. Novoa, P. Constans, M. Whangbo. On the Strength of the CH···O Hydrogen Bond and the Eclipsed Arrangement of the Methyl Group in a Tricyclic Orthoamide Trihydrate. *Angew. Chrm. Int. Ed. EngI.,* 32, 4, 1993.

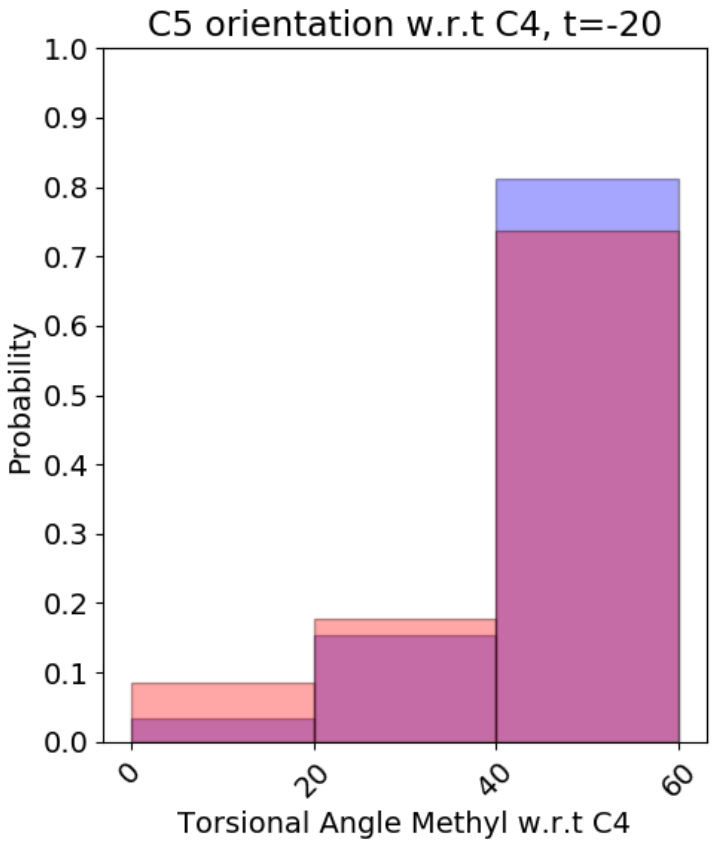
43. S. Horowitz, L.M.A. Dirk, J.D. Yesselman, J.S. Nimtz, U. Adhikari, R.A. Mehl, S.Scheiner, R.L. Houtz, H.M. Al–Hashimi, R.C. Trievel. Conservation and Functional Importance of Carbon–Oxygen Hydrogen Bonding in AdoMet–Dependent Methyltransferases. *J. Am. Chem. Soc.,* 135, 41, 15536–15548, 2013.

44. S. Horowitz. The Functions and Importance of CH···O Bonds in SET Domain Methyltransferases. Doctoral Dissertation, University of Michigan, 2013.

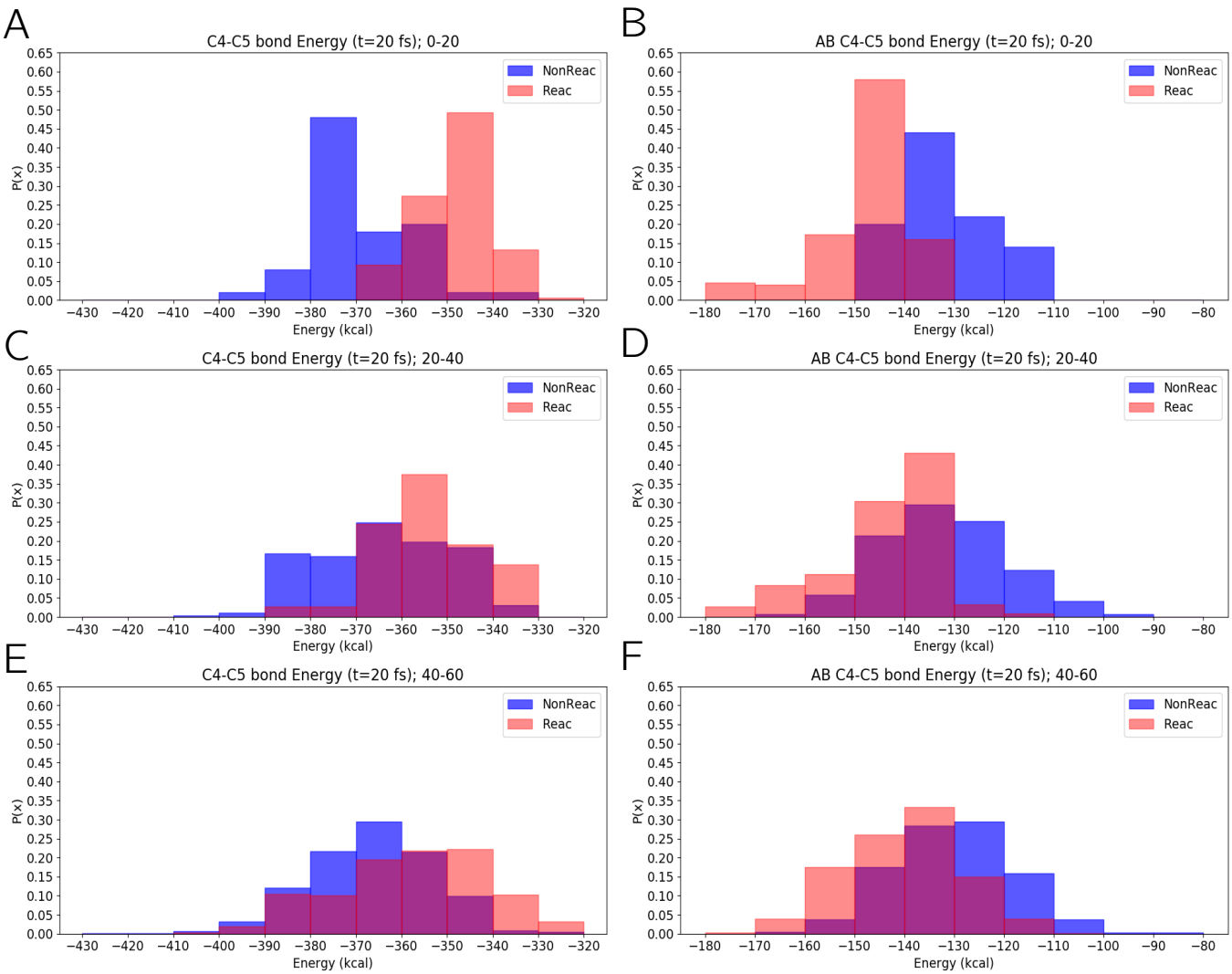
45. S. Horowitz, J.D. Yesselman, H.M. Al–Hashimi, R.C. Trievel. Direct Evidence for Methyl Group Coordination by CH···O Hydrogen Bonds in SET Domain Methyltransferases. *J Biol. Chem*, 286(21):18658–63, 2011.

46. E. D. Glendening, C. R. Landis, F. Weinhold. NBO 7.0: New Vistas in Localized and Delocalized Chemical Bonding Theory. *J. Comput. Chem.* 40, 2234–2241, 2019.
47. C. R. Landis, F. Weinhold. "The NBO View of Chemical Bonding", in, G. Frenking and S. Shaik (eds.), The Chemical Bond: Fundamental Aspects of Chemical Bonding, *Wiley*, pp. 91–120, 2014.
48. F. Weinhold. Natural Bond Orbital Analysis: A Critical Overview of its Relationship to Alternative Bonding Perspectives. *J. Comp. Chem.* 33, 2363–2379, 2012.
49. E. D. Glendening, F. Weinhold. Natural Resonance Theory. I. General Formulation. *J. Comp. Chem.* 19, 593–609, 1998.
50. E. D. Glendening, F. Weinhold. Natural Resonance Theory. II. Natural Bond Order and Valency. *J. Comp. Chem.* 19, 610–627, 1998.
51. E. D. Glendening, F. Weinhold. Natural Resonance Theory. III. Chemical Applications. *J. Comp. Chem.* 19, 628–646, 1998.

3.7 Supplementary Information

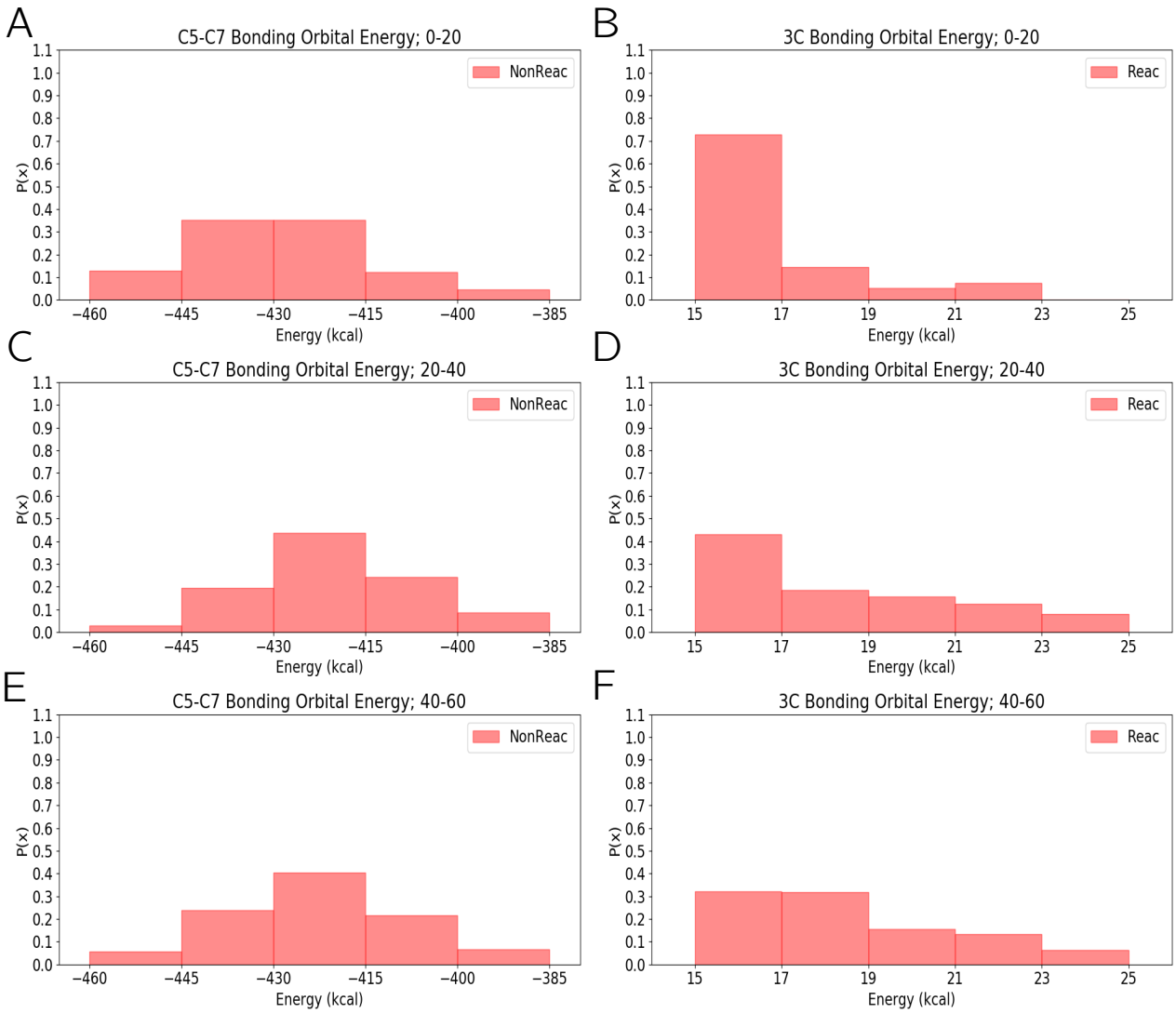


Supplementary Figure S1: Histograms of the reactive (red) and non–reactive(blue) ensembles as a function of torsion angle of C5 with respect to C4 at t = –20 fs in the reactant basin, slightly prior to crossing the barrier. The reactive ensemble had nearly 5% more trajectories that were in the eclipsed "0–20" degree orientation than the non–reactive ensemble. In contrast, the non–reactive ensemble possessed about 8% more staggered "40–60" degree orientation trajectories than the reactive ensemble. Both ensembles had roughly equivalent populations of the intermediate "20–40" degree orientation.

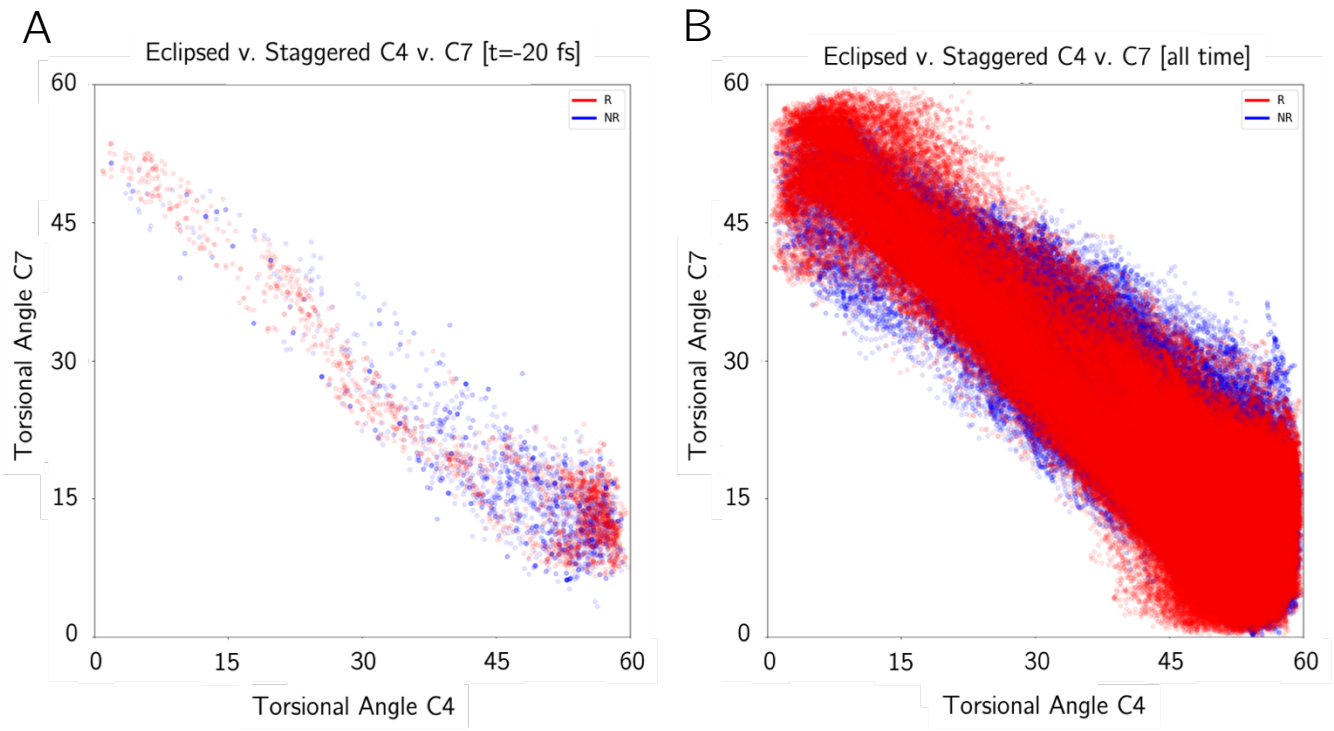


Supplementary Figure S2: Histograms of the C4–C5 σ–bond orbital energy (A, C, E) and C4–C5 σ* antibonding orbital energy (B, D, F) at t = 20 fs, roughly the time when the 3C bond first began to form in the reactive ensemble simulations. (A) C4–C5 σ–bond orbital energy for eclipsed "0–20" orientations of the C5 methyl. The reactive ensemble was –347 ± 8 kcal (mean ± std. dev) versus the non–reactive ensemble was –369 ± 11 kcal. (B) C4–C5 σ*–antibond orbital energy for eclipsed "0–20" orientations. The reactive ensemble energy was –148 ± 9 kcal (mean ± std. dev) versus the non–reactive ensemble was –132 ± 10 kcal. (C) C4–C5 σ–bond orbital energy for intermediate "20–40" orientation; the reactive ensemble mean/std.dev was –354 ± 11 kcal versus

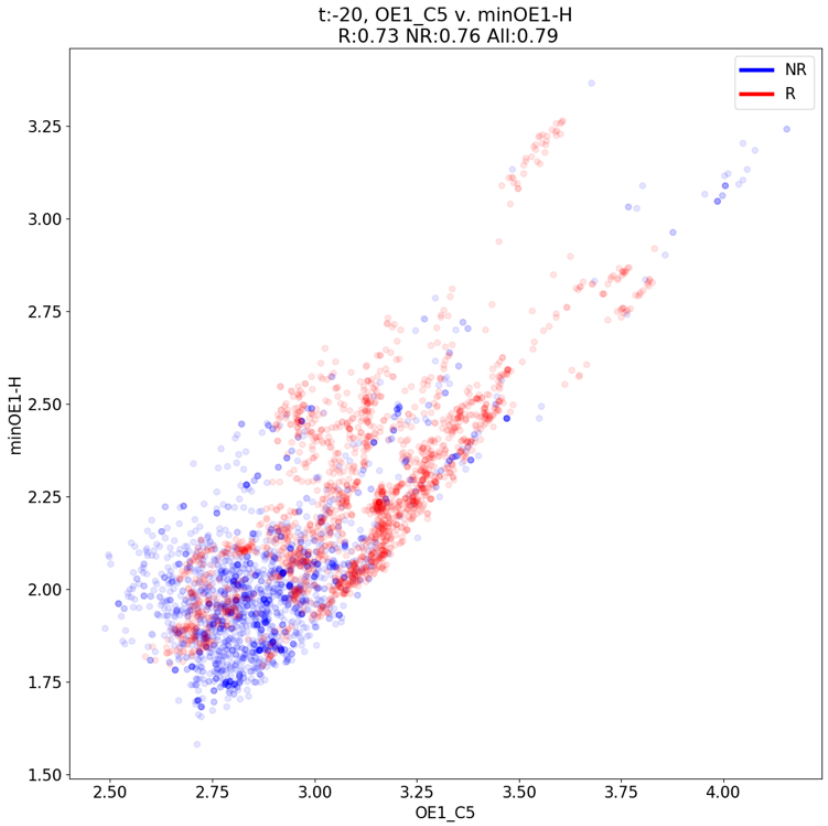
non–reactive with –364 ± 15 kcal. (D) C4–C5 σ*–antibond orbital energy for intermediate "20–40" orientation; the reactive ensemble mean/std.dev was –143 ± 11 kcal versus non–reactive with –132 ± 13 kcal. (E) C4–C5 σ–bond orbital energy for staggered "40–60" orientation; the reactive ensemble mean/std.dev was –357 ± 16 kcal versus non–reactive with –366 ± 13 kcal. (F) C4–C5 σ*–antibond orbital energy for staggered "40–60" orientation; the reactive ensemble mean/std.dev was –140 ± 12 kcal versus non–reactive with –130 ± 12 kcal.



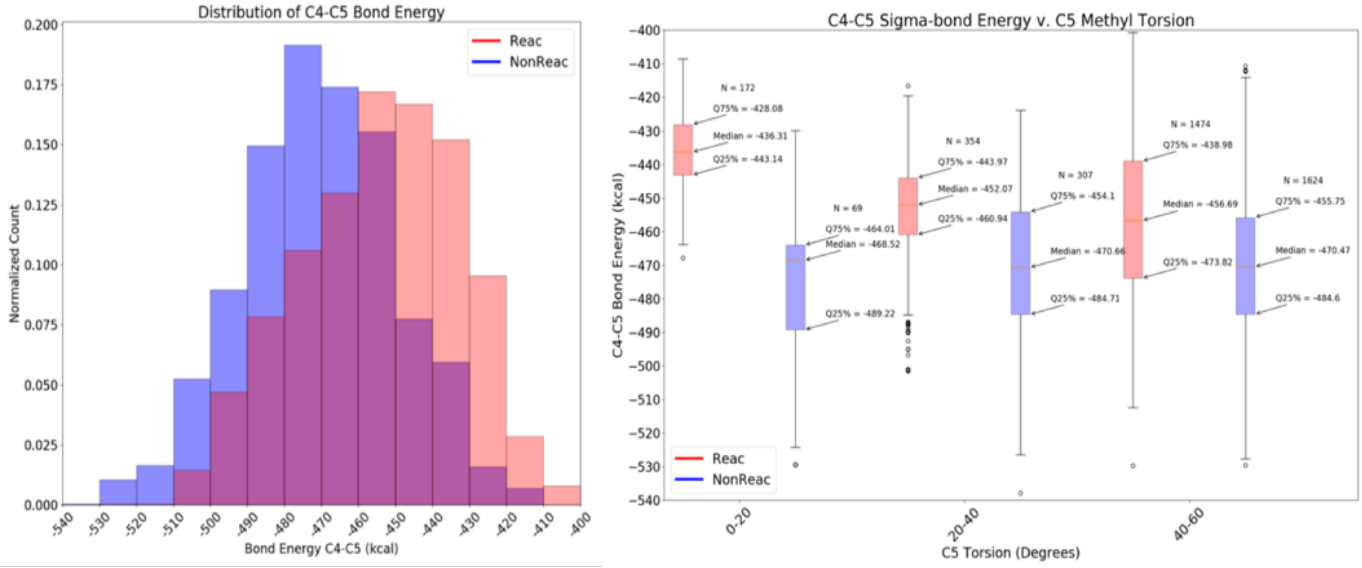
Supplementary Figure S3: Histograms of the C5–C7 σ–bond orbital energy (A, C, E) and 3C bonding orbital energy (B, D, F) for the stratified torsion populations at the time they first were defined in the simulations for the reactive ensemble. (A) Simulations that left the reactant basin in the eclipsed orientation possessed a C5–C7 σ–bond orbital energy at –428 ± 15 kcal (mean ± std. dev). (B) The orbital energy of the 3C bond formed between C4, C5, and C7 for the eclipsed orientation was 17 ± 2 kcal (mean ± std. dev). (C) The intermediate torsional orientation possessed C5–C7 σ–bond orbital energy at –420 ± 15 kcal (mean ± std. dev). (D) The intermediate torsion angle distribution had a 3C bond orbital energy at 19 ± 3 kcal (mean ± std. dev). (E) The staggered orientation torsion angle population had a C5–C7 σ–bond orbital energy at –422 ± 16 kcal (mean ± std. dev). (F) The staggered orientation torsion angle population had a 3C bond orbital energy at –19 ± 2 kcal (mean ± std. dev).



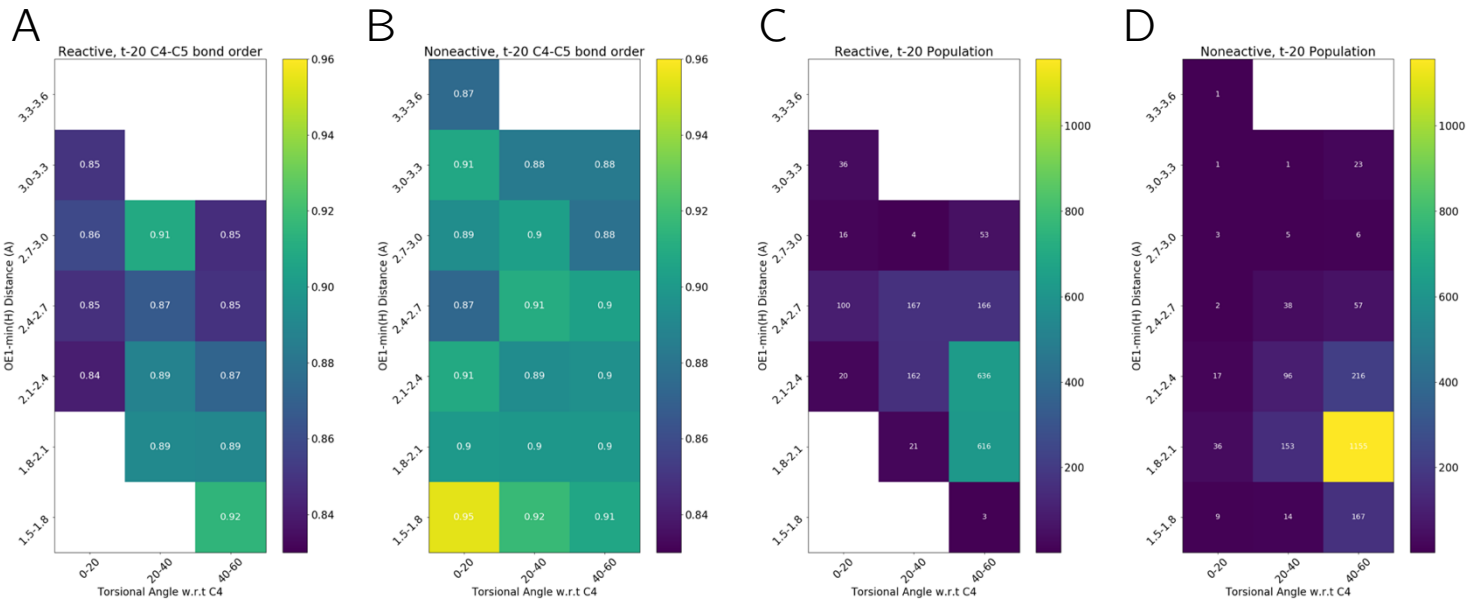
Supplementary Figure S4: Scatter plot of the torsional orientation of methyl C5 with respect to C4 versus methyl C5 with respect to carbon C7 (constituents C4, C9, O8). (A) Scatter plot at time t = –20 fs, where each point represents one time point on each member of the ensemble; the torsion angle with respect to C4 was anticorrelated compared to C5. (B) Scatter plot of torsional angles where each point represents a timepoint from a trajectory for all timepoints in the duration of 150 fs per simulation. The anti–correlated relationship between the torsional angles with respect to C4 and C7 were observed to be time invariant.



Supplementary Figure S5: Association between the E319–OE1—C5 methyl carbon distance and E319–OE1—minH distance, where minH represents the proton on the C5 methyl group closest to the OE1 oxygen on E319, for both Reactive (R) and Non–reactive (NR) populations. Each point represents a trajectory sampled at t = –20 fs. The Pearson correlation coefficient between the distance of the OE1–C5 distance and the OE1–H distance for the non–reactive ensemble is r = 0.76, whereas for the reactive ensemble, r = 0.73. Combining both ensembles, the correlation is r = 0.79. Carbon–oxygen distances of up to 3.7 Å are safely categorized as CH–O type bonds [39].



Supplementary Figure S6: The distribution of the energy (kcal), as calculated by NBO, between the C4–C5 σ–bond for reactive and non–reactive simulations (A) and the energy of this C4–C5 σ–bond, as stratified by methyl torsional angle (B) for both ensembles, at t = –20 fs. (A) Distribution of the orbital energy of the C4–C5 σ bond for the reactive and non–reactive simulations. Reactive simulations were higher in energy than non–reactive simulations. For reactive simulations, the σ–bond energy between C4–C5 was –454 ± 21 kcal (mean ± std. dev), whereas for non–reactive simulations, it was –470 ± 21 kcal. (B) Distribution of C4–C5 σ–bond energy as a function of the methyl torsional conformation. For methyl torsional angles that were between 0–20 degrees, reactive simulations exhibited an increase in energy in the C4–C5 σ bond, relative to non–reactive simulations, of 32 kcal/mol (difference in medians).



Supplementary Figure S7: 2D histogram showing the dependence of C4–C5 bond order and sample count on the (x–axis) torsional angle of C4–C5, and (y–axis) OE1–min(proton on C5) distance, for reactive and non–reactive ensembles. (A) Reactive bond–order of C4–C5, averaged for the population binned between the (x,y) axis. (B) Non–reactive bond–order of C4–C5, averaged for the population binned between the (x,y) axis. (C) Number of reactive trajectories identified to each given (x,y) bin. (D) Number of non–reactive trajectories identified to each given (x,y) bin.The reactive–only distribution in panel (A) shows a clear trend where low torsional angle (0–20 or "eclipsed" status) and high OE1–H distances correspond to generally lower bond–order. Inspecting the medium torsional angle (20–40) and large torsional angle (40–60 or 'staggered'), there is a trend that indicates that if OE1–H is farther away, the bond–order between C4–C5 underscores this. For the non–reactive ensemble in panel (B), this trend does not exist, which underscores the findings in Figure 15B and Figure 15D.

Chapter 4: Dynamic drivers of catalytic strategy in OMPDC and mutants

Natasha Seelam[1,2], Bruce Tidor[2,3,4]

5. Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge MA

6. Computer Science and Artificial Intelligence Laboratory (CSAIL), Massachusetts Institute of Technology, Cambridge MA

7. Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge MA

8. Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge MA

4.1 Abstract

Prior literature has hypothesized the importance of 'near–attack conformations' (NACs) as drivers of enzyme activity [2–7]. This concept emphasizes the role of the detailed conformational substructure of the enzyme–substrate (ES) ensemble, as opposed to the transition–state (TS) based theories that prioritize the relative energetics of the enzyme–TS compared to the enzyme–bound substrate. The current work investigates how the chemical environment, provided by the active site of the enzyme, influences reactive pathways, using wildtype (WT) orotidine 5′-monophosphate decarboxylase (OMPDC) and two empirically–characterized, catalytically–impaired mutants, S127A and V155D, as a case study. Transition path sampling (TPS) methodologies of the three enzymes were used to generate simulations leading to productive decarboxylation of the orotidine 5′-monophosphate (OMP) substrate. Analysis of the TPS ensembles across the three enzyme systems revealed two distinct catalytic strategies of decarboxylation: one that favored decarboxylation occurring "simultaneously" with the positioning of a catalytically important residue, K72, in close proximity, and second pathway in which decarboxylation occurred uncoupled from K72's position in a "stepwise" manner. To investigate the differences in the reactive pathways, we posed a classification problem for supervised machine learning methods to predict the 'simultaneous' or 'stepwise decarboxylation from these simulations.

Supervised machine learning methods were used to identify key descriptors ("features") that distinguished between the two pathways using data from six time points, three in advance of any reactivity, and three after committing to cross the barrier, but not explicitly reaching the product. Classifiers demonstrated that several pairs of geometric features were capable of predicting the catalytic strategy with over 80% testing accuracy and ROC AUC for all time points.

Moreover, numerous model–selected features corresponded to catalytically insightful geometries, despite the model having no prior knowledge of chemical mechanism.

Top predictive features highlighted the importance of a hydrogen bond between D75* and the 2' – hydroxyl group of the OMP substrate, and the proximity between the carboxylate group of D70 and the OMP carboxylate group; both features have been previously hypothesized to play a strong role in catalysis [20–27]. These features were also correlated to the degree of distortion exhibited by the orotidyl ring prior to reaction. Stratified analysis of the mechanisms by protein ensemble revealed that the V155D ensemble formed longer, and likely weaker, interactions with the aforementioned features that may answer why its reactive profile differed from WT and S127A.

These results of this work support the hypothesis that multiple catalytic strategies may exist toward facilitating successful catalysis, and that changing the local environment can result in switching amongst alternative means in achieving reactivity.

4.2 Introduction

The chemical environment in which a reactive center is embedded is crucial for catalysis. For enzymes, the active site provides an environment with conditions that greatly favor difficult reactions compared to pure solvent [1]. Extensive experimental and theoretical studies have been carried out to identify how components within an active site promote reactivity, particularly within the context of ground–state destabilization (GSD), transition–state stabilization (TSS), and near–attack conformations (NACs) [2–7]. In prior work, our group identified the role of an enhanced reactivity zone, representing a set of conformations of the enzyme–substrate (ES) complex more likely to lead to successful catalytic events, described by geometric features of the active–site [8]. This suggests that preorganization of the active site, prior to crossing the barrier, may highlight the key strategies used in catalysis.

Chapter 2 of this thesis presented a simulation analyses of the enzyme orotidine 5′-monophosphate decarboxylase (OMPDC) and two catalytically–impaired mutants, S127A and V155D. With both energetics (via potential of mean force "PMF") and dynamics (via path sampling calculations), our studies found different distributions of reactive pathways taken across this set of enzymes. The reactive paths of WT and mutant systems were described earlier in the context of two parameters: a decarboxylation coordinate corresponding to the stretching of the breaking bond, C6–CX; and a proton–transfer coordinate, involving a breaking bond between a proton and the side–chain amide nitrogen of the catalytically conserved K72 residue, and a forming bond between that proton and the ring carbon, C6 (Figure 4.1B). We observed that the majority of WT and S127A mutants decarboxylated in a "simultaneous" manner – the proximity of K72 with the OMP ring shortened concurrent with decarboxylation. In contrast, the V155D mutant demonstrated two distinct pathways dynamically: one that followed a mechanism similar to WT

and S127A, but an additional decarboxylation strategy in which decarboxylation was independent of proton transfer. Here, we investigated how changes in the geometric details of the active–site binding led to differences between the reactive pathways of OMPDC, and the chemical insight these features offered about the catalytic strategies.

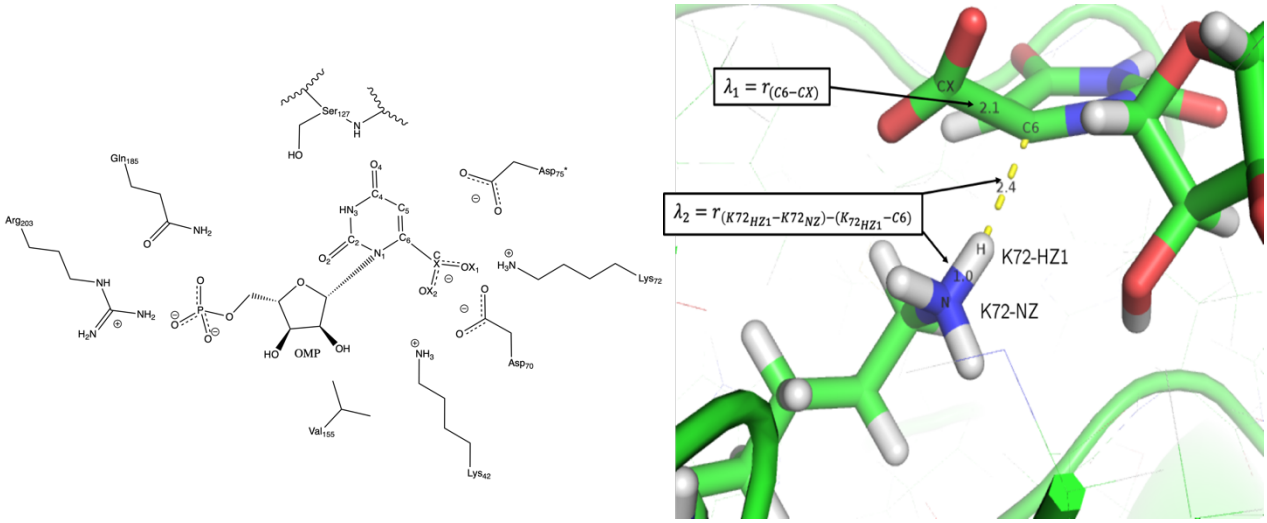


Figure 4.1: (A) The active–site of OMPDC including residues experimentally verified to affect the catalysis of the substrate. Geometric features were constructed from non–hydrogen atoms of the following residues K42, D70, K72, D75*, Q185, and R203, and non–hydrogen atoms of the orotate substrate. Starred nomenclature refers to the other monomer, as OMPDC is a homodimer [14]. Detailed characterization of the mechanism and the roles of these residues can be found in the Introduction section of Chapter 2. For a three–dimensional structure with relative placements of the residues, see Supplemental Figure 4.S17. (B) A schematic of the two order parameters that defined reactive pathways (See Methods 4.3.2); the decarboxylation coordinate (λ_1) and the proton–transfer coordinate, (λ_2).

To explore differences in the reactive pathways, we posed a classification problem between the "simultaneous" and "stepwise" mechanism across any protein system. Because of the inherent complexity of trajectories approaching the reaction barrier, we chose to use machine learning to identify the key descriptors ("features") that could distinguish between the two pathways. Models were trained to select five pairs of features from a large set of 620 geometric descriptors, involving

interactions of non–hydrogen atoms of catalytically important residues and the substrate, orotidine 5′-monophosphate (OMP; Figure 4.1A) [12–17, 23]. We performed this classification task across six time points, considering times while in the reactant basin up until 30 fs into crossing the barrier. Predictive performance (testing accuracy) was over 80% for all six times tested, suggesting pathways were distinguishable while still in the reactant basin and with relatively few features required.

Analysis of the top predictive features, specifically while in the reactant basin, illustrated the model's capacity to select chemically meaningful descriptors. Our model reported on the experimentally hypothesized hydrogen bond between D75* and the 2'–hydroxyl of the ribophosphate of OMP, thought to assist with binding of the substrate into the active site [22, 25–27]. The other key feature the model reported underscored the role of ground–state destabilization through the form of a distance between D70's carboxylate group oxygens and OMP's carboxylate oxygens [17, 18, 21]. Both features were shown to influence the distortion of the carboxylate group off the OMP orotidyl ring. Stratifying the ensembles by protein systems revealed at the WT and S127A shared similar feature distributions when compared to the either the simultaneous V155D or stepwise ensembles. Remarkably, the model was able to identify these features without *a priori* information of the chemical mechanism except for pathway labels.

This work does not intend to explicitly answer how such features give rise to reactivity, or what electronic differences are insinuated by the geometries, as no explicit electronic calculations were made. Extensions of this work by NBO analyses may directly attribute the magnitude and effect of the geometric features with regard to catalytic strategy of decarboxylation. With design in mind, this work may offer insight into how conformations of catalytically critical residues may influence reactivity.

4.3 Methods

4.3.1 Structure preparation, ensemble generation, and time alignment

TPS ensembles for OMPDC WT and mutants were generated and described in Chapter 2. Briefly, quantum mechanical/molecular mechanical (QM/MM) simulations were used to create the TPS ensembles using CHARMM version 39. The SQUANTUM module of CHARMM was used at the AM1 level of theory to treat the orotidyl ring and the side–chains of K42, D70, K72, D75* [9]. The generalized hybridized orbital (GHO) methods was used to treat the QM/MM interface across the $C_\alpha - C_\beta$ bond for side chains and across the ribonucleotide–orotidyl bond for the substrate [10].

The reactive pathway ensembles were generated using Transition Path Sampling (TPS) methods [See Chapter 2 Methods for further discussion; 11]. The TPS order parameter explicitly monitored the breaking bond distance:

$$\lambda = distance(C_6 - C_X)$$

While this order parameter did not specifically monitor the proton transfer, the QM description permitted proton transfer from neighboring lysines to the orotidyl ring. The reactant basin interface was set at $\lambda_0 = 1.7$ Å. To collect reactive pathways, trajectories were harvested from the [3.45 Å, 5 Å] TPS window for the WT, S127A, and V155D systems. A total of 6391 WT trajectories, 13453 S127A trajectories, and 13488 V155D trajectories were accumulated.

To provide a reference point for the progress of dynamical features toward reactivity, trajectories were temporally aligned such that time $t = 0$ fs corresponded to the simulation leaving the reactant basin, λ_0, for the last time (crossing the $\lambda_0 = 1.7$ Å interface; Figure 4.2, Supplemental Figure 4.S1).

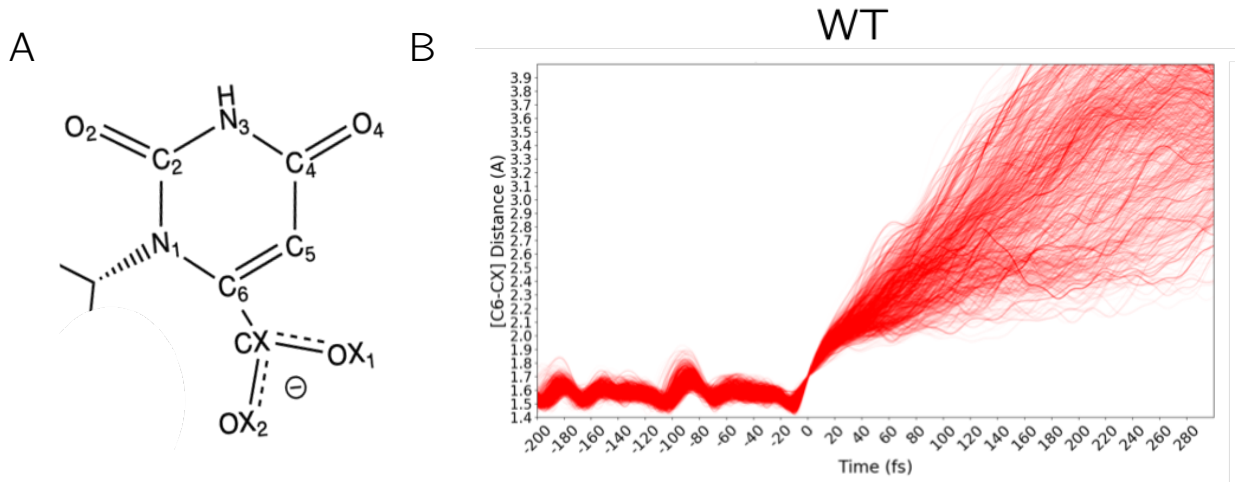


Figure 4.2: (A) Schematic of the orotidyl ring of OMP truncated across the bond with the ribophosphate. The decarboxylation coordinate for time–alignment is defined as the distance between atoms C6 and CX. (B) The decarboxylation coordinate versus time for a time–aligned ensemble of WT trajectories. The simulations leave the reactant basin and begin to cross the reaction barrier for $t > 0$ fs.

4.3.2 Pathway labeling

Pathway labeling used two parameters: the decarboxylation coordinate order parameter below:

$$\lambda_1 = distance(C_6 - C_X)$$

And an additional proton–transfer coordinate with K72, which considers the closest amide proton to the C6 atom of the ring:

$$\lambda_2 = \max_{i \in \{1,2,3\}} \left[distance\left(NZ_{K72} - HZ_{i_{K72}}\right) - distance\left(C6 - HZ_{i_{K72}}\right) \right]$$

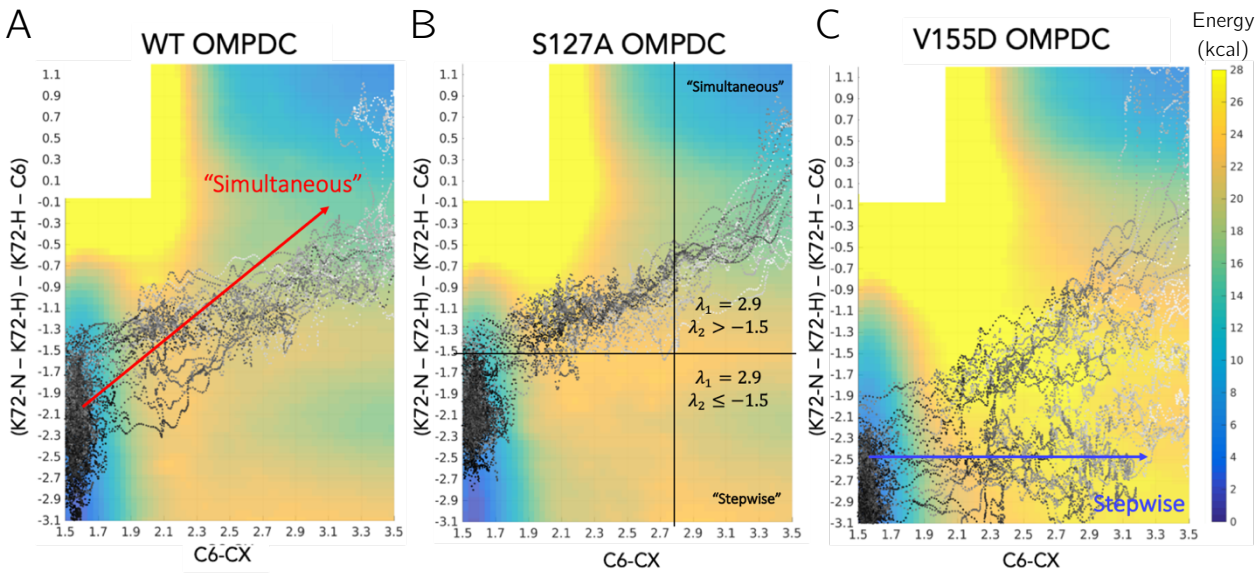


Figure 4.3: Trajectories overlaid on the PMFs of Chapter 2. (A) A random sampling of simultaneous trajectories on the WT PMF described by the decarboxylation coordinate (x) and the proton–transfer coordinate (y) (B) The boundary used for both order parameters to determine a simultaneous versus stepwise trajectory on the S127A PMF (C) A random sampling of both simultaneous and stepwise trajectories on the V155D PMF.

In order to label reactive pathways, a cutoff was drawn using the decarboxylation and proton transfer coordinates. If a given trajectory's proton–transfer coordinate, λ_2, was less than or equal to –1.5 Å at the time the decarboxylation distance, λ_1, was 2.9 Å, the trajectory was labeled as stepwise (Figure 4.3). This boundary roughly corresponded to the high energy region separating the lower–right energy basin and the upper–right energy basin of the PMF landscapes of Chapter 2. By this definition, 6296 WT trajectories were classified as simultaneous, versus 95 stepwise. For S127A, 13397 trajectories were simultaneous versus 56 stepwise, and finally 8198 V155D trajectories were simultaneous and 5290 were stepwise (Supplemental Figure 4.S2).

4.3.3 Feature construction

Prior literature and hypothesized mechanisms have directly implicated several residues in playing an active role in the catalysis of OMPDC; this subset includes the catalytic tetrad K42, D70, K72, D75*, Q185, and R203 [12–18]. Distances, planar angles, and torsions were constructed using atoms within the active–site that corresponded to non–hydrogen atoms of the side–chains of the aforementioned residues, the ribophosphate group of OMP, and the orotidyl ring of OMP with exception of the 6 descriptors characterizing the K72 proton (HZ1, HZ2, HZ3) – C6 distance and K72 amine nitrogen (NZ)–proton distance, as these features comprised the proton–transfer coordinate. Angles and torsions were specifically constructed to describe mechanistic hypotheses. Of the 620 total geometric features, the composition of each feature type was 546 distance features, 36 angle features, and 38 torsional features. For each protein, every feature was computed for six timepoints ($t = -20$ fs, $t = -10$ fs, $t = 0$ fs, $t = 10$ fs, $t = 20$ fs, and $t = 30$ fs). These time points span the duration of the reaction before reaching the product, starting in the reactant basin ($t \leq 0$ fs) to starting to cross the barrier ($t > 0$ fs). Features were normalized as follows: for each protein at a particular timepoint, the mean and standard deviation were calculated by combining the simultaneous and stepwise ensembles together. Each feature was then individually mean–centered and standard–deviation scaled.

4.3.4 Machine learning

The simultaneous and stepwise ensembles were pooled from all protein systems resulting in 27891 simultaneous simulations and 5,441 stepwise simulations. From the total set, the six pairs of datasets (a simultaneous and stepwise ensemble pair for each of the six timepoints) were each subjected to supervised machine learning in a variety of modes. In each experiment, logistic

regression classification models were trained to use a subset of features, described above, to distinguish whether the simulation crossed the barrier through a simultaneous or a stepwise mechanism.

In one set of experiments, an exhaustive search of all pairs of features was performed to find the pair of features for each classification task, the second–best (non–overlapping) pair, and so forth. These sequences of "best pairs" were combined to produce cumulative feature sets that were trained on their respective data set to produce cumulative–feature models as opposed to pair–feature models. We used testing accuracy, averaged across all 10 sets, to determine predictive performance and rank order features.

To identify the set of best performing pairs, 10 training/testing splits were created by randomly choosing 1691 trajectories from each ensemble, ensuring an equal representation of both pathways in a training/testing split. The training/testing ratios were 80%/20% respectively. Performances (accuracy and ROC AUC), coefficients, and biases were averaged across all 10 models and reported in the Results and Discussion sections (Tables 1–4, Supplementary Tables 1–6 for coefficients and bias, Supplemental Figures 4.S3–S8 for schematics of features). Models were constructed using the *LogisticRegression* module of Scikit–learn [19]. Pathway labels were binarized such that "simultaneous" and "stepwise" corresponded to 1 or 0 respectively.

4.3.5 Calculation of the orotidyl ring (N1) improper

Distortion of the π–network on the orotidyl ring of OMP can be measured by several angles. We computed the improper dihedral angle centered around atom N1 of the orotidyl ring using constituent atoms C1', C6, and C2. The dihedral angle is measured from [0, 360] to account

for in–plane or out–of–plane orientations. We defined an "absolute distortion" angle as follows to measure the net deviation from planarity of N1:

$$Absolute\ Angle = \left| N_{1_{dihedral\ angle}} - 180 \right|$$

For geometry optimized OMP, this angle is 0.25 degrees (Supplemental Figure 4.S9).

4.3.6 Calculation of the C2–N1–C6–CX and C4–C5–C6–CX angles

The C2–N1–C6–CX and C4–C5–C6–CX angles are proper dihedral angles from [0, 360] degrees depending on clockwise or counterclockwise positioning. We compared their absolute deviation from a *trans*–angle (180°) as follows:

$$Absolute\ Angle = |\theta - 180|$$

Where θ can represent either the C2–N1–C6–CX and C4–C5–C6–CX dihedral. For geometry optimized OMP, this angle is 0.5 and 4.5 degrees respectively (Supplemental Figures 4.S9, 4.S16, 4.S15).

4.4 Results and Discussion

4.4.1 Several feature pairs equally distinguish between pathways with high performance

Feature selection calculations identified the top unique predictive pairs across all pair models comprised of the available 620 features (See Methods). We ranked all two–feature models for their predictive performance (testing accuracy), and selected the top five pairs with unique features for each time point. We also combined all unique pairs, per time point, into a cumulative, ten–feature model, and calculated the predictive performance across the training/testing splits.

For each timepoint and across all ten randomized splits of the data, all top performing pair–feature models were capable of achieving over 80% testing accuracy and AUC, with standard error at most ± 1.5% for both metrics (Table 1). The high predictive performance across these features suggested that there was redundancy across geometric features, further underscored by the similarity of the features, such as multiple descriptors involving bonded atoms of the same residue with the same target atom (Table 2; i.e. D75*/CG – OMP/O2' and D75*/CB – OMP/O2', where CG and CB are bonded atoms). The cumulative classifier for all time points except $t \geq 20$ fs did not significantly outperform any given pair–feature model in the top five pairs. This may be due to the fact that no "new" chemical information was available to improve predictivity, as features within the cumulative model were redundant.

	t=-20 fs		t=-10 fs		t=0 fs		t=10 fs		t=20 fs		t=30 fs	
Performance	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
Model												
Set 1	0.84	0.89	0.83	0.89	0.83	0.89	0.84	0.89	0.84	0.89	0.85	0.9
Set 2	0.84	0.89	0.83	0.89	0.83	0.88	0.84	0.89	0.84	0.89	0.84	0.88
Set 3	0.84	0.89	0.83	0.87	0.84	0.89	0.83	0.89	0.84	0.89	0.84	0.9
Set 4	0.84	0.88	0.84	0.89	0.83	0.88	0.84	0.89	0.84	0.89	0.84	0.88
Set 5	0.84	0.89	0.83	0.88	0.83	0.89	0.83	0.88	0.84	0.89	0.84	0.9
Cumulative	0.85	0.91	0.84	0.9	0.84	0.9	0.85	0.91	0.86	0.91	0.87	0.92

Table 1: Pair–feature and cumulative feature testing performance (accuracy and ROC AUC) for reactive pathway prediction task. Each column indicates the accuracy (classification prediction of simultaneous versus stepwise) and AUC for a given timepoint among all six timepoints tested, averaged across 10 randomized training/testing splits. For all reported performance metrics across systems and timepoints, the standard error was less than 1.5 % unless otherwise noted. Pairwise classifiers obtain comparable performances within a given timepoint, and compared to the cumulative classifier which uses all ten features from the pairwise models. A slight increase in predictive performance occurred after time $t = 20$ fs.

Timepoint	Feature ID	t = -20 fs	t = -10 fs	t = 0 fs	t= 10 fs	t = 20 fs	t = 30 fs
Pair 1	1	(ASP 75 CG)-(OMP 1 O2')	(OMP 1 CX)-(OMP 1 O2')	(ASP 70 OD2)-(OMP 1 C1')	(ASP 70 OD2)-(OMP 1 C1')	(ASP 75 OD1)-(OMP 1 C1')	(ASP 70 OD2)-(OMP 1 C6)
	2	(LYS 42 NZ)-(LYS 72 CE)	(ASP 70 CG)-(OMP 1 OX1)	(LYS 42 NZ)-(LYS 72 CE)	(LYS 42 NZ)-(LYS 72 CE)	(LYS 42 NZ)-(LYS 72 CE)	(LYS 72 HZ3)-(LYS 72 NZ)-(OMP 1 C6)
Pair 2	3	(LYS 72 CE)-(ASP 75 OD1)	(ASP 75 CB)-(OMP 1 O2')	(OMP 1 CX)-(OMP 1 O2')	(ASP 70 CG)-(OMP 1 OX1)	(OMP 1 OX1)-(OMP 1 O2')	(OMP 1 O4')-(OMP 1 PA)
	4	(OMP 1 OX2)-(OMP 1 O2')	(LYS 42 NZ)-(LYS 72 CE)	(ASP 70 CG)-(OMP 1 OX1)	(OMP 1 OX1)-(OMP 1 O2')	(ASP 70 OD2)-(OMP 1 OX1)	(LYS 72 NZ)-(OMP 1 OX1)
Pair 3	5	(OMP 1 CX)-(OMP 1 O2')	(OMP 1 C4')-(OMP 1 O2')	(OMP 1 OX1)-(OMP 1 O2')	(ASP 70 OD2)-(OMP 1 CX)	(LYS 72 CE)-(ASP 75 CB)	(OMP 1 OX1)-(OMP 1 O2'
	6	(ASP 70 OD2)-(OMP 1 OX1)	(ASP 70 OD2)-(OMP 1 C1')	(ASP 70 OD2)-(OMP 1 OX1)	(OMP 1 OX2)-(OMP 1 O2')	(ASP 70 OD1)-(OMP 1 O2')	(ASP 70 OD2)-(OMP 1 OX1)
Pair 4	7	(OMP 1 C4')-(OMP 1 O2')	(OMP 1 OX1)-(OMP 1 O2')	(ASP 70 OD2)-(OMP 1 CX)	(OMP 1 CX)-(OMP 1 O2')	(ASP 70 OD2)-(OMP 1 C6)	(ASP 70 CG)-(OMP 1 O2')
	8	(ASP 70 OD2)-(OMP 1 C1')	(ASP 70 OD2)-(OMP 1 OX1)	(OMP 1 OX2)-(OMP 1 O2')	(ASP 70 OD2)-(OMP 1 OX1)	(LYS 72 HZ3)-(LYS 72 NZ)-(OMP 1 C6)	(LYS 72 CE)-(ASP 75 CB)
Pair 5	9	(ASP 75 CB)-(OMP 1 O2')	(OMP 1 OX2)-(OMP 1 O2')	(LYS 72 CE)-(OMP 1 O2')	(OMP 1 O4')-(OMP 1 PA)	(ASP 70 OD2)-(OMP 1 CX)	(LYS 72 CE)-(OMP 1 N1)
	10	(LYS 72 CE)-(ASP 75 CB)	(ASP 70 OD2)-(OMP 1 C6)	(ASP 75 CB)-(OMP 1 O2')	(ASP 70 OD2)-(OMP 1 C6)	(OMP 1 OX2)-(OMP 1 O2')	(LYS 72 CE)-(OMP 1 C2')

Table 2: Features for top 5 pairwise models to classify reactive pathway at the six timepoints selected. Distances are represented by a pair of atom identities, angles by a triplet of atoms, and dihedrals/improper angles by a set of 4 atoms. Feature ID references the coefficient label ID for the supplementary tables 1–6. Structural depictions of the top features are indicated in Supplementary Figures 3–8.

Aggregating features across all six timepoints revealed only 23 unique features were identified across all pairs of features. This suggested that important features remained predictive from the reactant basin until sometime after the system started to cross the reaction barrier; this is further underscored by the fact that 14 of these features were employed in at least 2 classifiers (Table 2, 3). Of the unique features, 18 out of the 23 features explicitly reported on catalytic–tetrad residue–substrate interactions. Inspection of the individual features revealed several chemically relevant findings – despite access to nearly 620 possible features, the models identified meaningful

geometric descriptors that match prior hypotheses of reactivity [2, 3, 13–18]. In particular, seven features directly reported on the proximity of the D70 carboxylate group atoms (CG, OD1, OD2) with the orotidine 5′-monophosphate (OMP), of which four were explicitly with regard to OMP's leaving carboxylate group (CX, OX1, OX2) and orotidyl ring carbon (C6; Supplementary Table 7, Supplementary Figures 4.S3–4.S8). The residue D70 is considered as a ground–state destabilizing residue, where close proximity of the carboxylate group with another negatively charged carboxylate group from the orotidyl ring creates a repulsive interaction that encourages reactivity [17, 18, 23, 28]. Similarly, 5 features reported on the proximity of K72 with OMP, with at least 2 features directly reporting on the vicinity of K72 with the carboxylate leaving group or orotidyl ring carbon (Supplementary Table 7, Supplementary Figures 4.S3–4.S8). The residue K72 has been characterized as a transition–state stabilizing residue, where the positively charged amine group stabilizes the carbanion formed through decarboxylation [9, 10, 14, 18]. Moreover, pathway labeling required the use of a proton–transfer coordinate derived from K72. While explicit K72–proton to C6 distances were not part of the model–selected features, there were two features that reported on the amine group in relation to the orotidyl ring: the angle K72/HZ3–K72/NZ–OMP/C6, the orientation/angle of one of K72's protons to the C6 ring carbon, and K72/NZ–OMP/OX1, the proximity of the amine nitrogen to one of the leaving group carboxylate oxygens (Supplementary Table 7). The models identified these features, despite no explicit knowledge of chemical mechanism, and only binarized labels of the reactive pathways across all three protein systems.

Curiously, late–time classifiers ($t \geq 20$ fs) possessed more geometric features related to K72 and OMP (Table 2, 3). In contrast, many of the features reporting on interactions between D70 and the substrate carboxylate group appeared across all time points among the six tested. The

growing appearance of the K72 features is interesting, as the times these features appeared corresponded to climbing the reaction barrier to approach the transition state as opposed to being in the reactant basin.

Type	Features	Timepoints					
		t = -20 fs	t = -10 fs	t = 0 fs	t = 10 fs	t = 20 fs	t = 30 fs
Distance	(ASP 70 OD2)-(OMP 1 OX1)	X	X	X	X	X	X
Distance	(OMP 1 OX2)-(OMP 1 O2')	X	X	X	X	X	
Distance	(LYS 42 NZ)-(LYS 72 CE)	X	X	X	X	X	
Distance	(OMP 1 OX1)-(OMP 1 O2')		X	X	X	X	X
Distance	(OMP 1 CX)-(OMP 1 O2')	X	X	X	X		
Distance	(ASP 70 OD2)-(OMP 1 C1')	X	X	X	X		
Distance	(ASP 70 OD2)-(OMP 1 C6)		X		X	X	X
Distance	(ASP 75 CB)-(OMP 1 O2')	X	X	X			
Distance	(LYS 72 CE)-(ASP 75 CB)	X				X	X
Distance	(ASP 70 CG)-(OMP 1 OX1)		X	X	X		
Distance	(ASP 70 OD2)-(OMP 1 CX)			X	X	X	
Distance	(OMP 1 C4')-(OMP 1 O2')	X	X				
Distance	(OMP 1 O4')-(OMP 1 PA)				X		X
Angle	(LYS 72 HZ3)-(LYS 72 NZ)-(OMP 1 C6)					X	X
Distance	(ASP 75 CG)-(OMP 1 O2')	X					
Distance	(LYS 72 CE)-(ASP 75 OD1)	X					
Distance	(LYS 72 CE)-(OMP 1 O2')			X			
Distance	(ASP 70 OD1)-(OMP 1 O2')					X	
Distance	(ASP 75 OD1)-(OMP 1 C1')					X	
Distance	(LYS 72 NZ)-(OMP 1 OX1)						X
Distance	(ASP 70 CG)-(OMP 1 O2')						X
Distance	(LYS 72 CE)-(OMP 1 C2')						X
Distance	(LYS 72 CE)-(OMP 1 N1)						X

Table 3: Unique features from all classifiers, listed in order of frequency of appearance. Out of 23 unique features, 14 features appeared in at least 2 timepoints. There were 7 features reporting on D70–substrate interactions, 5 features on K72–substrate interactions, 5 intra–substrate features, 3 D75*–substrate interactions, and 3 features between the catalytic tetrad alone.

4.4.2 Features from machine learning models are linked to distortions in the planarity of the OMP orotidyl ring, and influence the carboxylate distortion

We identified two model–selected features that were important for pathway prediction while in the reactant basin (Figure 4.4). One feature was the distance between a catalytically conserved aspartate D75* and the 2' – hydroxyl group of the ribophosphate of OMP (D75*/CG –

OMP/O2'), which appeared in the first predictive pair for the $t = -20$ fs time point. The other feature included the distance between the carboxylate oxygen on residue D70, an amino acid implicated in ground–state destabilization, and the carboxylate leaving group of the OMP ring (D70/OD2 – OMP/OX1) that appeared within the top five predictive pairs for all six time points tested.

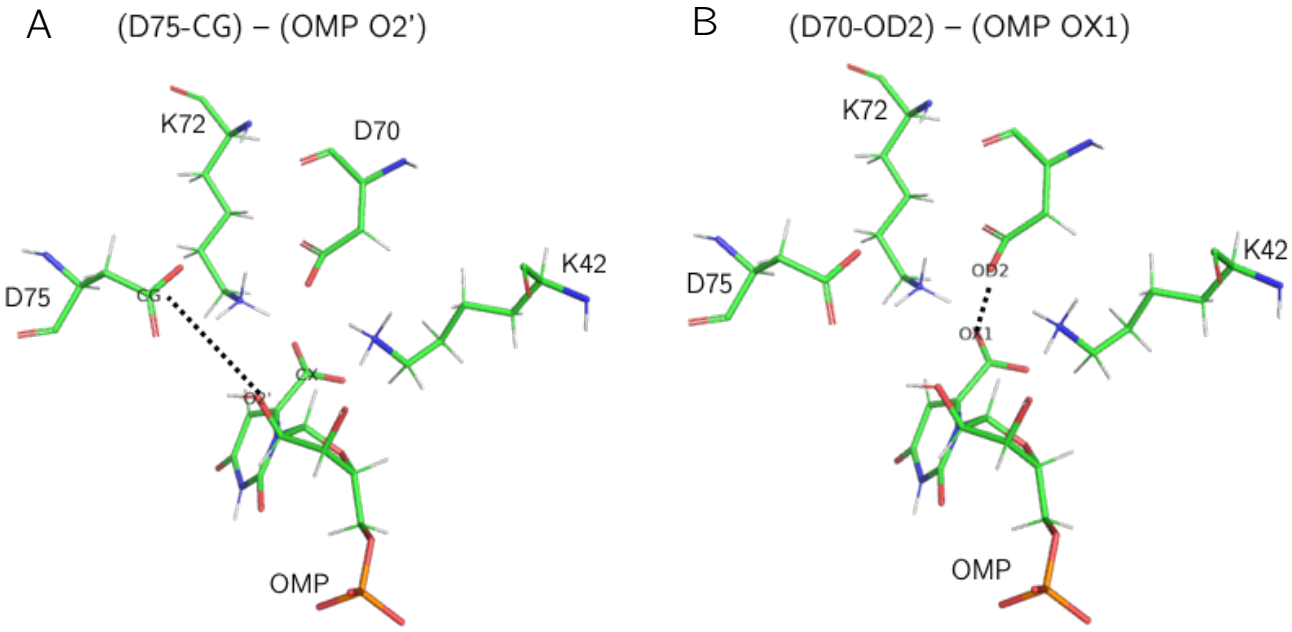


Figure 4.4: (A) Schematic of the (D75*/CG) – (OMP/O2') distance in the active–site of OMPDC. (B) Schematic of the (D70/OD2) – (OMP/OX1) distance in the active–site of OMPDC.

Analysis of their coefficients indicated that these features had negative coefficients in all pairwise classifiers, particularly the D70/OD2–OMP/OX1 feature had negative coefficients across all six time points (Table 4). Temporal analysis suggested that the coefficient, at least considered within the pair–feature model, could have reported on the shorter distances between D75* and D70 to OMP that persisted for the simultaneous ensemble across all six time points tested (Figure 4.5; Supplementary Figures 4.S10–S13).

Feature	t=-20 fs	t=-10 fs	t = 0 fs	t = +10 fs	t = +20 fs	t = +30 fs
(D75-CG) - (OMP-O2')	-1.32	-	-	-	-	-
(D70-OD2) - (OMP-OX1)	-1.11	-1.76	-1.84	-1.36	-1.99	-2.09

Table 4: Coefficients for the (D75*/CG – OMP/O2') and (D70/OD2 – OMP/OX1) machine learning features for the pairwise classifiers only. Coefficients provided are averaged across models trained on 10 randomized splits of the data. Full classifier information is provided in Supplemental Tables 1–6, including the cumulative model coefficients. Dashes indicate when the feature was not present in classifiers at other time points. The pair–feature model coefficients for both features were negative at the time points tested.

Comparison of the single–feature distributions of the simultaneous and stepwise ensemble for the (D75*/CG – OMP/O2')' distance revealed that the median of the simultaneous ensemble was nearly 1.6 Å shorter than the stepwise ensemble (Figure 4.5A). Separating the simultaneous distribution on the three protein systems (WT, S127A, and V155D) showed that the WT and S127A ensemble were nearly 1.6 Å shorter than the stepwise ensemble, but the V155D simultaneous ensemble was only 0.2 Å shorter than the stepwise ensemble (Figure 4.5B). The shorter distances between D75*'s carboxylate oxygen and the OMP 2'–hydroxyl oxygen implied the existence of a hydrogen bond between these atoms, which is thought to assist the binding of the substrate into the active site, as hypothesized in the literature [22, 25–27].

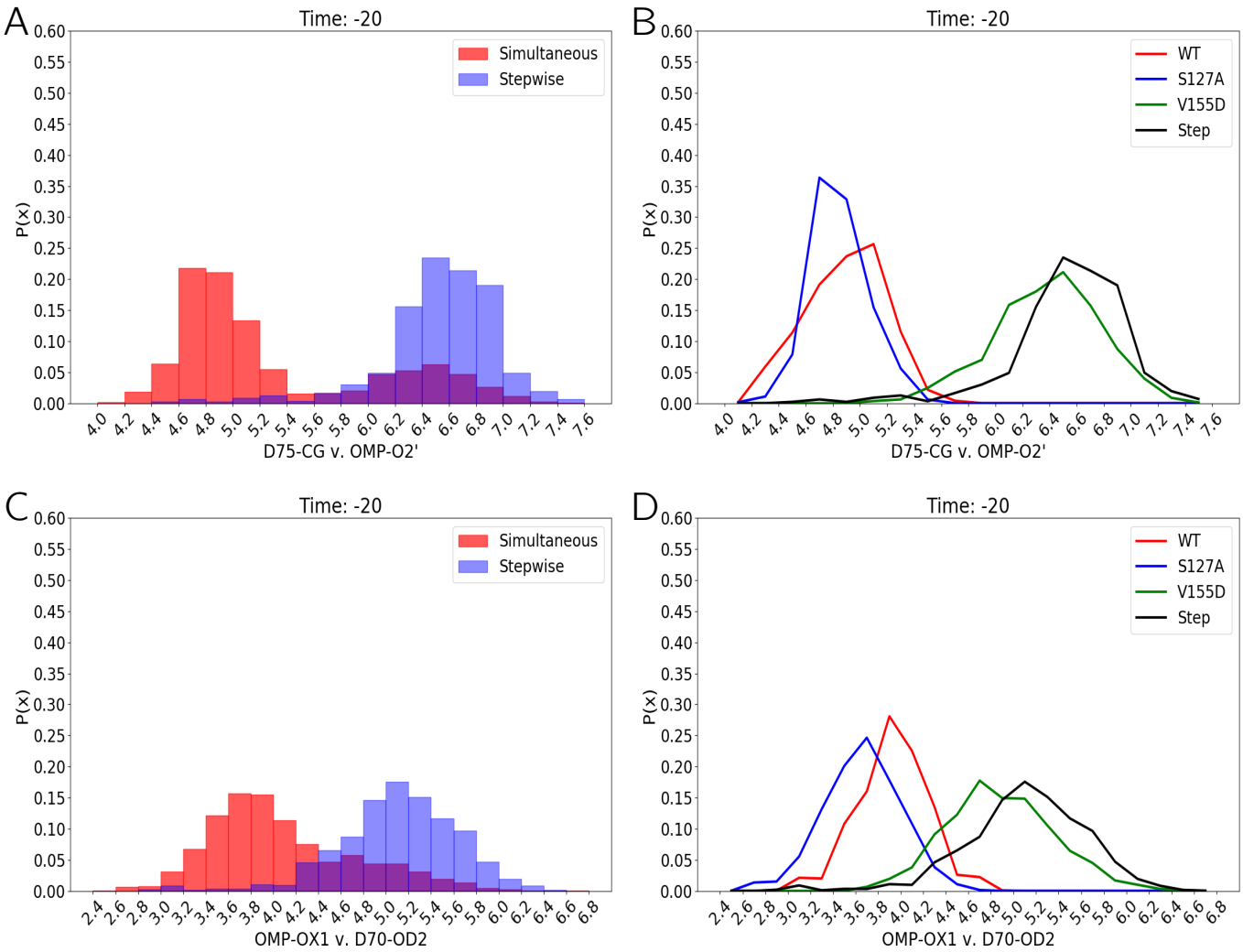


Figure 4.5: Distributions of model–selected features at time t = –20 fs. (A) The simultaneous (red) and stepwise (blue) distributions for the (D75*/CG – OMP/O2') distance. The simultaneous IQR (25% and 75% quantile) spanned [4.8 Å, 6.0 Å], with median 5.0 Å and standard deviation of 0.75 Å. The stepwise ensemble IQR was [6.4 Å, 6.8 Å] with median at 6.6 Å and standard deviation of 0.42 Å. (B) Distributions of (D75*/CG – OMP/O2') distance, stratified on simultaneous ensemble WT (red), S127A (blue), and V155D (green) versus the stepwise ensemble (black). The WT IQR was [4.7 Å, 5.1 Å] with median 4.9 Å and std. dev. 0.29 Å. The S127A ensemble was [4.7 Å, 5.0 Å] with median 4.8 Å and standard deviation 0.21 Å. The V155D ensemble was [6.2 Å, 6.6 Å] with median 6.4 Å and standard deviation 0.40 Å. The stepwise ensemble was the same as in (A). (C) The simultaneous (red) and stepwise (blue) distributions for the (D70/OD2 – OMP/OX1) distance. The simultaneous IQR (25% and 75% quantile) spanned [3.6 Å, 4.5 Å], with median 3.9 Å and standard deviation of 0.65 Å. The stepwise ensemble IQR was [4.8 Å, 5.5 Å] with median at 5.1 Å and standard deviation of 0.53 Å. (D) Distributions of (D70/OD2 – OMP/OX1) distance, stratified on simultaneous ensemble WT (red), S127A (blue), and V155D (green) versus the stepwise ensemble (black). The WT IQR was [3.7 Å, 4.1 Å] with median 4.0 Å and std. dev. 0.31 Å. The S127A ensemble was [3.4 Å, 3.9 Å] with median 3.7 Å and standard deviation 0.30 Å. The V155D ensemble was [4.6 Å, 5.2 Å] with median 4.9 Å and standard deviation 0.48 Å. The stepwise ensemble was the same as in (C).

Analysis of the distance between the hydroxyl proton (H2) and the closest carboxylate oxygen on D75* revealed that the simultaneous ensemble was nearly 1.6 Å closer than the stepwise ensemble, suggesting that the (D75*/CG – OMP/O2') feature acted as a proxy for the hydrogen bond (Supplementary Figure 4.S14). Comparison of the simultaneous V155D ensemble versus the simultaneous WT and S127A ensemble suggested that V155D was either less capable or did not form a hydrogen bond due to the elongated distances exhibited by the V155D ensemble. This is further highlighted by the fact that the stepwise ensemble possessed slightly longer distances than the V155D simultaneous ensemble, and was primarily comprised of V155D trajectories.

Similarly, comparison of the simultaneous and stepwise ensemble (D70/CG – OMP/OX1) distance distributions revealed that the simultaneous ensemble median was nearly 1.2 Å shorter than the stepwise ensemble, suggesting closer proximity between D70 and the leaving carboxylate group. Comparison across the simultaneous distributions, stratified by protein system, revealed that the WT and S127A simultaneous distributions were 0.9 Å and 1.2 Å shorter than the V155D simultaneous ensemble, and that the V155D simultaneous ensemble was only 0.2 Å shorter than the stepwise ensemble. Closer distances would imply greater electrostatic stress between the two negatively charged carboxylate groups. This electrostatic stress is thought to be a large contributor toward reactivity due to ground–state destabilization [17, 18, 21].

These two features were linked to distortions of the planarity of the orotidyl ring, namely via the N1 improper angle (Figure 4.4, Supplementary Figure 4.S9, See Methods for absolute improper angle definition). Prior experimental studies characterized the orotidyl ring's carboxylate motif bends out of plane, and that this distortion can assist the vinylic C6 carbanion to move in closer proximity to cationic residues like K72 [20–24]. Our simulations observed that the median distortion angle of the simultaneous ensemble was nearly 6.3 degrees more contorted than the

median angle of the stepwise ensemble, and that both ensembles were more distorted than the geometry optimized ground–state structure that had a distortion angle of 0.25 degrees (See Methods, Figure 4.6A). Similarly, the WT and S127A simultaneous ensembles were more likely to have larger distortions than the V155D simultaneous or stepwise ensemble (Figure 4.6B). This effect was also pronounced on the dihedrals C2–N1–C6–CX and C4–C5–C6–CX which reported on the orientation of the leaving carboxylate group of OMP (Supplementary Figures 4.S15 and 16).

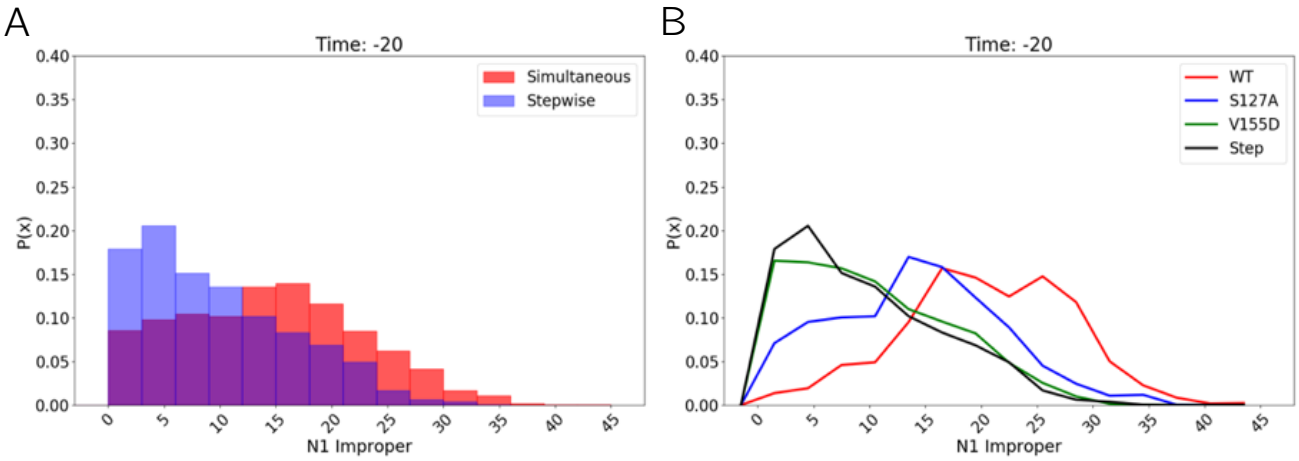


Figure 4.6: Distributions of the simultaneous and stepwise reactive pathway N1 absolute improper angle at $t = -20$ fs. (A) The simultaneous distribution IQR (25% and 75% quantile respectively) was [7.9°, 20.1°] with median 14.5° and std. dev. of 8.2°. The stepwise ensemble had IQR [3.7°, 14.3°] with median 8.2° and std. dev. of 6.9°. (B) The distributions of absolute N1 improper angle, stratified on protein system for WT simultaneous, S127A simultaneous, and V155D simultaneous versus stepwise. The WT IQR was [15.4°, 25.7°] with median 20.2° and std.dev. of 7.5 °. The S127A IQR was [8.4°, 19.2°] with median 14.5° and std.dev. of 7.4°. The V155D ensemble IQR was [4.5°, 15.4°] with median 9.2° and std.dev. of 7°. The stepwise ensemble was the same as in (A).

Stratifying the N1 improper angle as a function of these two features across both ensembles showed that shorter distances, for either feature, corresponded to greater angular distortion of the orotidyl ring (Figure 4.7). Shorter distances for the (D75*/CG – OMP/O2') feature (4.0 – 4.7 Å) and (D70/OD2 – OMP/OX1) feature (2.4 – 3.5 Å) had median distortion angles of 15.4° and 16.2°,

whereas the longer distances (6.8 – 7.5 Å and 5.5 – 6.7 Å respectively) possessed median angles nearly half the value, at 8.7° and 7.1° respectively.

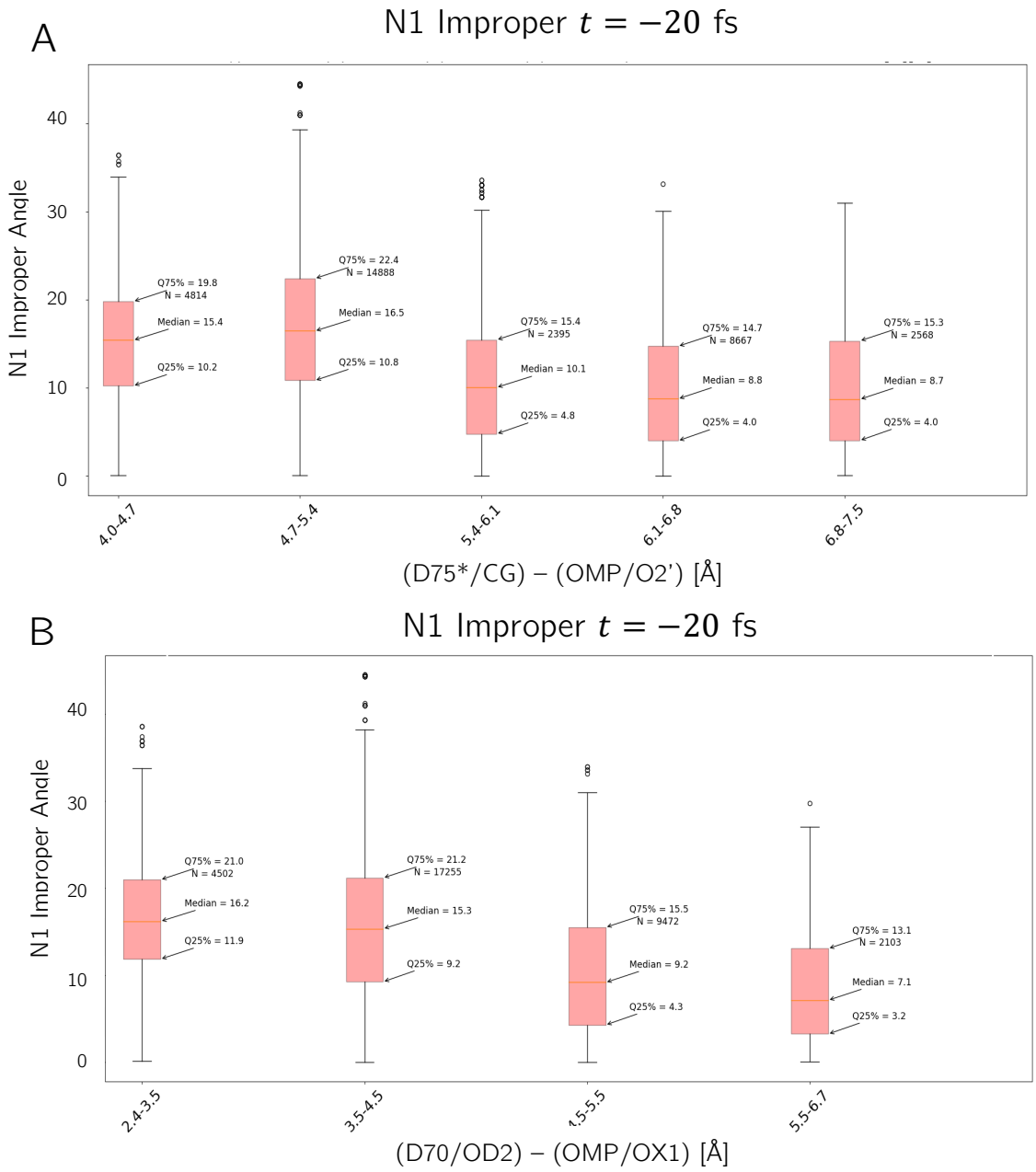


Figure 4.7: N1 improper angle stratified on the two different model–selected features for all trajectories across the simultaneous and stepwise ensembles combined. (A) The N1 improper angle as a function of the (D75*/CG) – (OMP/O2') distance revealed that close proximities (4.0 – 4.7 Å) had a median distortion angle of 15.4°, whereas the largest distances (6.8 – 7.5 Å) had a median distortion angle of 8.7°. (B) The N1 improper angle as a function of (D70/OD2) – (OMP/OX1) distance was observed to have a median distortion angle of 16.2° for short distances (2.4 – 3.5 Å) and a median distortion angle of 7.1° for larger distances (5.5 – 6.7 Å).

Taken together, this suggested the model–selected features were able to identify important "stand–in" features to signal important chemistry relevant to successful catalysis. The (D75*/CG – OMP/O2') distance suggested that the hydrogen bond formed between D75* and the 2'–hydroxyl of the ribophosphate group of OMP influenced substrate binding. Additionally, the (D70/OD2 – OMP/OX1) distance highlighted directly on ground–state destabilization by reporting on the proximity of two negatively charged groups within the active site. Both these features were observed to influence the planarity of the orotidyl ring prior to decarboxylation, as seen by considering the N1 improper angle as a function of stratified groups of these two features.

Further inspection of these features with regard to the specific protein distributions revealed that the WT and S127A ensemble were more effective at positioning the D75* and D70 residues in closer proximity to the OMP substrate when compared to either the V155D simultaneous ensemble versus the stepwise ensemble. This may also explain why the WT and S127A reactive pathways were so similar, as opposed to the V155D which exhibited more diversity in its decarboxylation strategy.

4.5 Conclusion and future directions

Enzyme active sites provide favorable conditions to allow for difficult reactions to occur. The following work investigated reactive pathways in decarboxylation performed by OMPDC and two catalytically hindered mutants: S127A and V155D. Prior analyses characterizing the energetics and dynamics of the reaction revealed that there were two reactive pathways distinguishable: a simultaneous and stepwise mechanism defined by the proximity of residue K72 as the decarboxylation proceeded. This study explored the relevant geometric features that were able to predict between these mechanisms, and probed their influence on reactivity.

Pair–feature classifiers across several time points, representing the system in the reactant basin up until committing 30 fs into crossing the barrier, achieved over 80% predictive performance for any time point. Interestingly, cumulative models did not outperform the pair–feature model. Inspection of the classifiers revealed many features alluding to prior experimental and mechanistic details discussed in the literature, despite access to hundreds of varied features [2, 3, 13–18, 23]. Two of these features were further discussed in the context of the chemical phenomena they highlighted. For both features, the (D75*/CG – OMP/O2') distance and the (D70/OD2 – OMP/OX1) distance, the simultaneous ensemble exhibited shorter interactions than the stepwise ensemble. These features, respectively, pointed to a hydrogen bond formed by D75* and the 2'–hydroxyl of OMP, and to explicit ground–state destabilization from the proximity of D70's carboxylate group with OMP's carboxylate group. Moreover, both these features were associated to increased distortion of the OMP ring planarity for shorter distances, suggesting the model was capable of identifying features that represented crucial chemistry important for catalysis. Stratifying the ensembles by protein system revealed that the WT and S127A ensembles were more likely to have shorter distances for both feature distributions than the simultaneous

V155D ensemble, or the stepwise ensemble (mostly comprised of V155D), underscoring how V155D may hinder ground–state destabilization.

A natural extension of this work is to provide an electronic description of how these features influence reactivity. While the N1 improper angle provides a view for how the π–network of the orotidyl ring is disrupted, *ab initio* quantum mechanical methods could quantify the degree of distortion of the π–character of the ring [28–31]. Despite possessing lower distortion angles, the stepwise ensemble is still capable of facilitating decarboxylation of OMPDC, suggesting there are other catalytic strategies to assist reactivity. Characterizing the underlying electronic state for the WT, S127A, and V155D could also highlight if there are fundamental differences between the two reactive pathways, and how the charged residue introduced by the V155D mutant interferes with the favorable conditions of the WT.

Additional inquiry into how models predict reactivity within each protein system, as opposed to the mechanism, could also reveal differences across how these systems cross the reaction barrier. Path sampling methods can generate simulations that attempt to cross the barrier but fail to complete the reaction (deemed "non–reactive") [8]. Comparing the reactive ensembles studied in this work to non–reactive ensembles generated for the WT and mutant systems can show how features promote reactivity, and what criteria are required in order to cross the reaction barrier successfully.

4.6 References

1. X. Zhang, K.N. Houk. Why Enzymes are Proficient Catalysts: Beyond the Pauling Paradigm. *Acc. Chem. Res.* 38, 5, 379–385, 2005.
3. K.K. Chan, B.M. Wood, A.A. Fedorov, E.V. Fedorov, H.J. Imker, T.L. Amyes, J.P. Richard, S.C. Almo, J.A. Gerlt. Mechanism of the orotidine 5'– monophosphate decarboxylase–catalyzed reaction: evidence for substrate destabilization. *Biochemistry* 48, 5518–553, 2009.
4. A. Warshel, M. Strajbl, J. Villa, J. Florian. Remarkable rate enhancement of orotidine 5‘–monophosphate decarboxylase is due to transition–state stabilization rather than to ground–state destabilization. *Biochemistry*. 39: 14728–14738, 2000.
5. G. Kiss, N. Celebi–Olcum, R. Moretti, D. Baker, K.N. Houk. Computational Enzyme Design. *Angew Chem Intl. Ed.* 52, 5700–5725, 2013.
6. D.A. Kraut, K.S. Carroll, D. Herschlag, D. Challenges in enzyme mechanism and energetics. *Annu. Rev. Biochem*. 72, 517− 571, 2003.
7. S. Hur, T.C. Bruice. The near attack conformation approach to the study of chorismite to prephenate reaction. *Proc. Natl. Acad. Sci. USA,* 100(21) 12015–12020, 2003.
8. Schramm, VL. Enzymatic Transition States and Transition–State Analog Design. *Annu. Rev. Biochem*. 693–720, 1998.
9. B.M. Bonk, J. Weis, B. Tidor. Machine Learning Identifies the Chemical Characteristics That Promote Enzyme Catalysis. *J. Am. Chem. Soc.* 141, 4108–4118, 2019.
10. B.R. Brooks, C.L. Brooks, III, A.D. MacKerell, Jr., L. Nilsson, R.J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A.R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R.W. Pastor, C.B. Post, J.Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock,

X. Wu, W. Yang, D.M. York, and M. Karplus. CHARMM: The Biomolecular Simulation Package. *J. Comput. Chem*, 30(10): 1545–1614, 2009.

11. J. Gao, P. Amara, C. Alhambra, M.J. Field. A generalized hybrid orbital (GHO) Method for the treatment of boundary atoms in combined QM/MM calculations. *J. Phys. Chem. A*, 102 (24), 4714–4721, 1998.
12. P.G. Bolhuis, D. Chandler, C. Dellago, P.L. Geissler. Transition Path Sampling: Throwing Ropes Over Rough Mountain Passes, in the Dark. *Annu. Rev. Phys. Chem*. 53: 291–318, 2002.
13. J. Yuan, A.M. Cardenas, H.F. Gilbert, T. Palzkill. Determination of the amino acid sequence requirements for catalysis by the highly proficient orotodine 5′-monophosphate decarboxylase. *Protein Sci*. 20; 1891–1906, 2011.
14. T.L. Amyes, S.A. Ming, L.M Goldman, B.M Wood, B.J. Desai, J.A. Gerlt, J.P. Richard. Orotidine 5′–monophosphate decarboxylase: transition–state stabilization from remote protein–phosphodianion interactions. *Biochemistry*. 51(23), 4630–4632, 2012.
15. A.A Federov, E.V. Federov, B.M Wood, J.A. Gerlt, S.C Almo. Conformational changes in orotidine 5′-monophosphate decarboxylase: "remote" residues that stabilize the active conformation. *Biochemistry*. 49; 3514–3516, 2010.
16. T.W. Traut, B.R. Temple. The chemistry of the reaction determines the invariant amino acids during the evolution and divergence of orotidine 5′–monophosphate decarboxylase. *J. Biol. Chem.* 275: 28675–81, 2000
17. B.J. Desai, M. Wood, A.A. Federov, E.V. Federov, B. Goryanova, T.L. Amyes, J.P. Richard, S.C. Almo, J.A. Gerlt. Conformational changes in orotidine 5′-monophosphate decarboxylase: A structure–based explanation for how the 5'–phosphate group activates the enzyme. *Biochemistry*. 51, 43, 8665–8678, 2012.

18. N. Wu, Y. Mo, J. Gao, E.F Pai. Electrostatic stress in catalysis: Structure and mechanism of the enzyme orotidine monophosphate decarboxylase. *Proc. Natl. Acad. Sci. USA* 97, 2017–2022, 2000

19. V. Iiams, B.J. Desai, A.A Fedorov, E.V. Fedorov, S.C Almo, J.A. Gerlt. Mechanism of the orotidine 5′–monophosphate decarboxylase–catalyzed reaction: Importance of residues in the orotate binding site. *Biochemistry*. 50(39): 8497–8507, 2011.

20. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay. Scikit–learn: Machine learning in python. *JLMR*, 12, 2825–2830, 2011.

21. B.P. Callahan, A.F. Bell, P.J. Tonge, R. Wolfenden. A Raman–active competitive inhibitor of OMP decarboxylase. *Bioinorg. Chem.*, 34, 59–65, 2006.

22. N. Wu, W. Gillon, E.F. Pai. Mapping the Active–Site Ligand Interactions of Orotidine 5′-monophosphate Decarboxylase by Crystallography. *Biochemistry*. 41 (12), 4002–4011, 2002.

23. B.G. Miller, G. L. Butterfoss, S.A. Short, R. Wolfenden. Role for enzyme–ribofuranosyl contacts in the ground state and transition state for orotidine 5′–phosphate decarboxylase: a role for substrate destabilization? *Biochemistry*, 40, 6227–6232, 2001.

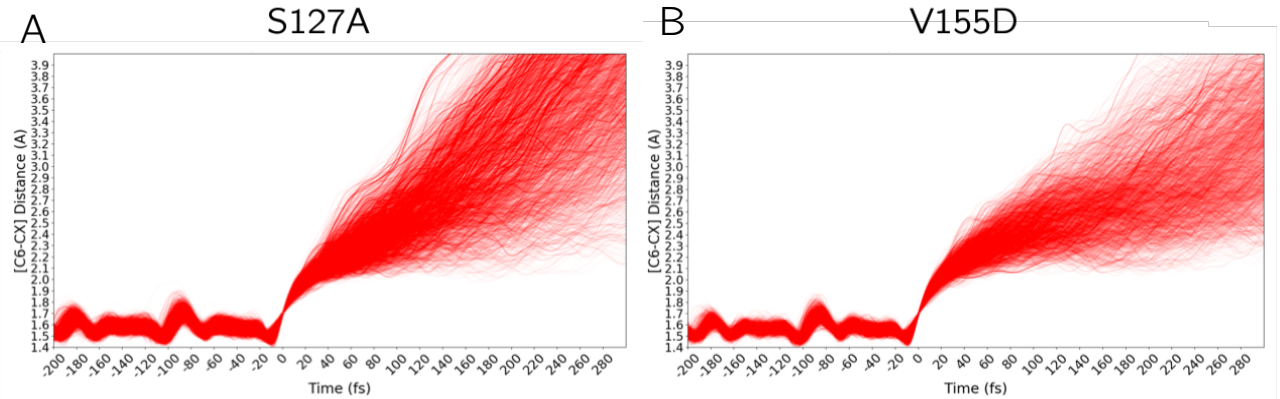
24. B.G. Miller, M.J. Snider, S.A. Short, R. Wolfenden. Dissecting a charged network at the active site of orotidine–5'–phosphate decarboxylase. *J. Biol. Chem.* 276 15174–15176, 2001.

25. B.P. Callahan, R. Wolfenden. Charge Development in the Transition State for Decarboxylations in Water: Spontaneous and Acetone–Catalyzed Decarboxylation of Aminomalonate. *J. Am. Chem. Soc.,* 126 (2004) 4514–4515.

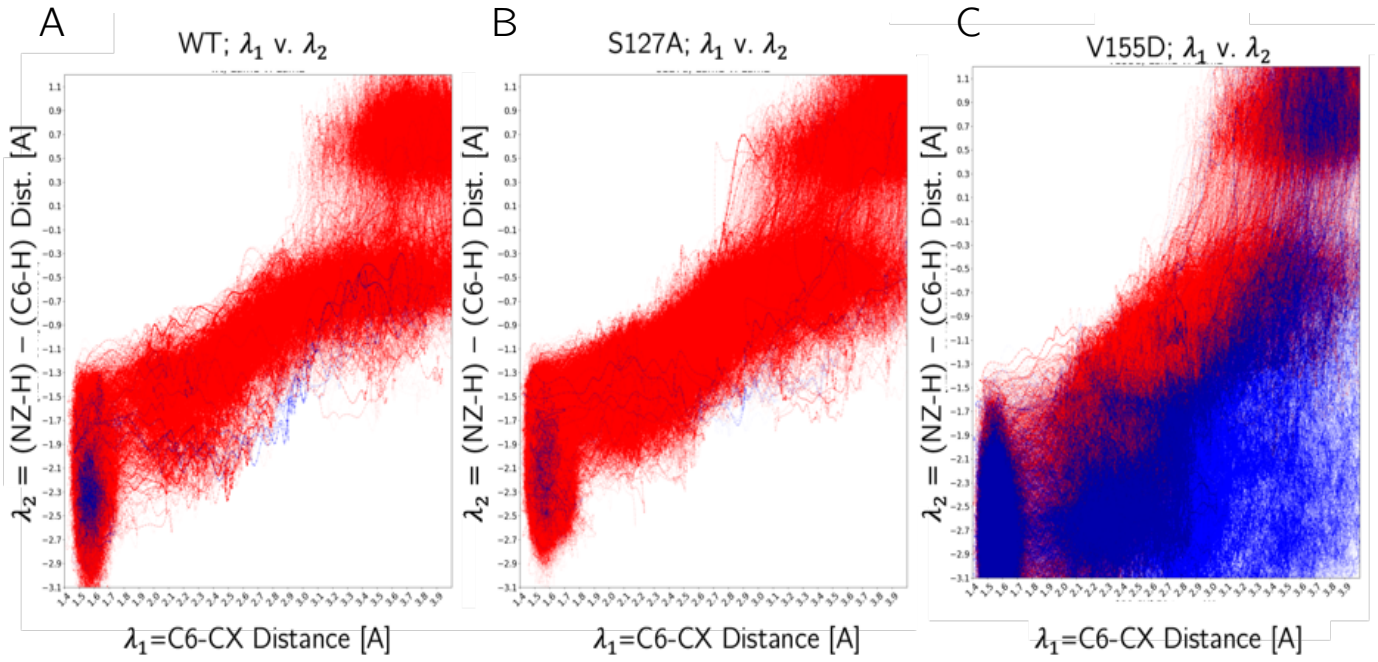
26. H. Hu, A. Boone, W. Yang. Mechanism of OMP decarboxylation in orotidine 5′–monophosphate decarboxylase. *J. Am. Chem. Soc.,* 130(44), 14493–503, 2008.

27. M. Fujihashi, T. Ishida, S. Kuroda, L.P. Kotra, E.F. Pai, K. Miki. Substrate distortion contributes to the catalysis of orotidine 5′-monophosphate. *J. Am. Chem. Soc.,* 134 (46), 17432–17443, 2013.
28. J. Gao. Catalysis by enzyme conformational change as illustrated by orotidine 5′–monophosphate decarboxylase. *Curr. Opin. Struct. Biol.* 13, 184–192, 2003
29. P. Hohenberg, W. Kohn. Inhomogeneous Electron Gas. *Phys. Rev.* 136, B864, 1964.
30. W. Kohn, L. J. Sham. Self–Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.* 140, A1133 (1965).
31. C. R. Landis, F. Weinhold. "The NBO View of Chemical Bonding", in, G. Frenking and S. Shaik (eds.), The Chemical Bond: Fundamental Aspects of Chemical Bonding. *Wiley*, pp. 91–120, 2014.
32. F. Weinhold. Natural Bond Orbital Analysis: A Critical Overview of its Relationship to Alternative Bonding Perspectives. *J. Comp. Chem.* 33, 2363–2379, 2012.
33. J.M. MacLeod, F. Rosei. Comprehensive Nanoscience and Technology. Editors: D.L. Andrews, G.D.Scholes, G.P. Wiederrecht. *Academic Press*, ISBN: 9780123743909, 2010.

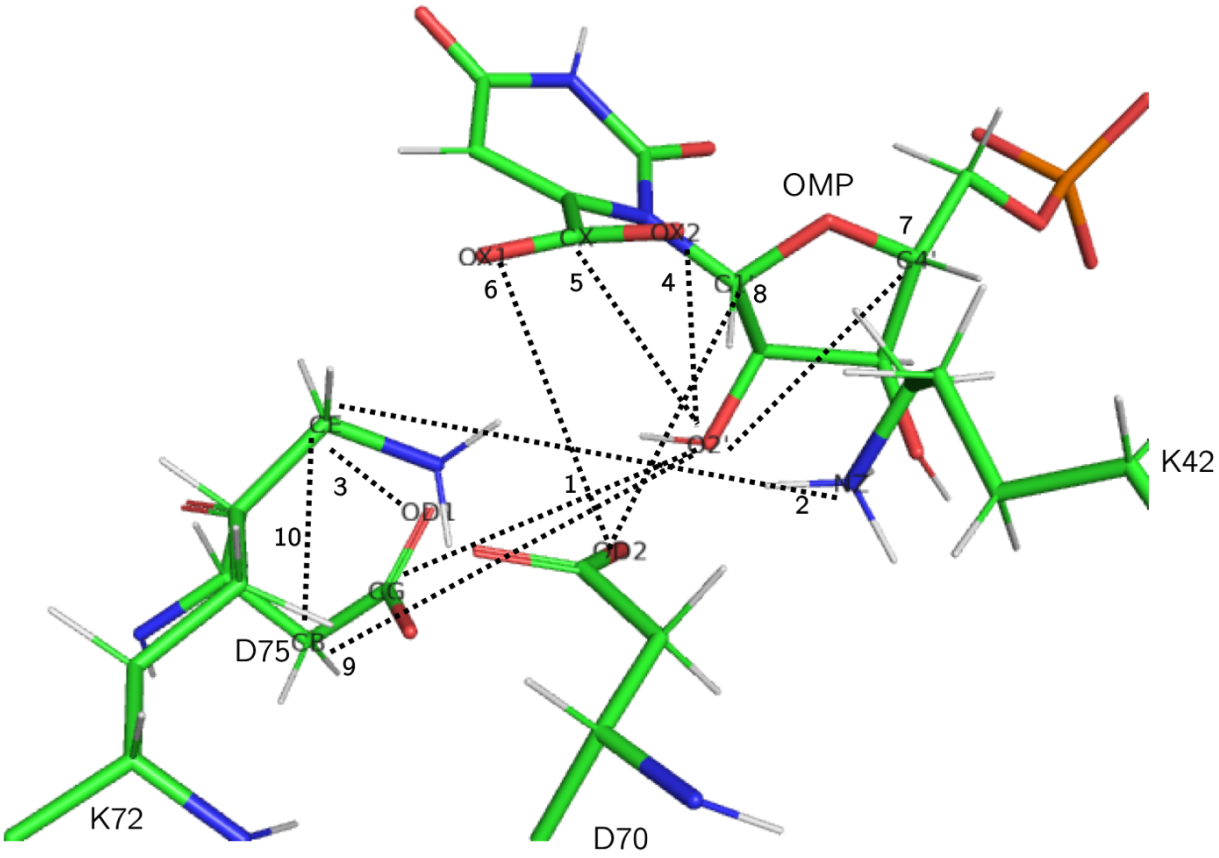
4.7 Supplementary Information



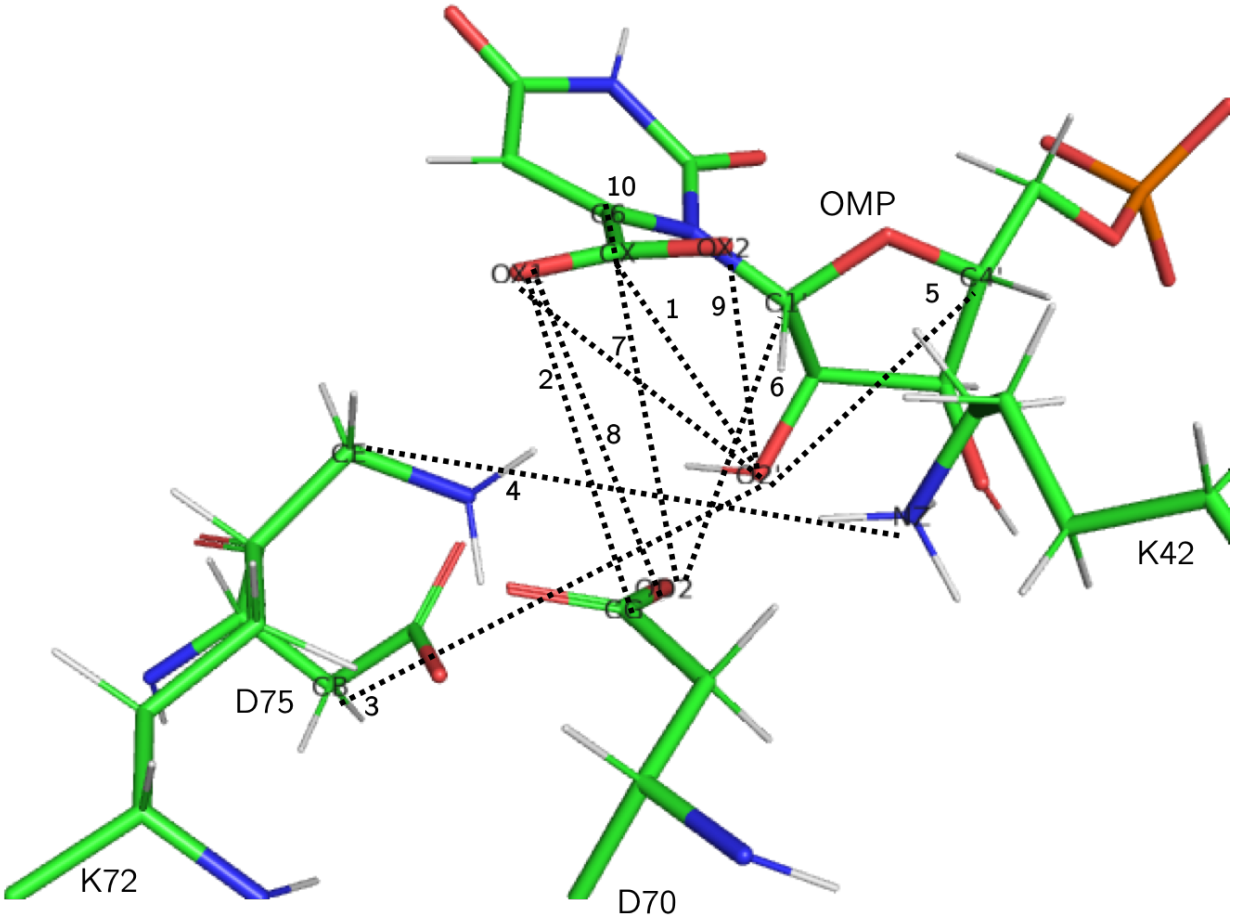
Supplementary Figure 4.S1: Additional time–aligned trajectories of the decarboxylation coordinate from the (A) S127A ensemble and (B) V155D ensemble. The time $t = 0$ fs corresponded to the last time a simulation was in the reactant basin before committing to crossing the reaction barrier. All trajectories across the three protein systems successfully decarboxylated the OMP substrate.



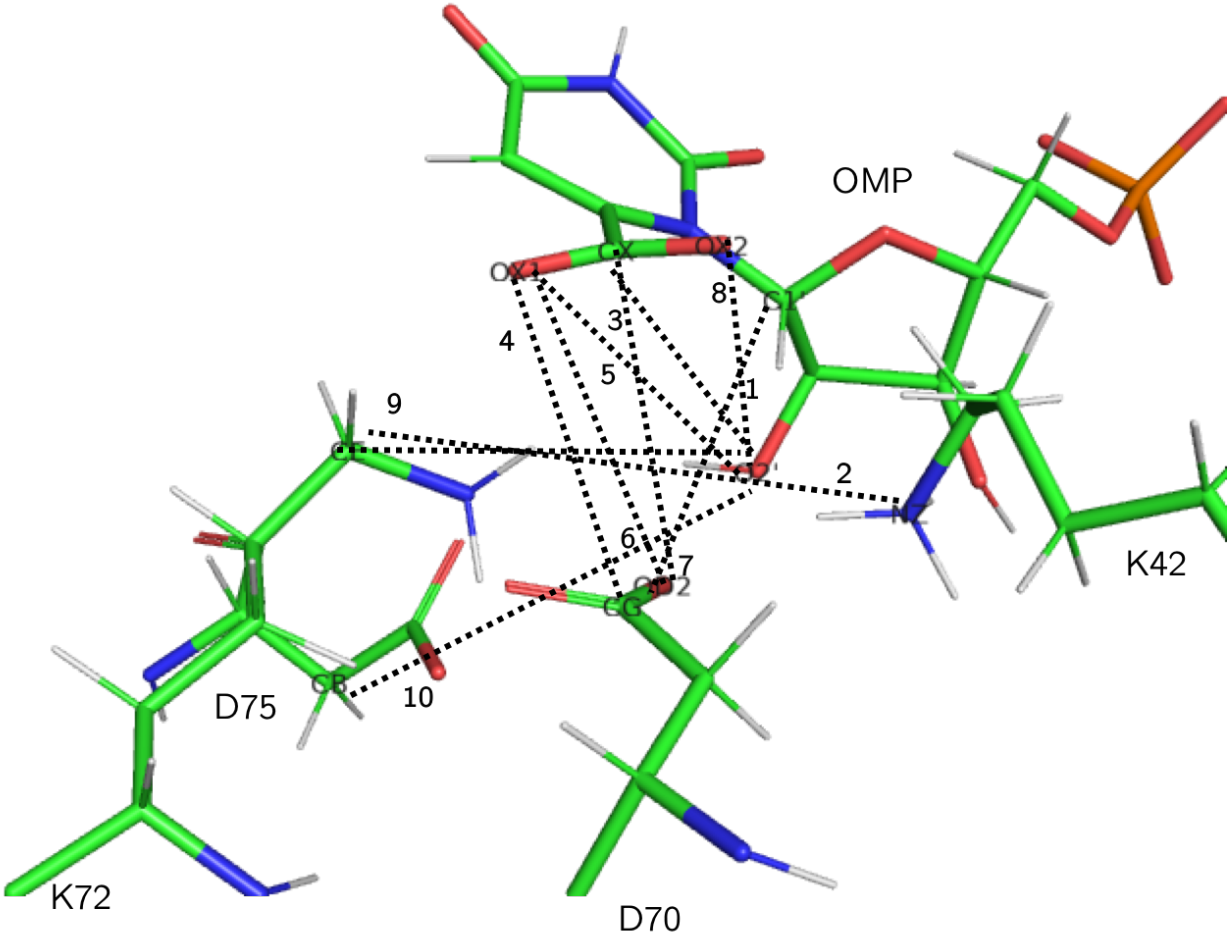
Supplementary Figure 4.S2: Simulations of the three protein systems as a function of two order parameters, the decarboxylation coordinate (C6–CX) and the proton–transfer coordinate (K72/NZ–K72/H – OMP/C6–K72/H). Each time point in the simulation is represented as a single marker, for all time points across the simulation. Simultaneous trajectories are colored red, and stepwise trajectories are indicated in blue. (A) WT decarboxylation pathway; most trajectories within the WT ensemble decarboxylated in a simultaneous manner. (B) S127A decarboxylation pathway; most trajectories in the S127A ensemble also decarboxylated with the simultaneous mechanism. (C) V155D decarboxylation pathway; by far, the V155D ensemble had the most stepwise trajectories compared to the other two mutants. The V155D mutant decarboxylated in both the simultaneous manner, akin to WT and S127A, but also in the stepwise pathway.



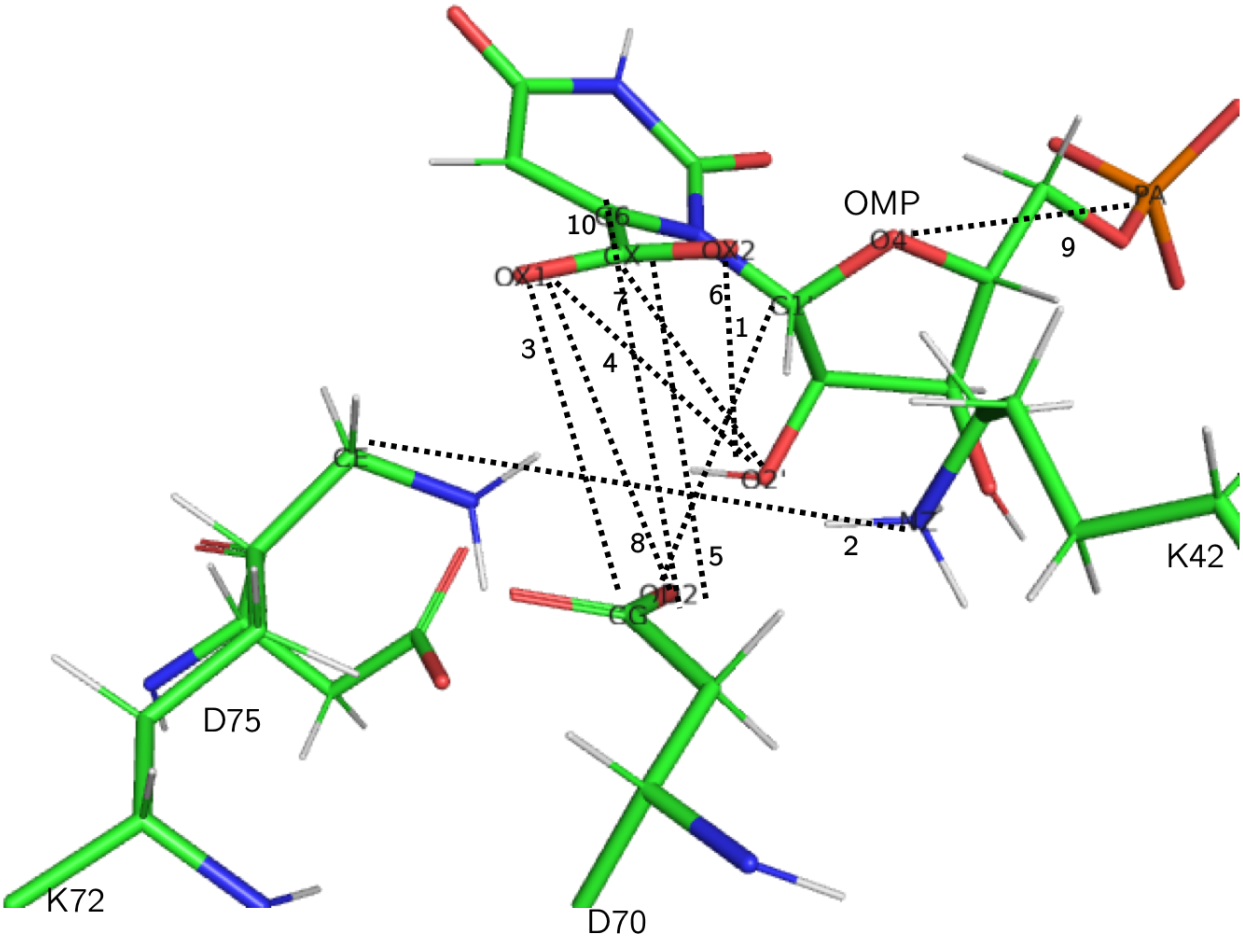
Supplementary Figure 4.S3: Schematic of top performing feature pairs for classifiers from timepoint $t = -20$ fs. The features are listed in the order indicated from Table 2. All features were distances.



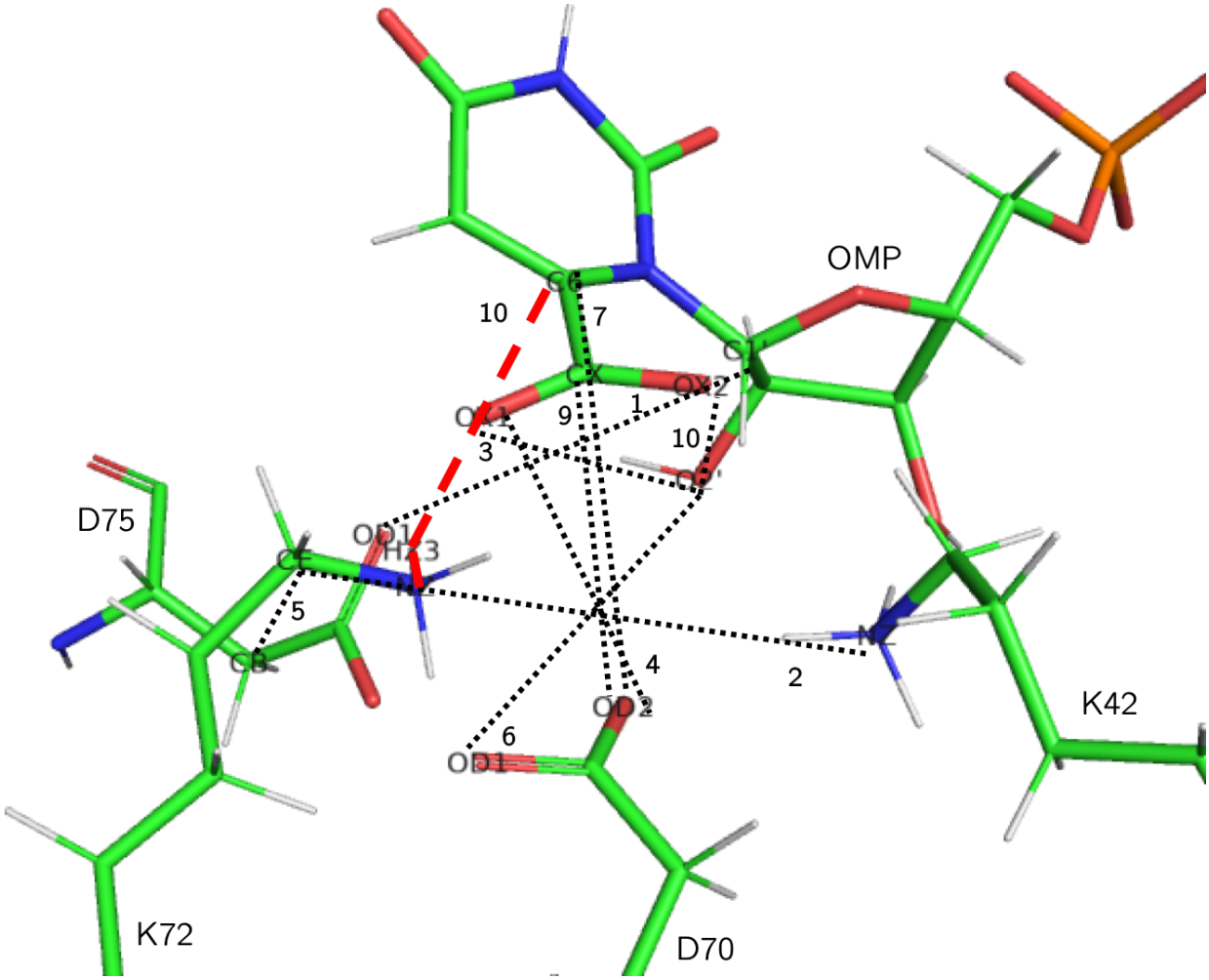
Supplementary Figure 4.S4: Schematic of top performing feature pairs for classifiers from timepoint $t = -10$ fs. The features are listed in the order indicated from Table 2. All features were distances.



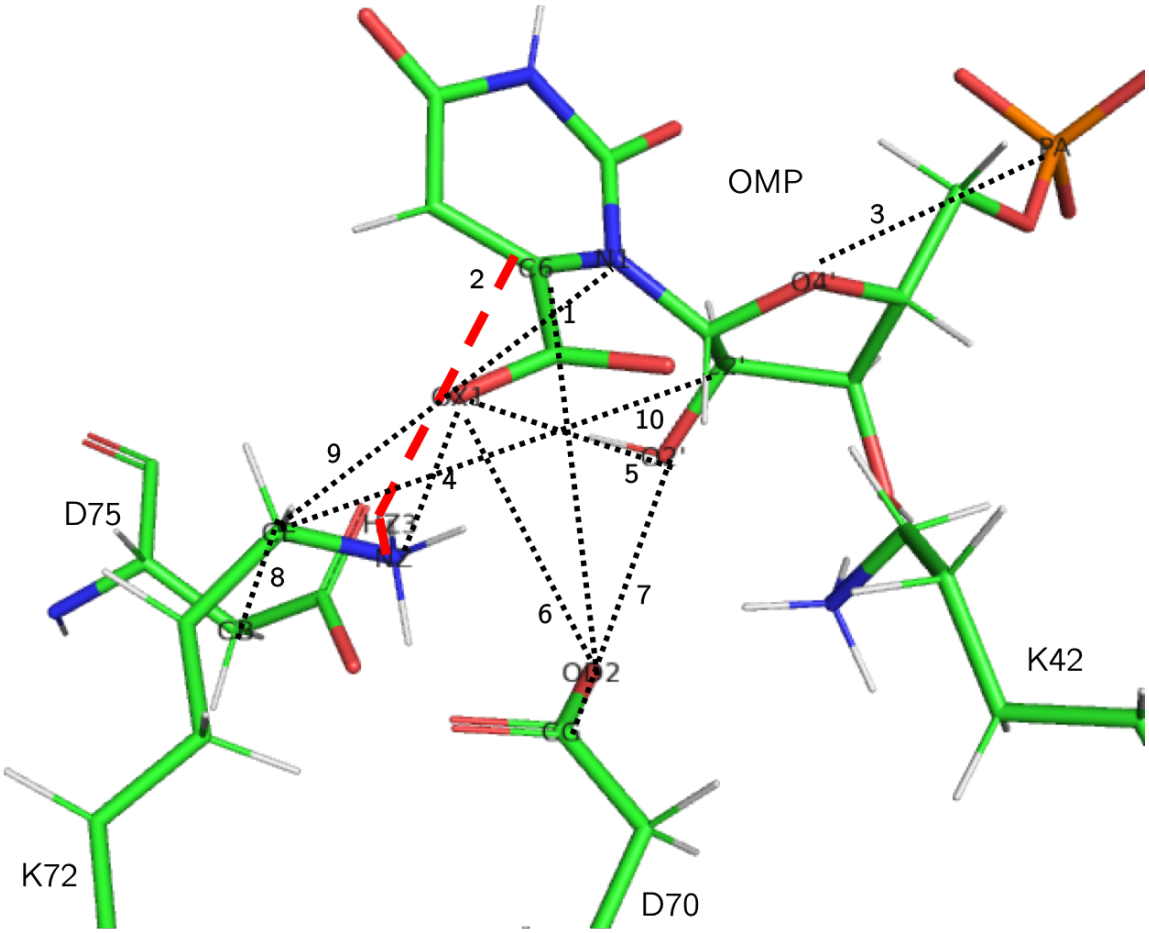
Supplementary Figure 4.S5: Schematic of top performing feature pairs for classifiers from timepoint $t = 0$ fs. The features are listed in the order indicated from Table 2. All features were distances.



Supplementary Figure 4.S6: Schematic of top performing feature pairs for classifiers from timepoint $t = 10$ fs. The features are listed in the order indicated from Table 2. All features were distances.



Supplementary Figure 4.S7: Schematic of top performing feature pairs for classifiers from timepoint $t = 20$ fs. The features are listed in the order indicated from Table 2. Black features indicate distances and the red feature is an angle.



Supplementary Figure 4.S8: Schematic of top performing feature pairs for classifiers from timepoint $t = 30$ fs. The features are listed in the order indicated from Table 2. Black features indicate distances and the red feature is an angle.

t=–20 fs	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
Set 1	0.43	–1.32	1.25	–	–	–	–	–	–	–	–
Set 2	0.23	–	–	–0.61	2.11	–	–	–	–	–	–
Set 3	0.33	–	–	–	–	1.43	–1.11	–	–	–	–
Set 4	0.49	–	–	–	–	–	–	1.65	–0.92	–	–
Set 5	0.39	–	–	–	–	–	–	–	–	–1.94	–0.62
Cumulative	0.55	–1.0	0.69	–0.41	–0.37	0.98	–0.48	0.18	–0.21	0.35	0.03

Supplementary Table 1: Feature coefficients for the top 5 pairwise models and the cumulative model, for time $t = -20$ fs. The order of the coefficients matches the order indicated by the Feature ID in Table 2 of the text, by order of appearance. The term β_0 refers to the value of the bias.

t=–10 fs	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
Set 1	0.35	1.39	–1.20	–	–	–	–	–	–	–	–
Set 2	0.49	–	–	–1.39	1.22	–	–	–	–	–	–
Set 3	0.50	–	–	–	–	1.63	–0.92	–	–	–	–
Set 4	0.40	–	–	–	–	–	–	1.08	–1.76	–	–
Set 5	0.29	–	–	–	–	–	–	–	–	1.25	–1.10
Cumulative	0.61	1.98	–1.33	–1.11	1.07	–0.22	–0.32	–0.12	0.35	–1.67	0.42

Supplementary Table 2: Feature coefficients for the top 5 pairwise models and the cumulative model, for time $t = -10$ fs. The order of the coefficients matches the order indicated by the Feature ID in Table 2 of the text, by order of appearance. The term β_0 refers to the value of the bias.

t= 0 fs	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
Set 1	0.31	–0.85	1.66	–	–	–	–	–	–	–	–
Set 2	0.35	–	–	1.29	–1.29	–	–	–	–	–	–
Set 3	0.40	–	–	–	–	1.0	–1.84	–	–	–	–
Set 4	0.35	–	–	–	–	–	–	–1.30	1.20	–	–
Set 5	0.34	–	–	–	–	–	–			0.77	–1.90
Cumulative	0.59	–0.25	0.98	2.04	–1.45	0.35	0.78	0.06	–1.59	0.14	–0.92

Supplementary Table 3: Feature coefficients for the top 5 pairwise models and the cumulative model, for time $t = 0$ fs. The order of the coefficients matches the order indicated by the Feature ID in Table 2 of the text, by order of appearance. The term β_0 refers to the value of the bias.

t=10 fs	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
Set 1	0.31	–0.80	1.69	–	–	–	–	–	–	–	–
Set 2	0.44	–	–	–1.86	1.07	–	–	–	–	–	–
Set 3	0.46	–	–	–	–	–1.38	1.15	–	–	–	–
Set 4	0.33	–	–	–	–	–	–	1.15	–1.37	–	–
Set 5	0.42	–	–	–	–	–	–	–	–	1.15	–1.29
Cumulative	0.57	–0.07	1.07	–1.37	0.24	–1.19	0.03	0.72	1.35	–0.03	0.33

Supplementary Table 4: Feature coefficients for the top 5 pairwise models and the cumulative model, for time $t = 10$ fs. The order of the coefficients matches the order indicated by the Feature ID in Table 2 of the text, by order of appearance. The term β_0 refers to the value of the bias.

t= 20 fs	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
Set 1	0.49	–1.14	1.54	–	–	–	–	–	–	–	–
Set 2	0.41	–	–	0.90	–1.99	–	–	–	–	–	–
Set 3	0.54	–	–	–	–	–0.79	2.06	–	–	–	–
Set 4	0.14	–	–	–	–	–	–	–2.24	0.42	–	–
Set 5	0.40	–	–	–	–	–	–	–	–	–1.55	1.07
Cumulative	0.56	–0.35	0.68	0.60	–0.47	–0.39	–0.25	0.43	0.44	–1.11	0.36

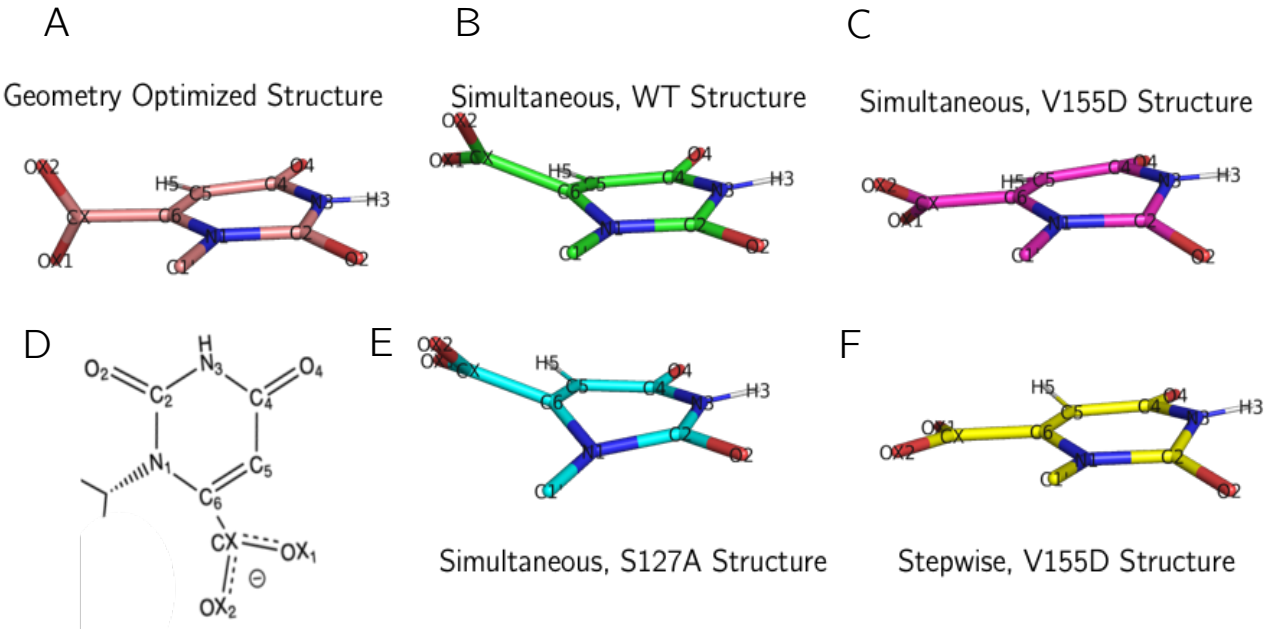
Supplementary Table 5: Feature coefficients for the top 5 pairwise models and the cumulative model, for time $t = 20$ fs. The order of the coefficients matches the order indicated by the Feature ID in Table 2 of the text, by order of appearance. The term β_0 refers to the value of the bias.

t=30 fs	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
Set 1	0.15	–2.30	0.47	–	–	–	–	–	–	–	–
Set 2	0.32	–	–	1.48	–0.99	–	–	–	–	–	–
Set 3	0.40	–	–	–	–	0.76	–2.09	–	–	–	–
Set 4	0.65	–	–	–	–	–	–	2.07	–0.81	–	–
Set 5	0.28	–	–	–	–	–	–	–	–	–0.92	2.31
Cumulative	0.51	–0.03	0.43	0.23	0.10	0.30	–1.18	–0.01	–0.69	–0.45	0.84

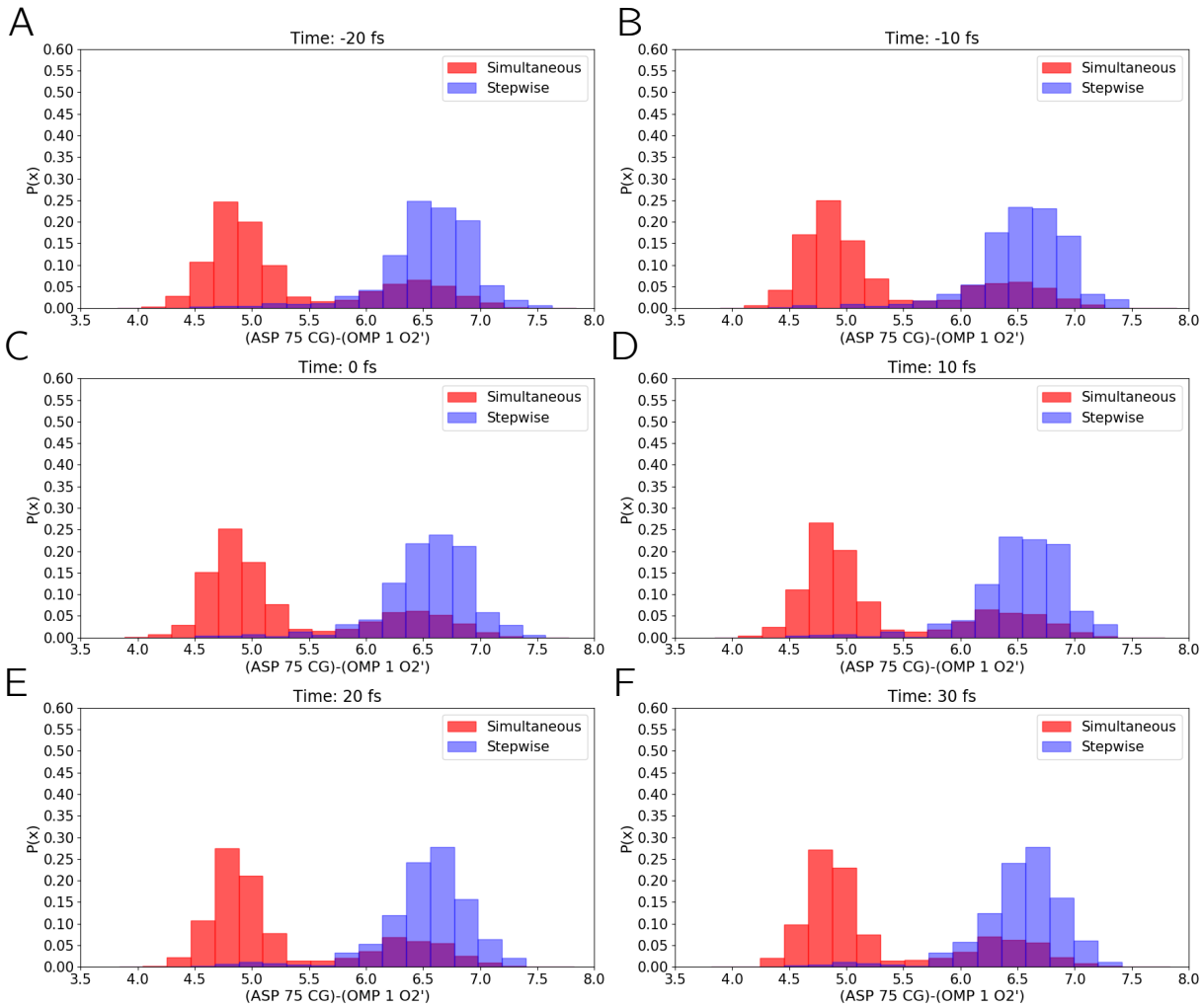
Supplementary Table 6: Feature coefficients for the top 5 pairwise models and the cumulative model, for time $t = 30$ fs. The order of the coefficients matches the order indicated by the Feature ID in Table 2 of the text, by order of appearance. The term β_0 refers to the value of the bias.

Features	Feature Type
(ASP 70 OD2)-(OMP 1 OX1)	**D70-substrate**
(OMP 1 OX2)-(OMP 1 O2')	**Intra-substrate**
(LYS 42 NZ)-(LYS 72 CE)	**Catalytic Tetrad**
(OMP 1 OX1)-(OMP 1 O2')	**Intra-substrate**
(OMP 1 CX)-(OMP 1 O2')	**Intra-substrate**
(ASP 70 OD2)-(OMP 1 C1')	**D70-substrate**
(ASP 70 OD2)-(OMP 1 C6)	**D70-substrate**
(ASP 75 CB)-(OMP 1 O2')	**D75-substrate**
(LYS 72 CE)-(ASP 75 CB)	**Catalytic Tetrad**
(ASP 70 CG)-(OMP 1 OX1)	**D70-substrate**
(ASP 70 OD2)-(OMP 1 CX)	**D70-substrate**
(OMP 1 C4')-(OMP 1 O2')	**Intra-substrate**
(OMP 1 O4')-(OMP 1 PA)	**Intra-substrate**
(LYS 72 HZ3)-(LYS 72 NZ)-(OMP 1 C6)	**K72-substrate**
(ASP 75 CG)-(OMP 1 O2')	**D75-substrate**
(LYS 72 CE)-(ASP 75 OD1)	**Catalytic Tetrad**
(LYS 72 CE)-(OMP 1 O2')	**K72-substrate**
(ASP 70 OD1)-(OMP 1 O2')	**D70-substrate**
(ASP 75 OD1)-(OMP 1 C1')	**D75-substrate**
(LYS 72 NZ)-(OMP 1 OX1)	**K72-substrate**
(ASP 70 CG)-(OMP 1 O2')	**D70-substrate**
(LYS 72 CE)-(OMP 1 C2')	**K72-substrate**
(LYS 72 CE)-(OMP 1 N1)	**K72-substrate**

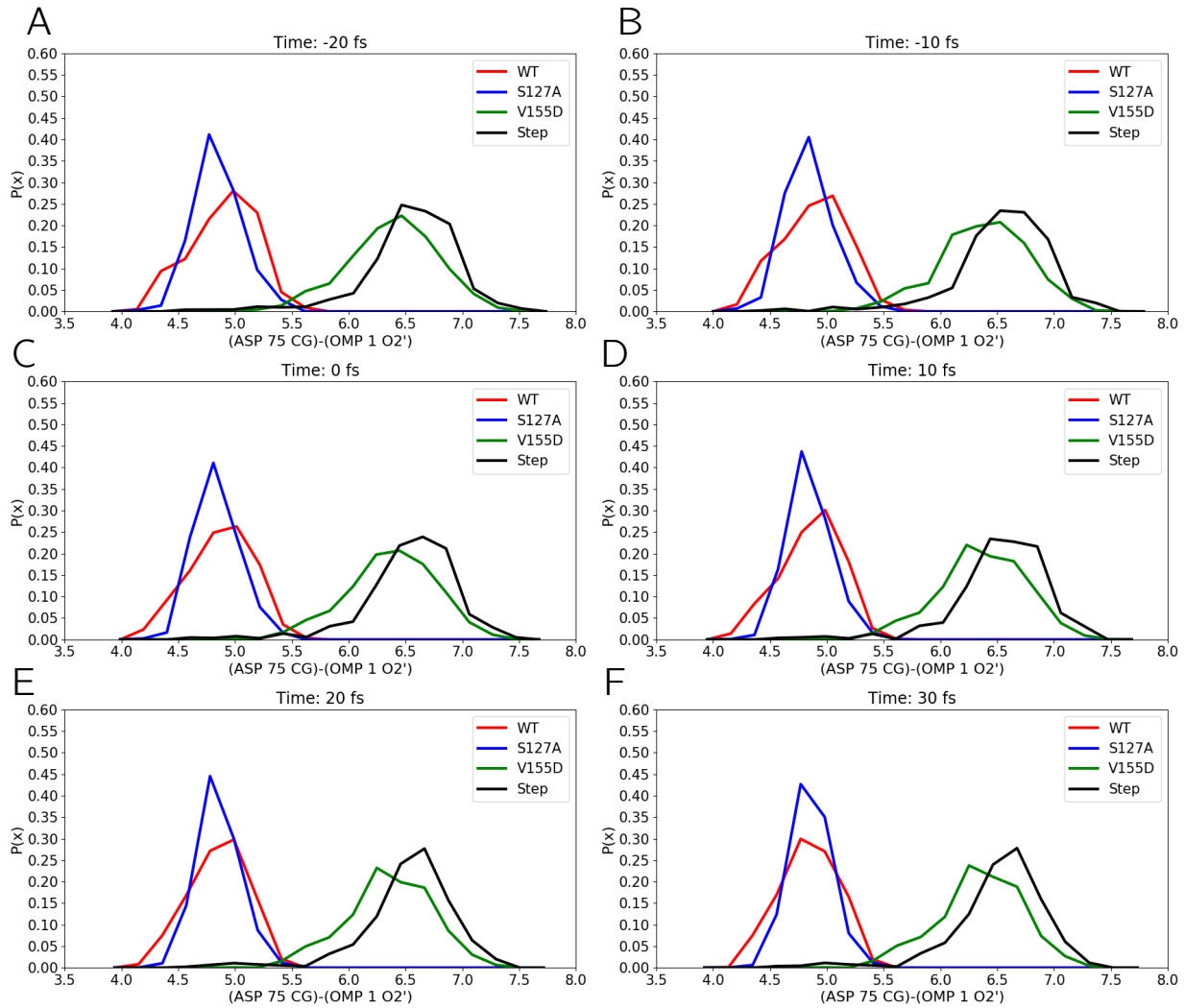
Supplementary Table 7: The 23 unique features identified across all six time points for the simultaneous/stepwise classification task, labeled by feature type.



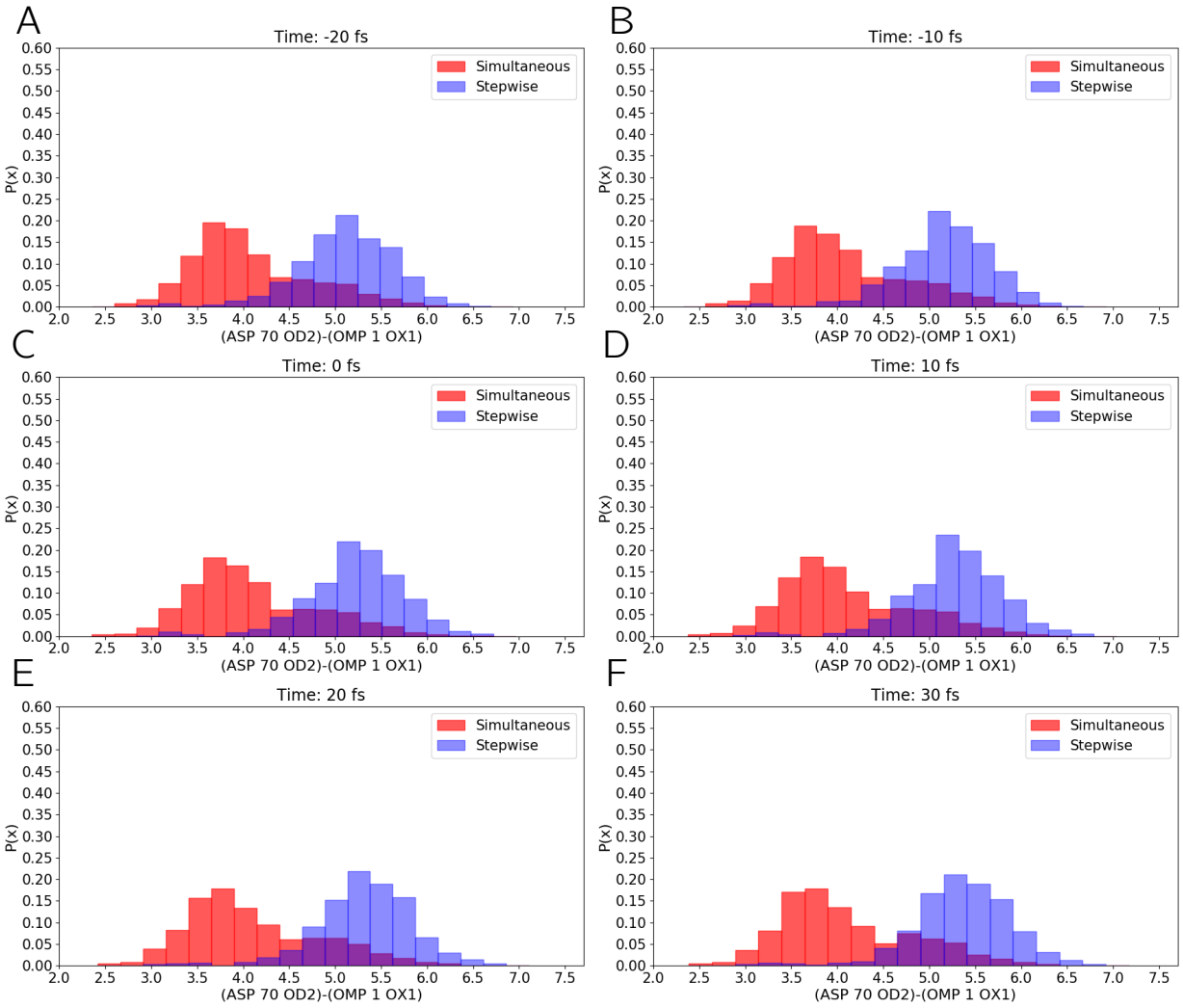
Supplementary Figure 4.S9: (A) An example of the orotidyl ring via RHF/6–31G* optimization. The C6–CX angle lies planar to the orotidyl ring. (B) An example of the orotidyl ring for a WT simultaneous trajectory at time $t = -20$ fs. (C) Example of the orotidyl ring for a V155D simultaneous trajectory at time $t = -20$ fs. (D) A schematic of the orotidyl ring with labeled atoms. (E) Example of the orotidyl ring for a S127A simultaneous trajectory at time $t = -20$ fs. (F) Example of an orotidyl ring in a stepwise (V155D) trajectory at time $t = -20$ fs.



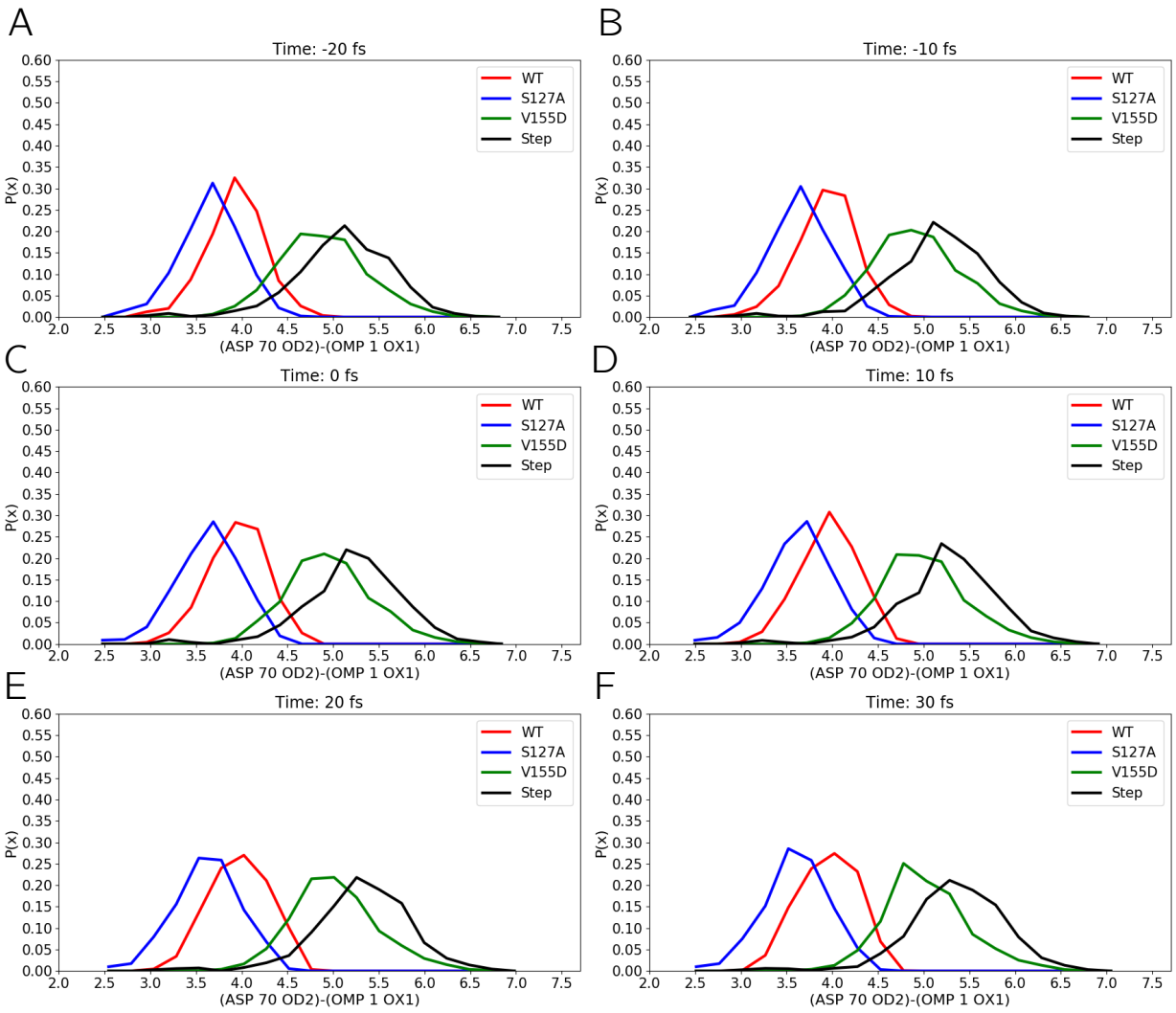
Supplementary Figure 4.S10: Simultaneous (red) and stepwise (blue) distributions of the (D75*–CG) – (OMP–O2') distance across the six time points tested. (A) Distributions for time t = –20 fs (B) t = –10 fs (C) t = 0 fs (D) t = 10 fs (E) t = 20 fs (F) t = 30 fs. For all time points, the simultaneous ensemble distribution was shifted nearly 2 angstroms shorter than the stepwise ensemble. Marginal overlap existed between the simultaneous distribution and the stepwise distribution between 5.5 – 7.5 Å.



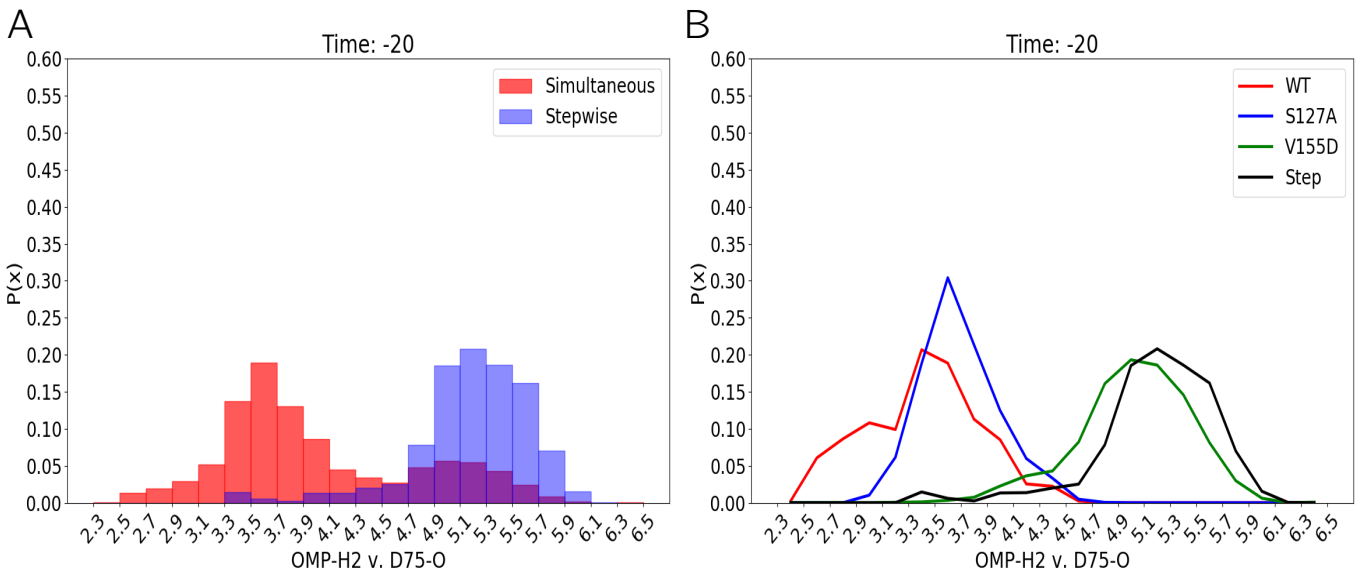
Supplementary Figure 4.S11: Distributions of the (D75*–CG) – (OMP–O2') for the WT, S127A, V155D simultaneous ensembles versus the total stepwise ensemble (indicated in the red, blue, green, and black lines respectively). (A) Distributions for time t = –20 fs (B) t = –10 fs (C) t = 0 fs (D) t = 10 fs (E) t = 20 fs (F) t = 30 fs. The WT and S127A ensemble distributions possessed virtually no overlap with the V155D and stepwise ensemble and were roughly 2.0 Å close for all time points, suggesting the D75*/OMP O2' ribose interaction was stronger in the WT/S127A ensemble. The V155D simultaneous ensemble is slightly shifted toward the WT/S127A ensemble compared to the stepwise ensemble.



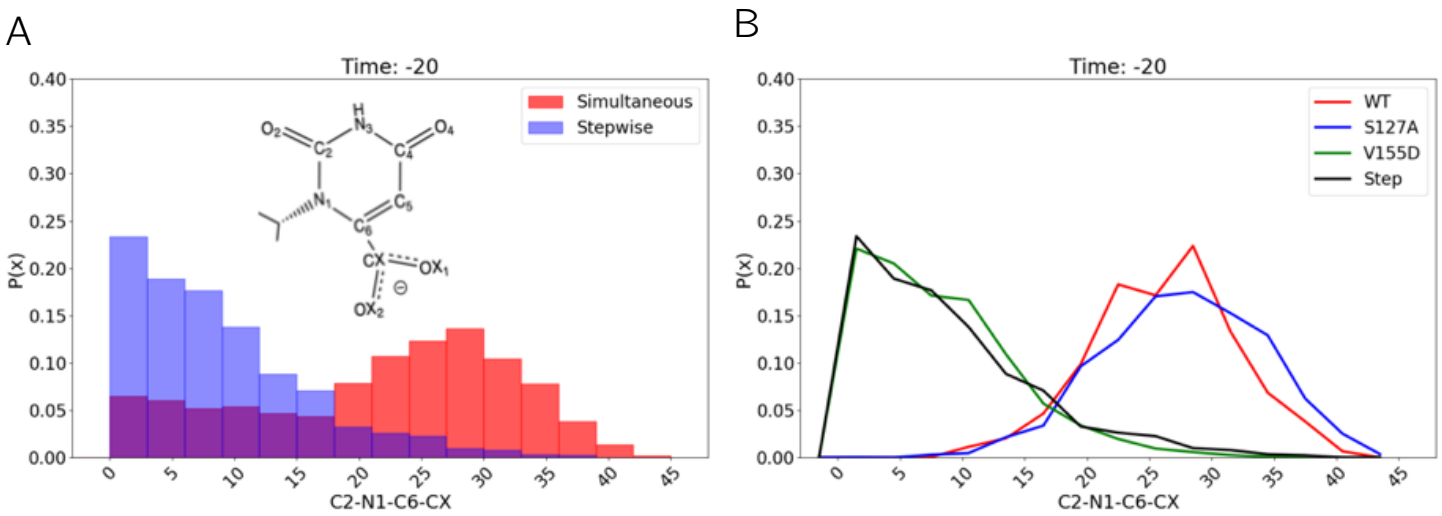
Supplementary Figure 4.S12: Simultaneous (red) and stepwise (blue) distributions of the (D70 – OD2) – (OMP – OX1) distance across the six time points tested. (A) Distributions for time t = –20 fs (B) t = –10 fs (C) t = 0 fs (D) t = 10 fs (E) t = 20 fs (F) t = 30 fs. For all time points, there was small overlap between the distributions near the 4.0 – 6.0 Å.



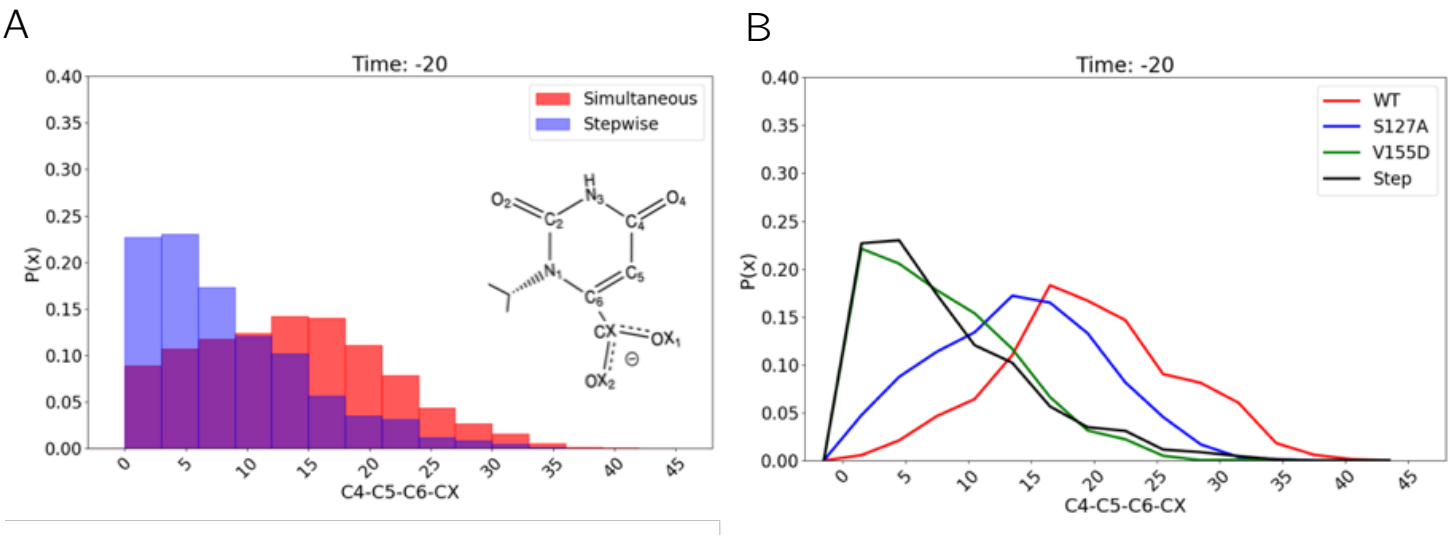
Supplementary Figure 4.S13: Distributions of the (D70 – OD2) – (OMP – OX1) for the WT, S127A, V155D simultaneous ensembles versus the total stepwise ensemble (indicated in the red, blue, green, and black lines respectively). (A) Distributions for time t = –20 fs (B) t = –10 fs (C) t = 0 fs (D) t = 10 fs (E) t = 20 fs (F) t = 30 fs. Small overlap existed across time between 4.0 – 5.0 Å. The WT/S127A ensembles formed shorter contact between D70's carboxylate oxygen and the leaving group's, OMP, carboxylate oxygen. The simultaneous V155D ensemble distribution formed shorter contacts, with a distribution shifted left, than the stepwise ensemble.



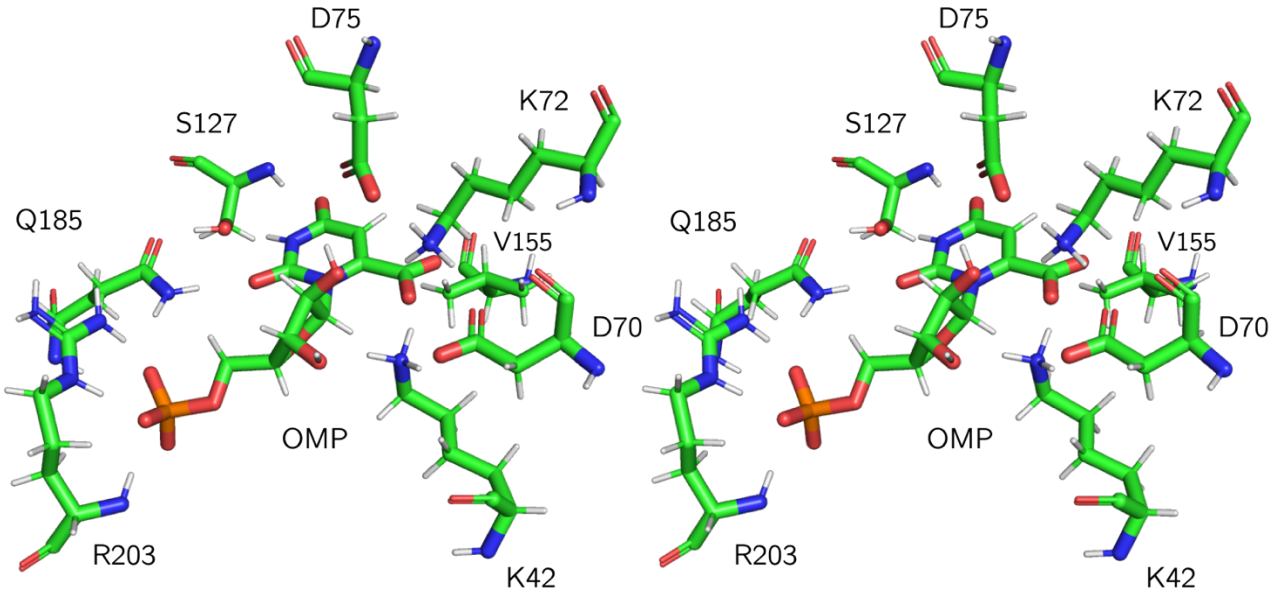
Supplementary Figure 4.S14: Distributions of the closest oxygen of D75* with respect to the 2' hydroxyl proton of the ribose ring on OMP, reporting on a potential hydrogen bond between D75* and the 2' hydroxyl group of OMP [33]. (A) Simultaneous (red) and stepwise (ensemble) distributions at time t = – 20 fs; the simultaneous IQR (at the 25% and 75% quantile respectively) was [3.5 Å, 3.8 Å] with median 3.6 Å and standard deviation 0.75 Å. The stepwise ensemble IQR was [5.0 Å, 5.5 Å] with median 5.2 Å and standard deviation 0.45 Å. (B) Distributions stratified on the simultaneous WT, S127A, and V155D ensembles versus the stepwise ensemble. The WT IQR is [3.1 Å, 3.7 Å] with median 3.4 Å and std. dev. 0.43 Å. The S127A IQR was [3.5 Å, 3.9 Å] with median 3.7 Å and std. dev. 0.3 Å. The V155D IQR was [4.8 Å, 5.4 Å] with median 5.0 Å and std. dev. 0.43 Å. The stepwise ensemble was the same as (A).



Supplementary Figure 4.S15: Distributions of the C2–N1–C6–CX angle between the reactive pathways for $t = -20$ fs. (A) Distribution comparison for the simultaneous and stepwise pathways. The simultaneous distribution IQR (in order of the 25%, 50%, and 75% quantile) is [13.2, 23.9, 29.7] degrees with std. dev 10.6 degrees and the stepwise distribution IQR is [3.2, 7.2, 12.3] degrees with std. dev 7.1 degrees. (B) Distribution comparison stratified on the WT, S127A, and V155D simultaneous ensembles versus the stepwise ensemble. The WT IQR is [22.3, 26.5, 29.9] degrees with std. dev 5.7 degrees; the S127A IQR is [23.2, 27.7, 32.3] degrees with std. dev. 6.4 degrees; the V155D IQR is [3.4, 7.3, 11.7] degrees with std. dev 6.0 degrees; the stepwise ensemble IQR is [3.2, 7.2, 12.3] degrees with std. dev of 7.1 degrees.



Supplementary Figure 4.S16: Distributions of the C4–C5–C6–CX angle between the reactive pathways for $t = -20$ fs. (A) Distribution comparison for the simultaneous and stepwise pathways. The simultaneous distribution IQR (in order of the 25%, 50%, and 75% quantile) is [7.5, 13.3, 18.8] degrees with std. dev 7.6 degrees and the stepwise distribution IQR is [3.3, 6.6, 12.0] degrees with std. dev 8.2 degrees. (B) Distribution comparison stratified on the WT, S127A, and V155D simultaneous ensembles versus the stepwise ensemble. The WT IQR is [15, 19.5, 24.2] degrees with std. dev 7.0 degrees; the S127A IQR is [9.1, 14.0, 18.7] degrees with std. dev. 6.6 degrees; the V155D IQR is [3.5, 7.2, 11.9] degrees; the stepwise distribution IQR is [3.3, 6.6, 12.0] degrees with std. dev 8.2 degrees



Supplementary Figure 4.S17: PyMOL structure of OMPDC active–site, including residues K42, D70, K72, D75, S127, V155, Q185, R203.

Chapter 5:

General conclusions and future outlook

Natasha Seelam[1,2]

1. Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge MA

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

5.1 Thesis Overview

This thesis investigated the atomistic drivers of reactivity using novel methods and applications of computational tools. Enzymes are capable of performing difficult reactions at ambient conditions; thus, harnessing their exquisite specificity and selectivity would unlock massive potential for customized chemical reactions [1–3]. As illustrated in earlier chapters, enzymes leverage a variety of catalytic strategies in order to facilitate complex chemistries [4–13]. The theoretical approaches and methods applied in this work focused on harnessing the richer atomic details required for a more complete picture of enzyme catalysis.

Chapter 2 presented a theoretical study of the experimentally characterized enzyme orotidine 5′-monophosphate decarboxylase (OMPDC) and two catalytically impaired mutants [14]. This enzyme performs a remarkable 10^{17}–fold rate enhancement of the decarboxylation of its substrate, without the use of co–factors to assist its catalytic proficiency [15]. A combination of potential-of-mean-force (PMF) and transition path sampling (TPS) methods were used to characterize the energetic landscape of the decarboxylation and to collect full dynamic trajectories of the reactions, respectively; both methods provided relative rate constant estimates that matched empirical evidence [14, 16, 17]. PMFs of the wildtype (WT) and mutant enzymes revealed differences in the energetic landscape of reactivity that suggested two potential catalytic strategies toward decarboxylation: a simultaneous strategy where decarboxylation is coordinated with a shortening distance between a neighboring residue, K72, and a stepwise strategy that is independent of K72's position. Dynamic paths, constructed from path sampling ensembles, revealed that the WT and S127A mutant preferentially decarboxylated using the simultaneous pathway, but the V155D mutant decarboxylated both simultaneously and in a stepwise manner.

Our work demonstrated the rich detail path sampling ensembles can offer in studies of kinetic processes.

The work of Chapter 2 primarily compared the WT to two mutants, S127A and V155D. These mutants hinder the catalytic strategies employed by OMPDC (transition-state stabilization "TSS" and ground-state destabilization "GSD", respectively) [4–7, 14]. Extending this work to include a broader panel of empirically-characterized mutants would offer refined details into the reactive pathways OMPDC leverages to decarboxylate the substrate. Several studies have considered single and/or double mutants to probe the effect of residues D20, K42, D70, K72, S127, Q185, R203, and hydrophobic–pocket residues I96, L123, and V155 [14, 26–31]. A larger panel of mutants could underscore whether certain mechanisms are preferentially adopted when different catalytic mechanisms are affected, or if there exists more pathways than currently considered.

In the first part of Chapter 3, quantum mechanical techniques explored the electronic underpinnings of the methyl transfer reaction performed by ketol-acid reductoisomerase (KARI). The study investigates a spinach variant that facilitates isomerization with two magnesium ion co–factors and NADPH [18–20]. Natural Bonding Orbital analyses of two ensembles of simulations, one that successfully catalyzes methyl transfer and another that failed to cross the reaction barrier, revealed a three-center-two-electron bond (3C) transition–state, and a catalytic strategy where carbonyl formation and lone-pair formation with the adjacent oxygens on the substrate occurred simultaneously.

In the second part of Chapter 3, machine learning methods showed that a subset of 10 geometric or 6 electronic features predict reactivity with high performance. Dynamic analyses revealed that the geometric feature, E319/OE1–C5 distance, and electronic feature, the C4–C5 bond index, were associated with the torsional orientation of the methyl prior to exiting the reactant

basin. Reactive simulations' methyl groups were more likely to be eclipsed than in the non-reactive simulations; this eclipsed orientation was shown to be correlated with both destabilized C4–C5 breaking bond orbital energies and stabilized 3C bond orbital energies during the course of the reaction. Similarly, reactive simulations with eclipsed orientations had larger E319–methyl distance and smaller C4–C5 bond index populations. Lastly, we extended this work to identify subsets of the top 10 geometric features in predicting the 6 model-selected electronic features, as the electronic features spanned different components of the characterized mechanism. We found that small subsets of geometric features were capable of reporting on the electronic features with similar predictive performance to the entire cumulative geometric classifier.

A limitation of this work is the narrow regime of time used to predict reactivity; while these features were identified in the reactant basin, identifying conformations over longer periods of time that improve reactivity can provide insight into the potential design strategies. Possible extensions to investigate this hypothesis include testing timepoints further back in time to see how feature selection could diverge as the system crosses the barrier. An additional option, leveraging the power of path sampling, is to condition the accepted ensemble to only accept simulations that are predicted as 'reactive' for some fraction of the time the system spent in the reactant basin. This approach modifies prior work by conditioning on multiple time points as opposed to one, which alone was shown to remarkably improve the rate [25]. Alternative modeling approaches could also consider features that are temporally linked; viable models with these properties include Hidden Markov Models (HMMs) or Recurrent Neural Networks (RNNs) [22–24].

While few features may be required to identify reactivity, it does not exclude the possibility that multiple catalytic strategies exist. Prior work has identified that even within the same set of features, clusters may exist that employ these features differentially [25]. A follow–up study could

consider what minimal subset of conformational changes to the non-reactive ensemble could permit successful barrier crossing, demonstrating how subtleties in dynamics may encourage reactivity. Given there are many non-reactive trajectories that cross relatively far up the barrier, an appropriate perturbation (e.g., if possible, changing the orientation of the methyl prior to reacting) on high–energy states that is then integrated forwards and/or backwards in time could reveal how dynamic catalytic strategies influence reactivity.

Finally, in Chapter 4, we synthesized the methodologies and findings of the prior chapters to study geometric descriptors that were indicative of the reactive pathways of OMPDC. Across three protein systems (WT, S127A, and V155D), we posed a classification problem, leveraging machine learning to select up to 5 predictive pairs of features that distinguished between the pathways from times starting in the reactant basin until 30 fs after the reaction had proceeded. Model-selected features, despite no prior knowledge of chemical mechanism, were able to identify several sets of catalytically related geometries involving residues in direct contact with the orotidine 5′-monophosphate (OMP) substrate. In particular, two features signaled a weakened hydrogen bond between D75* and the 2'-hydroxyl of OMP, and diminished repulsion by extended distances between the carboxylate of D70, a ground-state destabilizing residue, and OMP's carboxylate group. Both structural features were linked to distortions in the planarity of the orotidyl ring, a feature not explicitly provided to the models, and whose distortions are thought to be beneficial toward reactivity [32–34].

An interesting study would contrast how reactivity differs between these protein systems and whether overlap between the features that predict pathway also influence the ability to cross the reaction barrier. New ensembles of "non-reactive" simulations that attempt to catalyze decarboxylation but fail to do so can be constructed by path sampling methods for each protein

system. These ensembles can be then used to compare the feature distributions that are most relevant to catalysis, versus those that distinguish between decarboxylation pathways. It may also be possible to train classifiers on the wild type (WT) and employ them on the mutant systems; as the WT is the most catalytically proficient enzyme across the three proteins, predictive performance of the reactive conformations on mutants may reveal whether the mutants use the same conformations as WT, just less effectively, or if there are different ones altogether [14].

A noble goal of theoretical work is to empower and offer insight for experimental design. Given the increasing progress in interpretability within machine learning, and its powerful capacity to learn subtle correlates within data, dissecting the nuances within path sampling ensembles will become more tractable and can be used as a generalized tool toward enzyme catalysis [35–39]. With this larger perspective into the details of enzyme catalysis, we can probe refined details of mechanism, and construct models that identify variants that perform desired reactions.

5.2 References

1. X. Zhang, K.N. Houk. Why Enzymes are Proficient Catalysts: Beyond the Pauling Paradigm. *Acc. Chem. Res.* 38, 5, 379–385, 2005.
2. J.M. Choi, S.S. Han, H.S. Kim. Industrial applications of enzyme biocatalysis: current status and future aspect. *Biotechnol Adv.* 33:1443–1454, 2015.
3. D.M. Quinn, R.S. Sikorski. Enzymatic Rate Enhancements. In: *eLS John Wiley & Sons, Ltd*: Chicester, 2014. DOI: 10.1002/9780470015902.a0000717.pub3
4. W.P. Jencks. Catalysis in Chemistry and Enzymology, 2nd ed., *Dover*, New York, 1987.
5. Warshel, A.; Sharma, P. K.; Kato, M.; Xiang, Y.; Liu, H.; Olsson, M. H. M. Electrostatic Basis for Enzyme Catalysis. *Chem. Rev.* 2006, 106, 3210−3235.
6. Warshel. Electrostatic Origin of the Catalytic Power of Enzymes and the Role of Preorganized Active Sites. *J. Biol. Chem.* 273, 27035−27038, 1998.
7. S.J. Benkovic, S.A. Hammes–Schiffer. Perspective on Enzyme Catalysis. *Science*. 301(5637), 1196–11202, 2003.
8. K. Zinovjev, I. Tunon. Quantifying the limits of transition state theory in enzymatic catalysis. *Proc. Natl. Acad. Sci USA*. 114:12390–12395, 2017.
9. S. Hur, T.C. Bruice. The near attack conformation approach to the study of chorismite to prephenate reaction. *Proc. Natl. Acad. Sci. USA,* 100(21) 12015–12020, 2003.
10. J.R. Knowles. To build an enzyme. *Philos. Trans. R. Soc*., B 332, 115−121, 1991
11. J.R. Knowles. Enzyme catalysis: Not different, just better. *Nature.* 350, 121−124, 1991.
12. G.G. Hammes, S.J. Benkovic, S. Hammes–Schiffer. Flexibility, diversity, and cooperativity: Pillars of enzyme catalysis. *Biochemistry* 50, 10422−10430, 2011.

13. L.D. Andrews, T.D. Fenn, D. Herschlag. Ground–State Destabilization by Anionic Nucleophiles Contributes to the Activity of Phosphoryl Transfer Enzymes. *PLOS Biology*, 11:7, 1–18, 2013.
14. V. Iiams, B.J. Desai, A.A Fedorov, E.V. Fedorov, S.C Almo, J.A. Gerlt. Mechanism of the orotidine 5′–monophosphate decarboxylase–catalyzed reaction: Importance of residues in the orotate binding site. *Biochemistry*. 50(39): 8497–8507, 2011.
15. A. Radzicka, R. Wolfenden R. A proficient enzyme. *Science*. 267, 90–93, 1995.
16. G.M Torrie, J.P Valleau. Nonphysical sampling distributions in Monte Carlo free–energy estimation: umbrella sampling. *J. Comput. Phys.,* 23, 187–199, 1977.
17. P.G. Bolhuis, D. Chandler, C. Dellago, P.L. Geissler. Transition Path Sampling: Throwing Ropes Over Rough Mountain Passes, in the Dark. *Annu. Rev. Phys. Chem*. 53: 291–318, 2002.
18. V. Biou, R. Dumas, C. Cohen–Addad, R. Douce, D. Job, E. Pebay–Peyroula. The crystal structure of plant acetohydroxy acid isomeroreductase complexed with NADPH, two magnesium ions and a herbicidal transition state analog determined at 1.65 Å resolution. *EMBO J,* 16(12): 3405–3415, 1997.
19. R. Dumas, M.C. Butikofer, D. Job, R. Douce. Evidence for two catalytically different magnesium– binding sites in acetohydroxy acid isomeroreductase by site–directed mutagenesis. *Biochemistry*, 34, 6026–6036, 1995.
20. R. Tyagi, Y. Lee, L.W. Guddat, R.G. Duggleby. Probing the mechanism of the bifunctional enzyme ketol–acid reductoisomerase by site–directed mutagenesis of the active site. *FEBS Journal*, 272, 593–602, 2005.
21. C. R. Landis, F. Weinhold. "The NBO View of Chemical Bonding", in, G. Frenking and S. Shaik (eds.), The Chemical Bond: Fundamental Aspects of Chemical Bonding, *Wiley*, pp. 91–120, 2014.

22. Cappe, O., Moulines, E., Ryden, T., "Inference in Hidden Markov Models" *Springer Series in Statistics.* 2005; DOI: 10.1007/0–387–2898–8

23. Gers, F., Schraudolph, N., Schmidhuber, J., "Learning precise timing with LSTM recurrent networks" (2002). *Journal of Machine Learning Research*. 3, 115–143.

24. Salaun, A., Petetin, Y., Desbouvries, F., "Comparing the Modeling Powers of RNN and HMM" (2019) *ICMLA*, DOI: 10.1109/ICMLA.2019.00246.

25. B.M. Bonk, J. Weis, B. Tidor. Machine Learning Identifies the Chemical Characteristics That Promote Enzyme Catalysis. *J. Am. Chem. Soc.* 141, 4108–4118, 2019.

26. R.B. Silverman, M.P. Groziak. Model chemistry for a covalent mechanism of action of orotidine 5'–phosphate decarboxylase. *J. Am. Chem. Soc.* 104 (23): 6434–6439, 1982.

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

28. J. Yuan, A.M. Cardenas, H.F. Gilbert, T. Palzkill. Determination of the amino acid sequence requirements for catalysis by the highly proficient orotodine 5′-monophosphate decarboxylase. *Protein Sci*. 20; 1891–1906, 2011.

29. A.A Federov, E.V. Federov, B.M Wood, J.A. Gerlt, S.C Almo. Conformational changes in orotidine 5′-monophosphate decarboxylase: "remote" residues that stabilize the active conformation. *Biochemistry*. 49; 3514–3516, 2010.

30. P. Harris, J.N. Poulsen, K.F. Jensen, S. Larsen. Structural basis for the Catalytic Mechanism of a Proficient Enzyme: Orotidine 5′-monophosphate Decarboxylase. *Biochemistry*. 39, 4217–4224, 2000.

31. K.N. Houk, J.K Lee, D.J. Tantillo, S. Bahmanyar, B.N Hietbrink. Crystal structures of orotidine monophosphate decarboxylase: does the structure reveal the mechanism of nature's most proficient enzyme? *Chem BioChem.* 2, 113–118, 2001.

32. H. Hu, A. Boone, W. Yang. Mechanism of OMP decarboxylation in orotidine 5′–monophosphate decarboxylase. *J. Am. Chem. Soc.,* 130(44), 14493–503, 2008.

33. M. Fujihashi, T. Ishida, S. Kuroda, L.P. Kotra, E.F. Pai, K. Miki. Substrate distortion contributes to the catalysis of orotidine 5′-monophosphate. *J. Am. Chem. Soc.,* 134 (46), 17432–17443, 2013.

34. J. Gao. Catalysis by enzyme conformational change as illustrated by orotidine 5′–monophosphate decarboxylase. *Curr. Opin. Struct. Biol.* 13, 184–192, 2003.

35. L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter and L. Kagal, Explaining Explanations: An Overview of Interpretability of Machine Learning, *2018 IEEE 5th International Conference on Data Science and Advanced Analytics* (DSAA), Turin, Italy, 2018, pp. 80–89

36. Z.C. Lipton. The Mythos of Model Interpretability. 2016 *ICML Workshop on Human Interpretability in Machine Learning* (WHI 2016), New York, NY

37. E.C. Alley, G. Khimulya, S. Biswas, M. AlQuraishi, G.M. Church. Unified rational protein engineering with sequence–based deep representation learning. *Nat Methods*, 16, 1315–1322, 2019.

38. K.K. Yang, Z. Wu, F.H. Arnold. Machine–learning–guided directed evolution for protein engineering. *Nat. Methods*, 16(8):687–694, 2019

39. J. Ingraham, V. Garg, R. Barzilay, T. Jaakola. Generative models for graph–based protein design. *Advances in Neural Information Processing Systems* 32 (NeurIPS 2019).