DISSERTATION

MODELING CONFORMATIONAL HETEROGENEITY IN BIOMOLECULES

Submitted by

Heidi Klem

Department of Chemistry

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Summer 2023

Doctoral Committee:

Advisor: Robert Paton
Co-Advisor: Martin McCullagh

Nancy Levinger
Alan Kennan
Brian Geiss

ABSTRACT

MODELING CONFORMATIONAL HETEROGENEITY IN BIOMOLECULES

Regulation of biocatalytic cascades is essential for biological processes but has yet to be exploited in real-world applications. Allostery is a prime example, where binding of an effector molecule alters function in a remote location of the same biomolecule. V-type allostery is especially fascinating, as the reaction rate can be either increased or decreased in response to effector binding. Determining how conformational changes affect the reaction rate is challenging due to the disparity of timescales between the underlying molecular processes. Experimental methods, such as X-ray crystallography, can help to capture large-scale conformational change. However, the resulting structures are not guaranteed to correspond to the biophysical state relevant to the research questions being addressed. Structural changes that occur during the chemical reaction are particularly elusive to this approach. To understand the connection between conformational change and catalytic consequence, a description of the reaction mechanism and relevant configurations is needed. Quantum mechanical (QM) methods can be used to propose enzyme reaction mechanisms by modeling femtosecond motions of forming and breaking bonds. Large-scale conformational changes take place over much longer timescales that cannot be simulated at the QM level, therefore requiring classical simulation techniques.

This dissertation focuses on the challenges posed by conformational change in the field of computational biocatalysis. The first chapter examines the prevalence of conformational change in enzymes, its relationship to catalysis, and the difficulties it presents. The second chapter looks at the influence of active site structural features on reaction rates in the allosteric enzyme IGPS

using QM approaches and energy decomposition schemes. The third chapter covers the development of methods that use molecular dynamics (MD) simulations to analyze relevant structural states from simulation data and identify long-range communication pathways in biomolecules. The fourth chapter presents a Python code, enzyASM, that automates the generation of QM-based truncated active site models and discusses ongoing developments that will aid reproducibility and standardization in this field of research. The fifth and final chapter summarizes the implications of this Thesis work in computational biocatalysis and envisions how remaining challenges can be addressed to maximize potential to solve real-world problems.

ACKNOWLEDGEMENTS

I am indebted to my PhD advisors Martin McCullagh and Robert Paton for making me a better scientist. Martin, your fierce drive to push projects to the finish line and high expectations, though intimidating at first, motivated me throughout graduate school. Rob, your ability to meld ideas from different fields of science and foresight of research direction made me a more well-rounded researcher. I sincerely thank you both for agreeing to jointly advise me and giving me significant autonomy and independence in the research I pursued. I also want to thank all the members of the McCullagh group, Paton group, and Theory Suite. To my past and present committee members, Brian Geiss, Grzegorz Szamel, Alan Kennan, and Nancy Levinger; thank you for your support and constructive feedback. I especially want to thank Nancy, for joining my committee on short notice, and introducing me to the NRC Fellowship.

I would not have pursued this path without the influence of Maurice Godfrey, Julie Soukup, and Brad Parsons. Maurice, it was because of your encouragement that I sought out research opportunities at Creighton. Julie, you warmly welcomed me into your research group before the start of my freshman year and have been a constant mentor and friend/family to me ever since. Words cannot express how grateful I am to have been a member of your group. Brad, your P. Chem course was the most challenging and rewarding course I have ever taken. Your teaching efforts and confidence in my ability to master the course material are significant factors that led me to pursue a Ph.D. in physical chemistry, despite nearly failing first semester P. Chem.

I want to thank my mom and dad for supporting me throughout my education, and never imposing their aspirations on me. Thank you to my sisters Chrissy and Katie for their ambitious drive that motivates me to keep up with them. I am particularly grateful for Chrissy and her

husband, my brother-in-law, Nathan, who housed me for 3 months during the shelter-in-place period of the Covid-19 Pandemic. I should also thank their dog, Bailey, cat, Oliver, and turtle, Squirt for providing emotional support as I prepared for my preliminary examination; and my rabbit Rorschach, who joined me on this journey my first year of graduate school.

I am grateful for my surrogate grandparents (and neighbors), Mike and Connie Scasny and Tom and Marsha Hernon. A special thanks to Mike and Tom for being my best friends in what was often a lonely childhood. Whether they know it or not, they contributed significantly to my insatiable work ethic by always letting me do yard work for money and assembling a makeshift headlamp out of a hat, flashlight, and duct tape so I could stay out late on the back deck working on homework in the company of their antics; times I will cherish forever.

During graduate school I joined a bowling league for fun, however, this hobby quickly turned into a competitive passion of mine. I want to thank the CSU Monday League for being my escape from Ph.D. stressors and putting up with me when I was bowling bad. Special thanks go to my league team, Stranger Danger (season champions), and 2023 USBC State Open Tournament team, Two Blondes and a Six Pack (2nd place in current unofficial standings), with Stacey and Isaiah Rairdon, Brandon Scherlagt, Melinda Myers and Brad Gayler.

Finally, I thank all of my close friends who have provided so much laughter and joy over the years: Hanna Ritland; Krystal Hopkins; Rae and Jake Anderson; Ryan Weber; Curtis Fochte; Colin Richards; Noah Yoshida (I'm sorry I missed your wedding); Brianna Callahan; Louis de Lescure; Guilian Luchini; Raul Perez-Soto; and especially Trey Pittman who always reminds me of my strengths in times of weakness; even though I know you would insist against it, I would not have accomplished all I have this last year with you.

TABLE OF CONTENTS

CHAPTER 1. INTRODUCTION[§]

1.1 Background

On evolutionary timescales, biological systems have developed enzymes to harness thermochemical processes in physiological conditions. Enzyme catalytic proficiency has motivated groundbreaking experimental and theoretical research to determine underlying design principles with atomic detail. The qualitative understanding of enzymes has expanded drastically from Emil Fischer's 1894 seminal lock-and-key hypothesis that the shape complementary between a rigid active site (lock) and its native substrate (key) dictates molecular interactions and, consequently, activity.[1] In 1958, Dan Koshland modified this model of enzyme-substrate mechanics by accounting for enzyme flexibility in his induced fit model.[2] The scientific community was skeptical of the induced fit model at first.[3] However, concomitant developments in macromolecular crystallography led to evidence of Koshland's hypothesis by demonstrating conformational differences in substrate unbound (apo) and bound (holo) structures.[4] Our understanding of the importance conformation has on biomolecular function has only grown from there.

Protein motions governed by thermodynamical principles invoke the existence of energetically accessible, interconverting conformations, termed the conformational ensemble. Contemporary models of biological mechanisms elaborate on this intrinsic dynamism of proteins to incorporate critical contributions of the functionally relevant substates sampled in the ensemble.[5–8] This framework has catapulted the field from a structure-function to an ensemble-

[§]This chapter is composed of selected content and reproduced with permission from Klem, H., McCullagh, M., & Paton, R. S. Modeling Catalysis in Allosteric Enzymes: Capturing Conformational Consequences. *Topics in Catalysis* (2022) **65**:165–186.

function paradigm, and motivated techniques to probe the conformational ensemble such as room-temperature (RT) and multi-state X-ray crystallography,[9–11] nuclear magnetic resonance (NMR),[12–15] and cryogenic electron microscopy (cryoEM).[16–19] Despite tremendous experimental advances, determining the relationship between the protein ensemble and function remains a challenge; one that is best tackled by combining experimental and theoretical approaches.

Three classes of potentials are commonly employed in theoretical/computational studies of enzymes: quantum mechanics (QM) only, hybrid QM and molecular mechanics (MM), and fully MM calculations. QM potentials are typically used to study stationary points on the potential energy surface, while computationally less-expensive MM potentials can be used to explore enzyme structure and motions across different timescales, such as in molecular dynamics (MD) simulations. QM-only approaches (Section 1.4.1), notably the cluster approach, are now routinely used to study reactions using a reduced active site model consisting of up to 200-300 atoms. The protein environment is accounted for by a continuum dielectric model, assuming the surrounding can be approximated as a homogeneous polarizable medium with a constant dielectric.[20] High levels of quantitative accuracy have been demonstrated with reduced active site models: for example, Himo and co-workers have successfully modeled competing enzymatic pathways leading to enantiomeric products (i.e., asymmetric biocatalysis), where energy differences on the order of 1-2 kcal/mol must be captured.[21–23] Cluster models have similarly been used to study how enzyme active sites control regioselectivity, for example, in facilitating intrinsically disfavored epoxide-opening pathways.[24] A related QM-only method, the theozyme approach, models a theoretical enzyme active site as a selection of functional groups directly involved in catalysis.[25] Houk and co-workers have used theozyme models to illustrate how side-chain motions are

minimized in the multistep serine esterase catalytic cycle[26] and in the computational design of enzymes for abiological reactions.[27]

To study enzymes where long-range interactions (e.g., electrostatic interactions in particular) play a fundamental role in the catalytic mechanism, the substrate and active site residues directly involved in the reaction can be described with QM, and the remaining protein and solvent can be modeled with classical (MM) force fields (Section 1.4.3). This QM/MM hybrid approach has yielded high quantitative accuracy of enzymatic reaction barriers. For example, Mulholland, Thiel, Werner, and co-workers have shown that systematic improvements to the QM level of theory, using LCCSD(T0),[28] provide near-quantitative results for the activation enthalpies and free energies of the reactions catalyzed by chorismate mutase and *para*-hydroxybenzoate hydroxylase.[29] MD simulations employing a QM/MM potential are more expensive than with classical force fields, although sampling times of nanoseconds are now attainable. While such timescales are insufficient to explore large and slow enzyme conformational changes, these may be addressed with classical simulations, as discussed in Section 1.5. As with the selection of a QM cluster model, the choice of QM/MM boundary and description of MM partial point charges at the boundary validation of the choices is often necessary to support conclusions.[30]

Classical MD (MD applied to classical force fields) can be used to study whole enzyme mechanics on timescales ranging from the nanosecond-microseconds regime. There may be many thermally accessible conformations sampled by a protein, one or more of which that are essential to the catalytic mechanism. MD simulation provides one of the few ways to identify and quantify their involvement in atomistic detail. The ability of proteins to redistribute conformational populations to influence function in response to perturbations is a leading hypothesis in the fields of structural allostery[31–33] and, more recently, in directed evolution.[34–36] Classical MD is well

suited for sampling "local" conformational diversity in enzymes such as different torsional angles adopted by amino acid side- and main-chains. However, some conformational changes necessary to achieve a catalytically competent active site, such as dimer interface, loop and helix motions, occur on the millisecond timescale.[37,38]

A single enzyme conformation taken from an X-ray crystal structure might be an excellent starting point for the computational study of enzyme catalysis. However, it is vital to consider the assumptions being made. Mainly, that the single conformation is catalytically relevant, and no other conformations are important for catalysis.[39] The validity of these assumptions is difficult to test unless *apriori* knowledge or hypotheses regarding the catalytic mechanism exist. In Section 1.3, I discuss myriad enzymatic reactions now known for which these simplifying assumptions do not hold, and in CHAPTER 2 I present an investigation of the influence active site conformational change has on elementary reaction steps. These systems illustrate the necessity for computational biocatalysis to incorporate rigorous analyses of conformational ensembles alongside high-accuracy potentials to study reaction mechanisms. Improvements in conformational sampling techniques, accurate force field parameters, and quantum mechanical treatments have accelerated progress towards this goal. The explicit consideration of multiple enzyme conformations to support mechanistic conclusions and calculate barrier heights has now been used in multiple studies.[38,40–43]

1.2 Enzyme Catalysis: The Reaction Coordinate and Protein Dynamics

A fundamental understanding of enzyme mechanics has long been the goal of many chemical and biological scientists. Accomplishing this goal requires answering *how* the enzyme performs its function. To begin formulating an answer to this question, a simple enzyme mechanism can be considered:

$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} ES \overset{k_2}{\rightarrow} E + P, \tag{1.1}$$

where *E* indicates the enzyme, *S* the free substrate, *ES* the enzyme-substrate complex, and *P* the product after its release from the enzyme. In the steady-state approximation, where the substrate concentration is saturating and therefore negligible, there are two kinetic parameters used to describe enzyme performance: the maximum rate of product formation, k_{cat}, and the Michaelis-Menten constant, K_m. From Equation 1.1, the steady-state kinetic parameters are defined as $k_{cat} = k_2$ and $K_m = (k_2 + k_{-1})/k_1$.

However, a more realistic enzyme model to that presented above accounts for additional elementary mechanistic steps such as the following:

$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} ES \overset{k_2}{\rightarrow} EX \overset{k_3}{\rightarrow} E + P, \tag{1.2}$$

where *EX* is an intermediate state distinguished by a unique chemical species (i.e., involving a change in bonding relative to *ES*) or a kinetically significant conformation distinct from *ES*. With the expanded mechanistic scheme in Equation 1.2, the steady-state parameters are defined as $k_{cat} = k_2k_3/(k_2 + k_3)$ and $K_m = (k_2 + k_{-1})k_3/(k_2 + k_3)k_1$. Although the expressions for k_{cat} and K_m differ between Equations 1.1 and 1.2, laboratory observations of steady-state kinetics cannot distinguish between these two mechanistic scenarios. If the goal is to determine an enzyme's substrate specificity, then simplification to Equation 1.1 is sufficient. However, computational analysis of individual reaction steps (including all relevant reactants, intermediates, and products) provides the basis to understand the enzyme's mechanism and the atomistic factors influencing rate and selectivity. Importantly, a measured k_{cat} often reflects several elementary rate constants in the overall mechanism. Although, for example, in Equation 1.2, if $k_2 >> k_3$,

simplification to k_{cat}= k_2 is justified. It is important to note that k_{cat} is not necessarily solely defined by "chemical" steps, such as with product inhibition mechanisms.

One critical enzyme attribute missing from Equations 1.1 and 1.2 is conformational heterogeneity. Much of the focus of enzyme catalysis in the 21st century has been on the role of conformational changes involving both experimental and computational expertise.[5,8,37,44,45] The free energy landscape, a multi-dimensional construction of the intermediate and transition states available to an enzyme before, during, and after catalysis, is particularly useful to conceptualize the relationship between the conformational ensemble and the chemical reaction coordinate (Figure 1.1).[46]

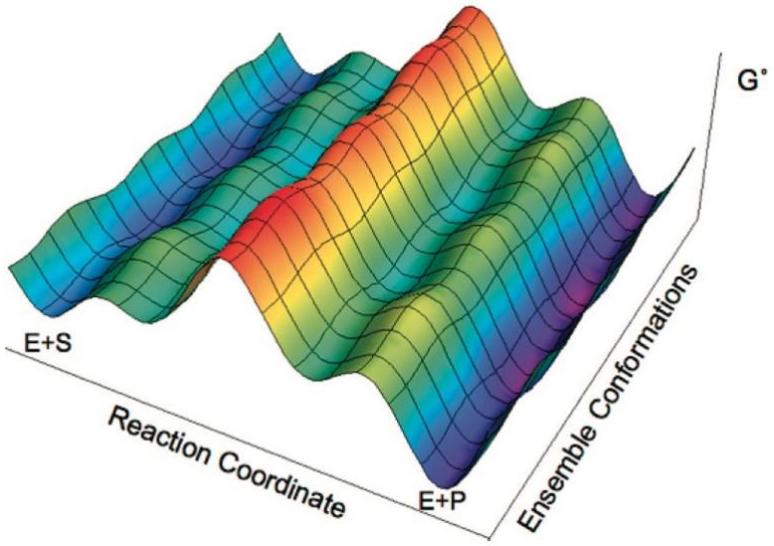


Figure 1.1 Enzyme free energy landscape projected onto the reaction and conformational coordinates. Adapted with permission from.[46] Copyright 2021 American Chemical Society.

Equations 1.1 and 1.2 do not account for the possibility of interconverting conformers (e.g., of the free enzyme, *E*) that may possess different reactivities along parallel reaction pathways, as described by Hammes-Schiffer and co-workers.[46] Consequently, quantitative agreement of computed reaction barriers with experiment may require explicit consideration of the enzyme's conformational ensemble. In contrast to the weak coupling of the reaction and conformational

coordinates depicted in Figure 1.1, strong coupling can arise where even qualitative agreement with experiment requires consideration of the ensemble.[37,38] Construction of the free energy landscape for large systems is a laborious, if not prohibitive, task. Such studies require computational methods in classical simulation techniques, while the breaking and forming of bonds requires *ab initio* (i.e., QM) calculations. As a result, a combination of distinct computational methodologies and expertise is often required.

1.2.1 The enzymatic reaction coordinate

Catalytic mechanisms are often illustrated by (Gibbs) energy profiles, characterized by relative free energies of intermediate and transition state (TS) structures that culminate in the transformation of reactant into product within the enzyme complex (ES and EP, respectively), as illustrated in Figure 1.2. Multiple factors can influence the quality of the Gibbs energy profile and its mechanistic interpretation.[47] While computations can be used to provide evidence in favor of or against a particular reaction pathway, a reaction mechanism can never be conclusively proven, only experimentally corroborated.[48] With much success, contemporary QM approaches have been used to describe chemical reactions by calculating observables, such as energy barriers, kinetic isotope effects, and product selectivities that can be validated with experiments.[49]

In comparing the computed catalytic (Gibbs) energy with experiment, the concepts and language introduced by Kozuch and Shaik are both illuminating and influential.[51] In contrast to using rate constants to define a catalytic cycle (the *k*-representation), computational studies generate state energies (the *E*-representation). The apparent activation barrier is then represented by the energetic span of the catalytic cycle, defined by the difference ($\Delta G^{\ddagger}$) between the lowest energy, turnover determining intermediate (TDI), and the highest energy, turnover determining TS

(TDTS) (Figure 1.2). The value of $\Delta G^{\ddagger}$ can then be used to calculate a theoretical rate constant based on Eyring's transition state theory equation:

$$k_{cat} = \frac{\kappa k_B T}{h} exp\left(\frac{-\Delta G^{\ddagger}}{RT}\right), \tag{1.3}$$

where κ is the transmission coefficient, k_B is Boltzmann's constant, T is temperature, h is Planck's constant, and R is the gas constant. It is common to set κ =1, but more advanced techniques have been developed to evaluate non-equilibrium effects and the contributions of recrossing and tunneling to enzymatic rate constants.[52,53]

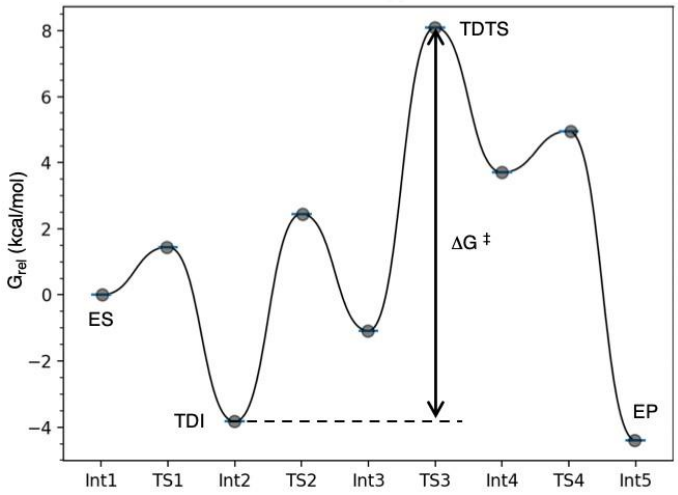


Figure 1.2 Example reaction coordinate, created with GoodVibes,[50] for a multistep enzymatic chemical reaction showing several minima and transition structures proceeding from enzyme-substrate (ES) to enzyme-product complex (EP). Turnover determining intermediate (TDI), Int2, and transition state (TDTS), TS3, define the energetic span and apparent activation energy of the cycle, $\Delta G^{\ddagger}$.

In the case of enzyme models, computed activation barriers for catalytic cycles have reached impressive levels of accuracy. For example, by using relatively large QM regions optimized at the DFT (B3LYP-D3(BJ)/TZVP) level of theory and performing DLPNO-CCSD(T) single-point energy calculations, Neese and co-workers have demonstrated accuracies within 1 kcal/mol of experimental enzyme-catalyzed barriers. Even without high-level coupled-cluster corrections, B3LYP-D3 provides qualitatively correct results.[54] As cautioned by Kozuch and

Shaik, care must be taken when calculating $\Delta G^{\ddagger}$ to be used in Equation 1.3. In some situations, the internal energy can be a fair approximation for this value, but in other cases either additional methods or thermal and entropic corrections should be made.

Computational studies of enzyme catalysis are routinely performed using 1) QM-only approaches, which include cluster models[55] and theozyme[25] approaches, and 2) mixed QM/MM studies. These approaches are surveyed in **Section 1.4**. More detailed technical summaries of these approaches can be found elsewhere, e.g., in reference.[39] Herein, I focus on basic concepts to emphasize what information can be gained and how QM approaches can supplement a multiscale computational integration to model enzyme catalysis, particularly in the context of a conformational ensemble.

1.2.2 The conformational coordinate

Introductions to biochemistry emphasize that structure leads to function. However, this is an oversimplification, and there are many macromolecules where the interconversion between different structures influences overall biological function. Proteins are known to populate multiple metastable structures (the conformational ensemble) under typical physiological conditions.[56] Mounting evidence suggests that the ensemble nature of proteins is intricately tied to their function.[57,58] From a statistical mechanics perspective, it is the probability and properties of the microscopic structural states that dictate the macroscopic properties of the protein. Consequently, the determination of the conformational ensemble, specifically the distribution of microscopic structural states, of enzymes is of utmost importance. One of the challenges in this field is the variety in timescales associated with motions, which ultimately dictates the most appropriate method used to study the underlying dynamics (Figure 1.3).

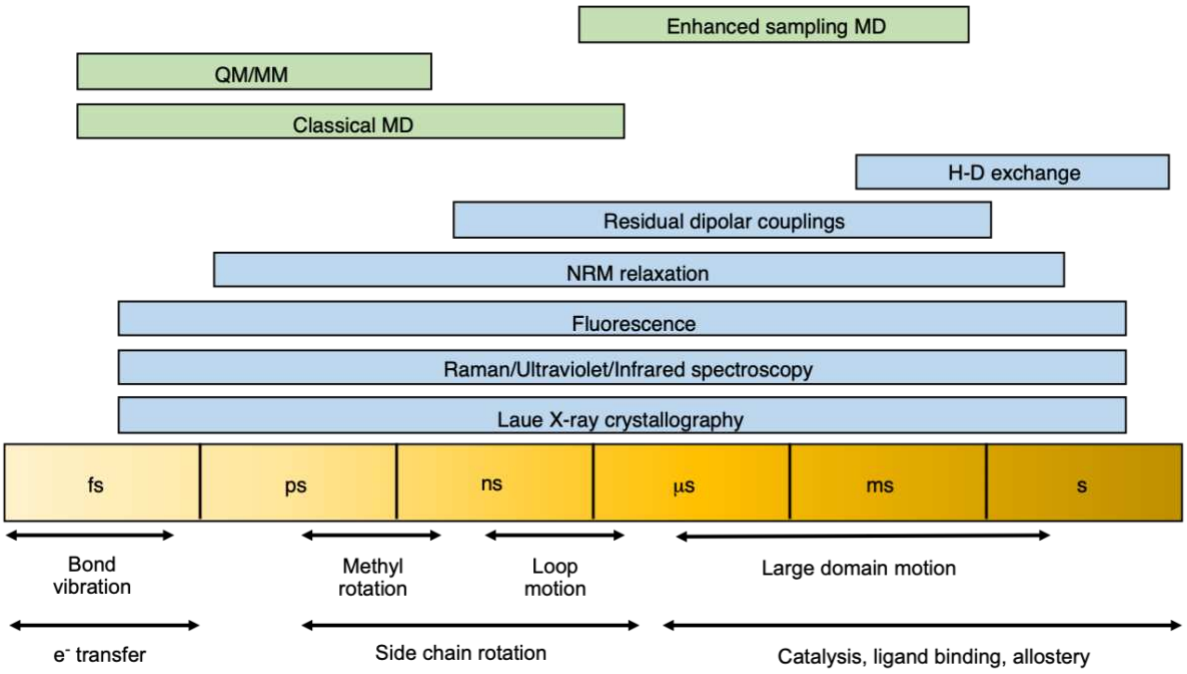


Figure 1.3 Methods for studying dynamic processes in proteins ranging from femtosecond to second timescales. Experimental and computational techniques are shown in blue and green, respectively.

Experimental techniques used to probe protein ensembles include X-ray crystallography, nuclear magnetic resonance (NMR),[58] small-angle X-ray scattering (SAXS), atomic force microscopy (AFM),[59–61] and more recently, cryogenic electron microscopy (cryoEM).[16–19] Each of these methods has its own set of advantages and disadvantages. X-ray crystallography, for example, typically determines a single structure that represents a minimum energy structure under crystallization conditions. Recent advances in the field include the advent of room temperature crystallography[9] and recognition that a given crystal likely has multiple structures in the unit cell.[10,62] Protein NMR is inherently an ensemble measurement that can be done under solution conditions. However, the measurement timescale of NMR dictates that the resulting values, for example, chemical shifts[12] and dipolar couplings,[13] are ensemble averages over multiple metastable states. NMR has successfully been coupled to SAXS,[14,15] spectroscopic techniques, and molecular simulations[63–65] to tie the average values to the microscopic ensemble. AFM can provide

structural trajectories of proteins but does not have atomic resolution. CryoEM stands out as one of the most promising approaches, with recent advances in image capturing hardware and software driving the resolution down to the atomic scale.[16–19] CryoEM is, however, devoid of temporal information.

Molecular modeling and *in silico* simulations can provide atomic-level protein ensemble data to complement experiment. A robust approach is to use all-atom force fields to model the protein and the solvent environment. MD simulations provide time-dependent trajectories of the system of interest. In the theoretical limit of infinite sampling and accurate force fields, these data would represent a complete atom-level picture of the protein ensemble. Practically, improvements to both force fields and sampling protocols continue to improve agreement between simulation and experiment, yet room for improvement still exists.[66] Other than reparameterization of standard functional forms, new directions in force field development include polarizable[67] and machine-learned force fields.[68] Regardless of the specific force field details, sufficient ensemble sampling is also an ongoing concern in the field.

Typical timescales for conventional molecular-dynamics (cMD) on medium-sized proteins are in the tens of microseconds range. While this can adequately sample the conformational ensemble of small globular proteins near their native state, it is insufficient for large proteins and processes such as protein folding. Techniques to overcome this shortcoming include the development of specialized hardware[69] or enhanced sampling techniques. The latter is a more approachable solution and includes techniques such as replica exchange,[70,71] metadynamics,[72] and adaptive sampling.[73–75] A recent study of the conformational ensemble of an intrinsically disordered protein demonstrated that cMD was adequate to reproduce the NMR chemical shifts

(an ensemble property) but not the SAXS data (influenced by the distribution of molecular sizes). Enhanced sampling was shown to improve the agreement with SAXS data.[76]

Characterizing the conformational ensemble is essential to fully understand the relationship between enzyme structure, dynamics, and the catalytic mechanism. Recently, consideration of the conformational ensemble has been pivotal in determining how the sampling of distinct conformations influences different catalytic properties in the directed evolution of various enzymes,[34,35,77–79] and a direct correlation between reaction rate and active conformation population influenced by different allosteric ligands and enzyme mutations has been quantified.[38]

1.3 Allosteric Regulation

Allosteric enzymes exemplify the importance of conformational ensembles.[2] Allostery is a biological phenomenon that occurs when a perturbation at a site distant from the primary active site modulates a protein's function. As the simplest case, we can consider a two-state model (Figure 1.4), characterized by an active and inactive conformation. Although a variety of perturbations may influence the free energy landscape, such as mutations in directed evolution, in the context of allostery we consider the binding of a small molecule distal from the enzyme's primary active site as the perturbation. Upon binding the allosteric ligand, the energy landscape can reshape, and the resulting populations of inactive and active conformations are altered. This conformational shift phenomenon is often used in discussions of structural allostery but is not limited to such enzymes.

The long-ranged coupling of sites in an allosteric system can be explained through short-range interactions linking the distant sites.[80] This mode of regulation has been referred to as the second secret of life, behind the central dogma that describes information transfer between DNA, RNA, and proteins.[81] Biological systems harness allostery to respond to changes in their environment. This is recognized in various biological processes such as signal transduction,[82,83]

transcriptional regulation[84,85] and metabolism.[86] It has been proposed that any system can be allosterically regulated; it becomes a matter of how to probe the interactions that couple binding sites.[83] In 2011, Huang and co-workers created the Allosteric Database (ASD) to provide a comprehensive collection of allosteric data.[87] Presently, the ASD contains 1,949 allosteric proteins and 82,070 allosteric modulators.[88,89]

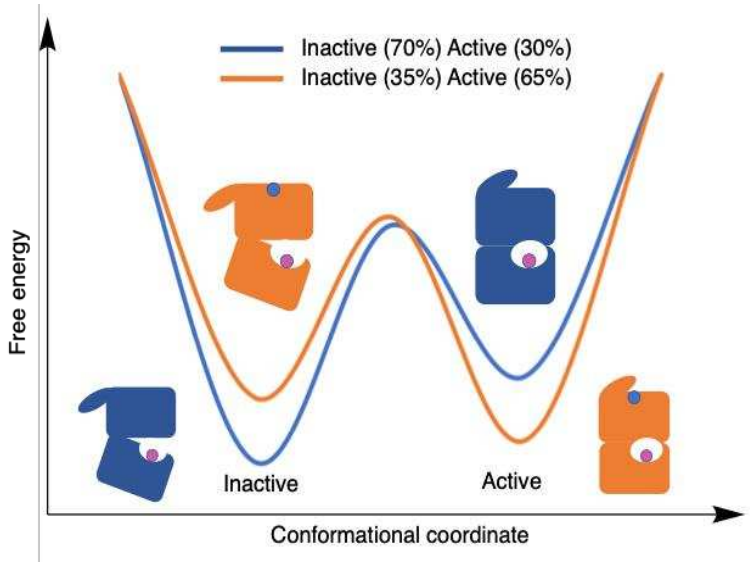


Figure 1.4 A free energy profile projected onto a two-state conformational coordinate. The blue line and toy enzyme model illustrate a system without the allosteric effector bound. Orange corresponds to a system with the allosteric effector bound. The legend shows relative populations of the active and inactive states that result from different ligand states.

1.3.1 Categorizing allostery

In allosteric regulation, the binding of an effector molecule alters an enzyme's activity towards its natural substrate. This is illustrated by the thermodynamic cycle shown in Figure 1.5. In the absence of the allosteric effector, *X*, the basal substrate-binding dissociation/affinity constant K_{ia} and rate constant k_{cat} are observed. Enzyme allostery is classified by the process of activity alteration. In K-type allostery, the allosteric response to effector binding is a change in the affinity, K_{ia}, for the substrate, *A*. This system is the most commonly studied, and an effective allosteric coupling constant, defined by the ratio of substrate binding affinity in the absence versus presence

of the effector $K_{ix}/K_{ix/a}$, or equivalently $K_{ia}/K_{ia/x}$ as defined in Figure 1.5, has been developed to quantify the allosteric effect of a K-type system.[81] There exist experimental and computational techniques to measure the allosteric coupling constant in a K-type system. In V-type allostery, binding of the effector causes a change in the catalytic activity, k_{cat}. An allosteric coupling metric analogous to a K-type system could be based on k_{cat} with and without the effector bound ($k_{cat}/k_{cat/x}$). Although there exist computational methods to measure k_{cat}, as previously mentioned, this is widely unexplored in the context of allostery.

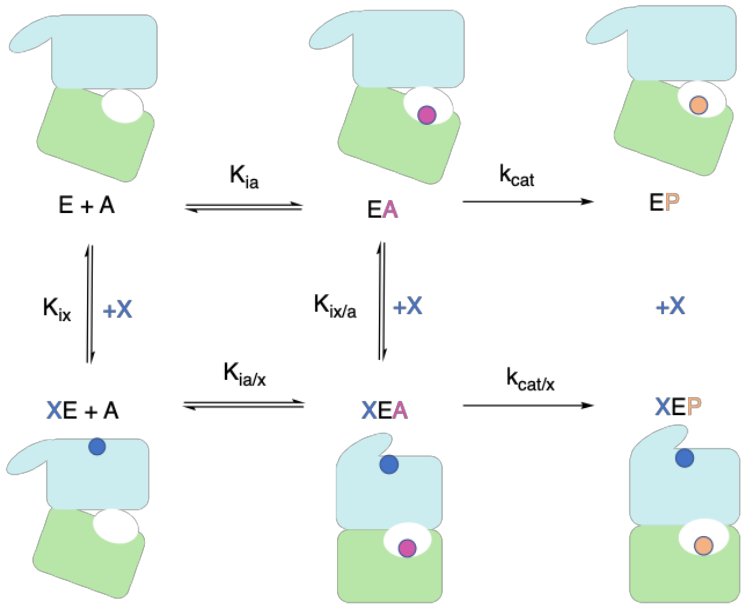


Figure 1.5 Complete thermodynamic cycle of an allosteric enzyme, E. The top row represents the enzyme in the presence of only substrate, A. The bottom row represents the enzyme in the presence of the substrate and allosteric effector, X. Binding of X modifies the activity of E by either altering the binding affinity of E for A, as in K-type allostery, or influencing the rate-determining step, as in V-type allostery. Nomenclature from Ref.[81] is used here for consistency.

1.3.2 The ensemble model of allostery

The so-called concerted Monod, Wyman and Changeux (MWC) model of allostery, established in 1965, states that most allosteric proteins are oligomers involving multiple identical protomers and thus have an axis of symmetry; the quaternary structures of such systems are altered by allosteric interactions; there exist at least two states that differ by the distribution and/or energy

of interactions at the protomer interface(s); these interactions alter the affinity of the binding site towards its corresponding ligand, and the effector is not chemically identical to the substrate.[90] These statements were developed from observations of known allosteric systems at that time, with hemoglobin serving as the hallmark allosteric system, given that an X-ray structure was established for this protein.[91] It was observed that in apo-hemoglobin (absence of effector), there exists an equilibrium between two states, historically known as the constrained and relaxed states. The MWC model claims that effector binding shifts the equilibrium. In 1966, Koshland, Nemethy and Filmer (KNF) composed an alternative model.[92] This model favors a "induced fit" mechanism, whereby the apo-protein binds the effector, which induces a conformational transition to the holo-protein. The MWC and KNF model share a static view of allostery, dependent upon noticeable conformational differences in apo- and holo-structures.

A dynamic model of allostery was proposed in 1984 by Cooper and Dryden, which argues that large-scale conformational changes are not a requisite for allosteric regulation.[93] Instead, changes in thermodynamic fluctuations could mediate the coupling of binding sites. More recently, evidence has shown the importance of functional states unrelated to the so-called "tense" or "relaxed" states that belong to the MWC model.[94] In this more dynamics-driven view of allostery, a protein exists as an ensemble of states, and binding of an effector results in a global redistribution of protein fluctuations and thus alters the relative entropy of the ensembles. This ensemble model of allostery is favored in the literature today.[33,83,86,95,96] Allostery research, which traditionally emphasized static comparisons, is currently faced with the challenge of leveraging the ensemble nature of allostery.[97] Although this challenge is now considered fundamental in the field of allostery, it applies to enzymes in a general way. More recently, enzymes engineered through

directed evolution have achieved higher catalytic efficiency by redistributing the energy landscape.[35]

Restructuring of the energy landscape in response to allosteric ligand binding or a relevant mutation is expected to be the driving force that alters enzyme function, meaning multiple conformations and their populations are crucial to consider. Population shifts have been found to directly influence functional change in enzymes.[38,93] The influence of the enzyme conformational ensemble on the chemical rate in V-type allosteric systems is particularly intriguing in this regard.

Enzyme catalytic activity is typically compared against the background reaction rate in (aqueous) solution to investigate catalytic origins.[98] Alternatively, one could compare the same reaction in enzymes with different activities. Such an approach is adopted to evaluate a V-type allosteric enzyme in Chapter 2. To evaluate the source of an allosteric effect in catalysis, one must consider the reaction with and without the allosteric effect. In many cases, this means considering the enzyme in the presence and absence of the allosteric ligand and comparing the relative energy barriers resulting from the two systems. From a computational perspective, this is an attractive comparison of relative rather than absolute barrier heights. This, in addition to the biological relevance of allostery, makes allosteric systems ideal to explore and evaluate computational approaches for modeling conformationally heterogenous enzymes.

1.4 Quantum Mechanics for Enzyme Catalysis

1.4.1 The cluster approach

The quantum chemical cluster approach explicitly models the critical features of a biologically relevant active site using QM while typically accounting for the remainder of the protein with homogeneous dielectric continuum models.[55] This approach has been pioneered in the study of biologically relevant metal centers by Siegbahn and Blomberg, with model sizes

around 60-70 atoms,[99] and by Himo, who has used cluster models that can be much larger, even surpassing 300 atoms.[20] For more focused reviews on the quantum chemical cluster approach, we urge the interested reader to explore several excellent reviews.[20,55,100] The steps taken in building a quantum chemical cluster model are summarized below.

Model Selection. Once a biological target has been chosen, a deep dive into the literature is worthwhile to explore what is known, unknown, and theorized about the system. Most cluster models are designed from crystal structure coordinates. There may be attributes of the crystal structure that do not align with the system desired to be modeled, such as residue mutations, and alternative substrates bound. In some cases, it may be necessary to manually alter the structure to match the intended system of study, which will require longer MD simulation to allow the structure to properly relax. The primary literature may describe artifacts in the crystallographic model and whether specific residues are essential for catalytic activity (e.g., through the experimental study of enzyme mutants). Depending on the research goal, multiple crystal structures, sometimes with different ligand states, are important to consider. Additionally, the protonation states of some residues, particularly Glu, Asp, and His may be important to consider. If this information is not already available, it may be necessary to consider all the possibilities.[101] If crystallographic waters or ions are present in the active site, they may need to be explicitly included in the model as well.[102] It is typical to test various sizes of clusters, ranging from 100-300 atoms.[20] Convergence studies suggest that QM-cluster models give reliable energetics when the model size is large enough.[55,103–107] Including additional residues beyond those immediately in contact with the substrate may be crucial to avoid unrealistic, extensive conformational reorganization of the active site following geometry optimization,[108] however, it has been proposed that informed residue selection should take priority over a simple distance cutoff.[109]

Model Truncation. The cluster approach requires cuts or truncations to be made since only a subset of the protein's atoms will be included. The most common approach is to residues at the alpha carbon by removing all peptide bond atoms and capping the alpha carbon with hydrogens. The methyl-capping approach is performed when the N-Cα and Cα-CO peptide bonds on either side of the R chain are cut. There may be situations where peptide bond atoms are involved in the reaction, such as forming H-bonds to the substrate. In such cases, residues will be cut at either the N-terminus (N'), N-Cα bond or C-terminus (C'), Cα-CO bond, and the truncated ends will need an additional hydrogen to achieve saturation ("hydrogen-capping"). During this step, important considerations include atoms that influence catalysis to ensure the atom caps are neither artificially influencing the mechanism nor significantly altering the electronic structure, such as changing atomic hybridization or cutting across a highly polar bond. A brief study on the effects of different truncation schemes is presented in CHAPTER 4.

Coordinate Locking. In most cluster models, it is necessary to lock or freeze certain atoms to preserve the active site geometry, maintain sidechain rotamer states along the reaction coordinate, and limit the model from accessing geometries that would not be possible inside the protein environment. As the structure is not fully optimized, evaluating the full QM partition function, including vibrational effects to thermochemistry, is challenging. Therefore, potential energies, rather than Gibbs energies, are often reported. There are methods to approximate entropic effects, such as projecting out the frozen coordinates from the Hessian.[110] The effects of coordinate locking have been explored in Ref.[111], where the authors studied phosphotriesterase using a cluster model of 82 atoms. The authors found locking induced significant strain, altering some geometric parameters. However, these did not influence the conclusions regarding the reaction mechanism and only altered the calculated barrier by 2 kcal/mol. The authors also noted that in this particular

application, the truncation method resulted in a model that was too rigid. For example, a His residue in the cluster was modeled only as an imidazole ring, where one atom from the ring was locked, significantly hindering the motion of that group. However, the strain induced by coordinate locking is expected to subside with larger models.

Freezing the alpha carbons of each truncated residue is a good choice. However, in some models, this might allow too many degrees of freedom resulting in inconsistent side-chain conformations along the pathway. Locking up to two hydrogens of the alpha carbon to restrict more degrees of freedom[20] or not locking residues that directly interact with the substrate[112] may be necessary. The more frozen atoms, the more rigid the model, which may influence energetics. Therefore, the coordinate locking scheme must be balanced, achieved through trial and error, between maintaining a reliable structure and allowing flexibility for energy minimization during geometry optimization.

Model Chemistry. In common with contemporary QM studies of organic and inorganic reactivity, dispersion-corrected density functionals such as B3LYP-D3 or ωB97XD are now commonly used for geometry optimizations of cluster models. In many cases, dispersion effects are expected to influence cluster geometry, such that Grimme's zero-damped (D3) and Becke-Johnson damped (D3(BJ)) corrections are recommended with typical generalized gradient approximation (GGA) or hybrid-GGA density functionals such as B3LYP.[113] While valence double-zeta basis sets are often suitable for geometry optimization, single-point energy corrections with solvent models and larger basis sets are generally employed to account for electrostatic effects and approach more accurate energies. A dielectric constant close to 4 (e.g. diethyl ether, for which $\varepsilon = 4.24$) is expected to mimic the relatively hydrophobic protein interior, although examples in the literature have evaluated cluster model energetics with multiple dielectrics to investigate their

sensitivity to this value.[55,104] Significant changes in relative energies at different dielectric values may indicate that the cluster is too small, and more residues or active site water molecules should be included.

1.4.2 QM/MM Approaches

Standing as an alternative to the truncation schemes discussed above, hybrid QM/MM approaches enable a small portion of the enzyme, typically in the active site region, to be treated by QM, while the remaining larger part of the system is described by MM. Compared to QM cluster models, the QM/MM approach explicitly represents the steric effects (e.g., via mechanical embedding) of an inhomogeneous protein, alongside noncovalent interactions such as long-range electrostatics between the QM and MM subsystems (e.g., via electronic embedding). QM/MM methods are widely-used across organic and organometallic chemistry, and are particularly well-established in the study of enzymatic reaction mechanisms.[114–123]

The explicit inclusion of the extended protein environment can quantitatively impact relative energetics, or indeed, fundamentally alter a computed energy profile. For example, electrostatic effects contribute several additional kcal/mol in QM/MM studies of dioxygen binding energies in JMJD2A.[124] Further, the role of protein reorganization along the reaction coordinate, which can be consequential, may also be captured and quantified.[125] Readers are referred to detailed reviews on the status of QM/MM approaches in modeling enzymatic catalysis.[115,122,126,127] QM/MM models, which typically also include explicit solvent molecules as well as the enzyme and substrate, are much larger than those used in QM-only models. Geometry optimizations become more challenging for large system sizes and so may be performed with (partially) fixed surrounding atoms. Even so, more energy minima are possible with these larger models, and the importance of extensive configurational sampling becomes paramount.

For example, using energy minimization to generate potential QM/MM energy surfaces for an enzymatic reaction while starting from a single protein structure has been shown to produce severe errors in activation barrier heights and binding free energies. Averaging results over several protein configurations generated by long MD simulations has been recommended.[115,128,129] The computational demands of QM/MM simulations can be influenced by choice of the QM region size and by using low-cost semi-empirical QM (SQM) schemes in place of more expensive DFT calculations. In the first respect, many of the chemical and physical considerations relevant to the choice of cluster models stand; the QM region should be evaluated for convergence,[130–133] and attention to the region definition is also required.[134–137] Although, compared to QM-only models, Warshel has demonstrated a relative lack of sensitivity to including more distal groups in the QM region.[138] The use of SQM methods generally requires careful benchmarking or parameterization against more expensive QM results.[139] Nevertheless, typical QM/MM simulations are run for timescales on the order of nanoseconds. Enhanced sampling techniques have been applied to compute the free energy profiles of reactions in QM regions, but, even in these cases, sampling of the rest of the protein is limited to the nanosecond timescale.[140–142] While this may be sufficient to sample side-chain rotamer preferences, protein conformational changes required to reach a catalytically active state (e.g., involving loop and helix motions) routinely take place on timescales several orders of magnitude slower.[143] Classical MD simulations over microsecond timescales, inaccessible while employing QM/MM potentials, have been necessary to observe binding site formation, for example, in bromodomains.[144]

1.4.3 QM Approaches Using Multiple Conformations

In contrast to using an X-ray structure as a starting point for QM studies, a short MD simulation can be used (500-1000 ps) to generate an initial geometry.[145] This will relax the

structure towards a local minimum, potentially reducing crystalline artifacts. However, this approach is unlikely to help when the X-ray structure is sufficiently far from the catalytically relevant conformation(s). QM/MM approaches often incorporate multiple conformations as initial geometries for modeling, usually taken as a series of random or evenly spaced snapshots from an MD simulation.

The question of how to combine the results from several QM or QM/MM models has been addressed by various approaches. An assortment of averaging techniques have been explained elsewhere;[146] however, questions continue to be raised, such as what is required to converge multistate energy data[103] and how to account for conformations with different reactivities.[44,147,148] Alternatively, the Boltzmann ensemble can be evaluated with a combination of extended MD and structural clustering, which is further outlines in Chapter 5. In such an approach, the average structures and populations can be used to generate and weight energy barriers resulting from either the QM-only or QM/MM calculations.

1.5 Structural Clustering with Molecular Dynamics

Structural clustering is a necessary and often overlooked component of describing the conformational ensemble of a protein. To illustrate the need for structural clustering, we consider a set, **X**, of N configurations of a protein system, $\mathbf{X} = \{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_N\}$, where $\vec{x}_i$ is the ith configuration of the system. The expected value of an observable, M, can be computed as an average over this set,

$$\langle M \rangle = \frac{\sum_i^N M(\vec{x}_i) P_i}{\sum_i P_i} \tag{1.4}$$

where P_i denotes the probability of configuration i and is equivalent to $\frac{1}{N}$ in conventional MD sampling. The observable denotes any property of interest that depends on the conformation of the system, such as the electrostatic potential at the active site or the rate constant of a chemical

reaction. If N is small and $M(\vec{x}_i)$ is cheap to compute, Equation 1.4 can be used directly to estimate the property of interest. When N gets large and/or $M(\vec{x}_i)$ is expensive to compute, clustering approaches are employed to approximate Equation 1.4. Clustering is particularly necessary and challenging for MD descriptions of protein ensembles due to the sheer size of the data sets: it is not uncommon to have millions of protein configurations in a trajectory[69] or combined trajectories.[149] Given a clustering of a conformational ensemble into $K < N$ clusters with well-defined average (or mediod) structures, observable M can be estimated as

$$\langle M \rangle \approx \frac{\sum_j^K M(\langle \vec{x} \rangle_j) P_j}{\sum_j P_j} \tag{1.5}$$

where $\langle \vec{x} \rangle_j$ is the average structure and P_j is the probability of cluster j. The accuracy of Equation 1.5 depends on several factors, including the conformational heterogeneity of clusters and the sensitivity of the observable to conformational heterogeneity.

The goal of structural clustering is to combine configurations into clusters that have similar values of the observable of interest. In the context of enzyme catalysis, the active site is an obvious focal point. Therefore, consideration of only active site residues for structural clustering may be sufficient. The results of clustering are a set of K macrostates, made from MD snapshots, or microstates, and their associated populations, $\{P_j\}$. There are numerous clustering methods employed in the field and, unfortunately, no consistent best choice. Most approaches, however, share a common protocol described by the following five steps: (1) choosing a coarse-grained description of the protein, (2) picking features to describe the coarse-grained protein, (3) dimensionality reduction of the features, (4) clustering in reduced dimensions and (5) analysis of the results. While this is a typical order of the steps, it is possible, for example, to cluster first and then perform dimensionality reduction (swapping steps 3 and 4).

Step 1: Choose a coarse-grained description of the protein. Atomistic descriptions of a protein and its environment must be coarse-grained to make the conformational ensemble a tractable object. In a typical aaMD simulation of an averaged-sized protein there are ∼100K total atoms and thus 300,000 degrees of freedom. It is intractable to consider each degree of freedom of the system, even with millions of frames. Luckily, many of these degrees of freedom are of the solvent and thus only implicitly important in determining the protein conformational ensemble. Thus, a natural coarse-graining of the system is to ignore the solvent degrees of freedom and only analyze the atoms of the protein. Even this yields exceptionally high-dimensional data sets that are intractable to analyze. Additional coarse-graining of a protein is almost always performed. Common examples of coarse-grained protein descriptions are backbone atoms, Cα atoms, center-of-mass (COM) of residues, and/or secondary structural elements.

A good choice of the coarse-grained description of the protein depends on the application. For example, if one is interested in protein folding, the coarse-grained description should include all protein residues. Typical coarse-graining for this application includes all backbone atoms, only Cα atoms, or COM of residues. A depiction of this type of coarse-graining is given in the first two panels of 7. The top-left panel depicts the folded state of a small, fast-folding protein (Trp-cage) in a box of water. The protein is coarse-grained in the top-middle panel by only choosing the Cα atoms of each residue. The choice of all Cα atoms is tractable for this small protein but will become intractable for larger proteins. In the case of enzymes will well-defined active sites, it is natural to choose a more detailed description of the active site while ignoring protein residues far away from the active site.

Step 2: Choosing features. With a coarse-grained mapping chosen, one must then choose features or coordinates to describe these coarse-grained structures. A good set of features can

discern between the important metastable conformations. For example, the features in a two-state protein folding problem must differentiate between folded and unfolded states. There are two natural choices for features: internal coordinates and positions. Each has its advantages and disadvantages that will be described below. There are also combinations of internal coordinates or positions called collective variables (CVs, e.g., helical content, RMSD of substructures, etc.) that we will not discuss in detail. Generally, these CVs have similar advantages and disadvantages to the internal coordinates or positions used to compute them. Internal coordinates of a macromolecule include distances (two-body terms), angles (three-body terms), dihedral angles (four-body terms), and higher-order terms. Within the context of all-atom simulations of proteins, most force fields only include up to four body terms, and thus it is uncommon to consider terms higher than four bodies. Internal coordinates have the advantage of being the natural coordinates of the Hamiltonian of the system: they are rotationally invariant and can uniquely describe each structure. The downside of these features is feature space size, and the mixing of features can limit the types of clustering algorithms that can be employed. The choice of all pairwise distances for the Trp-cage example is depicted using blue lines between Cα atoms in the top-right panel of Figure 1.7. With 20 residues (and thus Cα atoms) in the Trp-cage protein, there are 190 pairwise distances to consider. Particle positions are also a natural choice of features to cluster trajectory data of macromolecular systems. Advantages of these features include no significant over-determination of the system, coordinates are directly output by simulation software, and coordinates can differentiate between all configurations observed. A significant limitation of particle positions is that they are defined in the lab frame and thus are not immediately rotationally invariant. This leads to equivalences in particle positions that make them challenging to deal with.[150]

Step 3: Dimensionality reduction. Despite an initial particle coarse-graining (Step 1), the remaining degrees of freedom after featurization are often too large to consider in full-dimensional space. Take, for example, the folding and unfolding of an alpha helix coarse-grained into 12 beads. If we choose all pairwise distances as our features (Step 2), there are 66 degrees of freedom to describe each simulation frame. To make any sense of these dimensions that are often coupled, dimensionality reduction is performed. The goal is to determine a small subset of degrees of freedom that retain the essential information of the full-dimensional space.

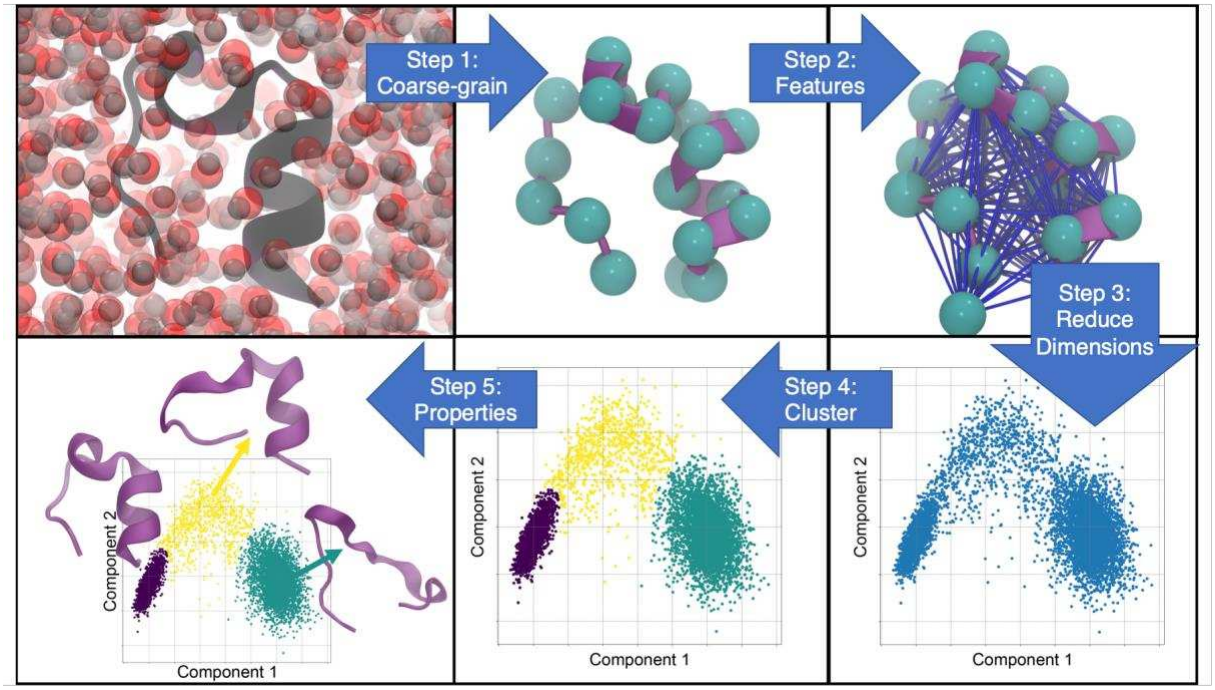


Figure 1.7 Five-step schematic for structural clustering from all-atom molecular dynamics trajectories. Starting from the top left, the first step is to choose a particle coarse-grained mapping and apply it to an explicit solvent all-atom trajectory. From there, a set of features to describe the system, such as all pairwise distances, are chosen. A trajectory of these features is then passed through an unsupervised dimensionality reduction algorithm (e.g., principal component analysis) and projected onto reduced dimensions. A clustering algorithm is applied in reduced dimensionality space to classify each frame of the trajectory. Finally, the structural clusters are analyzed (bottom left).

Multiple dimensionality reduction methods have been used on MD data. These methods differ based on what they define as *important* information. Principal component analysis (PCA), for example, determines the linearly independent coordinates that retain the most variance in the

data. It is not uncommon to see trajectory data projected onto the first two principal components with the largest variances. Time-lagged independent component analysis (tICA) determines the linear combination of coordinates that rank the timescale of collective motions.[151] tICA is often coupled with Markov State Modeling (MSM) to determine rates of transitions between metastable states.[152] Other dimensionality reduction methods have been employed, such as Sketch Map[153] and UMAP,[154] to achieve specific properties in the reduced space. The outcome of any dimensionality reduction technique will be a description of each trajectory frame in reduced dimensions.

Step 4: Clustering algorithm. Clustering algorithms are typically applied following dimensionality reduction methods to assign each frame to a conformational state. A variety of clustering algorithms have been applied in this context. The types of algorithms can be broken down into two categories: hierarchical and non-hierarchical. Both types have been used though more recent efforts have focused on non-hierarchical clustering algorithms.

Hierarchical clustering algorithms start with each frame in its own cluster and iteratively group similar frames together. This results in a dendrogram structure of the clustering. Different distances can be used to define similarity. Methods differ in how they join clusters together and the resulting distance used between clusters. Advantages of hierarchical methods include that the methods are fast, can be used to determine macroclusters and subclusters, and are easy to implement. The major disadvantage of hierarchical methods is that they are greedy: once clusters have been formed, they will not be broken apart to find a more global minimum. Examples of hierarchical methods include Ward, average linkage, and minimum linkage.

Non-hierarchical clustering methods are most common for MD trajectory data. The most common example is k-means clustering and its variants. k-means determines the means and trajectory partitioning that minimize the summed distances between frames and associated means.

This method works well for well-separated spherical distributions of points. Other density-based fitting algorithms include Gaussian mixture modeling (GMM) and DBSCAN, both of which have been employed to trajectory data projected onto reduced dimensional space with reasonable success.[155] These methods are not as greedy as hierarchical methods: frames can move from one cluster to another during the iterations until a (local) minimum in the algorithm metric is reached. A significant drawback to such methods is the computational cost and the difficulty in determining the appropriate number of clusters.

No single best method works for all trajectory data. An often-overlooked aspect, however, is to consider the underlying assumptions of the clustering algorithm. GMMs, for example, work under the assumption that the probability of the density can be represented as a sum (mixture) of Gaussian functions. In a recent study of clustering algorithms, Westerlund and Delemotte showed that GMMs do not work well when the density in reduced dimensional space is non-Gaussian.[155] It is not immediately clear whether we should expect MD data to be Gaussian in these spaces and almost assuredly depends on the features chosen and dimensionality reduction technique employed. In CHAPTER 3, a GMM clustering algorithm is developed for direct use with particle position features, the probability density of which is approximately Gaussian.

Analysis of Resulting Data. The result of a clustering algorithm is an assignment of each frame into a conformational state. With this information in hand, one can analyze different aspects of the clusters. The specific analysis employed will depend on the application. For example, one might build an MSM based on conformational clustering and determine the rates and associated structural mechanisms if one cares about folding rates. If one cares about catalysis, it is relevant to compute the relative reaction rate in each conformational state. The implicit assumption is that the value of a macroscopic property does not vary significantly within a given cluster.

1.6 Thesis Overview

The subsequent chapters describe my scientific contributions that address challenges in modeling conformational heterogeneity in biomolecules. CHAPTER 2 presents a DFT investigation of the relationship between active site structural change and reaction pathway energetics in glutamine hydrolysis performed by the allosterically regulated glutamine amidotransferase (GAT) imidazole glycerol phosphate synthase (IGPS). This work is the first atomic-level description of the reaction mechanism in IGPS and provides fundamental insights into the influence of local and subtle geometric variations on stabilizing the rate limiting chemical step. CHAPTER 3 details the development and application of a biomolecular structural clustering procedure, shapeGMM. Specifically, the underlying framework of using particle positions as clustering features is explained and unified with a Gaussian Mixture Model fitting algorithm. In doing so, limitations of existing approaches are overcome, such as reliance on transforming the data into internal coordinates (e.g., dihedrals, pair-wise distances), which often over-determines the system and might not discriminate between functionally relevant states. My contributions to the field of biomolecular structural clustering address the challenge of quantifying and characterizing functionally relevant conformational substates with a broadly applicable and interpretable algorithm. In CHAPTER 4, a Python code, enzyASM, that aids automation and high-throughput calculations is described. In CHAPTER 5 a computational workflow unifying concepts from CHAPTERS 2, 3 and 4 is proposed. These scientific contributions are placed in the grander perspective of what challenges remain in computational biocatalysis and critical next steps to harness the power of computational modeling in biocatalysis are outlined.

REFERENCES

(1) Fischer, E. Einfluss Der Configuration Auf Die Wirkung Der Enzyme. *Berichte Dtsch. Chem. Ges.* **1894**, *27* (3), 2985–2993.

(2) Koshland, D. E. Application of a Theory of Enzyme Specificity to Protein Synthesis. *Proc. Natl. Acad. Sci.* **1958**, *44* (2), 98–104.

(3) Koshland, D. E. Crazy, but Correct. *Nature* **2004**, *432* (7016), 447–447.

(4) Gerstein, M.; Lesk, A. M.; Chothia, C. Structural Mechanisms for Domain Movements in Proteins. *Biochemistry* **1994**, *33* (22), 6739–6749.

(5) Ramanathan, A.; Savol, A.; Burger, V.; Chennubhotla, C. S.; Agarwal, P. K. Protein Conformational Populations and Functionally Relevant Substates. *Acc. Chem. Res.* **2014**, *47* (1), 149–156.

(6) Tsai, C.-J.; Kumar, S.; Ma, B.; Nussinov, R. Folding Funnels, Binding Funnels, and Protein Function. *Protein Sci.* **1999**, *8* (6), 1181–1190.

(7) Henzler-Wildman, K.; Kern, D. Dynamic Personalities of Proteins. *Nature* **2007**, *450* (7172), 964–972.

(8) Benkovic, S. J. A Perspective on Enzyme Catalysis. *Science* **2003**, *301* (5637), 1196–1202.

(9) Fraser, J. S.; van den Bedem, H.; Samelson, A. J.; Lang, P. T.; Holton, J. M.; Echols, N.; Alber, T. Accessing Protein Conformational Ensembles Using Room-Temperature X-Ray Crystallography. *Proc. Natl. Acad. Sci.* **2011**, *108* (39), 16247–16252.

(10) Fenwick, R. B.; van den Bedem, H.; Fraser, J. S.; Wright, P. E. Integrated Description of Protein Dynamics from Room-Temperature X-Ray Crystallography and NMR. *Proc. Natl. Acad. Sci.* **2014**, *111* (4), E445–E454.

(11) Yabukarski, F.; Doukov, T.; Pinney, M. M.; Biel, J. T.; Fraser, J. S.; Herschlag, D. Ensemble-Function Relationships to Dissect Mechanisms of Enzyme Catalysis. *Sci. Adv.* **2022**, *8* (41), eabn7738.

(12) Mulder, F. A. A.; Filatov, M. NMR Chemical Shift Data and Ab Initio Shielding Calculations: Emerging Tools for Protein Structure Determination. *Chem Soc Rev* **2010**, *39* (2), 578–590.

(13) Tolman, J. R.; Ruan, K. NMR Residual Dipolar Couplings as Probes of Biomolecular Dynamics. *Chem. Rev.* **2006**, *106* (5), 1720–1736.

(14) Sibille, N.; Bernadó, P. Structural Characterization of Intrinsically Disordered Proteins by the Combined Use of NMR and SAXS. *Biochem. Soc. Trans.* **2012**, *40* (5), 955–962.

(15) Schwalbe, M.; Ozenne, V.; Bibow, S.; Jaremko, M.; Jaremko, L.; Gajda, M.; Jensen, M. R.; Biernat, J.; Becker, S.; Mandelkow, E.; Zweckstetter, M.; Blackledge, M. Predictive Atomic Resolution Descriptions of Intrinsically Disordered hTau40 and α-Synuclein in Solution from NMR and Small Angle Scattering. *Structure* **2014**, *22* (2), 238–249.

(16) Bai, X.; McMullan, G.; Scheres, S. H. W. How Cryo-EM Is Revolutionizing Structural Biology. *Trends Biochem. Sci.* **2015**, *40* (1), 49–57.

(17) Glaeser, R. M. How Good Can Cryo-EM Become? *Nat. Methods* **2016**, *13* (1), 28–32.

(18) Nogales, E. The Development of Cryo-EM into a Mainstream Structural Biology Technique. *Nat. Methods* **2016**, *13* (1), 24–27.

(19) Bonomi, M.; Vendruscolo, M. Determination of Protein Structural Ensembles Using Cryo-Electron Microscopy. *Curr. Opin. Struct. Biol.* **2019**, *56*, 37–45.

(20) Himo, F. Recent Trends in Quantum Chemical Modeling of Enzymatic Reactions. *J. Am. Chem. Soc.* **2017**, *139* (20), 6780–6786.

(21) Lind, M. E. S.; Himo, F. Quantum Chemistry as a Tool in Asymmetric Biocatalysis: Limonene Epoxide Hydrolase Test Case. *Angew. Chem. Int. Ed.* **2013**, *52* (17), 4563–4567.

(22) Lind, M. E. S.; Himo, F. Theoretical Study of Reaction Mechanism and Stereoselectivity of Arylmalonate Decarboxylase. *ACS Catal.* **2014**, *4* (11), 4153–4160.

(23) Lind, M. E. S.; Himo, F. Quantum Chemical Modeling of Enantioconvergency in Soluble Epoxide Hydrolase. *ACS Catal.* **2016**, *6* (12), 8145–8155.

(24) Hotta, K.; Chen, X.; Paton, R. S.; Minami, A.; Li, H.; Swaminathan, K.; Mathews, I. I.; Watanabe, K.; Oikawa, H.; Houk, K. N.; Kim, C.-Y. Enzymatic Catalysis of Anti-Baldwin Ring Closure in Polyether Biosynthesis. *Nature* **2012**, *483* (7389), 355–358.

(25) Tantillo, D. J.; Jiangang, C.; Houk, K. N. Theozymes and Compuzymes: Theoretical Models for Biological Catalysis. *Curr. Opin. Chem. Biol.* **1998**, *2* (6), 743–750.

(26) Smith, A. J. T.; Müller, R.; Toscano, M. D.; Kast, P.; Hellinga, H. W.; Hilvert, D.; Houk, K. N. Structural Reorganization and Preorganization in Enzyme Active Sites: Comparisons of Experimental and Theoretically Ideal Active Site Geometries in the Multistep Serine Esterase Reaction Cycle. *J. Am. Chem. Soc.* **2008**, *130* (46), 15361–15373.

(27) Kiss, G.; Çelebi-Ölçüm, N.; Moretti, R.; Baker, D.; Houk, K. N. Computational Enzyme Design. *Angew. Chem. Int. Ed.* **2013**, *52* (22), 5700–5725.

(28) Schütz, M. Low-Order Scaling Local Electron Correlation Methods. III. Linear Scaling Local Perturbative Triples Correction (*T*). *J. Chem. Phys.* **2000**, *113* (22), 9986–10001.

(29) Claeyssens, F.; Harvey, J. N.; Manby, F. R.; Mata, R. A.; Mulholland, A. J.; Ranaghan, K. E.; Schütz, M.; Thiel, S.; Thiel, W.; Werner, H.-J. High-Accuracy Computation of Reaction Barriers in Enzymes. *Angew. Chem.* **2006**, *118* (41), 7010–7013.

(30) Lin, H.; Truhlar, D. G. Redistributed Charge and Dipole Schemes for Combined Quantum Mechanical and Molecular Mechanical Calculations. *J. Phys. Chem. A* **2005**, *109* (17), 3991–4004.

(31) Yu, E. W.; Koshland, D. E. Propagating Conformational Changes over Long (and Short) Distances in Proteins. *Proc. Natl. Acad. Sci.* **2001**, *98* (17), 9517–9520.

(32) Nussinov, R.; Tsai, C.-J. Unraveling Structural Mechanisms of Allosteric Drug Action. *Trends Pharmacol. Sci.* **2014**, *35* (5), 256–264.

(33) Nussinov, R. Introduction to Protein Ensembles and Allostery. *Chem. Rev.* **2016**, *116* (11), 6263–6266.

(34) Otten, R.; Liu, L.; Kenner, L. R.; Clarkson, M. W.; Mavor, D.; Tawfik, D. S.; Kern, D.; Fraser, J. S. Rescue of Conformational Dynamics in Enzyme Catalysis by Directed Evolution. *Nat. Commun.* **2018**, *9* (1), 1314.

(35) Otten, R.; Pádua, R. A. P.; Bunzel, H. A.; Nguyen, V.; Pitsawong, W.; Patterson, M.; Sui, S.; Perry, S. L.; Cohen, A. E.; Hilvert, D.; Kern, D. How Directed Evolution Reshapes the Energy Landscape in an Enzyme to Boost Catalysis. *Science* **2020**, eabd3623.

(36) Broom, A.; Rakotoharisoa, R. V.; Thompson, M. C.; Zarifi, N.; Nguyen, E.; Mukhametzhanov, N.; Liu, L.; Fraser, J. S.; Chica, R. A. Ensemble-Based Enzyme Design Can Recapitulate the Effects of Laboratory Directed Evolution in Silico. *Nat. Commun.* **2020**, *11* (1), 4808.

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

(38) Wurm, J. P.; Sung, S.; Kneuttinger, A. C.; Hupfeld, E.; Sterner, R.; Wilmanns, M.; Sprangers, R. Molecular Basis for the Allosteric Activation Mechanism of the Heterodimeric Imidazole Glycerol Phosphate Synthase Complex. *Nat. Commun.* **2021**, *12* (1), 2748.

(39) Lonsdale, R.; Harvey, J. N.; Mulholland, A. J. A Practical Guide to Modelling Enzyme-Catalysed Reactions. *Chem. Soc. Rev.* **2012**, *41* (8), 3025.

(40) Jiménez-Osés, G.; Osuna, S.; Gao, X.; Sawaya, M. R.; Gilson, L.; Collier, S. J.; Huisman, G. W.; Yeates, T. O.; Tang, Y.; Houk, K. N. The Role of Distant Mutations and Allosteric Regulation on LovD Active Site Dynamics. *Nat. Chem. Biol.* **2014**, *10* (6), 431–436.

(41) Maria-Solano, M. A.; Romero-Rivera, A.; Osuna, S. Exploring the Reversal of Enantioselectivity on a Zinc-Dependent Alcohol Dehydrogenase. *Org. Biomol. Chem.* **2017**, *15* (19), 4122–4129.

(42) Maria-Solano, M. A.; Iglesias-Fernández, J.; Osuna, S. Deciphering the Allosterically Driven Conformational Ensemble in Tryptophan Synthase Evolution. *J. Am. Chem. Soc.* **2019**, *141* (33), 13049–13056.

(43) Nussinov, R.; Tsai, C.-J. Allostery without a Conformational Change? Revisiting the Paradigm. *Curr. Opin. Struct. Biol.* **2015**, *30*, 17–24.

(44) Maria-Solano, M. A.; Serrano-Hervás, E.; Romero-Rivera, A.; Iglesias-Fernández, J.; Osuna, S. Role of Conformational Dynamics in the Evolution of Novel Enzyme Function. *Chem. Commun.* **2018**, *54* (50), 6622–6634.

(45) Glowacki, D. R.; Harvey, J. N.; Mulholland, A. J. Taking Ockham's Razor to Enzyme Dynamics and Catalysis. *Nat. Chem.* **2012**, *4* (3), 169–176.

(46) Benkovic, S. J.; Hammes, G. G.; Hammes-Schiffer, S. Free-Energy Landscape of Enzyme Catalysis †. *Biochemistry* **2008**, *47* (11), 3317–3321.

(47) Eisenstein, O.; Ujaque, G.; Lledós, A. What Makes a Good (Computed) Energy Profile? In *New Directions in the Modeling of Organometallic Reactions*; Lledós, A., Ujaque, G., Eds.; Topics in Organometallic Chemistry; Springer International Publishing: Cham, 2020; Vol. 67, pp 1–38.

(48) Yoon, T. Commentary: Reviewer Comments A Discussion of "Can Reaction Mechanisms Be Proven?" *Chemical Education Today* **2009**, 1.

(49) Peng, Q.; Duarte, F.; Paton, R. S. Computing Organic Stereoselectivity – from Concepts to Quantitative Calculations and Predictions. *Chem. Soc. Rev.* **2016**, *45* (22), 6093–6107.

(50) Luchini, G.; Alegre-Requena, J. V.; Funes-Ardoiz, I.; Paton, R. S. GoodVibes: Automated Thermochemistry for Heterogeneous Computational Chemistry Data. *F1000Research* **2020**, *9*, 291.

(51) Kozuch, S.; Shaik, S. How to Conceptualize Catalytic Cycles? The Energetic Span Model. *Acc. Chem. Res.* **2011**, *44* (2), 101–110.

(52) Masgrau, L.; Truhlar, D. G. The Importance of Ensemble Averaging in Enzyme Kinetics. *Acc. Chem. Res.* **2015**, *48* (2), 431–438.

(53) Garcia-Viloca, M.; Gao, J.; Karplus, M.; Truhlar, D. G. How Enzymes Work: Analysis by Modern Rate Theory and Computer Simulations. *Science* **2004**, *303* (5655), 186–195.

(54) Bistoni, G.; Polyak, I.; Sparta, M.; Thiel, W.; Neese, F. Toward Accurate QM/MM Reaction Barriers with Large QM Regions Using Domain Based Pair Natural Orbital Coupled Cluster Theory. *J. Chem. Theory Comput.* **2018**, *14* (7), 3524–3531.

(55) Siegbahn, P. E. M.; Himo, F. The Quantum Chemical Cluster Approach for Modeling Enzyme Reactions. *WIREs Comput. Mol. Sci.* **2011**, *1* (3), 323–336.

(56) Frauenfelder, H.; Sligar, S. G.; Wolynes, P. G. The Energy Landscapes and Motions of Proteins. *Science* **1991**, *254* (5038), 1598.

(57) Frauenfelder, H.; McMahon, B. Dynamics and Function of Proteins: The Search for General Concepts. *Proc. Natl. Acad. Sci.* **1998**, *95* (9), 4795–4797.

(58) Baldwin, A. J.; Kay, L. E. NMR Spectroscopy Brings Invisible Protein States into Focus. *Nat. Chem. Biol.* **2009**, *5* (11), 808–814.

(59) Ando, T.; Kodera, N.; Takai, E.; Maruyama, D.; Saito, K.; Toda, A. A High-Speed Atomic Force Microscope for Studying Biological Macromolecules. *Proc. Natl. Acad. Sci.* **2001**, *98* (22), 12468–12472.

(60) Junker, J. P.; Ziegler, F.; Rief, M. Ligand-Dependent Equilibrium Fluctuations of Single Calmodulin Molecules. *Science* **2009**, *323* (5914), 633–637.

(61) Kodera, N.; Yamamoto, D.; Ishikawa, R.; Ando, T. Video Imaging of Walking Myosin V by High-Speed Atomic Force Microscopy. *Nature* **2010**, *468* (7320), 72–76.

(62) Qin, H.; Lim, L.; Song, J. Protein Dynamics at Eph Receptor-Ligand Interfaces as Revealed by Crystallography, NMR and MD Simulations. *BMC Biophys.* **2012**, *5* (1), 2.

(63) Rivalta, I.; Sultan, M. M.; Lee, N.-S.; Manley, G. A.; Loria, J. P.; Batista, V. S. Allosteric Pathways in Imidazole Glycerol Phosphate Synthase. *Proc. Natl. Acad. Sci.* **2012**, *109* (22), E1428–E1436.

(64) Lisi, G. P.; Loria, J. P. Allostery in Enzyme Catalysis. *Curr. Opin. Struct. Biol.* **2017**, *47*, 123–130.

(65) Negre, C. F. A.; Morzan, U. N.; Hendrickson, H. P.; Pal, R.; Lisi, G. P.; Loria, J. P.; Rivalta, I.; Ho, J.; Batista, V. S. Eigenvector Centrality for Characterization of Protein Allosteric Pathways. *Proc. Natl. Acad. Sci.* **2018**, *115* (52), E12201–E12208.

(66) Bonomi, M.; Heller, G. T.; Camilloni, C.; Vendruscolo, M. Principles of Protein Structural Ensemble Determination. *Curr. Opin. Struct. Biol.* **2017**, *42*, 106–116.

(67) Jing, Z.; Liu, C.; Cheng, S. Y.; Qi, R.; Walker, B. D.; Piquemal, J.-P.; Ren, P. Polarizable Force Fields for Biomolecular Simulations: Recent Advances and Applications. *Annu. Rev. Biophys.* **2019**, *48* (1), 371–394.

(68) Unke, O. T.; Chmiela, S.; Sauceda, H. E.; Gastegger, M.; Poltavsky, I.; Schütt, K. T.; Tkatchenko, A.; Müller, K.-R. Machine Learning Force Fields. *Chem. Rev.* **2021**, acs.chemrev.0c01111.

(69) Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y.; Wriggers, W. Atomic-Level Characterization of the Structural Dynamics of Proteins. *Science* **2010**, *330* (6002), 341–346.

(70) Hansmann, U. H. E. Parallel Tempering Algorithm for Conformational Studies of Biological Molecules. *Chem. Phys. Lett.* **1997**, *281* (1–3), 140–150.

(71) Sugita, Y.; Okamoto, Y. Replica-Exchange Molecular Dynamics Method for Protein Folding. *Chem. Phys. Lett.* **1999**, *314* (1–2), 141–151.

(72) Laio, A.; Parrinello, M. Escaping Free-Energy Minima. *Proc. Natl. Acad. Sci.* **2002**, *99* (20), 12562–12566.

(73) Zhou, T.; Caflisch, A. Free Energy Guided Sampling. *J. Chem. Theory Comput.* **2012**, *8* (6), 2134–2140.

(74) Bacci, M.; Vitalis, A.; Caflisch, A. A Molecular Simulation Protocol to Avoid Sampling Redundancy and Discover New States. *Biochim. Biophys. Acta BBA - Gen. Subj.* **2015**, *1850* (5), 889–902.

(75) Zimmerman, M. I.; Bowman, G. R. FAST Conformational Searches by Balancing Exploration/Exploitation Trade-Offs. *J. Chem. Theory Comput.* **2015**, *11* (12), 5747–5757.

(76) Shrestha, U. R.; Smith, J. C.; Petridis, L. Full Structural Ensembles of Intrinsically Disordered Proteins from Unbiased Molecular Dynamics Simulations. *Commun. Biol.* **2021**, *4* (1), 243.

(77) Campbell, E.; Kaltenbach, M.; Correy, G. J.; Carr, P. D.; Porebski, B. T.; Livingstone, E. K.; Afriat-Jurnou, L.; Buckle, A. M.; Weik, M.; Hollfelder, F.; Tokuriki, N.; Jackson, C. J. The Role of Protein Dynamics in the Evolution of New Enzyme Function. *Nat. Chem. Biol.* **2016**, *12* (11), 944–950.

(78) Hong, N.-S.; Petrović, D.; Lee, R.; Gryn'ova, G.; Purg, M.; Saunders, J.; Bauer, P.; Carr, P. D.; Lin, C.-Y.; Mabbitt, P. D.; Zhang, W.; Altamore, T.; Easton, C.; Coote, M. L.; Kamerlin, S. C. L.; Jackson, C. J. The Evolution of Multiple Active Site Configurations in a Designed Enzyme. *Nat. Commun.* **2018**, *9* (1), 3900.

(79) Romero-Rivera, A.; Garcia-Borràs, M.; Osuna, S. Role of Conformational Dynamics in the Evolution of Retro-Aldolase Activity. *ACS Catal.* **2017**, *7* (12), 8524–8532.

(80) Lake, P. T.; Davidson, R. B.; Klem, H.; Hocky, G. M.; McCullagh, M. Residue-Level Allostery Propagates through the Effective Coarse-Grained Hessian. *J. Chem. Theory Comput.* **2020**, *16* (5), 3385–3395.

(81) Fenton, A. W. Allostery: An Illustrated Definition for the 'Second Secret of Life.' *Trends Biochem. Sci.* **2008**, *33* (9), 420–425.

(82) Changeux, J.-P. Allosteric Mechanisms of Signal Transduction. *Science* **2005**, *308* (5727), 1424–1428.

(83) Gunasekaran, K.; Ma, B.; Nussinov, R. Is Allostery an Intrinsic Property of All Dynamic Proteins? *Proteins Struct. Funct. Bioinforma.* **2004**, *57* (3), 433–443.

(84) Li, J.; White, J. T.; Saavedra, H.; Wrabl, J. O.; Motlagh, H. N.; Liu, K.; Sowers, J.; Schroer, T. A.; Thompson, E. B.; Hilser, V. J. Genetically Tunable Frustration Controls Allostery in an Intrinsically Disordered Transcription Factor. *eLife* **2017**, *6*, e30688.

(85) Davidson, R. B.; Hendrix, J.; Geiss, B. J.; McCullagh, M. Allostery in the Dengue Virus NS3 Helicase: Insights into the NTPase Cycle from Molecular Simulations. *PLOS Comput. Biol.* **2018**, *14* (4), e1006103.

(86) Wodak, S. J.; Paci, E.; Dokholyan, N. V.; Berezovsky, I. N.; Horovitz, A.; Li, J.; Hilser, V. J.; Bahar, I.; Karanicolas, J.; Stock, G.; Hamm, P.; Stote, R. H.; Eberhardt, J.; Chebaro, Y.; Dejaegere, A.; Cecchini, M.; Changeux, J.-P.; Bolhuis, P. G.; Vreede, J.; Faccioli, P.; Orioli, S.; Ravasio, R.; Yan, L.; Brito, C.; Wyart, M.; Gkeka, P.; Rivalta, I.; Palermo, G.; McCammon, J. A.; Panecka-Hofman, J.; Wade, R. C.; Di Pizio, A.; Niv, M. Y.; Nussinov, R.; Tsai, C.-J.; Jang, H.; Padhorny, D.; Kozakov, D.; McLeish, T. Allostery in Its Many Disguises: From Theory to Applications. *Structure* **2019**, *27* (4), 566–578.

(87) Huang, Z.; Zhu, L.; Cao, Y.; Wu, G.; Liu, X.; Chen, Y.; Wang, Q.; Shi, T.; Zhao, Y.; Wang, Y.; Li, W.; Li, Y.; Chen, H.; Chen, G.; Zhang, J. ASD: A Comprehensive Database of Allosteric Proteins and Modulators. *Nucleic Acids Res.* **2011**, *39* (Database), D663–D669.

(88) Huang, Z.; Mou, L.; Shen, Q.; Lu, S.; Li, C.; Liu, X.; Wang, G.; Li, S.; Geng, L.; Liu, Y.; Wu, J.; Chen, G.; Zhang, J. ASD v2.0: Updated Content and Novel Features Focusing on Allosteric Regulation. *Nucleic Acids Res.* **2014**, *42* (D1), D510–D516.

(89) Liu, X.; Lu, S.; Song, K.; Shen, Q.; Ni, D.; Li, Q.; He, X.; Zhang, H.; Wang, Q.; Chen, Y.; Li, X.; Wu, J.; Sheng, C.; Chen, G.; Liu, Y.; Lu, X.; Zhang, J. Unraveling Allosteric Landscapes of Allosterome with ASD. *Nucleic Acids Res.* **2019**, gkz958.

(90) Monod, J.; Wyman, J.; Changeux, J.-P. On the Nature of Allosteric Transitions: A Plausible Model. **1965**, 31.

(91) Perutz, M. F.; Rossmann, M. G.; Cullis, A. F.; Muirhead, H.; Will, G.; North, A. C. T. Structure of Hæmoglobin: A Three-Dimensional Fourier Synthesis at 5.5-Å. Resolution, Obtained by X-Ray Analysis. *Nature* **1960**, *185* (4711), 416–422.

(92) Koshland, D. E.; Némethy, G.; Filmer, D. Comparison of Experimental Binding Data and Theoretical Models in Proteins Containing Subunits [*]. *Biochemistry* **1966**, *5* (1), 365–385.

(93) Cooper, A. Protein Fluctuations and the Thermodynamic Uncertainty Principle. *Prog. Biophys. Mol. Biol.* **1984**, *44* (3), 181–214.

(94) Hilser, V. J.; Wrabl, J. O.; Motlagh, H. N. Structural and Energetic Basis of Allostery. *Annu. Rev. Biophys.* **2012**, *41* (1), 585–609.

(95) Wei, G.; Xi, W.; Nussinov, R.; Ma, B. Protein Ensembles: How Does Nature Harness Thermodynamic Fluctuations for Life? The Diverse Functional Roles of Conformational Ensembles in the Cell. *Chem. Rev.* **2016**, *116* (11), 6516–6551.

(96) Motlagh, H. N.; Wrabl, J. O.; Li, J.; Hilser, V. J. The Ensemble Nature of Allostery. *Nature* **2014**, *508* (7496), 331–339.

(97) Schueler-Furman, O.; Wodak, S. J. Computational Approaches to Investigating Allostery. *Curr. Opin. Struct. Biol.* **2016**, *41*, 159–171.

(98) Warshel, A.; Sharma, P. K.; Kato, M.; Xiang, Y.; Liu, H.; Olsson, M. H. M. Electrostatic Basis for Enzyme Catalysis. *Chem. Rev.* **2006**, *106* (8), 3210–3235.

(99) Siegbahn, P. E. M.; Blomberg, M. R. A. Density Functional Theory of Biologically Relevant Metal Centers. *Annu. Rev. Phys. Chem.* **1999**, *50* (1), 221–249.

(100) Siegbahn, P. E. M.; Himo, F. Recent Developments of the Quantum Chemical Cluster Approach for Modeling Enzyme Reactions. *JBIC J. Biol. Inorg. Chem.* **2009**, *14* (5), 643–651.

(101) Liao, R.-Z.; Yu, J.-G.; Raushel, F. M.; Himo, F. Theoretical Investigation of the Reaction Mechanism of the Dinuclear Zinc Enzyme Dihydroorotase. *Chem. - Eur. J.* **2008**, *14* (14), 4287–4292.

(102) Kazemi, M.; Sheng, X.; Kroutil, W.; Himo, F. Computational Study of *Mycobacterium Smegmatis* Acyl Transferase Reaction Mechanism and Specificity. *ACS Catal.* **2018**, *8* (11), 10698–10706.

(103) Ryde, U. How Many Conformations Need To Be Sampled To Obtain Converged QM/MM Energies? The Curse of Exponential Averaging. *J. Chem. Theory Comput.* **2017**, *13* (11), 5745–5752.

(104) Siegbahn, P. E. M.; Li, X. Cluster Size Convergence for the Energetics of the Oxygen Evolving Complex in PSII. *J. Comput. Chem.* **2017**, *38* (25), 2157–2160.

(105) Sumowski, C. V.; Schmitt, B. B. T.; Schweizer, S.; Ochsenfeld, C. Quantum-Chemical and Combined Quantum-Chemical/Molecular-Mechanical Studies on the Stabilization of a Twin Arginine Pair in Adenovirus Ad11. *Angew. Chem. Int. Ed.* **2010**, *49* (51), 9951–9955.

(106) Hu, L.; Söderhjelm, P.; Ryde, U. Accurate Reaction Energies in Proteins Obtained by Combining QM/MM and Large QM Calculations. *J. Chem. Theory Comput.* **2013**, *9* (1), 640–649.

(107) Liao, R.-Z.; Thiel, W. Convergence in the QM-Only and QM/MM Modeling of Enzymatic Reactions: A Case Study for Acetylene Hydratase. *J. Comput. Chem.* **2013**, n/a-n/a.

(108) Raugei, S.; Seefeldt, L. C.; Hoffman, B. M. Critical Computational Analysis Illuminates the Reductive-Elimination Mechanism That Activates Nitrogenase for N_2 Reduction. *Proc. Natl. Acad. Sci.* **2018**, *115* (45), E10521–E10530.

(109) Hu, L.; Eliasson, J.; Heimdal, J.; Ryde, U. Do Quantum Mechanical Energies Calculated for Small Models of Protein-Active Sites Converge? *J. Phys. Chem. A* **2009**, *113* (43), 11793–11800.

(110) Siegbahn, P. E. M. Model Calculations Suggest That the Central Carbon in the FeMo-Cofactor of Nitrogenase Becomes Protonated in the Process of Nitrogen Fixation. *J. Am. Chem. Soc.* **2016**, *138* (33), 10485–10495.

(111) Chen, S.-L.; Fang, W.-H.; Himo, F. Technical Aspects of Quantum Chemical Modeling of Enzymatic Reactions: The Case of Phosphotriesterase. *Theor. Chem. Acc.* **2008**, *120* (4–6), 515–522.

(112) Blomberg, M. R. A.; Borowski, T.; Himo, F.; Liao, R.-Z.; Siegbahn, P. E. M. Quantum Chemical Studies of Mechanisms for Metalloenzymes. *Chem. Rev.* **2014**, *114* (7), 3601–3658.

(113) Becke, A. D.; Johnson, E. R. A Density-Functional Model of the Dispersion Interaction. *J. Chem. Phys.* **2005**, *123* (15), 154101.

(114) Ahmadi, S.; Herrera, L. B.; Chehelamirani, M.; Hostaš, J.; Jalife, S.; Salahub, D. R. Multiscale Modeling of Enzymes: QM-Cluster, QM/MM, and QM/MM/MD: A Tutorial Review. *Int. J. Quantum Chem.* **2018**, *118* (9), e25558.

(115) Magalhães, R. P.; Fernandes, H. S.; Sousa, S. F. Modelling Enzymatic Mechanisms with QM/MM Approaches: Current Status and Future Challenges. *Isr. J. Chem.* **2020**, *60* (7), 655–666.

(116) Ramos, M. J.; Fernandes, P. A. Computational Enzymatic Catalysis. *Acc. Chem. Res.* **2008**, *41* (6), 689–698.

(117) Monard, G.; Merz, K. M. Combined Quantum Mechanical/Molecular Mechanical Methodologies Applied to Biomolecular Systems. *Acc. Chem. Res.* **1999**, *32* (10), 904–911.

(118) Gao, J.; Truhlar, D. G. QUANTUM MECHANICAL METHODS FOR ENZYME KINETICS. *Annu. Rev. Phys. Chem.* **2002**, *53* (1), 467–505.

(119) Rosta, E.; Klähn, M.; Warshel, A. Towards Accurate Ab Initio QM/MM Calculations of Free-Energy Profiles of Enzymatic Reactions. *J. Phys. Chem. B* **2006**, *110* (6), 2934–2941.

(120) Lin, H.; Truhlar, D. G. QM/MM: What Have We Learned, Where Are We, and Where Do We Go from Here? *Theor. Chem. Acc.* **2007**, *117* (2), 185.

(121) Warshel, A.; Levitt, M. Theoretical Studies of Enzymic Reactions: Dielectric, Electrostatic and Steric Stabilization of the Carbonium Ion in the Reaction of Lysozyme. *J. Mol. Biol.* **1976**, *103* (2), 227–249.

(122) Senn, H. M.; Thiel, W. QM/MM Methods for Biomolecular Systems. *Angew. Chem. Int. Ed.* **2009**, *48* (7), 1198–1229.

(123) Gao, J.; Ma, S.; Major, D. T.; Nam, K.; Pu, J.; Truhlar, D. G. Mechanisms and Free Energies of Enzymatic Reactions. *Chem. Rev.* **2006**, *106* (8), 3188–3209.

(124) Cortopassi, W. A.; Simion, R.; Honsby, C. E.; França, T. C. C.; Paton, R. S. Dioxygen Binding in the Active Site of Histone Demethylase JMJD2A and the Role of the Protein Environment. *Chem. – Eur. J.* **2015**, *21* (52), 18983–18992.

(125) Walker, R. C.; de Souza, M. M.; Mercer, I. P.; Gould, I. R.; Klug, D. R. Large and Fast Relaxations inside a Protein: Calculation and Measurement of Reorganization Energies in Alcohol Dehydrogenase. *J. Phys. Chem. B* **2002**, *106* (44), 11658–11665.

(126) van der Kamp, M. W.; Mulholland, A. J. Combined Quantum Mechanics/Molecular Mechanics (QM/MM) Methods in Computational Enzymology. *Biochemistry* **2013**, *52* (16), 2708–2728.

(127) Senn, H. M.; Thiel, W. QM/MM Studies of Enzymes. *Curr. Opin. Chem. Biol.* **2007**, *11* (2), 182–187.

(128) Sousa, S. F.; Ribeiro, A. J. M.; Neves, R. P. P.; Brás, N. F.; Cerqueira, N. M. F. S. A.; Fernandes, P. A.; Ramos, M. J. Application of Quantum Mechanics/Molecular Mechanics Methods in the Study of Enzymatic Reaction Mechanisms. *WIREs Comput. Mol. Sci.* **2017**, *7* (2), e1281.

(129) Klähn, M.; Braun-Sand, S.; Rosta, E.; Warshel, A. On Possible Pitfalls in Ab Initio Quantum Mechanics/Molecular Mechanics Minimization Approaches for Studies of Enzymatic Reactions. *J. Phys. Chem. B* **2005**, *109* (32), 15645–15650.

(130) Sumowski, C. V.; Ochsenfeld, C. A Convergence Study of QM/MM Isomerization Energies with the Selected Size of the QM Region for Peptidic Systems. *J. Phys. Chem. A* **2009**, *113* (43), 11734–11741.

(131) Blank, I. D.; Sadeghian, K.; Ochsenfeld, C. A Base-Independent Repair Mechanism for DNA Glycosylase—No Discrimination Within the Active Site. *Sci. Rep.* **2015**, *5* (1), 10369.

(132) Roßbach, S.; Ochsenfeld, C. Influence of Coupling and Embedding Schemes on QM Size Convergence in QM/MM Approaches for the Example of a Proton Transfer in DNA. *J. Chem. Theory Comput.* **2017**, *13* (3), 1102–1107.

(133) Hu, L.; Söderhjelm, P.; Ryde, U. On the Convergence of QM/MM Energies. *J. Chem. Theory Comput.* **2011**, *7* (3), 761–777.

(134) Solt, I.; Kulhánek, P.; Simon, I.; Winfield, S.; Payne, M. C.; Csányi, G.; Fuxreiter, M. Evaluating Boundary Dependent Errors in QM/MM Simulations. *J. Phys. Chem. B* **2009**, *113* (17), 5728–5735.

(135) Kulik, H. J.; Zhang, J.; Klinman, J. P.; Martínez, T. J. How Large Should the QM Region Be in QM/MM Calculations? The Case of Catechol *O* -Methyltransferase. *J. Phys. Chem. B* **2016**, *120* (44), 11381–11394.

(136) Lyman, E.; Pfaendtner, J.; Voth, G. A. Systematic Multiscale Parameterization of Heterogeneous Elastic Network Models of Proteins. *Biophys. J.* **2008**, *95* (9), 4183–4192.

(137) Das, S.; Nam, K.; Major, D. T. Rapid Convergence of Energy and Free Energy Profiles with Quantum Mechanical Size in Quantum Mechanical–Molecular Mechanical Simulations of Proton Transfer in DNA. *J. Chem. Theory Comput.* **2018**, *14* (3), 1695–1705.

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

(139) Govender, K.; Gao, J.; Naidoo, K. J. AM1/d-CB1: A Semiempirical Model for QM/MM Simulations of Chemical Glycobiology Systems. *J. Chem. Theory Comput.* **2014**, *10* (10), 4694–4707.

(140) McCullagh, M.; Saunders, M. G.; Voth, G. A. Unraveling the Mystery of ATP Hydrolysis in Actin Filaments. *J. Am. Chem. Soc.* **2014**, *136* (37), 13053–13058.

(141) Yagi, K.; Ito, S.; Sugita, Y. Exploring the Minimum-Energy Pathways and Free-Energy Profiles of Enzymatic Reactions with QM/MM Calculations. *J. Phys. Chem. B* **2021**, *125* (18), 4701–4713.

(142) Sun, R.; Sode, O.; Dama, J. F.; Voth, G. A. Simulating Protein Mediated Hydrolysis of ATP and Other Nucleoside Triphosphates by Combining QM/MM Molecular Dynamics with Advances in Metadynamics. *J. Chem. Theory Comput.* **2017**, *13* (5), 2332–2341.

(143) Quinn, T. R.; Steussy, C. N.; Haines, B. E.; Lei, J.; Wang, W.; Sheong, F. K.; Stauffacher, C. V.; Huang, X.; Norrby, P.-O.; Helquist, P.; Wiest, O. Microsecond Timescale MD Simulations at the Transition State of *Pm* HMGR Predict Remote Allosteric Residues. *Chem. Sci.* **2021**, *12* (18), 6413–6418.

(144) Cortopassi, W. A.; Kumar, K.; Duarte, F.; Pimentel, A. S.; Paton, R. S. Mechanisms of Histone Lysine-Modifying Enzymes: A Computational Perspective on the Role of the Protein Environment. *J. Mol. Graph. Model.* **2016**, *67*, 69–84.

(145) Quesne, M. G.; Borowski, T.; de Visser, S. P. Quantum Mechanics/Molecular Mechanics Modeling of Enzymatic Processes: Caveats and Breakthroughs. **2016**, *22*, 2562–2581.

(146) Cooper, A. M.; Kästner, J. Averaging Techniques for Reaction Barriers in QM/MM Simulations. *ChemPhysChem* **2014**, *15* (15), 3264–3269.

(147) Romero-Téllez, S.; Cruz, A.; Masgrau, L.; González-Lafont, À.; Lluch, J. M. Accounting for the Instantaneous Disorder in the Enzyme–Substrate Michaelis Complex to Calculate the Gibbs Free Energy Barrier of an Enzyme Reaction. *Phys. Chem. Chem. Phys.* **2021**, *23* (23), 13042–13054.

(148) von der Esch, B.; Dietschreit, J. C. B.; Peters, L. D. M.; Ochsenfeld, C. Finding Reactive Configurations: A Machine Learning Approach for Estimating Energy Barriers Applied to Sirtuin 5. *J. Chem. Theory Comput.* **2019**, *15* (12), 6660–6667.

(149) Rhee, Y. M.; Sorin, E. J.; Jayachandran, G.; Lindahl, E.; Pande, V. S. Simulations of the Role of Water in the Protein-Folding Mechanism. *Proc. Natl. Acad. Sci.* **2004**, *101* (17), 6456–6461.

(150) Sittel, F.; Stock, G. Perspective: Identification of Collective Variables and Metastable States of Protein Dynamics. *J. Chem. Phys.* **2018**, *149* (15), 150901.

(151) Pérez-Hernández, G.; Noé, F. Hierarchical Time-Lagged Independent Component Analysis: Computing Slow Modes and Reaction Coordinates for Large Molecular Systems. *J. Chem. Theory Comput.* **2016**, *12* (12), 6118–6129.

(152) Pérez-Hernández, G.; Paul, F.; Giorgino, T.; De Fabritiis, G.; Noé, F. Identification of Slow Molecular Order Parameters for Markov Model Construction. *J. Chem. Phys.* **2013**, *139* (1), 015102.

(153) Ceriotti, M.; Tribello, G. A.; Parrinello, M. Simplifying the Representation of Complex Free-Energy Landscapes Using Sketch-Map. *Proc. Natl. Acad. Sci.* **2011**, *108* (32), 13023–13028.

(154) Trozzi, F.; Wang, X.; Tao, P. UMAP as a Dimensionality Reduction Tool for Molecular Dynamics Simulations of Biomacromolecules: A Comparison Study. *J. Phys. Chem. B* **2021**, *125* (19), 5022–5034.

(155) Westerlund, A. M.; Delemotte, L. InfleCS: Clustering Free Energy Landscapes with Gaussian Mixtures. *J. Chem. Theory Comput.* **2019**, *15* (12), 6752–6759.

CHAPTER 2. THE CATALYTIC EFFECTS OF ACTIVE SITE CONFORMATIONAL CHANGE IN THE ALLOSTERIC ACTIVATION OF IMIDAZOLE GLYCEROL PHOSPHATE SYNTHASE

2.1 Introduction

Imidazole glycerol phosphate synthase (IGPS) is a glutamine amidotransferase (GAT) vital to purine and histidine biosynthetic pathways in bacteria, fungi, and plants, making it an attractive antimicrobial therapeutic target.[1] IGPS from *Thermotoga maritima* is a heterodimer composed of the HisH and HisF subunits (HisFH) (Figure 2.1).[2] HisH performs glutamine (*L*-Gln) hydrolysis to form glutamate (*L*-Glu) and ammonia using a catalytic triad of *h*Cys84, *h*His178, and *h*Glu180 (*h* and *f* prefixes indicate if the residue belongs to HisH or HisF, respectively), characterizing it as a class-I GAT.[3,4] The free ammonia is shuttled across the dimer interface to react with N^1-(5'-phosphoribosyl)-formimino-5-aminoimidazole-4 carboxamide ribonucleotide (PrFAR) in the HisF active site over 25 Å away to form imidazole glycerol phosphate (IGP) and 5'-(5-aminoimidazole-4-carboxamide) ribonucleotide (AICAR).[5] Binding of PrFAR to HisF elicits a V-type allosteric effect that enhances glutamine hydrolysis efficiency in HisH by approximately 4500-fold, primarily influencing the rate of glutamine turnover.[2,6]

The hydrolysis reaction mechanism has not yet been studied at the atomic level in IGPS, to the best of our knowledge. However, similar well-studied enzymes, such as other class-I GATs and cysteine/serine proteases provide a useful theoretical foundation.[3,4,7–13] These reactions are proposed to occur via two stages: acylation and deacylation. In the acylation stage, the nucleophilic *h*Cys84 attacks glutamine to form a glutamyl thioester, first structurally observed in another class-I GAT, carbamoyl phosphate synthetase,[10] via a tetrahedral oxyanion intermediate. In the

deacylation stage, nucleophilic attack by a water molecule breaks down the covalent enzyme-substrate intermediate thioester, to yield the glutamate product. The rate-limiting step is proposed to occur during acylation in IGPS,[14] as well as related class-I GATs, carbamyl phosphate synthetase, and anthranilate synthase.[3,11,15]

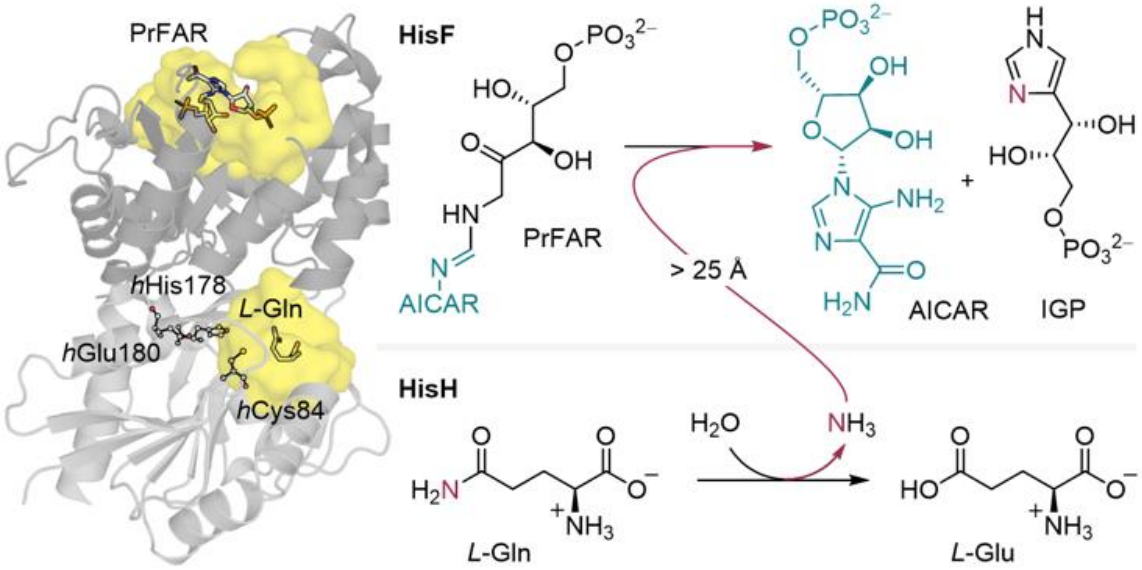


Figure 2.1. General IGPS scheme of the coupled L-Gln hydrolysis and PrFAR cyclization reactions performed in HisH. and HisF, respectively. The L-Gln and PrFAR binding sites are highlighted in yellow.

IGPS serves as a paradigmatic allosteric enzyme for experimental and methodological developments.[16–27] An abundance of dynamical information contribute to our understanding of the allosteric mechanism in IGPS at the molecular level. A sequence of residues (*h*Pro49—*h*Gly50—*h*Val51—*h*Gly52 in IGPS) comprise the oxyanion strand, a structural motif common to class-I GATs that positions a backbone amide hydrogen to stabilize the formation of an oxyanion throughout the reaction.[28–32] Increased flexibility in the IGPS oxyanion strand upon PrFAR binding was observed in NMR experiments and molecular dynamics (MD) simulations, supporting its mechanistic involvement.[25,33,34] An interfacial hydrogen bond between *f*Pro10 and the backbone N–H of *h*Val51 is weakened in MD simulations with PrFAR,[26,35] and explains the enhanced

flexibility of the oxyanion strand observed in NMR experiments.[25,34] PrFAR also reduces the opening angle of the interface, which is expected to influence the hydrolysis reaction, although it is unclear if there is a direct effect on the reaction mechanism.[6,19,33,36]

A leading hypothesis to explain the allosteric rate effect in IGPS is that PrFAR enables a conformational change in the *h*Val51 backbone that lowers the rate of glutamine hydrolysis via a catalytically competent oxyanion hole.[14,35,37,38] Contrary to this hypothesis, the *h*Val51 amide C=O group points into the active site and the N–H away in all but one crystallographic conformations of IGPS in various ligand bound states.[5,6,39–41] The anamolous conformation is observed in a recently deposited structure 7ac8, chains E and F with bound allosteric ligand and Gln substrate.[14] The catalytic *h*Cys84 was mutated to alanine in this structure to disable glutamine turnover and capture IGPS in the presumably active conformation. Osuna and coworkers provide additional support for this hypothesis through MD simulations employing a biasing potential to sample the *h*Val51 dihedral flip transition. The energetic barrier of this process was estimated to be lower in simulations with PrFAR present (approx. 8 kcal/mol) compared to without PrFAR present (approx. 22 kcal/mol).[35]

Exploration of an alternative activation mechanism is warranted for a few reasons. Extensive (10 μs) unbiased MD simulations of IGPS only reproduced the *h*Val51 dihedral transition when neither Gln nor PrFAR were present, contrary to what was expected.[24] The authors noted the novel IGPS conformation revealed in the 7ac8 crystallographic model could be an artifact of the loss of function *h*Cys84Ala mutation rather than intrinsic to the allosteric mechanism. An alternative activation mechanism that has yet to be evaluated in IGPS involves repositioning of the oxyanion strand relative to the Gln substrate. This activation hypothesis has been proposed for another class-I GAT, aminodeoxychorismate synthase (ADCS) since the presence of two prolines

in its oxyanion strand (Pro51—Gly52—Pro53) inhibits Gly52 backbone rotation.[42] All Gln bound IGPS structures show a hydrogen bond between the *h*Gly52 N–H and Gln carbonyl; however, Gln is presumed to be too far from *h*Cys84 to facilitate the nucleophilic attack. The increased oxyanion strand flexibility observed in the presence of PrFAR could facilitate the oxyanion strand reorganization necessary to stabilize the substrate after nucleophilic attack.

Despite various X-ray structures, mutagenesis studies, kinetic experiments, and MD simulations, a connection between the allosteric effect and the glutamine hydrolysis mechanism remains hypothetical. This work targets two essential questions regarding IGPS activation*: How do local structural aspects of the active site influence the reaction energetics? Is the hVal51 backbone flip required for rate enhancement?* To address these questions, we present a dispersion-corrected Density Functional Theory (DFT) study on large (237 atoms) active site cluster models of IGPS in various active and inactive conformations. The multistep reaction leading to thioester formation and ammonia production is investigated and energy decomposition analyses are performed to evaluate the relationship between active site geometry and reaction stabilization. Our results are valuable for future endeavors in modeling the allosteric behavior of this prototypical system by providing clear insights into the atomistic changes required for rate enhancement.

Although rigorous workflows have been devised to consider active site multi-state effects on catalysis,[43–46] this work affords a simplified approach given the available crystallographic data and scope of knowledge from preceding investigations. In doing so, our application of the quantum chemical cluster approach[47] to investigate an allosteric effect and compare the catalytic impact of active site conformational change traverses a challenge in the field, as recently noted by Himo and de Visser.[48,49]

2.2 Results and Discussion

2.2.1 Active Site Models

Positions of the active site residues and the Gln substrate are highly conserved in crystallographic models except for the HisFH dimer conformation composed of chains E and F from PDB 7ac8 (Figure 2.2). There are four distinct geometric differences observed in this conformation: 1) an interfacial residue, *f*Gln123, interacts with the bound Gln as a result of subunit closure; 2) the *h*Val51 amide N–H points toward the substrate carbonyl; 3) the oxyanion strand is positioned above the catalytic thiol; and 4) the reactive carboxamide of Gln is preorganized for acylation.

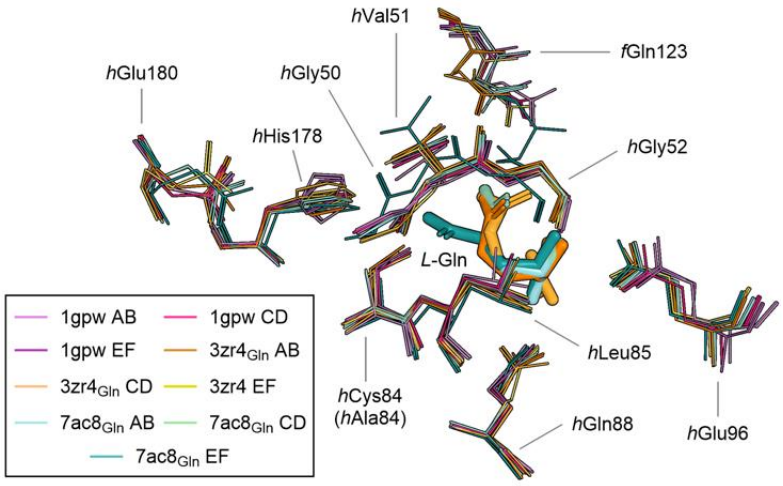


Figure 2.2. Overlap of HisH active-site geometries across multiple crystalized structures.

Two truncated active site models were created from the 7ac8[14] crystallized unit of IGPS using atomic positions from chains E/F and chains C/D; the deposition authors refer to these complexes as active and inactive conformations, respectively. Residue *h*Ala84 in chains F and D was reverted back to wild-type *h*Cys84 with PyMol[50] by selecting the backbone-dependent side chain rotamer with the least steric clash. We will refer to these truncated models as Active and Inactive.

The size and residue components of the QM model are important to consider, and informed decisions based on selection criteria remains an active area of development.[50–55] Residues were selected based on interactions with the substrate, and relevance indicated in the literature. The ligand interaction diagrams in Figure 2.3 illustrate how the C/D and E/F conformations yield different interactions with the glutamine substrate. Since the main objective of this work is to evaluate the catalytic impact of active site structural changes, we focused on building models suitable for direct comparison. Therefore, the same atoms were included in each model from the following residues: *f*Gly121, *f*Ser122, *f*Gln123 and *f*Ala124 from the HisF subunit, *h*Gly50, *h*Val51, *h*Gly52, *h*His53, *h*Cys84, *h*Leu85, *h*Gln88, *h*Glu96, *h*Val140, *h*His141, *h*Thr142, *h*Tyr143, *h*His178 and *h*Glu180 from the HisH subunit, the Gln substrate in zwitterionic form and two conserved crystallographic waters. Here on all residues without a prefix are assumed to be from the HisH monomer since they are the model majority, but the HisF prefix will be kept for clarity. Amino acid side and main chains were truncated according to potential involvement in the elementary reaction steps and interactions with the substrate (see Table S1 and text for a detailed list of included atoms, truncation scheme, and geometry optimization constraints). Each model contains 237 atoms, an overall charge of –2 and 12 constrained carbon atoms (Figure 2.3). Geometry optimizations were performed with the range-separated, dispersion-corrected ωB97X-D[64] functional using the 6-31G*[65] basis set for all C, H, and N atoms and the 6-31+G*[66] basis set for S and O atoms. Stationary point electronic energies were further refined at the B3LYP-D3(BJ)/6-311+G(2d,2p) level of theory and the default IEF–PCM implicit solvent method with internal parameters for diethyl ether (ε=4.24).[67–72] Additional information is presented in the Computational Details section.

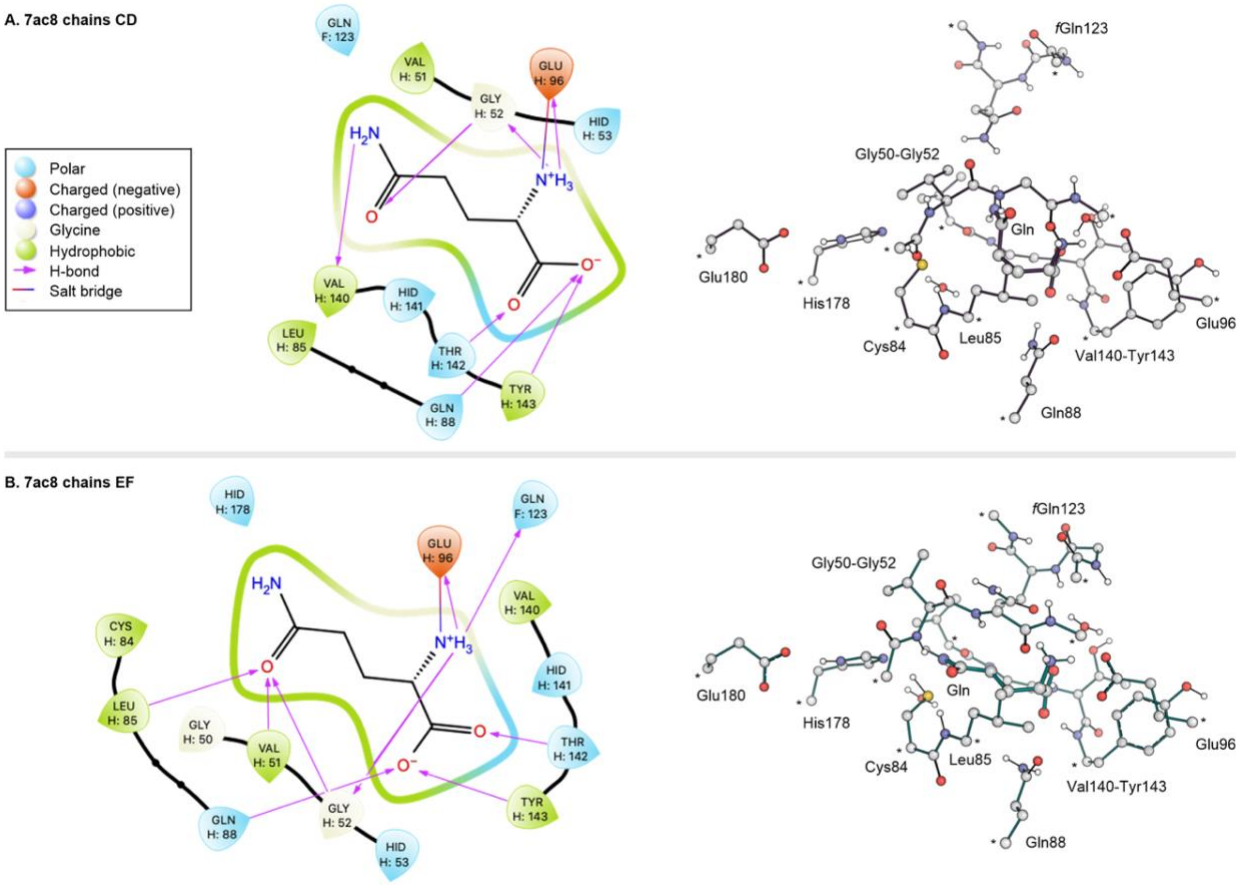


Figure 2.3. Interactions at the HisH active site in the inactive (left-top) and active (left-bottom) conformations and resulting QM cluster models (right). Asterisks indicate Cα atoms frozen during geometry optimizations.

2.2.2 Acylation Reaction

The thioester formation mechanism proposed in this work (Figure 2.4) begins from the enzyme-substrate (**ES**) complex with deprotonation of Cys84 by His178 to form the thiolate in **Int1**. The nucleophilic Cys84 attacks the substrate carbonyl forming a tetrahedral oxyanion intermediate (**Int2**). The His178 imidazolium transfers a proton to the substrate to form an ammonium in (**Int3**). Lastly, the tetrahedral intermediate collapses and the N–C bond breaks, forming the thioester acyl-intermediate and liberating ammonia (**Int4**). All optimized structures are shown in APPENDIX A.

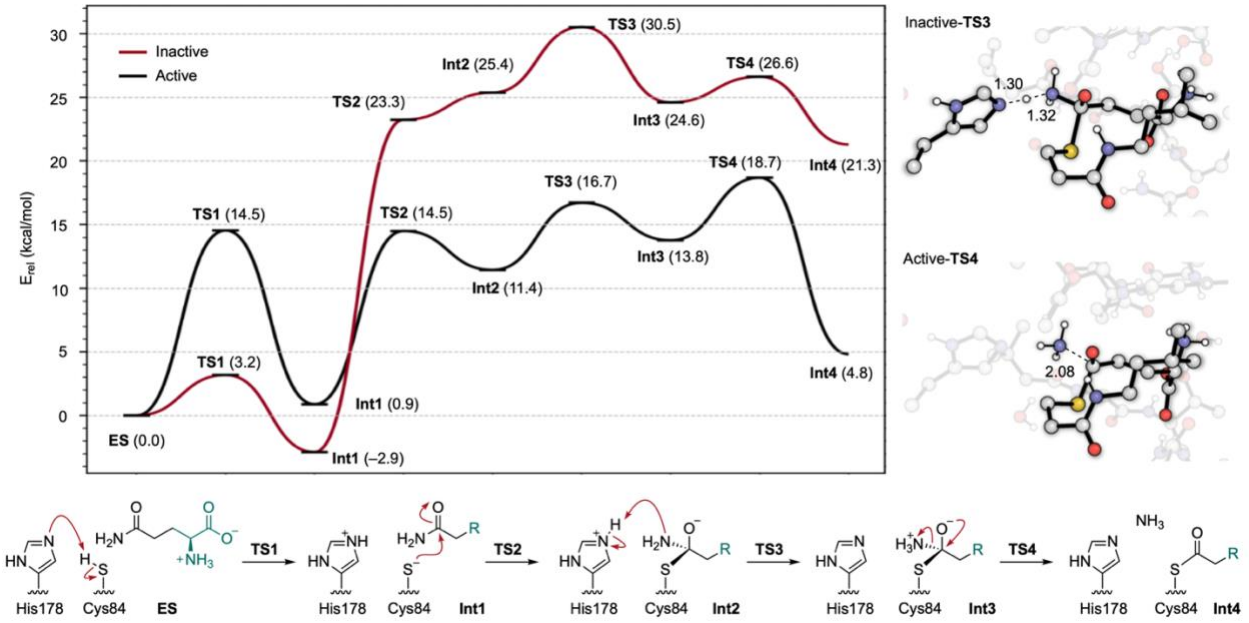


Figure 2.4. Energy profiles of Active (black) and Inactive (red) IGPS active site geometries showing ΔE relative to each model's enzyme substrate (**ES**) complex in kcal/mol. Zoomed in representations of the rate limiting steps corresponding to each model with TS bonds shown as dotted lines and distances in Å

The electronic energy profiles of Active and Inactive models (Figure 2.4) illustrate how differences in active site geometry substantially influence the reaction. Deprotonation of Cys84 in Active has a barrier of 14.5 kcal/mol, significantly higher than the 3.2 kcal/mol required for Inactive. This difference is due to the positioning of the substrate. In the optimized Active-**ES**, the substrate is observed in a near-attack conformation, with the carbonyl carbon 3.56 Å from the Cys84 sulfur. The carboxamide NH_2 group hydrogen bonds with the His178 Nε difference is due to the positioning of the substrate. In the optimized Active-**ES**, the substrate is observed in a near-attack conformation, with the carbonyl carbon 3.56 Å from the Cys84 sulfur. The carboxamide NH_2 group hydrogen bonds with the His178 Nε (2.01 Å). This interaction is disrupted in the Active-**TS1** when the His178 Nε accepts the proton from Cys84. Alternatively, in Inactive the substrate is initially positioned farther away (3.84 Å), which allows Cys84 to hydrogen bond with

His178 (1.91 Å) in the Inactive-**ES**. This is not possible in Active because the proximity of *f*Gln123 confines the substrate.

The rate of thioester formation in the Active geometry is determined by the collapse of the tetrahedral oxyanion intermediate, with a TS bond breaking distance of 2.08 Å (Active-**TS4**, Figure 2.4). This yields a computationally predicted barrier of $\Delta E^{\ddagger}$=18.7 kcal/mol, which agrees well with the experimental rate of *T. maritima* (0.67 ± 0.02 s^{-1} at 298.15 K, corresponding to a barrier of around 17.7 kcal/mol).[6] Isotope effects of the amide nitrogen on glutamine bound to a similar class-I GAT, carbamyl phosphate synthetase (CPS), support the release of ammonia as the rate-limiting step.[11] Thioester formation was also found to be rate-limiting in cysteine esterase.[57]

In the Inactive geometry, the formation of ammonium (Inactive-**TS3**, Figure 2.4) is highest in energy, yielding a barrier of $\Delta E^{\ddagger}$=30.5 kcal/mol for thioester formation. This barrier is larger than experimental measurements (21.1 and 22.7 kcal/mol).[6,58] Alternative mechanistic pathways were explored, such as concerted NH_3 formation and tetrahedral collapse, but were not located. Another possibility is that more energetically favorable structural rearrangements occur along the reaction coordinate to enable catalysis. In support of this, the barrier for the Val51 backbone flip in the absence of PrFAR is estimated by steered MD to be approximately 22 kcal/mol,[35] which aligns well with the experimental rate. Therefore, our current DFT results suggest that the enzymatic reaction rate in the absence of the allosteric effector is limited by the barrier for conformational interconversion rather than by the elementary bond-forming and breaking steps in the active site.

The elementary step most affected by an active site structural change is the formation of the tetrahedral acyl-enzyme via **TS2**. This transformation requires 13.6 kcal/mol from **Int1** in Active and nearly twice as much in Inactive (26.1 kcal/mol). This step demonstrates the importance

of IGPS active site geometry in stabilizing the incipient oxyanion. Although it has been proposed that the IGPS structure 3zr4 (structurally consistent with our Inactive model) is in a catalytically competent conformation with the Gly52 N–H completing the oxyanion hole, our data instead indicate that the energy associated with the Inactive geometry hinders thioester formation.

There are multiple structural differences between Active and Inactive models. These include the Val51 backbone flip and *f*Gln123 proximity, but also less apparent variations, such as the relative position of the oxyanion strand (APPENDIX A Figure A1). The capacity of Gly52 to serve as an H-bond donor along the reaction coordinate may depend on the subtle positioning of the oxyanion strand. Additionally, it is unclear from the energy profiles alone how proximity of *f*Gln123 influences the reaction. We therefore pursued additional computational studies to deconvolute the energetic effects due to specific structural variations in the IGPS acylation reaction.

2.2.3 Oxyanion Hole Contributions

Two additional models, Inactive *f*Gln123 and Inactive Val51, were manually constructed (see SI for construction details) to analyze separately how *f*Gln123 proximity, the Val51 backbone conformation, and the position of the oxyanion strand influence TS stabilization in the Active model. This approach is reminiscent of theozyme modeling,[60] where residues, or functional groups, are hypothetically arranged in a truncated active site to explore catalytic contributions of individual atomic interactions and predict optimal catalytic scaffolds. In our approach, the distinct models enable us to quantify the influence of each structural difference relative to the active model separately. The oxyanion hole contributions were evaluated in each TS by estimating the donor-acceptor orbital interaction energies of the Leu85, Gly52 and Val51 σ^* orbitals and the Gln Oε

lone pair electrons via Natural Bonding Orbital (NBO) analysis.[59] Calculation details are found in the Computational Details section.

As hypothesized, the Active oxyanion hole structure provides substantial stabilization (Figure 2.5A). Val51 is the predominant stabilizing component during oxyanion formation (**TS2**), with a favorable interaction energy of –14.7 kcal/mol. Contributions from Gly52 and Leu85 are much smaller (–4.0 kcal/mol and –2.6 kcal/mol, respectively). Surprisingly, when the oxyanion species breaks down in **TS4**, Val51 no longer interacts with the oxyanion. The N–H shifts direction to interact with the lone pair of electrons on the His178 Nε. To ensure the Active-**TS4** geometry optimization was not unintentionally biased by the initial geometry input, the oxyanion strand fragment in the optimized **TS4** structure was replaced with that from the optimized **TS2** structure. The optimization converged to the same **TS4** structure, suggesting this is the appropriate first-order saddle point.

Although the Val51 N–H points away from the active site in the Inactive model, the Val51 Cα–H stabilizes **TS2** by –5.9 kcal/mol and **TS4** by –4.9 kcal/mol (Figure 2.5D). Leu85 contributes –4.3 kcal/mol in Inactive-**TS2**, which is slightly more than its contributions in Active-**TS2** and Inactive *f*Gln123-**TS2** (Figure 2.5B), presumably to compensate for the weak interaction from Val51. However, during oxyanion breakdown in Inactive-**TS4**, the Gln:Oε–Leu85:H distance Inactive Val51 differs from Active only in the backbone dihedrals of Val51; the N–H points away from the active site, mimicking the inactive conformation, but the oxyanion strand position is consistent with Active. The Gly52 N–H contributes –10.8 kcal/mol in Inactive Val51-**TS2,** which is less than the contributions from Val51 in Active (–14.7 kcal/mol), but the difference in energy between **TS2** and **ES** in Inactive Val51 is smaller ($\Delta E^{\ddagger}$=9.8 kcal/mol, vs. ($\Delta E^{\ddagger}$=14.5 kcal/mol in Active) (Figure 2.5C). This results from a higher energy Inactive Val51-**ES** relative to Active-**ES**

(ΔE=5.3 kcal/mol), while the **TS2** structures have approximately the same energy (ΔE=0.7 kcal/mol). The atom positions frozen during geometry optimizations are the same between Inactive Val51 and Active; therefore, their structure energies can be directly compared, unlike the other models. The Inactive Val51 oxyanion hole stabilizes **TS4** more than in any other model (nearly twice that of Active-**TS4**). This is because a third lone pair of electrons is detected in the NBO analysis of this structure, whereas the other structures only have two electron lone pairs in **TS4**. Regardless, Inactive Val51-**TS4** is 3.4 kcal/mol higher in energy than Active-**TS4**.

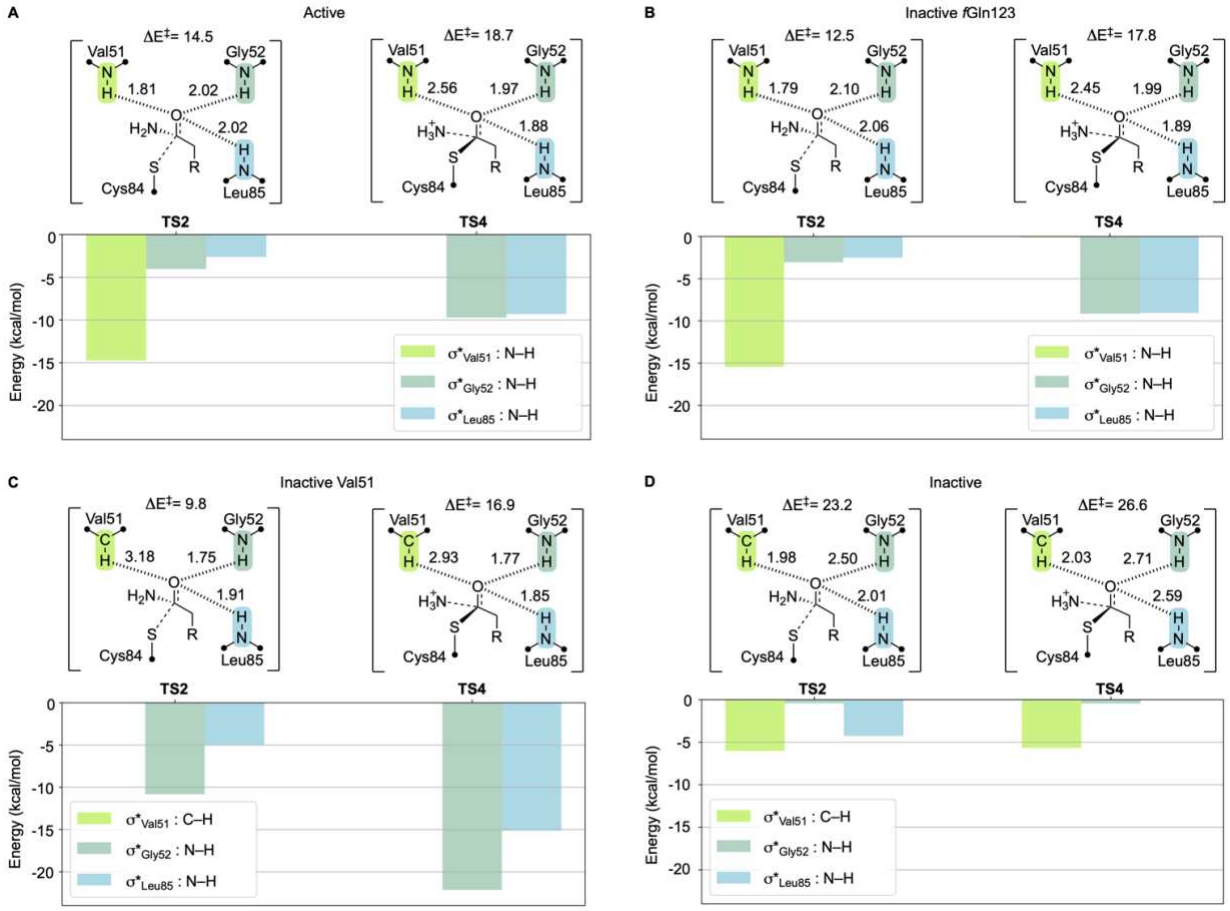


Figure 2.5. Energetic contributions to oxyanion stabilization during oxyanion formation (**TS1**) and oxyanion collapse (**TS3**). Top quadrants A and B correspond to models with the Val51 N-H pointing into the active site: Active and Inactive fGln123. Bottom quadrants C and D correspond to models with the Val51 N-H directed outwards: Inactive Val51 and Inactive.

In the initial Inactive Val51 model, the Val51 φ dihedral matches that of the optimized Inactive-**ES** structure (–138°) by construction. During geometry optimizations, the dihedral adjusts to –111° in Inactive Val51. This adjustment is consistent with the description of a third oxyanion strand conformation identified in MD simulations as an intermediate conformation of the Val51 dihedral flip.[35] The partial dihedral rotation introduces a weak interaction (less than 0.5 kcal/mol) between the Gln Oε lone pair electrons and the π* orbital of backbone Gly50 C=O. This type of n–π* interaction to stabilize an oxyanion tetrahedral intermediate has precedence in aspartic proteases[61] and, therefore, could be a catalytically relevant state in IGPS.

Although Gly52 does not contribute to oxyanion stabilization in the inactive conformation, as previously suggested, [41] repositioning of the oxyanion strand enables such stabilizing interactions, as shown in the Inactive Val51 model. In addition, the new crystallographic conformation of IGPS (7ac8 chains E/F) also results in a catalytically competent oxyanion hole geometry involving Val51. From the energetic perspective both the Active and Inactive Val51 models are consistent with experimental kinetic data.

2.2.4 Decomposition of *f*Gln123 interaction energy

The proximity of *f*Gln123 to the HisH active site is another structural feature influenced by the binding of the allosteric effector, PrFAR. It is expected that *f*Gln123 assists in recruiting and stabilizing the glutamine substrate, but it is unclear to what extent, if any, this proximity influences the chemical coordinate. The Inactive *f*Gln123 oxyanion hole contributions (Figure 2.5B) are nearly identical to Active, with an average difference of less than 0.2 kcal/mol per residue in each TS (all numerical values are listed in Table S2). Therefore, the oxyanion hole energetics are not influenced by the proximity of *f*Gln123.

To further investigate the role of *f*Gln123, the interaction energy between *f*Gln123 and the HisH active site is calculated via the ALMO–EDA procedure for each QM model (Computational Methods). The total interaction energy (Total INT) in **ES** is much more favorable when *f*Gln123 is closer to the active site. Active- and Inactive Val51-ES have similar *f*Gln123 interaction energies of –29.2 and –30.6 kcal/mol, which are much lower than the Inactive *f*Gln123 and Inactive values (–5.4 and –13.1 kcal/mol, respectively) (Figure 2.6). The difference in *f*Gln123 interaction energy between Inactive *f*Gln123 and Inactive can be reasoned by the presence of a hydrogen bond between the sidechain NH_2 group of *f*Gln123 and the backbone carbonyl oxygen of Val140 (APPENDIX A Figure A4) that is present in the Inactive model but absent in the Inactive *f*Gln123model.

Figure 2.6. Decomposition of interaction energy calculated with ALMO-EDA between fGln123 and the HisH active site in the optimized **ES** complex for each QM model.

The interaction energy in Inactive Val51 becomes slightly less favorable in **TS2**, **TS3**, and **TS4** because the Gly52 backbone shifts away from *f*Gln123 to interact more strongly with the Gln Oε. The Inactive model also shows interaction energy changes between **ES** and **TS2**, **TS3**, and **TS4**, because the Gln substrate moves farther from *f*Gln123 to react with Cys84. The Active and Inactive *f*Gln123 models have minimal structural rearrangements along the reaction coordinate; therefore, the *f*Gln123 interaction energy values are consistent across the evaluated structures.

2.3 Conclusions

This work provides new insights into local active site structural changes that can yield energetics consistent with experimentally measured allosteric rate acceleration in IGPS. The hypothesized Val51 backbone flip leads to a more favorable reaction pathway via oxyanion hole stabilization, supporting the longstanding hypothesis to explain the observed allosteric effect. However, this work reveals that this is not the only plausible configuration to yield the expected catalytic effect, as the Gly52 N–H is also a capable oxyanion hole contributor.

Oxyanion stabilization via Gly52 has been proposed previously,[41] however, most studies to date have focused on the Val51 dihedral flip. When the Val51 backbone of the Active model is manually modified to mimic the Inactive structure the Gly52 N–H becomes a suitable oxyanion hole donor. These results show that multiple active site conformations are catalytically competent and reveal the positioning of the oxyanion strand is a definitive structural change required for proper oxyanion hole formation.

Two other class-I GATs, pyridoxal 5'-phosphate (PLP, PDB: 2NV2[30]) synthase and carbamoyl phosphate synthetase (CPS, PDB: 1C3O[62]), have been crystallized in their catalytic forms with the Gln substrate bound. Upon alignment of the Gln substrate reactive carboxamide, the oxyanion strand residue involved in the preformed oxyanion holes in PLP synthase and CPS is more structurally consistent with the IGPS Gly52 than Val51 in Active-**ES**. Alignment with Inactive Val51-**ES** shows an even closer alignment, providing evidence that this manually constructed geometry of the IGPS glutaminase active site is feasible. Lastly, glutaminase activation via oxyanion strand repositioning has been proposed for another class-I GAT, aminodeoxychorismate synthase (ADCS), which displays an allosteric response similar to IGPS.[42]

The truncated active-site approach uniquely allowed for the modification of distinct structural features. Such modifications would be inaccessible or much more challenging with alternative modeling approaches (e.g., QM/MM) that explicitly treat the protein environment. This approach is generalizable to other systems in efforts to probe the energetic influence of local structural aspects and highlights the unique applications available to theozyme and cluster modeling.

2.4 Computational details

Geometry optimizations were performed using Gaussian 16, Revision C.01[63] with the range-separated, dispersion-corrected ωB97X-D[64] functional using the 6-31G*[65] basis set for all C, H, and N atoms and the 6-31+G*[66] basis set for S and O atoms. The α-carbons of selected protein residues were frozen during optimizations to conserve the active site geometry in the absence of the greater protein environment. All geometries were inspected for structural distortions that could have occurred during optimization. Vibrational frequencies were computed to confirm the nature of minima (all real normal modes) and transition structures (a single imaginary normal mode). Stationary point electronic energies were further refined at the B3LYP-D3(BJ)/6-311+G(2d,2p) level of theory and the default IEF–PCM implicit solvent method with internal parameters for diethyl ether (ε=4.24).[67–72] This procedure is commonly applied in enzymatic QM cluster modeling.[47,73,74] The GoodVibes software was used to generate electronic energy profiles.[75] The energetic span[76] between the lowest energy intermediate and the highest energy transition structure (TS) was used to estimate the activation energy. Due to the necessary frozen constraints during geometry optimizations, the vibrational partition function is not considered reliable. Therefore, we focus on the electronic energy profile (i.e., rather than Gibbs energy) throughout.

This approximation is common in truncated enzyme modeling and has shown success in mechanistic investigations.

The oxyanion hole strength in **TS2**, **TS3**, and **TS4** of every model was assessed using the Natural Bond Orbital (NBO) analysis program (version 7.0.5)[59]. NBO analysis transforms the optimized atomic orbital basis set into a localized, ideal Lewis structure basis. Second-order perturbation theory analysis of the Fock matrix in the NBO basis is used to estimate the stabilization energy associated with electron delocalization from a filled donor NBO to an unfilled acceptor NBO. The default interaction energy minimum threshold of 0.05 kcal/mol was used. The oxyanion hole interaction energies were investigated between Gln Oε lone pair electrons and the σ* orbitals of the backbone N–H of Leu85, Gly52, and Val51. Additionally, the Val51 Cα–H was investigated in Inactive Val51 and Inactive.

To better understand how *f*Gln123 influences the pathway energetics, we evaluated the interaction energy between the *f*Gln123 molecular fragment and the rest of the active site residues for the Active, Inactive *f*Gln123, Inactive Val51 and Inactive models. Absolutely Localized Molecular Orbitals based Energy Decomposition Analysis (ALMO–EDA) implemented in Q–Chem 5.4 was performed at the B3LYP-D3(BJ)/6-311+G(2d,2p) level of theory and CPCM with ε=4.24, to be consistent with the single-point energies used to produce the reaction profiles.[77,78] Within this framework, the interaction energy is defined as:

$$\Delta E_{INT} = \Delta E_{FRZ} + \Delta E_{POL} + \Delta E_{CT}.$$

E_{POL} and E_{CT} refer to the polarization and charge transfer energy components, respectively. The E_{FRZ} term combines the interaction energy contributions of the unrelaxed fragment densities, namely attractive dispersion (E_{DISP}), repulsive Pauli (E_{PAULI}) and permanent electrostatics (E_{ELEC}):

$$\Delta E_{FRZ} = \Delta E_{DISP} + \Delta E_{PAULI} + \Delta E_{ELEC}.$$

2.5 Associated Content

Raw data of all calculations and the corresponding software keywords and versions, aligned xyz coordinates of all evaluated structures, and the energy profiles of the Active and Inactive models calculated with 1) ωB97X-D/6-31G*(C,H,N);6-31+G*(S,O) 2) ωB97X-D/def2QZVPP and 3) B3LYP-D3(BJ)/6-311+G(2d,2p), are publicly available at **https://github.com/hklem/IGPS_QM_cluster_models**. Supporting information can be found in APPENDIX A: IGPS active site model components and protonation states; Construction of the Inactive Val52 and Inactive fGln123 models; NBO calculation details and data; ALMO-EDA calculation details and data; Absolute energies from geometry optimizations and single point corrections for all evaluated structures; Geometries and relevant atomic distances of each evaluated transition state; Structural comparison with PLP synthase and CPS.

REFERENCES

(1) Klem, T. J.; Davisson, V. J. Imidazole Glycerol Phosphate Synthase: The Glutamine Amidotransferase in Histidine Biosynthesis. *Biochemistry* **1993**, *32* (19), 5177–5186.

(2) Beismann-Driemeyer, S.; Sterner, R. Imidazole Glycerol Phosphate Synthase fromThermotoga Maritima. *J. Biol. Chem.* **2001**, *276* (23), 20387–20396.

(3) Massière, F.; Badet-Denisot, M.-A. The Mechanism of Glutamine-Dependent Amidotransferases. *Cell. Mol. Life Sci.* **1998**, *54* (3), 205–222.

(4) Zalkin, H.; Smith, J. L. Enzymes Utilizing Glutamine as an Amide Donor. In *Advances in Enzymology - and Related Areas of Molecular Biology*; Purich, D. L., Ed.; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 1998; pp 87–144.

(5) Douangamath, A.; Walker, M.; Beismann-Driemeyer, S.; Vega-Fernandez, M. C.; Sterner, R.; Wilmanns, M. Structural Evidence for Ammonia Tunneling across the $(\beta/\alpha)_8$ Barrel of the Imidazole Glycerol Phosphate Synthase Bienzyme Complex. *Structure* **2002**, *10*, 185–193.

(6) Kneuttinger, A. C.; Rajendran, C.; Simeth, N. A.; Bruckmann, A.; König, B.; Sterner, R. Significance of the Protein Interface Configuration for Allostery in Imidazole Glycerol Phosphate Synthase. *Biochemistry* **2020**, *59* (29), 2729–2742.

(7) Smith, A. J. T.; Müller, R.; Toscano, M. D.; Kast, P.; Hellinga, H. W.; Hilvert, D.; Houk, K. N. Structural Reorganization and Preorganization in Enzyme Active Sites: Comparisons of Experimental and Theoretically Ideal Active Site Geometries in the Multistep Serine Esterase Reaction Cycle. *J. Am. Chem. Soc.* **2008**, *130* (46), 15361–15373.

(8) Ma, S.; Devi-Kesavan, L. S.; Gao, J. Molecular Dynamics Simulations of the Catalytic Pathway of a Cysteine Protease: A Combined QM/MM Study of Human Cathepsin K. *J. Am. Chem. Soc.* **2007**, *129* (44), 13633–13645.

(9) Chittur, S. V.; Klem, T. J.; Shafer, C. M.; Davisson, V. J. Mechanism for Acivicin Inactivation of Triad Glutamine Amidotransferases. *Biochemistry* **2001**, *40* (4), 876–887.

(10) Thoden, J. B.; Miran, S. G.; Phillips, J. C.; Howard, A. J.; Raushel, F. M.; Holden, H. M. Carbamoyl Phosphate Synthetase: Caught in the Act of Glutamine Hydrolysis. *Biochemistry* **1998**, *37*, 8825--8831.

(11) Rishavy, M. A.; Cleland, W. W.; Lusty, C. J. ^{15} N Isotope Effects in Glutamine Hydrolysis Catalyzed by Carbamyl Phosphate Synthetase: Evidence for a Tetrahedral Intermediate in the Mechanism. *Biochemistry* **2000**, *39* (24), 7309–7315.

(12) Pina, A. F.; Sousa, S. F.; Cerqueira, N. M. F. S. A. The Catalytic Mechanism of Pdx2 Glutaminase Driven by a Cys-His-Glu Triad: A Computational Study. *ChemBioChem* **2022**, *23* (9), e202100555.

(13) van den Heuvel, R. H. H.; Curti, B.; Vanoni, M. A.; Mattevi, A. Glutamate Synthase: A Fascinating Pathway from L-Glutamine to L-Glutamate. *Cell. Mol. Life Sci.* **2004**, *61* (6), 669–681.

(14) Wurm, J. P.; Sung, S.; Kneuttinger, A. C.; Hupfeld, E.; Sterner, R.; Wilmanns, M.; Sprangers, R. Molecular Basis for the Allosteric Activation Mechanism of the Heterodimeric Imidazole Glycerol Phosphate Synthase Complex. *Nat. Commun.* **2021**, *12* (1), 2748.

(15) List, F.; Bocola, M.; Haeger, M. C.; Sterner, R. Constitutively Active Glutaminase Variants Provide Insights into the Activation Mechanism of Anthranilate Synthase. *Biochemistry* **2012**, *51* (13), 2812–2818.

(16) VanWart, A. T.; Eargle, J.; Luthey-Schulten, Z.; Amaro, R. E. Exploring Residue Component Contributions to Dynamical Network Models of Allostery. *J. Chem. Theory Comput.* **2012**, *8* (8), 2949–2961.

(17) Ribeiro, A. A. S. T.; Ortiz, V. Determination of Signaling Pathways in Proteins through Network Theory: Importance of the Topology. *J. Chem. Theory Comput.* **2014**, *10* (4), 1762–1769.

(18) Van Wart, A. T.; Durrant, J.; Votapka, L.; Amaro, R. E. Weighted Implementation of Suboptimal Paths (WISP): An Optimized Algorithm and Tool for Dynamical Network Analysis. *J. Chem. Theory Comput.* **2014**, *10* (2), 511–517.

(19) Negre, C. F. A.; Morzan, U. N.; Hendrickson, H. P.; Pal, R.; Lisi, G. P.; Loria, J. P.; Rivalta, I.; Ho, J.; Batista, V. S. Eigenvector Centrality for Characterization of Protein Allosteric Pathways. *Proc. Natl. Acad. Sci.* **2018**, *115* (52).

(20) Botello-Smith, W. M.; Luo, Y. Robust Determination of Protein Allosteric Signaling Pathways. *J. Chem. Theory Comput.* **2019**, *15* (4), 2116–2126.

(21) Gheeraert, A.; Pacini, L.; Batista, V. S.; Vuillon, L.; Lesieur, C.; Rivalta, I. Exploring Allosteric Pathways of a V-Type Enzyme with Dynamical Perturbation Networks. *J. Phys. Chem. B* **2019**, *123* (16), 3452–3461.

(22) Lake, P. T.; Davidson, R. B.; Klem, H.; Hocky, G. M.; McCullagh, M. Residue-Level Allostery Propagates through the Effective Coarse-Grained Hessian. *J. Chem. Theory Comput.* **2020**, *16* (5), 3385–3395.

(23) Maschietto, F.; Gheeraert, A.; Piazzi, A.; Batista, V. S.; Rivalta, I. Distinct Allosteric Pathways in Imidazole Glycerol Phosphate Synthase from Yeast and Bacteria. *Biophys. J.* **2022**, *121* (1), 119–130.

(24) Yao, X.-Q.; Hamelberg, D. Residue–Residue Contact Changes during Functional Processes Define Allosteric Communication Pathways. *J. Chem. Theory Comput.* **2022**, *18* (2), 1173–1187.

(25) Lipchock, J. M.; Loria, J. P. Nanometer Propagation of Millisecond Motions in V-Type Allostery. *Structure* **2010**, *18* (12), 1596–1607.

(26) Rivalta, I.; Sultan, M. M.; Lee, N.-S.; Manley, G. A.; Loria, J. P.; Batista, V. S. Allosteric Pathways in Imidazole Glycerol Phosphate Synthase. *Proc. Natl. Acad. Sci.* **2012**, *109* (22).

(27) Lisi, G. P.; Manley, G. A.; Hendrickson, H.; Rivalta, I.; Batista, V. S.; Loria, J. P. Dissecting Dynamic Allosteric Pathways Using Chemically Related Small-Molecule Activators. *Structure* **2016**, *24* (7), 1155–1166.

(28) Knöchel, T.; Ivens, A.; Hester, G.; Gonzalez, A.; Bauerle, R.; Wilmanns, M.; Kirschner, K.; Jansonius, J. N. The Crystal Structure of Anthranilate Synthase from *Sulfolobus Solfataricus* : Functional Implications. *Proc. Natl. Acad. Sci.* **1999**, *96* (17), 9479–9484.

(29) Tesmer, J. J. G.; Klem, T. J.; Deras, M. L.; Davisson, V. J.; Smith, J. L. The Crystal Structure of GMP Synthetase Reveals a Novel Catalytic Triad and Is a Structural Paradigm for Two Enzyme Families. *Nat. Struct. Biol.* **1996**, *3* (1), 74–86.

(30) Strohmeier, M.; Raschle, T.; Mazurkiewicz, J.; Rippe, K.; Sinning, I.; Fitzpatrick, T. B.; Tews, I. Structure of a Bacterial Pyridoxal 5′-Phosphate Synthase Complex. *Proc. Natl. Acad. Sci.* **2006**, *103* (51), 19284–19289.

(31) Morar, M.; Hoskins, A. A.; Stubbe, J.; Ealick, S. E. Formylglycinamide Ribonucleotide Amidotransferase from *Thermotoga Maritima:* Structural Insights into Complex Formation. *Biochemistry* **2008**, *47* (30), 7816–7830.

(32) Hongmin, L.; Ryan, T. J.; Chave, K. J.; Van Roey, P. Three-Dimensional Structure of Human -Glutamyl Hydrolase. *J. Biol. Chem.* **2002**, *277* (27), 24522–24529.

(33) Rivalta, I.; Lisi, G. P.; Snoeberger, N.-S.; Manley, G.; Loria, J. P.; Batista, V. S. Allosteric Communication Disrupted by a Small Molecule Binding to the Imidazole Glycerol Phosphate Synthase Protein–Protein Interface. *Biochemistry* **2016**, *55* (47), 6484–6494.

(34) Lipchock, J.; Loria, J. P. Millisecond Dynamics in the Allosteric Enzyme Imidazole Glycerol Phosphate Synthase (IGPS) from Thermotoga Maritima. *J. Biomol. NMR* **2009**, *45* (1–2), 73–84.

(35) Calvó-Tusell, C.; Maria-Solano, M. A.; Osuna, S.; Feixas, F. Time Evolution of the Millisecond Allosteric Activation of Imidazole Glycerol Phosphate Synthase. *J. Am. Chem. Soc.* **2022**, *144* (16), 7146–7159.

(36) Myers, R. S.; Amaro, R. E.; Luthey-Schulten, Z. A.; Davisson, V. J. Reaction Coupling through Interdomain Contacts in Imidazole Glycerol Phosphate Synthase. *Biochemistry* **2005**, *44* (36), 11974–11985.

(37) Chaudhuri, B. N.; Lange, S. C.; Myers, R. S.; Davisson, V. J.; Smith, J. L. Toward Understanding the Mechanism of the Complex Cyclization Reaction Catalyzed by Imidazole Glycerol Phosphate Synthase: Crystal Structures of a Ternary Complex and the Free Enzyme. *Biochemistry* **2003**, *42* (23), 7003–7012.

(38) Myers, R. S.; Jensen, J. R.; Deras, I. L.; Smith, J. L.; Davisson, V. J. Substrate-Induced Changes in the Ammonia Channel for Imidazole Glycerol Phosphate Synthase. *Biochemistry* **2003**, *42* (23), 7013–7022.

(39) Chaudhuri, B. N.; Lange, S. C.; Myers, R. S.; Chittur, S. V.; Davisson, V. J.; Smith, J. L. Crystal Structure of Imidazole Glycerol Phosphate Synthase: A Tunnel through a $(\beta/\alpha)_8$ Barrel Joins Two Active Sites. *Structure* **2001**, *10*, 987--997.

(40) Korolev, S.; Skarina, T.; Evdokimova, E.; Beasley, S.; Edwards, A.; Joachimiak, A.; Savchenko, A. Crystal Structure of Glutamine Amidotransferase fromThermotoga Maritima. *Proteins Struct. Funct. Genet.* **2002**, *49* (3), 420–422.

(41) List, F.; Vega, M. C.; Razeto, A.; Häger, M. C.; Sterner, R.; Wilmanns, M. Catalysis Uncoupling in a Glutamine Amidotransferase Bienzyme by Unblocking the Glutaminase Active Site. *Chem. Biol.* **2012**, *19* (12), 1589–1599.

(42) Semmelmann, F.; Straub, K.; Nazet, J.; Rajendran, C.; Merkl, R.; Sterner, R. Mapping the Allosteric Communication Network of Aminodeoxychorismate Synthase. *J. Mol. Biol.* **2019**, *431* (15), 2718–2728.

(43) Klem, H.; McCullagh, M.; Paton, R. S. Modeling Catalysis in Allosteric Enzymes: Capturing Conformational Consequences. *Top. Catal.* **2022**, *65* (1–4), 165–186.

(44) Glowacki, D. R.; Harvey, J. N.; Mulholland, A. J. Taking Ockham's Razor to Enzyme Dynamics and Catalysis. *Nat. Chem.* **2012**, *4* (3), 169–176.

(45) Ryde, U. How Many Conformations Need To Be Sampled To Obtain Converged QM/MM Energies? The Curse of Exponential Averaging. *J. Chem. Theory Comput.* **2017**, *13* (11), 5745–5752.

(46) Masgrau, L.; Truhlar, D. G. The Importance of Ensemble Averaging in Enzyme Kinetics. *Acc. Chem. Res.* **2015**, *48* (2), 431–438.

(47) Siegbahn, P. E. M.; Himo, F. The Quantum Chemical Cluster Approach for Modeling Enzyme Reactions: The Quantum Chemical Cluster Approach. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *1* (3), 323–336.

(48) Himo, F.; de Visser, S. P. Status Report on the Quantum Chemical Cluster Approach for Modeling Enzyme Reactions. *Commun. Chem.* **2022**, *5* (1), 29.

(49) Schrödinger, LLC and Warren Delano. PyMol, 2020.

(50) Summers, T. J.; Daniel, B. P.; Cheng, Q.; DeYonker, N. J. Quantifying Inter-Residue Contacts through Interaction Energies. *J. Chem. Inf. Model.* **2019**, *59* (12), 5034–5044.

(51) Vennelakanti, V.; Nazemi, A.; Mehmood, R.; Steeves, A. H.; Kulik, H. J. Harder, Better, Faster, Stronger: Large-Scale QM and QM/MM for Predictive Modeling in Enzymes and Proteins. *Curr. Opin. Struct. Biol.* **2022**, *72*, 9–17.

(52) Sumner, S.; Söderhjelm, P.; Ryde, U. Effect of Geometry Optimizations on QM-Cluster and QM/MM Studies of Reaction Energies in Proteins. *J. Chem. Theory Comput.* **2013**, *9* (9), 4205–4214.

(53) Liao, R.-Z.; Thiel, W. Convergence in the QM-Only and QM/MM Modeling of Enzymatic Reactions: A Case Study for Acetylene Hydratase. *J. Comput. Chem.* **2013**, 2389–2397.

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

(55) Kulik, H. J.; Zhang, J.; Klinman, J. P.; Martínez, T. J. How Large Should the QM Region Be in QM/MM Calculations? The Case of Catechol *O* -Methyltransferase. *J. Phys. Chem. B* **2016**, *120* (44), 11381–11394.

(56) Masson, P.; Bec, N.; Froment, M.-T.; Nachon, F.; Balny, C.; Lockridge, O.; Schopfer, L. M. Rate-Determining Step of Butyrylcholinesterase-Catalyzed Hydrolysis of Benzoylcholine and Benzoylthiocholine. Volumetric Study of Wild-Type and D70G Mutant Behaviour. *Eur. J. Biochem.* **2004**, *271* (10), 1980–1990.

(57) Lisi, G. P.; Currier, A. A.; Loria, J. P. Glutamine Hydrolysis by Imidazole Glycerol Phosphate Synthase Displays Temperature Dependent Allosteric Activation. *Front. Mol. Biosci.* **2018**, *5*, 4.

(58) Tantillo, D. J.; Jiangang, C.; Houk, K. N. Theozymes and Compuzymes: Theoretical Models for Biological Catalysis. *Curr. Opin. Chem. Biol.* **1998**, *2* (6), 743–750.

(59) Windsor, I. W.; Gold, B.; Raines, R. T. An $n \rightarrow \pi^*$ Interaction in the Bound Substrate of Aspartic Proteases Replicates the Oxyanion Hole. *ACS Catal.* **2019**, *9* (2), 1464–1471.

(60) Thoden, J. B.; Huang, X.; Raushel, F. M.; Holden, H. M. The Small Subunit of Carbamoyl Phosphate Synthetase: Snapshots along the Reaction Pathway. *Biochemistry* **1999**, *38* (49), 16158–16166.

(61) Gaussian 16,Revision C.01,M.J.Frisch, G.W.Trucks, H.B.Schlegel, G.E. Scuseria, M.A.Robb, J.R.Cheeseman, G.Scalmani, V.Barone, G.A. Petersson, H.Nakatsuji, X.Li,M.Caricato, A.V.Marenich, J.Bloino, B.G. Janesko, R.Gomperts, B.Mennucci, H.P.Hratchian, J.V.Ortiz, A.F. Izmaylov, J.L.Sonnenberg, D.Williams-Young, F.Ding,F.Lipparini, F. Egidi,J.Goings, B.Peng, A.Petrone, T.Henderson, D.Ranasinghe, V.G. Zakrzewski, J.Gao,N.Rega, G.Zheng, W.Liang, M.Hada, M.Ehara, K. Toyota, R.Fukuda, J.Hasegawa, M.Ishida, T.Nakajima, Y.Honda, O. Kitao,H.Nakai, T.Vreven, K.Throssell, J.A.Montgomery, Jr.,J.E.Peralta, F.Ogliaro, M.J.Bearpark, J.J.Heyd, E.N.Brothers, K.N.Kudin, V.N. Staroverov, T.A.Keith, R.Kobayashi, J.Normand, K.Raghavachari, A.P. Rendell, J.C.Burant, S.S.Iyengar, J.Tomasi, M.Cossi, J.M.Millam, M. Klene, C.Adamo, R.Cammi, J.W.Ochterski, R.L.Martin, K.Morokuma, O.Farkas, J.B.Foresman, DJ.Fox,Gaussian, Inc.,Wallingford CT, 2016.

(62) Chai, J.-D.; Head-Gordon, M. Long-Range Corrected Hybrid Density Functionals with Damped Atom–Atom Dispersion Corrections. *Phys. Chem. Chem. Phys.* **2008**, *10* (44), 6615.

(63) Hehre, W. J.; Ditchfield, R.; Pople, J. A. Self—Consistent Molecular Orbital Methods. XII. Further Extensions of Gaussian—Type Basis Sets for Use in Molecular Orbital Studies of Organic Molecules. *J. Chem. Phys.* **1972**, *56* (5), 2257–2261.

(64) Hariharan, P. C.; Pople, J. A. The Influence of Polarization Functions on Molecular Orbital Hydrogenation Energies. *Theor. Chim. Acta* **1973**, *28* (3), 213–222.

(65) Becke, A. D. Density-functional Thermochemistry. III. The Role of Exact Exchange. *J. Chem. Phys.* **1993**, *98* (7), 5648–5652.

(66) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti Correlation-Energy Formula into a Functional of the Electron Density. *Phys. Rev. B* **1988**, *37* (2), 785–789.

(67) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A Consistent and Accurate *Ab Initio* Parametrization of Density Functional Dispersion Correction (DFT-D) for the 94 Elements H-Pu. *J. Chem. Phys.* **2010**, *132* (15), 154104.

(68) Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the Damping Function in Dispersion Corrected Density Functional Theory. *J. Comput. Chem.* **2011**, *32* (7), 1456–1465.

(69) Luque, F. J.; López, J. M.; Orozco, M. Perspective on "Electrostatic Interactions of a Solute with a Continuum. A Direct Utilization of Ab Initio Molecular Potentials for the Prevision of Solvent Effects." *Theor. Chem. Acc. Theory Comput. Model. Theor. Chim. Acta* **2000**, *103* (3–4), 343–345.

(70) Cossi, M.; Barone, V.; Cammi, R.; Tomasi, J. Ab Initio Study of Solvated Molecules: A New Implementation of the Polarizable Continuum Model. *Chem. Phys. Lett.* **1996**, *255* (4–6), 327–335.

(71) Himo, F. Recent Trends in Quantum Chemical Modeling of Enzymatic Reactions. *J. Am. Chem. Soc.* **2017**, *139* (20), 6780–6786.

(72) Blomberg, M. R. A. How Quantum Chemistry Can Solve Fundamental Problems in Bioenergetics. *Int. J. Quantum Chem.* **2015**, *115* (18), 1197–1201.

(73) Luchini, G.; Alegre-Requena, J. V.; Funes-Ardoiz, I.; Paton, R. S. GoodVibes: Automated Thermochemistry for Heterogeneous Computational Chemistry Data. *F1000Research* **2020**, *9*, 291.

(74) Kozuch, S.; Shaik, S. How to Conceptualize Catalytic Cycles? The Energetic Span Model. *Acc. Chem. Res.* **2011**, *44* (2), 101–110.

(75) Glendening, E. D.; Landis, C. R.; Weinhold, F. *NBO 7.0* : New Vistas in Localized and Delocalized Chemical Bonding Theory. *J. Comput. Chem.* **2019**, jcc.25873.

(76) Horn, P. R.; Mao, Y.; Head-Gordon, M. Probing Non-Covalent Interactions with a Second Generation Energy Decomposition Analysis Using Absolutely Localized Molecular Orbitals. *Phys. Chem. Chem. Phys.* **2016**, *18* (33), 23067–23079.

(77) Epifanovsky, E.; Gilbert, A. T. B.; Feng, X. Software for the Frontiers of Quantum Chemistry: An Overview of Developments in the Q-Chem 5 Package. *J. Chem Phys.* **2021**, 155, 084801.

CHAPTER 3: SIZE-AND-SHAPE SPACE GAUSSIAN MIXTURE MODELS FOR STRUCTURAL CLUSTERING OF MOLECULAR DYNAMICS TRAJECTORIES[§]

Determining the optimal number and identity of structural clusters from an ensemble of molecular configurations continues to be a challenge. Recent structural clustering methods have focused on the use of internal coordinates due to the innate rotational and translational invariance of these features. The vast number of possible internal coordinates necessitates a feature space supervision step to make clustering tractable but yields a protocol that can be system type specific. Particle positions offer an appealing alternative to internal coordinates but suffer from a lack of rotational and translational invariance, as well as a perceived insensitivity to regions of structural dissimilarity. Here, we present a method, denoted shapeGMM, that overcomes the shortcomings of particle positions using a weighted maximum likelihood alignment procedure. This alignment strategy is then built into an expectation maximization Gaussian mixture model (GMM) procedure to capture metastable states in the free-energy landscape. The resulting algorithm distinguishes between a variety of different structures, including those indistinguishable by root-mean-square displacement and pairwise distances, as demonstrated on several model systems. Shape-GMM results on an extensive simulation of the fast-folding HP35 Nle/Nle mutant protein support a four-state folding/unfolding mechanism, which is consistent with previous experimental results and provides kinetic details comparable to previous state-of-the art clustering approaches, as measured

[§]Reproduced with permission from **Klem, H**.; Hocky, G. M.; McCullagh, M. Size-and-Shape Space Gaussian Mixture Models for Structural Clustering of Molecular Dynamics Trajectories. *J. Chem. Theory Comput.* **2022**, *18* (5), 3218–3230.

by the VAMP-2 score. Currently, training of shape-GMMs is recommended for systems (or subsystems) that can be represented by $\lesssim$200 particles and $\lesssim$100k configurations to estimate high-dimensional covariance matrices and balance computational expense. Once a shape-GMM is trained, it can be used to predict the cluster identities of millions of configurations.

3.1 Introduction

Structural clustering of molecular dynamics (MD) simulation data of macromolecules is often necessary to draw physical conclusions about molecular mechanisms from large amounts of simulation data. Consider, for example, the challenge of protein folding. At or near its folding temperature, a protein will populate both folded and unfolded states. Molecular simulations used to probe this process yield a trajectory of atomic positions that approximate the equilibrium conformational ensemble of the molecule. To study the mechanisms of folding and unfolding, it is useful to discretize the protein configurations that are observed within this trajectory. Which configurations are folded, which are unfolded? Are there multiple folded and/or unfolded states? Are there metastable states along the folding pathway(s)?

To address these, and other similar questions, it is helpful to employ automated structural clustering methods on the trajectory data. In time-continuous simulations, kinetic clustering, such as spectral clustering of the transition matrix,[1–3] can be employed (see refs[4–6] for comparisons between kinetic and structural clustering). Alternatively, clustering solely on the basis of structural data (structural clustering) can be applied to either time-continuous or disjoint simulation data. Here, we focus on developing a robust structural clustering procedure that can be readily applied to a broad range of systems and depends on only a small number of user-specified, readily understandable, parameter choices. Structural clustering requires the choice of both a clustering algorithm and the features on which to perform the clustering. We start our discussion with the

choice of features, which can be separated into two categories: internal coordinates and particle positions.

The internal coordinates of a macromolecule are an appealing choice of features because they are invariant to both translation and rotation. These coordinates include distances (two-body terms), angles (three-body terms), dihedral angles (four-body terms), and higher order terms not often considered for clustering purposes. The inclusion of even all two-body terms yields a feature space that drastically overdetermines the system and hinders the ability to cluster in this space. The possible solutions to the overdetermination problem include the use of a priori knowledge of the system to develop a hypothesis-driven selection of key internal coordinates, segmentation of the trajectory,[7] combination of both kinetic and structural information,[8] and unsupervised dimensionality reduction techniques. We note that applying dimensionality reduction to internal coordinates is nontrivial as projection errors can misconstrue data.[9] Recent efforts, even methods that combine kinetic and structural information, have focused on choosing a subset of features, for example, the backbone dihedral angles of proteins, that are able to distinguish between perceived important structural states of the macromolecule of interest.[10,11] This supervision step, however, yields a protocol that is specific to the system/problem of interest. Thus, it remains a challenge to use internal coordinates in a general structural clustering workflow.

Particle positions offer an appealing alternative choice of features because they do not drastically overdetermine a system. A major limitation of particle positions is that they are defined in the lab-frame and thus are not invariant to translation and rotation.[11] Previously, structural alignment based on root-mean-square displacement (RMSD) has been used to overcome rotational and translational invariance, and subsequent RMSD between structures has been used as a distance metric in clustering algorithms.[4,12] This has led to the conclusion that particle positions have a

second disadvantage: RMSD between structures is a poor estimate of similarity for systems in which there is regional heterogeneity in particle fluctuations.[4,12] Below, we will demonstrate that this conclusion is due to the assumption, implicit in RMSD alignment, that particles vary in independent, equivalent, and spherical distributions.

Numerous clustering algorithms have been applied to structural data from MD simulations, but the connection between the resulting clusters and the underlying assumptions of the algorithm are often obscure. Recent efforts have focused on the use of nonhierarchical algorithms applied to a reduced dimensional subspace of the original features. These algorithms include k-means,[13–15] Gaussian mixture models (GMMs),[16] and density-based clustering schemes.[17–20] Unsupervised dimensionality reduction has been achieved using principal component analysis, time-lagged independent component analysis, sketch map,[21] UMAP,[22] or a variety of other methods.[6] The combination of internal coordinates (e.g. distances, angles, and/or dihedrals) and dimensionality reduction schemes makes it challenging to predict the expected form of the distribution in these spaces. This has led to the popularity of density-based clustering schemes. The downside of these schemes are twofold: first, there is a perception that they can only be applied in fairly low-dimensional spaces, and, second, there is no strong theoretical connection between the clustering algorithm and the physics governing the resulting conformational clusters.

In this paper, we present shape-GMM, a method to cluster macromolecule trajectory data directly using particle positions. Previous shortcomings of RMSD clustering are overcome using a more general Mahalanobis distance and a weighted maximum likelihood (ML) alignment procedure.[23] The shapeGMM algorithm incorporates this rigorous alignment strategy into an expectation maximization (EM) GMM procedure. The GMM approach is conceptually appealing because it resembles a second-order approximation to the expected probability density of particle

positions. The ability of shape-GMM to correctly cluster molecular structures is first assessed using a set of elastic network model (ENM) systems with known "ground truths". Subsequently, the real-world applicability of this method is demonstrated on a challenging, well-studied protein folding system, and the results are shown to be consistent with previous experimental and theoretical studies. Combined, this work establishes shape-GMM as an appealing clustering approach applicable to a broad range of macromolecular structures.

3.2 Theory and Methods

3.2.1 Molecular size and shape

We consider a MD simulation of a macromolecule in solution. From a single frame, the macromolecule is represented by N atoms considered important (e.g., all proteins, all heavy atoms, just C_α atoms, *etc*.) or an N particle coarse-grained mapping (e.g., the centers of mass of each protein residue). These features are a point in $\mathbb{R}^3$ encoded by a matrix, $\boldsymbol{x}$, of order $N \times 3$. The Hamiltonian for any system considered here is independent of the lab-frame, and so the features of interest, $[\boldsymbol{x}_i]$, are the orbits of all possible rigid-body transformations of $\boldsymbol{x}_i$. This is written as an equivalence class

$$[\boldsymbol{x}_i] = \left\{\boldsymbol{x}_i \boldsymbol{R}_i + \mathbf{1}_N \vec{\xi}_i^T : \vec{\xi}_i \in \mathbb{R}^3, \boldsymbol{R}_i \in \mathrm{SO}(3)\right\}, \tag{3.1}$$

where $\vec{\xi}_i$ is a translation in $\mathbb{R}^3$, $\boldsymbol{R}_i$ is a rotation $\mathbb{R}^3 \rightarrow \mathbb{R}^3$, and $\mathbf{1}_N$ is the N × 1 vector of ones. The features, $[\boldsymbol{x}_i]$, exist as a point in size-and-shape space.[24] Size-and-shape space has dimensions $3N - 6$ and is defined as $\mathrm{S}\sum_N^3 = \mathbb{R}^{3N}/\mathrm{G}$, where $\mathrm{G} = \mathbb{R}^3 \times \mathrm{SO}(3)$ is the group of all rigid-body transformations for each frame with elements $\mathrm{g} = \left(\vec{\xi}, \boldsymbol{R}\right)$.

For a trajectory consisting of M frames or M molecular configurations, the set of recorded points $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$ is effectively regarded as a set of orbits

$$\{[\boldsymbol{x}_1], \ldots, [\boldsymbol{x}_M]\} = \{\mathrm{g_i}^{-1}\boldsymbol{x}_i : \mathrm{g_i} \in \mathrm{G}\}. \tag{3.2}$$

The group G is said to act freely, or G^M acts component-wise, one group element per frame.

3.2.2 Gaussian distribution of positions

If we consider the macromolecule to be in a single free-energy minimum, the approximate second-order approximation to the probability density is Gaussian in $\mathbb{R}^{3N}$. The normalized multivariate Gaussian distribution is given as

$$N(\boldsymbol{x}_i|\boldsymbol{\mu},\Sigma) = \frac{\exp\left[-\frac{1}{2}(\mathrm{g}_i^{-1}\boldsymbol{x}_i-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathrm{g}_i^{-1}\boldsymbol{x}_i-\boldsymbol{\mu})\right]}{\sqrt{(2\pi)^{(3N)}\det\boldsymbol{\Sigma}}}, \tag{3.3}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the mean and covariance ($3N \times 3N$), respectively, and the multiplication inside the exponent requires a flattening of the $(\mathrm{g}_i^{-1}\boldsymbol{x}_i - \boldsymbol{\mu})$ matrix. The estimation of well-defined average and covariance matrices from an MD trajectory requires determining the appropriate set of rigid body transformations, $(\mathrm{g}_1, \ldots, \mathrm{g}_M)$, for which we define a ML procedure.

3.2.3 ML Alignment. The following ML alignment is adapted from Theobald and Wuttke[25] in two ways. First, we generalize the allowed form of the covariance and, second, we do not enforce an inverse γ distribution of the eigenvalues of the covariance because we expect sufficient sampling of this matrix from the MD trajectory. The log likelihood of a trajectory alignment is given as

$$\ln(L) = \sum_{i=1}^{M} \ln\left(N(\boldsymbol{x}_i|\boldsymbol{\mu},\boldsymbol{\Sigma})\right). \tag{3.4}$$

There are two distinct components for the parameters in a single trajectory ML alignment:

1. the alignment parameters or group elements $(\mathrm{g}_1, \ldots, \mathrm{g}_M)$ and
2. the mean configuration, $\boldsymbol{\mu}$, and the covariance matrix $\boldsymbol{\Sigma}$.

The translation alignment parameters are easily resolved by removing the center of geometry from each configuration. Methods to determine optimal rotation matrices have only been identified for covariances of the form $\boldsymbol{\Sigma} \propto \boldsymbol{I}_{3N}$[26,27] or $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_N \otimes \boldsymbol{I}_3$,[23,28] where $\otimes$ is a Kronecker

product and $\boldsymbol{\Sigma}_N$ is a $N \times N$ covariance matrix. We note that a covariance of the form $\boldsymbol{\Sigma} \propto \boldsymbol{I}_{3N}$ assumes that particles oscillate in uncoupled, identical, spherical distributions. Here, we will employ the forms $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_N \otimes \boldsymbol{I}_3$("weighted") and $\boldsymbol{\Sigma} \propto \boldsymbol{I}_{3N}$ ("uniform").

The problem of finding the parameter values that maximize the likelihood can be solved by iterating between two steps until the log likelihood (Equation 3.4) has converged within some threshold.

1. Given $\mathrm{g}_1, \ldots, \mathrm{g}_M$, estimate $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ by

$$\widehat{\boldsymbol{x}}_i = \mathrm{g}_i^{-1}\boldsymbol{x}_i \tag{3.5}$$

$$\widehat{\boldsymbol{\mu}} = \sum_{i=1}^{M} \frac{\widehat{\boldsymbol{x}}_i}{M} \tag{3.6}$$

$$\widehat{\boldsymbol{\Sigma}}_N = \sum_{i=1}^{M} \frac{(\widehat{\boldsymbol{x}}_i - \widehat{\boldsymbol{\mu}})(\widehat{\boldsymbol{x}}_i - \widehat{\boldsymbol{\mu}})^T}{3(M-1)} \tag{3.7}$$

For the submodel in which $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}_{3N}$, it suffices to take $\hat{\sigma}^2 = Tr\left(\hat{\Sigma}_N\right)/N$.

2. Given $\widehat{\boldsymbol{\mu}}$, $\widehat{\boldsymbol{\Sigma}}$ and an appropriate inverse $\boldsymbol{W} = \boldsymbol{\Sigma}^{-1}$, estimate the alignments $\mathrm{g}_1, \ldots, \mathrm{g}_M$ by minimizing the Mahalanobis distance

$$\left\|\mathrm{g}_i^{-1}\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}\right\|^2 = \mathrm{Tr}\left((\mathrm{g}_i^{-1}\boldsymbol{x}_i - \boldsymbol{\mu})^T \boldsymbol{W} (\mathrm{g}_i^{-1}\boldsymbol{x}_i - \boldsymbol{\mu})\right) \tag{3.8}$$

with respect to g_i.

The second step is the generalized Procrustes problem, which can be solved either by using the singular value decomposition[23,25,26,28] or by quaternion methods.[27,29,30] For the case that $\boldsymbol{\Sigma} \propto \boldsymbol{I}_{3N}$ ("uniform"), Equation 3.8 simplifies to the mean-square displacement. Thus, our uniform covariance model is equivalent to an RMSD alignment and distance metric.

If $\boldsymbol{\Sigma}_N$ is unrestricted, the estimate $\widehat{\boldsymbol{\Sigma}}_N$ has rank $N - 1$ and kernel $\mathbf{1}$. In that case, $\boldsymbol{W}$ is the spectral inverse which also has kernel $\mathbf{1}$.

3.2.4 GMM for size-and-shape features

In a nonharmonic force field description of a globular macromolecule, multiple metastable configurations will be encountered. It is natural to consider the probability density in this feature space as a mixture of K multivariate Gaussians[31]

$$P(\boldsymbol{x}_i) = \sum_{j=1}^{K} \phi_j N\left(\boldsymbol{x}_i | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right), \quad (3.9)$$

where ϕ_j is the weight $\left(\sum_{j=1}^{K} \phi_j = 1\right)$, $\boldsymbol{\mu}_j$ is the mean, and $\boldsymbol{\Sigma}_j$ is the covariance of the *j*th Gaussian probability density. Equation 3.9 will be a good approximation for the probability density near local minima, the preferentially sampled configurations in conventional MD. This approximation will fail near transition states although adaptations to this formulation have been developed to account for this.[16]

Equation 3.9 is known as a GMM, and the goal of determining the optimal $\{\hat{\phi}_j\}$, $\{\hat{\boldsymbol{\mu}}_j\}$, and $\{\hat{\boldsymbol{\Sigma}}_j\}$ that fit observations has been achieved previously with algorithms including EM with a ML criterion or Bayesian approaches including variational inference or sampling techniques (*e.g.*, Markov chain Monte Carlo[32] or Gibbs sampling[33]). Any of these approaches can be applied to the trajectory features, $\{[\boldsymbol{x}_i]\}$, but care must be taken to correctly account for the feature equivalences.

GMM: Expectation Maximization. Within the EM framework, for a fixed number of Gaussians, *K*, the following steps are taken.[34] Note that in this work two choices are available for step 3, depending on the approximation to the form of $\boldsymbol{\Sigma}$ chosen, as described in the previous section.

1. Provide initial guesses for $\{\hat{\phi}_j\}$, $\{\hat{\boldsymbol{\mu}}_j\}$, and $\{\hat{\boldsymbol{\Sigma}}_j\}$.
2. **Expectation**: estimate the posterior distribution for the latent variable of mixture components, $Z_i \in \{1, \ldots, K\}$

$$\gamma_{Z_i}(j) = \frac{\hat{\phi}_j N(x_i | \hat{\mu}_j, \hat{\Sigma}_j)}{\sum_{j=1}^{K} \hat{\phi}_j N(x_i | \hat{\mu}_j, \hat{\Sigma}_j)} \quad (3.10)$$

3. **Maximization**: update $\{\hat{\phi}_j\}$, $\{\hat{\boldsymbol{\mu}}_j\}$, and $\{\hat{\boldsymbol{\Sigma}}_j\}$.

$$\hat{\phi}_j = \frac{\sum_{i=1}^{M} \gamma_{Z_i}(j)}{M} \tag{3.11}$$

$$\hat{\boldsymbol{\mu}}_j = \frac{\sum_{i=1}^{M} \gamma_{Z_i}(j) \mathrm{g}_{i,j}^{-1} \boldsymbol{x}_i}{\sum_{i=1}^{M} \gamma_{Z_i}(j)} \tag{3.12}$$

$$\hat{\boldsymbol{\Sigma}}_j = \begin{cases} \frac{\sum_{i=1}^{M} \gamma_{Z_i}(j) \langle \hat{\sigma}^2 \rangle_i}{\sum_{i=1}^{M} \gamma_{Z_i}(j)} \boldsymbol{I}_{3N} & \text{uniform} \\ \frac{\sum_{i=1}^{M} \gamma_{Z_i}(j) \langle \hat{\boldsymbol{\Sigma}}_N \rangle_i}{\sum_{i=1}^{M} \gamma_{Z_i}(j)} \otimes \boldsymbol{I}_3 & \text{weighted} \end{cases} \tag{3.13}$$

4. Iterate steps 2–3 until the log likelihood converges within some threshold. Log likelihood is defined as

$$\ln(L) = \sum_{i=1}^{M} \ln\left(\sum_{j=1}^{K} \hat{\phi}_j N\left(\boldsymbol{x}_i | \hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j\right)\right). \tag{3.14}$$

Shape-GMM Procedure. Shape-GMM adapts the EM algorithm for size-and-shape space by including trajectory alignments in both the expectation and maximization steps:

1. In the expectation step, the distribution $N\left(\boldsymbol{x}_i | \hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j\right)$ is determined for each frame after alignment (either uniform or weighted) to the respective average. This requires K noniterative trajectory alignments.
2. In the maximization step, the means and covariances are updated using the ML alignment described in Section 3.2.3. This requires K iterative trajectory alignments.

Model Initialization. The EM procedure must be initialized by providing guesses for $\{\hat{\phi}_j\}$, $\{\hat{\boldsymbol{\mu}}_j\}$, and $\{\hat{\boldsymbol{\Sigma}}_j\}$. This can be achieved in a variety of ways including dividing the trajectory into K parts, using a *k*-means algorithm, or randomly selecting initial frames as averages and assigning frames to their nearest (measured by RMSD) cluster center. All of these procedures are implemented in our code, but we use the random frame initialization for all examples unless explicitly stated otherwise.

Assigning Frames to Clusters. Within the context of clustering from a GMM, a frame (or data point) is assigned to the cluster in which it has the largest likelihood. This is no different in size-and-shape space.

Determining the Number of Clusters. The number of clusters, *K*, must be specified in order to perform shapeGMM. The best choice of K will be system-specific and is a challenge to determine for any clustering method. Here, we use the elbow method coupled with cross-validation (CV). The elbow method is a heuristic approach to identify a significant change in slope in the log likelihood (in the case of GMMs) as a function of number of clusters. CV is a method of applying the trained model on a validation (nontrained) data set. This is done as a function of number of clusters, and an elbow in this curve is used to pick the number of clusters. Examples of these plots are given for systems in the Results and Discussion section and Supporting Information.

For CV, we use a simple scheme to break up the data into a training set and a validation set. The training set is randomly (uniformly) chosen from the entire data set, and the remaining data are used for validation. The number of frames chosen for training is system-specific (but at least 1000 for all systems studied here) and is given in the Results and Discussion section for each system.

Implementation. The current implementation of shape-GMM is written in Python with acceleration/parallelization using the Numba package.[35] The shape-GMM package is written to mimic the interface of the Scikit-learn Gaussian mixture package.[36] The package can be easily installed using pip (pip install shapeGMM). All codes as well as examples of the full analyses performed below are available from https://github.com/mccullaghlab/GMM-Positions.

3.3 Results and Discussion

We explore the shape-GMM clustering method using a variety of examples. The first example we consider is a set of ENMs designed to probe the ability of both uniform and weighted covariance versions of shape-GMM to distinguish between structures for a known clustering. Subsequently, we examine the performance on one toy and one real example of models where molecular topology is free to change, to demonstrate the ability of weighted shape-GMM to estimate the number and identity of configurational clusters in the data. Full descriptions of MD procedures for each system are provided in Section 3.5.

3.3.1 Elastic network models

In order to rigorously compare and contrast clustering protocols, we create amalgamated trajectories of various ENMs. The force constants and topologies of the ENMs are chosen such that each ENM will populate a single conformation. Specifically, we consider a variety of ENM models with topologies motivated by the protein secondary structure elements. To further mimic the nature of the combined protein structural elements, we utilize an anisotropic network model (ANM). The Hamiltonian for our ANM model is given as

$$H(x) = \sum_{i,j>i} k_{ij}\left(\left|\vec{r}_{ij}\right| - \left|\vec{r}_{ij}^{0}\right|\right)^2 \tag{3.15}$$

where $\left|\vec{r}_{ij}\right| = \left|\vec{x}_i - \vec{x}_j\right|$ is the distance between particles i and j, $\left|\vec{r}_{ij}^{0}\right|$ is the distance between particles i and j in the reference geometry, and the force constants:

$$k_{ij} = \begin{cases} 100 & j = i+1 \\ 20 & \left|\vec{r}_{ij}^{0}\right| = 1.71 \text{ Å} \\ 10\, e^{-\left|\left|\vec{r}_{ij}^{0}\right| - 1.71\right|} & otherwise \end{cases} \tag{3.16}$$

where k_{ij} has units of kcal·mol^{-1}·Å^{-2}. The strongest force constant is reserved for primary sequence bonds; "hydrogen bonds" are given the second strongest force constant and are set at a separation distance of 1.7 Å; and all remaining pairwise interactions are given an exponentially decaying distance-dependent force constant. A variety of topologies and system sizes of these

types of ANMs are used in the subsequent subsections. Examples of the types of topologies are depicted in Figure 3.1A. We note that because shape-GMM relies only on positions, it applies to these coarse-grained models without modification, whereas many state-of-the art clustering methods designed for protein data either require temporal information or are applied on backbone dihedral angles, and hence we cannot compare to them here.

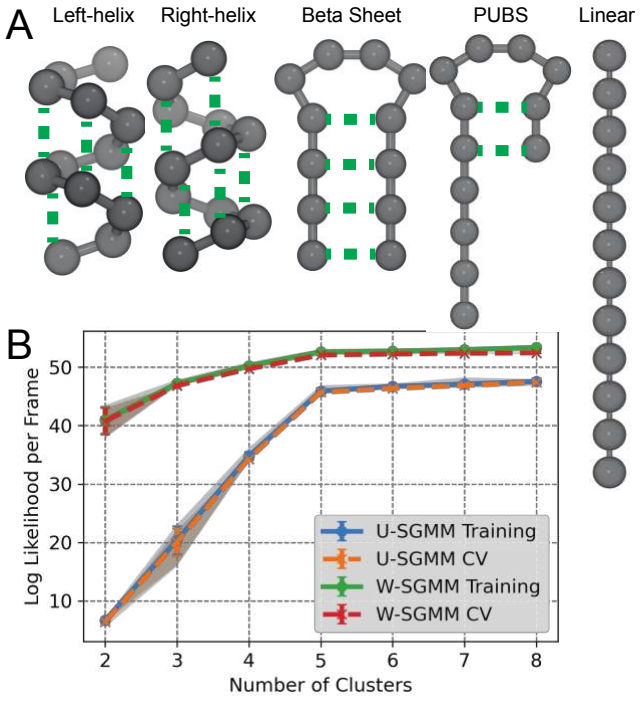


Figure 3.1. Clustering of an amalgamated trajectory of five ANMs of a 12-bead system. **A**) Schematic of the five topologies considered, including left-handed helix, right-handed helix, β-sheet-like, PUBS, and linear. **B**) Log likelihood per frame as a function of number of clusters for uniform shape-GMM and weighted shape-GMM. Each model has two corresponding curves: the log likelihood per frame of the training set and the log likelihood per frame from the CV set. Error bars are the standard deviation of sampling 10 different training sets for each model. Shaded regions denote the 90% confidence interval.

Clustering of Five ANMs Using Shape-GMM. To rigorously assess the ability of uniform and weighted shapeGMM to cluster molecular-like configurations, we have constructed a set of five ANMs, each with 12 particles depicted in Figure 3.1A. These five structures are motivated by protein secondary structural elements: a left-handed helix, a right-handed helix, a β-sheet-like structure with a four-bead hairpin, a partially unfolded β-sheet (PUBS) with both a structured

region and an unstructured region, and a linear chain that might represent an unfolded or an unstructured region of a protein. Equal-length simulations (10k frames each) of each model were run, and a single trajectory was created by concatenation. Clusterings can be compared to the ground truth (denoted by which ANM simulation the frame comes from) using correctly clustered frame pairs to avoid cluster index invariance.

Shape-GMM is able to correctly identify five clusters in the amalgamated trajectory of the five ANMs depicted in Figure 3.1A. To demonstrate this, we trained both uniform and weighted shape-GMMs for the number of clusters varying from two to eight on 2000 randomly selected frames; this process was repeated on five different training sets to demonstrate repeatability (and to estimate error), and the CV set was the remaining 48k frames for each set. The results for uniform and weighted shape-GMM are comparable, so we will only discuss the uniform results for clarity. The resulting log likelihood per frame of the uniform shape-GMM is plotted as a function of the number of clusters as the blue (training) and orange (CV) curves in Figure 3.1B. The log likelihoods of both the training set and the CV set rapidly increase from two to five clusters at which point the values plateau. The plateau in the log likelihood at five clusters indicates that increasing the number of clusters does not improve the fit of the model to the data. Additionally, a lack of measurable deviation between the training set and the CV set indicates no over- or underfitting of the model. Thus, uniform shape-GMM on this data suggests choosing five clusters. Both uniform and weighted five cluster shape-GMMs yield 100% agreement with the ground truth on the cluster assignments.

***Clustering of Five ANMs Using Internal Coordinates*.** Internal coordinates can be used to cluster these same five ANM structures, but an accurate clustering requires a combination of pairwise distances and backbone dihedrals. Internal coordinates do not comprise a vector space

and thus the distributions in these spaces will not be strictly Gaussian. Thus, we employ a density-based clustering algorithm, CLoNe, that has been recently applied to MD data.[20] CLoNe has a single free parameter, denoted pdc, that dictates the maximum distance to consider grouping points together. Altering this parameter can change the number and identity of clusters in the data. For all feature spaces investigated, we scanned this parameter and chose a value that optimized the clustering overlap with the ground truth.

CLoNe on pairwise positions or backbone dihedrals is insufficient to adequately cluster the five amalgamated ANM trajectories. In both cases, CLoNe predicts four major clusters (some frames are assigned as noise). CLoNe on all pairwise distances (pdc = 6; 66 total distances) unsurprisingly groups the two helix structures into one cluster, as pairwise distances take on the same values in the two clusters. This yields a clustering overlap with the ground truth of 91.3%. CLoNe on backbone dihedrals (pdc = 5.7; 18 values encoded by sin and cos of the nine dihedrals) divides the linear structure into two clusters due to the ability of each dihedral to take on a wide array of values. The resulting overlap with ground truth is 88.6%. CLoNe on a feature space composed of backbone dihedrals and pairwise distances (pdc = 4.1; 84 total features) predicts the proper five clusters and yields an overlap with ground truth of 99.9%.

While CLoNe on the combination of internal coordinates is able to properly cluster the five ANM models, this agreement requires two significant supervision steps. First is the choice of specific internal coordinates. It is common to choose one or the other of these coordinates, but there are some cases in which both are important. This could be a situation like the combined dissociation of tertiary and secondary structural elements during protein unfolding. The second supervision step is the choice of the pdc parameter. Here, we chose a pdc parameter that yielded an improved agreement with ground truth. Clearly, this is not feasible in all situations.

Uniform Versus Weighted Alignment in ShapeGMM. Uniform shape-GMM is unable to distinguish between subtle structural differences in systems with enhanced flexibility throughout the different regions of the molecule. This is because uniform shape-GMM relies on a RMSD-based alignment that assumes that particles fluctuate independently in equivalent, spherical distributions. To demonstrate this failing, we look at two partially unfolded β-sheet ANM models, named PUBS 1 and PUBS 2, and depicted in the insets in Figure 3.2. The only difference between the two structures is that two beads in the four-bead loop region of PUBS 1 are moved into the "hydrogen-bonding" distance in PUBS 2. The size of the "unfolded" region is varied from zero to eight beads in each system to assess the impact of changing the size of the highly varying region. Simulations of each of these were performed independently, and an amalgamated trajectory was created with 5000 frames from each simulation.

Uniform shape-GMM is unable to distinguish the two partially unfolded β-sheet structures when the unfolded region becomes comparable in size to the folded region. To quantify this, we performed uniform shape-GMM on the combined trajectories of PUBS 1 and PUBS 2 for varying sizes of the unfolded region (quantified by n in Figure 3.2). The number of clusters was fixed at two, and the model was initiated with the correct clustering (to avoid convergence to local maxima). The fraction of correctly clustered pairs of frames from the output clustering was then calculated as a function of unfolded region size (Figure 3.2). The uniform shape-GMM curve (blue) correctly clusters almost all frame pairs for unfolded region sizes from 0 to 3 beads. For $3 < n < 7$, the uniform shapeGMM overlap with the ground truth clustering rapidly decays to only getting 50% of the frame pairs correct, the minimum possible given that each frame must be clustered into one of the two clusters. Thus, uniform shape-GMM is unable to distinguish between configurations PUBS 1 and PUBS 2 for the unfolded regions of five or more beads, situations in

which the folded region and unfolded region are comparable in size. This result is consistent with previous findings based on pairwise RMSD clustering.

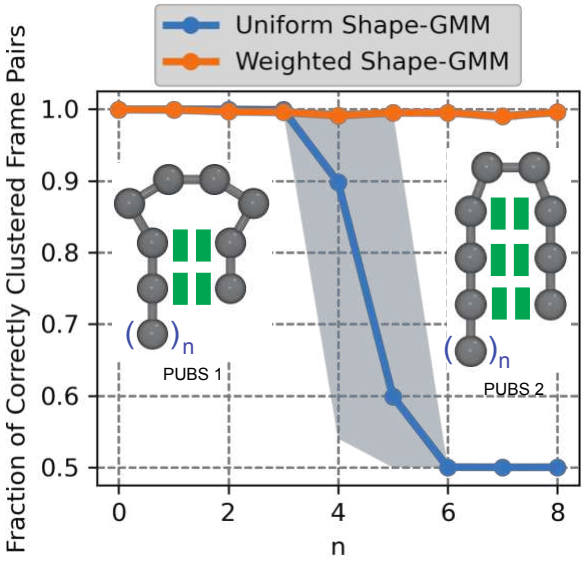


Figure 3.2. Clustering of two ANMs with varying length of an unfolded region. Fraction of correctly clustered pairs of frames as a function of the size of the unfolded region for uniform shape-GMM and weighted shape-GMM. Shaded regions denote the 90% confidence interval obtained by clustering with 10 different training sets. Representations of the two topologies of ANMs used in the clustering are provided as insets.

Weighted shape-GMM can distinguish between the two partially unfolded β-sheet structures for all unfolded regions sampled. Again, this is quantified by the clustering overlap with the ground truth as a function of unfolded region size in Figure 3.2B. The weighted shape-GMM curve, depicted in orange, demonstrates a fraction of correctly clustered frame pairs near 1.0 for all sampled unfolded region sizes. This result demonstrates that the covariance in the weighted model is a better match to the actual covariance from the PUBS 1 and PUBS 2 ANMs than the uniform covariance. Additionally, this result indicates that the shape-GMM procedure can correctly capture a subtly different behavior in heterogeneously fluctuating molecules such as the ones depicted here. This result exhibits how weighted shape-GMM overcomes a previous challenge that limited the applications of particle positions as features for clustering.

3.3.2 Beaded helix transitions

To assess the ability of the shape-GMM to cluster a nonharmonic system, we consider a 12-bead helix system with harmonic bonds along the backbone and Lennard-Jones (LJ) attractions between every fifth bead described in ref[37]. Depending on the strength of attraction between every $i, i+4$ bead (or nearly equivalently, the temperature), this model can be trapped in its starting helix or make transitions between the left- and right-handed helices. A schematic of this process is depicted in the inset of Figure 3.3B. We choose an attractive strength of $\varepsilon = 6k_BT$, in which case the helices are stabilized but some rare transitions can occur. We performed both uniform and weighted shape-GMM on a 10k frame trajectory to assess the ability of these models to cluster the data.

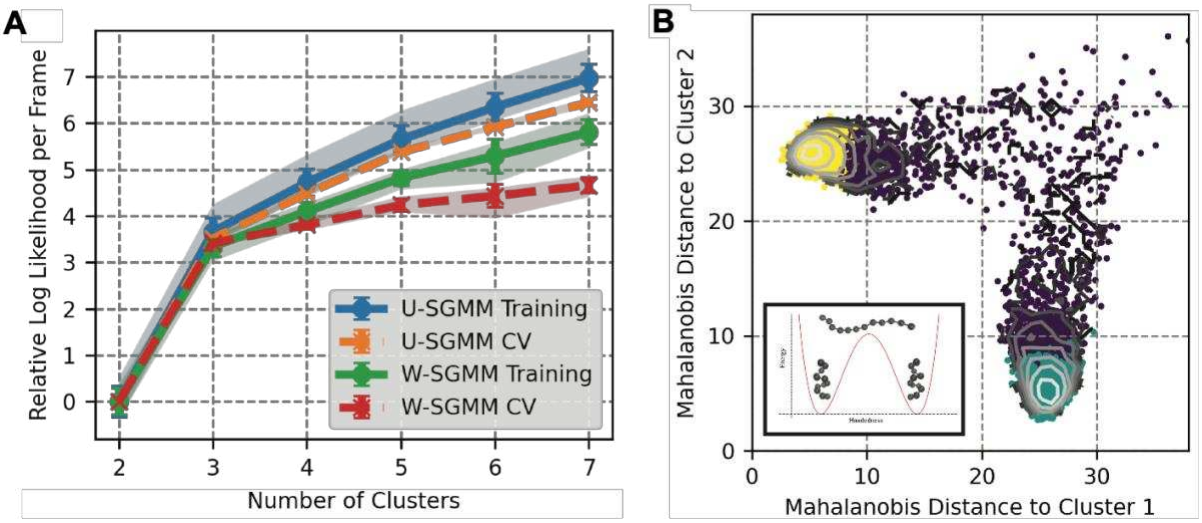


Figure. 3.3. Clustering of a 12-bead helix transition trajectory. **A)** Log likelihood as a function of number of clusters for uniform shapeGMM (U-SGMM) and weighted-shape-GMM (W-SGMM). Error bars denote standard deviation and shaded regions denote the 90% confidence interval obtained from clustering using five different training sets. **B)** Clustering results for three clusters from W-SGMM on 2D free-energy surface of weighted Mahalanobis distance from two cluster centers (inset: schematic of the potential energy diagram).

Both uniform and weighted shape-GMM support the choice of three clusters to represent the simulation data. A plot of the log likelihood per frame as a function of number of clusters is depicted in Figure 3.3A for both uniform and weighted models. In this scan, five randomly selected

training sets of 1000 frames are chosen. Uniform and weighted shape-GMMs are fit with the training set and then cross-validated on the remaining 9000 frames. Both uniform and weighted models have a significant change in slope at three clusters for both the training and CV curves. Additionally, the CV curve for weighted shape-GMM (red) starts to deviate significantly from the training curve (green) as the number of clusters is increased above three. This behavior indicates an overfitting of the weighted shapeGMM above three clusters. The overfitting behavior is less significant for the uniform shape-GMM but is still present, and, as seen in the previous section, a weighted shape-GMM is able to differentiate between heterogeneously fluctuating systems more readily than a uniform shape-GMM.

Weighted shape-GMM identifies three states corresponding to the left-handed helix, right-handed helix, and an intermediate configuration between the two. To demonstrate this, we plot the free energy of the simulation as a function of the Mahalanobis distance (Equation 3.8) from clusters 1 and 2 as a contour plot in Figure 3.3B. There are two distinct stable states in this 2D free-energy surface, each of which is centered around a small separation from either cluster 1 or cluster 2. These are the left- and right-handed helices. Cluster identities from the three-cluster model of weighted shape-GMM are indicated by three colors of points on the same plot. The cluster labeling does a good job distinguishing fully folded (yellow and teal) from partly folded (purple) configurations as well as distinguishing the left- and right-handed helical configurations picked out by the centers of two clusters. This result is encouraging because it demonstrates that shape-GMM is able to identify not only quasi-harmonic metastable states but also separate out and classify together intermediate structures.

3.3.3 Application to HP35

HP35, the C-terminal subdomain of villin, has been the focus of numerous experimental and theoretical studies of protein folding due to its ability to autonomously fold.[38,39] The 35-residue chicken villin headpiece protein contains three helices, with helix 1 composed of residues 4−10, helix 2 of residues 15−19, and helix 3 of residues 23−32. Wild-type HP35 has a folding rate of $(4.3 \pm 0.6\ \mu s)^{-1}$,[40] which was further observed in atomistic MD simulations with an implicit solvent.[41] Mutations were sought to enhance the folding rate[42] and ultimately yielded a fast-folding double mutant referred to as Nle/Nle (Nle—norleucine) with a measured folding rate of $(0.73 \pm 0.05\ \mu s)^{-1}$ at 300 K.[43] The ∼305 μs all-atom, explicit solvent, MD simulation of the HP35 Nle/Nle mutant from D. E. Shaw Research has been extensively studied and serves as a benchmark system for structural clustering.[7,9,17,44–51] Here, we use this simulation at 360 K to compare and contrast between uniform and weighted shape-GMM, compare weighted shape-GMM clusterings to previous results, and, finally, propose the structural folding mechanism most consistent with the weighted shape-GMM results.

Uniform Versus Weighted Shape-GMM for HP35. We hypothesize that uniform and weighted shape-GMM will give consistent clustering results for well-folded states but inconsistent results for partially and completely unfolded states. This hypothesis is consistent with the cluster results from the ANMs in Section 3.1.3. To test this hypothesis, we trained six-state uniform and weighted shape-GMMs on ∼61k frames of the HP35 trajectory. For reference, a reasonable training set size should include at least 25k frames to sample one of the observed 61 folding/unfolding events within the ∼1.5M frame trajectory[44]. A six-state model was chosen due to literature precedence and simplicity of comparison (results from a scan of cluster sizes are discussed in Section 3.3.2). Backbone atoms (C, CA, and N from residues 2 through 34, C from

residue 1, and N from residue 35) are selected for the feature set to remain consistent with previous backbone dihedral value clustering methods.[7,48] For each covariance approximation, 20 models with different randomly chosen starting conditions were trained, and the model with the largest log likelihood was used for subsequent analysis.

Uniform and weighted six-state shape-GMMs show consistency in separating the folded from unfolded states of HP35. To support this claim, we combine the clustering overlap between the two models, as assessed by a matching wheel (Figure 3.4A), and the stability of dihedral distributions of the residues in each cluster represented in ramacolor plots[17] (Figure 3.4B for uniform and Figure 3.4C for weighted). The backbone dihedral to color array mapping can be found in Figure S2. Starting with the weighted shape-GMM six-state ramacolor plot (Figure 3.4C), we observe that states I and II have bright coloration for all residues, indicating well-defined dihedral distributions in these states. State IV has the majority of residues in well-defined dihedral states, with the C-terminus starting to display some dihedral disorder, as evidenced by the muted color for residues 30−34. The remaining weighted shape-GMM states (III, V, and VI) have over half of the residues with muted ramacolor distributions indicating that these states are relatively disordered. Thus, we conclude that weighted shape-GMM states I, II, and IV are folded states and III, V, and VI are unfolded. The matching wheel indicates that the weighted shape-GMM folded states (I, II, and IV) are predominantly matched with the states 1, 2, 4, and 6 from uniform shape-GMM. The ramacolor plot for uniform shapeGMM indicates that states 1, 2, 4, and 6 are all fairly structured in their dihedral distributions and that the remaining clusters (3 and 5) are disordered. Thus, we conclude that both the uniform and weighted shape-GMM are able to separate folded from unfolded states.

Weighted shape-GMM separates the states by distinct regions of instability corresponding to the secondary structure elements in HP35. Considering the most unstable uniform states 3 and 5, indicated by muted colors across nearly all residues, there are no obvious differences between these two states indicated in Figure 3.4B. In contrast, the unstable weighted states III, V, and VI have clear structural differences (Figure 3.4C). The dihedral distributions of residues within the weighted state III closely resemble either uniform states 3 and 5, where all three helices show instability, but there is stability in turn 2 and in some subsequent residues. Alternatively, weighted state VI is entirely unstable but gets divided up evenly to define uniform states 3 and 5. Weighted state V has stability in helix 3 and helix 2 and is composed of frames that make uniform states 3, 4, and 5. Finally, weighted state IV is a very stable state besides the C-terminal end, which compares reasonably well with the uniform state 4 besides the weighted state showing more stability in the N-terminal helix. Altogether, these results confirm that weighted shape-GMM improves structural discrimination when it comes to flexibility in particular regions, whether it is a single residue or secondary structures. Therefore, for identifying the states in challenging biomolecular processes such as protein folding, a weighted shape-GMM is preferred.

The six-state weighted shape-GMM trained on backbone atoms captures several structural distinctions between unfolded, intermediate, and native states previously suggested by the backbone dihedral clustering schemes for 6- and 12state models.[7,9,17,48] For example, the ramacolor plot for this model in Figure 3.4C illustrates that states I and II differ in the conformation of Asp3. The helical conformation of Asp3 has previously been shown to distinguish between the native and intermediate states in the 6- and 12-cluster models.[17,48] That weighted shape-GMM predicts the native state to be larger in population than the intermediate state is inconsistent with some previous results[17] but consistent with other recent results.[7] A native-like state that is structurally

similar to the native state, but differing mainly in the dynamics due to an unlocked and partially unfolded helix 3, has been identified previously and is in line with state IV from weighted shapeGMM.[45,48,52–54] Alternatively, state V indicates the opposite trend: the N-terminus and helix 1 are highly fluctuating, and there is more stability from residue 17 leading into helix 3, with a slight increase in positional variance again at the C-terminus. This indicates that state V has an unfolded helix 1 but partial folding of helices 2 and 3. These structural attributes have also been discussed in previous models and used to define unfolded states in more detailed 12-state models.[17] The ability of our weighted shape-GMM to capture these features in a six-state model supports the high capacity for structural discrimination and the model's prioritization of these states as distinct.

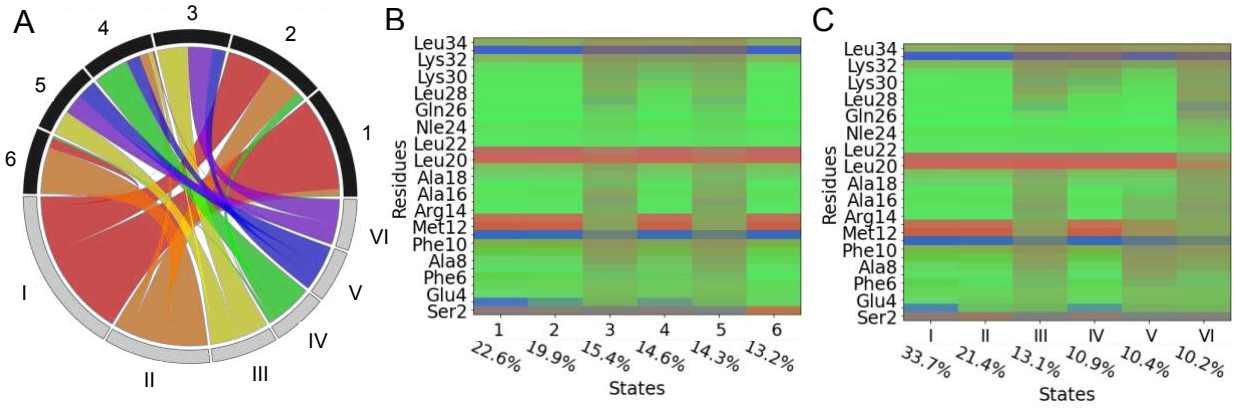


Figure 3.4. Uniform and weighted shape-GMM of six-state clustering results on HP35 Nle/Nle mutant. **A**) Matching wheel for cluster trajectories of uniform (top, 1−6) and weighted (bottom, I−VI). **B**) Ramacolor plot of uniform shape-GMM and **C**) weighted shape-GMM. Percentages correspond to cluster populations.

HP35 Folding: An Intuitive Four-State Model. Weighted shape-GMM results on HP35 are most consistent with a four-state model. A 10-fold CV scheme using a ~25k frame training set applied to a scan of cluster sizes shows a significant % change in the slope of the log likelihood at a cluster size of 4 (see Figure S1), which suggests that four clusters is an appropriate choice given the structural data used to fit the model. A cluster size of 8 could also be considered with this reasoning; however, the deviation of log likelihood between the training and prediction data is an

indication of model overfitting. It is possible that with this large cluster size, the training data size must be increased to build a model with representation in the training data for eight states to be sufficiently clustered.

The four-state weighted shape-GMM is composed of a native state, near-native state, intermediate state, and an unfolded state. The most populated cluster (53.41%) is characterized as the native state, N, with completely folded helices and the α_L-Asp3 conformation (Figure 3.5). The native-like state, N′, that resembles N, but differs in the partial unfolding of helix 3, is the least populated (13.95%). In contrast to the previous six-state model, a near-native intermediate state with completely folded helices and an α_R-Asp3 is not prioritized as a distinct state with a cluster size of 4. This model still identifies an intermediate state, I (17.46%) with an unfolded helix 1, and partial folding in helices 2 and 3, and a broadly disordered unfolded state, U (15.17%).

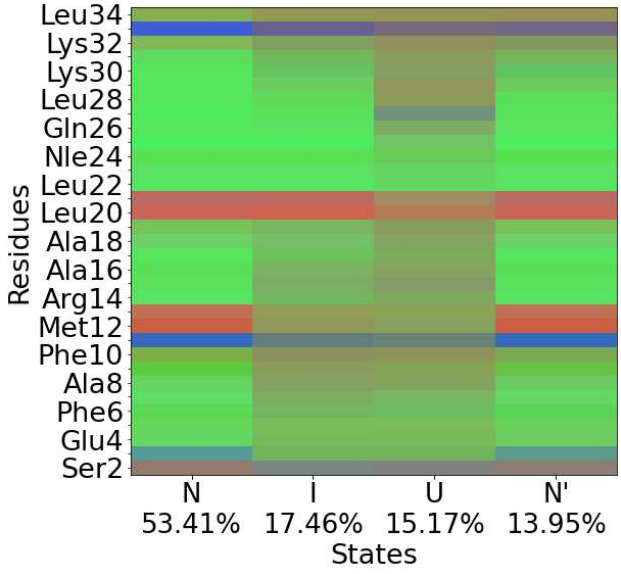


Figure 3.5. Ramacolor plot for the HP35 weighted shape-GMM four-state model prior to dynamical coring. Percentages correspond to the native (N), intermediate (I), unfolded (U) and near-native (N′) state populations.

A Markov state model (MSM) with a lag time of 2 ns, built from the four-state shape-GMM results using PyEMMA,[55] yields three dynamically distinct processes. The trajectory of

cluster identities was first fed through a dynamic coring procedure with a minimum window of 2 ns (10 frames).[51] This changed only 3.8% of frames, suggesting that the results from weighted shape-GMM were already dynamically stable. The four-state model yields a passing Chapman−Kolmogorov test (see Figure S3). Three implied timescales are observed, having values of 348, 40.9, and 19.1 ns. The processes associated with these three timescales are population shifts from states I and U to states N and N′, between states I and U, and between states N and N′, respectively. Additional results from this model are depicted in the network diagram presented in Figure 3.6.

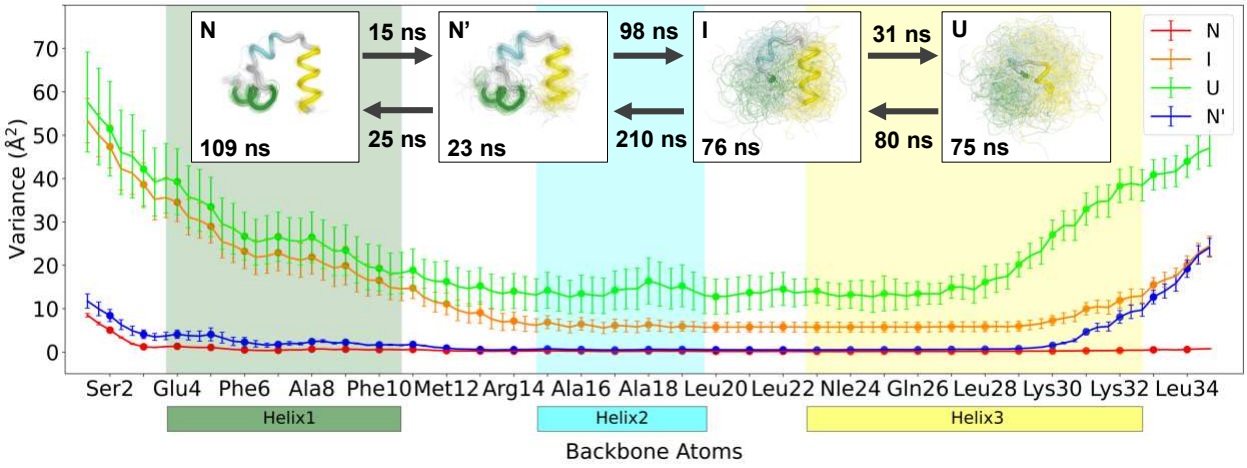


Figure 3.6. Variation in the backbone atomic displacement for each state in the four-state weighted shape-GMM results for HP35. Regions corresponding to helix 1, helix 2, and helix 3 are colored forest green, cyan, and yellow, respectively. Error bars represent the standard deviation of the values estimated from dividing the trajectory into 10 equal-sized continuous segments. Enlarged points represent the Cα atoms of each residue. Top: GMM centers in thick tube representation, with 100 random frames from each cluster in thin tube representation. State lifetimes are shown in each structural figure, with MFPTs between the connected states.

The folding mechanism of HP35 (Nle/Nle mutant) proceeds from the unfolded state, U, through an intermediate state, I, and near-native state, N′, to the native state, N. The mean first passage time (MFPT) to go from U to I is 80 ns, and this process is characterized by a partial structuring of the C-terminal end of helix 3 (residues 29−34). Subsequent folding of helices 1 and 2 (I to N′) is the rate-limiting step, requiring 210 ns. Once the N-terminal helix has formed, the final step is to completely fold the C-terminal helix, taking 0.12 μs. Previous models proposed a

native-like state that serves as a kinetic trap, only accessible through N;[47] however, our model shows that N′ is *en route* to N, which is consistent with experimental evidence.[52] The unfolding mechanism is similar to folding but in reverse. In both cases, the transition states are observed between the N′ and I states, as quantified by the committors.

The structural characteristics of the intermediate state support a mechanism that primarily folds helix 3 and 2 before helix 1. The order of helix formation during HP35 folding has been heavily debated, both experimentally and theoretically, with some evidence that the order is highly influenced by both physical and methodological factors, such as temperature, or the force field used.[48,50,52,56–58] A recent MSM of Nle/Nle HP35, built from a 57-state model, found that over 90% of all pathways formed helix 3 before helices 2 and 1.[51] The most preferred pathway appeared to fold in the order of helix 3, then 2, and lastly 1, which agrees well with our much more reduced model. Additionally, recent studies show evidence of Nle/Nle folding beginning with stabilization in the middle of helix 3 at residue 28, to the helix 3 N-terminus at residue 22, rather than a fully folded helix 3,[50,56] which aligns well with our results that show that even in the unfolded state the N-terminus of helix 3 has the lowest variance in the backbone atom positions (Figure 3.6). In going from U to I, the backbone atom variances decrease significantly for the residues in helices 3 and 2. Interestingly, residues 20−29 in the intermediate state appear fairly stable in that the variance is nearly the same for each backbone atom, yet the unchanging value is still high (∼20 $Å^2$). While these residues exist in stable conformations (also indicated in Figure 3.5), they may need to accommodate for the large fluctuations coming from the unfolded helix 1 and loop 1. This indicates that even though these regions are folded, their stability can be influenced by the rest of the protein. Furthermore, this information is a direct result of weighted shape-GMM, as the covariance is iteratively evaluated and used to optimally define each cluster.

A MSM based on our weighted shape-GMM 4-state discretization represents a respectable balance between structural resolution and kinetic detail. The mechanism of folding aligns well with previous models, even those based on much larger MSMs, and captures subtle but relevant state differences, such as the increased C-terminal variance in the native-like state. It is important to reiterate that this model was fit on ~4% of the total trajectory data; therefore, not all folding events were represented during training. It is becoming more evident that although HP35 is a small protein, the folding mechanism is complex and likely to result from multiple pathways with possible kinetic traps and misfolded states. If the goal was to model an ensemble of folding pathways, which have been suggested for this system, then we propose: (1) using a larger training size to invite more representations for potential pathways, such that larger cluster sizes (*e.g.*, 8) can be chosen without overfitting and (2) exploring additional features, such as side-chain center of mass, as it has been proposed that for HP35 the secondary structures form first, primarily in the unfolded basin, and then longer scale sidechain ordering must occur to achieve native contacts.[46,59] As a measure of MSM validation, VAMP-2 scores for various HP35 discretizations, with and without dynamical coring, are provided in Table S1 of the Supporting Information. The VAMP-2 scores of 4-state, 6-state, and 12-state shapeGMM results compare favorably with a well-informed 12-state model, following the most probable path clustering protocol designed by Nagel *et al.*,[51] especially when the clusterings are further refined using dynamical coring. This supports the viability of this structure-based method in yielding kinetically relevant information, even though the procedure is void of explicit dynamic information.

3.4 Practical considerations for shapeGMM

ShapeGMM is an appealing approach to structural clustering but is not without its limitations. There are three major points to consider when using this method: (1) difficulty in

convergence to global maxima, (2) estimating high-dimensional objects, and (3) the computational expense of fitting the models. The first of these limitations is true for any EM procedure. This concern can be alleviated by performing each maximization numerous times with different model initialization parameters and selecting the model with the highest resulting log likelihood, as suggested by others.[60] While there may be additional improvements that can be done to ensure global maxima convergence, we believe that the other two limitations are more specific to the application of shape-GMM to MD data and are discussed in more detail in the following subsections.

Clustering Convergence and Estimating N × N Covariances. The most challenging component of estimating a weighted shape-GMM are the multiple, high-dimensional, covariance matrices. To estimate a weighted shape-GMM for K clusters on a trajectory with N atoms, one must estimate K $N \times N$ covariance matrices. The minimum criteria is to make each matrix full rank, which requires N + 1 independent data points for each matrix. Shape-GMM is aided in this endeavor by two aspects of the model. First, there are three independent measures of the $N \times N$ covariance in each frame. Second, each frame can contribute to the covariance matrix of each cluster as indicated in Equation 3.13. The weight of each frame is determined by the posterior distribution (Equation 3.10) and can be vanishingly small. Practically speaking, one must have multiple (>10) independent samples of each covariance to have any confidence in their estimates. Thus, a reasonable estimate is $\alpha \frac{(N+1)}{3}$ frames with $\alpha > 10$ for a training set. The quality of the estimate of the covariance will be dictated by α, and the value needed to perform clustering will depend on the system being clustered. The PUBS 1 and PUBS 2 ANM models with $n = 8$ provide a challenging system to cluster because the difference in the two structures is heavily dependent on the difference in covariances. Thus, we investigate the convergence of weighted shape-GMM

on this data as a function of the training set size (Figure 3.7). We consider training sets of sizes varying from 100 to 2000 frames, with each training set picked randomly from the total of 10k frames. Each training set size was sampled 10 times, and the resulting models were used to predict clustering on the full 10k frame trajectory. The resulting clustering overlap with the ground truth and log likelihood per frame are depicted in Figure 3.7. For training sets of size <300 frames, we see that the clustering overlap with ground truth is low (≈50%) but that this overlap rapidly increases from training set sizes 300 to 700. From there, there is some oscillation of the average value, but overlaps are all near or above 90%. The log likelihood per frame of the entire clustering shows a similar but more steady increase and plateauing behavior. We note that the clustering overlap for models trained on even >1000 frames still shows significant deviation between 50 and 100% overlap with ground truth, as indicated by the shaded region. This indicates the importance of the particular training set chosen, especially for relatively small training sets. Therefore, it is recommended to fit shape-GMM on various training sets and use the model with the highest log likelihood on the entire data set.

As N gets larger, sampling of the covariance matrices will become more challenging. It has previously been estimated that RMSD-based clustering will start to fail for greater than 200 particles.[61] This analysis utilized both a standard RMSD measure as well as a weighted RMSD measure, but this weighted measure still equates to a diagonal covariance matrix, not the weighted form of the covariance we describe in this paper. Thus, it remains an open question as to the limit of N for our weighted shape-GMM clustering. All systems studied in this paper fall under the 200-particle threshold given for RMSD.

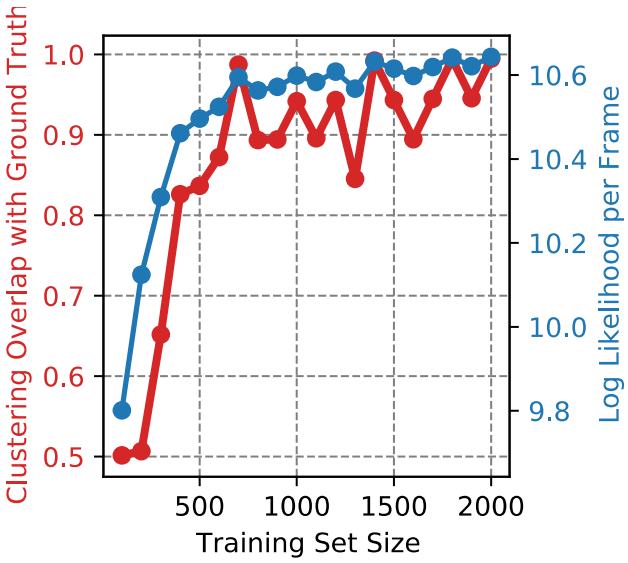


Figure 3.7. Clustering convergence for weighted shape-GMM on the PUBS 1 and PUBS 2 ANMs. The average clustering overlap with ground truth and log likelihood per frame are plotted as a function of the training set. Each training set was chosen randomly from the 10k total frames. This procedure was repeated 10 times to compute the average and 90% confidence interval (shaded region).

Computational Expense of shapeGMM. Computational expense of weighted shape-GMM scales linearly with the number of clusters. This can be observed in both plots depicted in Figure 3.8. We focus on the orange curve with a training set of 25k frames in Figure 3.8A, as this is the same training set size used for our cluster scan. We observe that for two clusters, a weighted shape-GMM takes approximately 10 min to optimize. For 14 clusters, it takes approximately 150 min to optimize. The large variance in optimization time, as indicated by 90% confidence intervals surrounding the solid lines, is a natural aspect of the EM procedure for GMMs. The time it takes to find a maximum in log likelihood depends dramatically on the starting conditions. Regardless, the trend in the average time as a function of number of clusters is clearly linear, as one would expect for a serial implementation of the EM GMM procedure.

Computational expense scales linearly with the number of frames in the training set. The four different lines in Figure 3.8A indicate different training set sizes. We start with 12.6k frames and then increase by multiples of 2 to include ~25k, 50k, and 75k frames. The slopes of the best

fit lines are 5.1, 10.4, 19.7, and 30.0 CPU min/cluster, respectively. These scale by the same multiplicative factor as the number of frames, thus indicating that the procedure scales linearly with the number of frames.

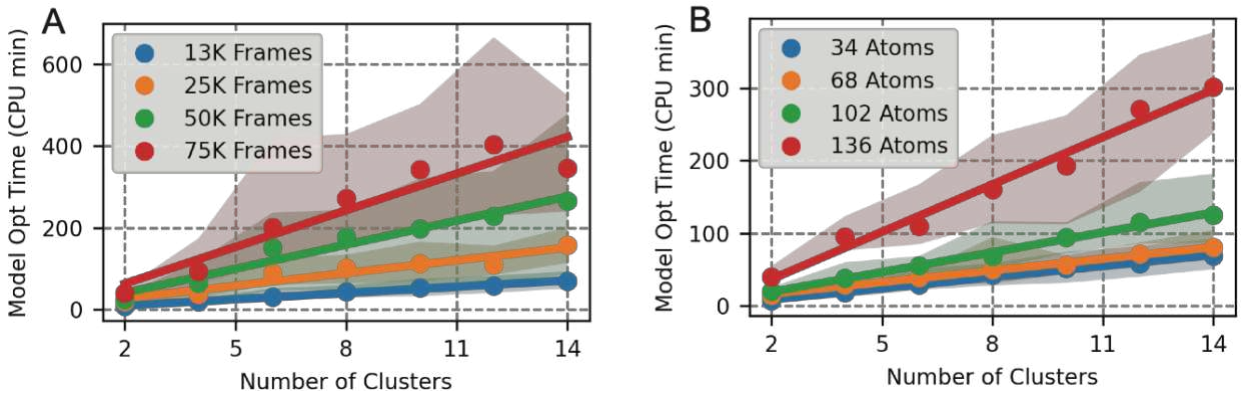


Figure 3.8. CPU time for weighted shape-GMM optimization for HP35. **A**) CPU time to optimize a single weighted shape-GMM on 34 atoms as a function of number of clusters for different training set sizes. **B**) CPU time to optimize a single weighted shape-GMM on ~13k frames as a function of number of clusters for different number of atoms in the feature space.

Computational expense scales quadratically with the number of atoms, beyond an initial sub quadratic region. The CPU time as a function of number of clusters is plotted for four different feature space sizes in Figure 3.8B. All of these were done with 12.6k frames in the training set. The behavior of different feature space sizes is not as clear as that of the training set size. The slopes of the best fit lines are 5.1 CPU min/cluster for 34 atoms, 5.3 CPU min/cluster for 68 atoms, 9.1 CPU min/ cluster for 102 atoms, and 21.8 CPU min/cluster for 136 atoms. The ratio of the slopes of the smallest feature space size (34 atoms) to that of the largest feature space size (136 atoms) is approximately 4, comparable to the ratio of the feature space sizes themselves. The two medium-sized feature spaces behave sub linearly compared to the smallest feature space size. Comparing the slopes of the CPU times of 68 and 136 atoms, we observe a quadratic relationship between time and feature space size. We expect CPU time to scale as N^2 due to the two dominant processes being the calculations of covariances.

It is possible to train weighted shape-GMMs with 100s of atoms on 10s of thousands of frames within a few hours. The computational time is compounded by the recommendation to run multiple iterations to increase the chances of converging to a global maximum, as well as the potential need to scan the number of clusters. That said, each of these runs is completely independent and thus can be run simultaneously on separate processors. Additionally, the clustering algorithm requires very few user parameters: feature space size, training set size, and number of clusters. This is in stark contrast to many other clustering protocols.

3.5 Conclusions and Outlook

A GMM on particle positions is a conceptually appealing approach to model the high-dimensional probability density of macromolecules. The difficulty in doing so has been to account for the ability of the molecule to freely rotate and translate in a way that properly accounts for particle−particle correlation. Here, we present the ML alignment and GMM procedures necessary to determine the optimal parameters for a given mixture size. This procedure can be used to discretize conformational space in a way that theoretically matches the intuition of molecules hopping between harmonic free-energy minima.

Two shape-GMMs are presented: uniform, a model in which the particle covariance of each mixture is presumed to be proportional to the identity matrix, and weighted, a model in which the particle covariance of each mixture is presumed to have the form $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_N \otimes \boldsymbol{I}_3$. Both uniform and weighted models are able to distinguish between five structurally distinct ENMs.

Weighted shape-GMM is able to distinguish between distinct structures for species with heterogeneous particle variation. This finding disproves a previously held belief that particle positions cannot distinguish between these types of structures and demonstrates an achievement of shape-GMM that will greatly benefit the field.

Weighted shape-GMM also provides globally distinct clusters for the folding/unfolding of HP35 from an all-atom MD trajectory. The identified clusters corroborate some previously identified features of the conformational ensemble of HP35 but are distinct in the global picture of the folding pathway. Specifically, a four-state model is the most consistent with our data and predicts that the native state comprises >50% of the entire trajectory and follows a C-terminal unfolding and then N-terminal unfolding pathway.

Weighted shape-GMM can also be used to compare clusterings from other, potentially faster procedures. The log likelihood of a particular clustering can be readily computed and compared between clusterings. The clustering with the largest log likelihood represents the best partitioning of the data under the given GMM.

The generality of our approach is strengthened by the ease of access and application. The package can be easily installed from PyPI (pip install shapeGMM) or directly from github (https://github.com/mccullaghlab/GMM-Positions). The code interface is designed to mimic the Scikit-learn Gaussian mixture package with object initialization, fitting (albeit with uniform and weighted versions), and prediction routines.

The shape-GMM algorithm is more computationally demanding than some other clustering procedures. While linear scaling is observed as a function of number of clusters and training set size, quadratic scaling is observed for feature space size. This will limit the system sizes that weighted shapeGMM can be readily applied to. We note, however, that there are some distinct advantages, other than conceptual ones, of the algorithm: few parameters are needed for clustering, and the method can be directly applied to any type of molecular system. Looking forward, we expect a GPU implementation of the algorithm to greatly improve the applicability for larger system sizes or data sets.

3.6 Simulation Details

3.6.1 ENM simulations

ENMs were simulated in the LAMMPS package.[62] Harmonic bonds were placed between each bead. Langevin dynamics simulations were performed at 300 K in the NVT ensemble with a damping coefficient of 10 fs^{-1}. An integration timestep of 2 fs was used. Simulations were run for 10 million steps, with frames written every 100 steps.

3.6.2 Beaded helix simulations

A12-beadmodel designed to have two equi-energetic ground states as left- and right-handed helices[37] was simulated in LAMMPS.[62] Eleven harmonic bonds between beads having rest length of 1.0 and spring constant of 100 form a polymer backbone. LJ interactions between every i, $i + 4$ pair of beads with $\varepsilon = 6.0$ and $\sigma = 1.5$ and a cutoff length of 3.0 give rise to the helical shape. All nonbonded i, $i + 2$ and farther also have a repulsive WCA interaction with $\varepsilon = 3.0$ and $\sigma = 3.0$ added to prevent overlap, with ε for i, $i + 2$ reduced by 50%. Simulations at temperature 1.0 were performed using "fix nvt" using a simulation timestep of 0.005 and a thermostat timestep of 0.5. A folding/unfolding trajectory of length 50,000,000 steps was generated and analyzed as above. Here, all parameters are in reduced (LJ) units.

3.6.3 HP35 Simulation

A 305 μs all-atom simulation of Nle/Nle HP35 at 360 K from Piana et al.[58] was analyzed. The simulation was performed using the Amber ff99SB*-ILDN force field and TIP3P explicit water model. Protein configurations were saved every 200 ps, resulting in ~1.5M frames.

3.7 Associated Content

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.1c01290. Log likelihood of weighted shape-GMM (S-

GMM) as a function of number of clusters; RGB values corresponding to phi−psi angle pairs used to assign a unique color value to a dihedral conformation; Chapman−Kolmogorov test for four-state MSM resulting from weighted shape-GMM clustering on HP35; and scores for a variety of clusterings of the HP35 trajectory.

REFERENCES

(1) Deuflhard, P.; Huisinga, W.; Fischer, A.; Schütte, C. Identification of Almost Invariant Aggregates in Reversible Nearly Uncoupled Markov Chains. *Linear Algebra Appl.* **2000**, *315* (1–3), 39–59.

(2) Deuflhard, P.; Weber, M. Robust Perron Cluster Analysis in Conformation Dynamics. *Linear Algebra Appl.* **2005**, *398* (1–3), 161–184.

(3) Kannan, D.; Sharpe, D. J.; Swinburne, T. D.; Wales, D. J. Optimal Dimensionality Reduction of Markov Chains Using Graph Transformation. *J. Chem. Phys.* **2020**, *153* (24), 244108.

(4) Keller, B.; Daura, X.; Van Gunsteren, W. F. Comparing Geometric and Kinetic Cluster Algorithms for Molecular Simulation Data. *The Journal of Chemical Physics* **2010**, *132* (7), 074110.

(5) Peng, J. H.; Wang, W.; Yu, Y. Q.; Gu, H. L.; Huang, X. Clustering Algorithms to Analyze Molecular Dynamics Simulation Trajectories for Complex Chemical and Biological Systems. *Chinese J. Chem. Phys.* **2018**, *31* (4), 404–420.

(6) Glielmo, A.; Husic, B. E.; Rodriguez, A.; Clementi, C.; Noé, F.; Laio, A. Unsupervised Learning Methods for Molecular Simulation Data. *Chem. Rev.* **2021**, *121* (16), 9722–9758.

(7) Damjanovic, J.; Murphy, J. M.; Lin, Y.-S. CATBOSS: Cluster Analysis of Trajectories Based on Segment Splitting. *J. Chem. Inf. Model.* **2021**, *61* (10), 5066–5081.

(8) Cocina, F.; Vitalis, A.; Caflisch, A. Sapphire -Based Clustering. *J. Chem. Theory Comput.* **2020**, *16* (10), 6383–6396.

(9) Sittel, F.; Filk, T.; Stock, G. Principal Component Analysis on a Torus: Theory and Application to Protein Dynamics. *J. Chem. Phys.* **2017**, *147* (24).

(10) Altis, A.; Otten, M.; Nguyen, P. H.; Hegger, R.; Stock, G. Construction of the Free Energy Landscape of Biomolecules via Dihedral Angle Principal Component Analysis. *J. Chem. Phys.* **2008**, *128* (24).

(11) Sittel, F.; Stock, G. Perspective: Identification of Collective Variables and Metastable States of Protein Dynamics. *J. Chem. Phys.* **2018**, *149* (15), 150901.

(12) Daura, X.; Gademann, K.; Jaun, B.; Seebach, D.; Van Gunsteren, W. F.; Mark, A. E. Peptide Folding: When Simulation Meets Experiment. *Angew. Chemie - Int. Ed.* **1999**, *38* (1–2), 236–240.

(13) Shao, J.; Tanner, S. W.; Thompson, N.; Cheatham, T. E. Clustering Molecular Dynamics Trajectories: 1. Characterizing the Performance of Different Clustering Algorithms. *J. Chem. Theory Comput.* **2007**, *3* (6), 2312–2334.

(14) Bowman, G. R.; Beauchamp, K. A.; Boxer, G.; Pande, V. S. Progress and Challenges in the Automated Construction of Markov State Models for Full Protein Systems. *J. Chem. Phys.* **2009**, *131* (12), 124101.

(15) Jain, A.; Stock, G. Identifying Metastable States of Folding Proteins. *J. Chem. Theory Comput.* **2012**, *8* (10), 3810–3819.

(16) Westerlund, A. M.; Delemotte, L. InfleCS: Clustering Free Energy Landscapes with Gaussian Mixtures. *J. Chem. Theory Comput.* **2019**, *15* (12), 6752–6759.

(17) Sittel, F.; Stock, G. Robust Density-Based Clustering to Identify Metastable Conformational States of Proteins. *J. Chem. Theory Comput.* **2016**, *12* (5), 2426–2435.

(18) Rodriguez, A.; Laio, A. Clustering by Fast Search and Find of Density Peaks. *Science (80-.).* **2014**, *344* (6191), 1492–1496.

(19) Melvin, R. L.; Godwin, R. C.; Xiao, J.; Thompson, W. G.; Berenhaut, K. S.; Salsbury, F. R. Uncovering Large-Scale Conformational Change in Molecular Dynamics without Prior Knowledge. *J. Chem. Theory Comput.* **2016**, *12* (12), 6130–6146.

(20) Träger, S.; Tamò, G.; Aydin, D.; Fonti, G.; Audagnotto, M.; Dal Peraro, M. CLoNe: Automated Clustering Based on Local Density Neighborhoods for Application to Biomolecular Structural Ensembles. *Bioinformatics* **2021**, *37* (7), 921–928.

(21) Ceriotti, M.; Tribello, G. A.; Parrinello, M. Simplifying the Representation of Complex Free-Energy Landscapes Using Sketch-Map. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108* (32), 13023–13028.

(22) Trozzi, F.; Wang, X.; Tao, P. UMAP as a Dimensionality Reduction Tool for Molecular Dynamics Simulations of Biomacromolecules: A Comparison Study. *J. Phys. Chem. B* **2021**, *125* (19), 5022–5034.

(23) Theobald, D. L.; Wuttke, D. S. Empirical Bayes Hierarchical Models for Regularizing Maximum Likelihood Estimation in the Matrix Gaussian Procrustes Problem. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103* (49), 18521–18527.

(24) Dryden, I. L.; Mardia, K. V. *Statistical Shape Analysis*; John Wiley & Sons: Chichester, 1998.

(25) Theobald, D. L.; Wuttke, D. S. Accurate Structural Correlations from Maximum Likelihood Superpositions. *PLoS Comput. Biol.* **2008**, *4* (2), 43.

(26) Kabsch, W. A Discussion of the Solution for the Best Rotation to Relate Two Sets of Vectors. *Acta Crystallogr. Sect. A* **1976**, *34* (5), 827–828.

(27) Horn, B. K. P. Closed-Form Solution of Absolute Orientation Using Unit Quaternions. *J. Opt. Soc. Am. A* **1987**, *4* (4), 629.

(28) Goodall, C. *Procrustes Methods in the Statistical Analysis of Shape*; 2; 1991; pp 285–339.

(29) Theobald, D. L. Rapid Calculation of RMSDs Using a Quaternion-Based Characteristic Polynomial. *Acta Crystallogr. Sect. A Found. Crystallogr.* **2005**, *61* (4), 478–480.

(30) Liu, P.; Agrafiotis, D. K.; Theobald, D. L. Rapid Communication Fast Determination of the Optimal Rotational Matrix for Macromolecular Superpositions. *J. Comput. Chem.* **2010**, *31* (7), 1561–1563.

(31) Frauenfelder; Hans; Parak, F.; Young, R. D. Conformational Substates in Proteins. *Annu. Rev. Biophys. Biomol. Struct.* **1988**, *17*, 451–479.

(32) Fong, Y.; Wakefield, J.; Rice, K. An Efficient Markov Chain Monte Carlo Method for Mixture Models by Neighborhood Pruning. *J. Comput. Graph. Stat.* **2012**, *21*, 197–216.

(33) Stephens, M. Gibbs Sampling for a mixture of normals. https://stephens999.github.io/five MinuteStats/gibbs2.html (accessed 9/1/ 2021).

(34) Bonakdarpour, M.; Stephens, M. Introduction to EM: Gaussian Mixture Models. https://stephens999.github.io/fiveMinuteStats/ intro_to_em.html (accessed 9/1/2021).

(35) Lam, S. K.; Pitrou, A.; Seibert, S. Numba. *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC*, 2015; pp 1−6.

(36) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.

(37) Hartmann, M. J.; Singh, Y.; Vanden-Eijnden, E.; Hocky, G. M. Infinite Switch Simulated Tempering in Force (FISST). *J. Chem. Phys.* **2020**, *152* (24), 244120.

(38) McKnight, C. J.; Doering, D. S.; Matsudaira, P. T.; Kim, P. S. A Thermostable 35-Residue Subdomain within Villin Headpiece. *J. Mol. Biol.* **1996**, *260* (2), 126–134.

(39) McKnight, C. J.; Matsudaira, P. T.; Kim, P. S. NMR Structure of the 35-Residue Villin Headpiece Subdomain. *Nat. Struct. Biol.* **1997**, *4* (3), 180–184.

(40) Kubelka, J.; Eaton, W. A.; Hofrichter, J. Experimental Tests of Villin Subdomain Folding Simulations. *J. Mol. Biol.* **2003**, *329* (4), 625–630.

(41) Zagrovic, B.; Snow, C. D.; Shirts, M. R.; Pande, V. S. Simulation of Folding of a Small Alpha-Helical Protein in Atomistic Detail Using Worldwide-Distributed Computing. *J. Mol. Biol.* **2002**, *323* (5), 927–937.

(42) Chiu, T. K.; Kubelka, J.; Herbst-Irmer, R.; Eaton, W. A.; Hofrichter, J.; Davies, D. R. High-Resolution x-Ray Crystal Structures of the Villin Headpiece Subdomain, an Ultrafast Folding Protein. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102* (21), 7517–7522.

(43) Kubelka, J.; Chiu, T. K.; Davies, D. R.; Eaton, W. A.; Hofrichter, J. Sub-Microsecond Protein Folding. *J. Mol. Biol.* **2006**, *359* (3), 546–553.

(44) Piana, S.; Lindorff-Larsen, K.; Shaw, D. E. Protein Folding Kinetics and Thermodynamics from Atomistic Simulation. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109* (44), 17845–17850.

(45) Banushkina, P. V.; Krivov, S. V. High-Resolution Free-Energy Landscape Analysis of α-Helical Protein Folding: HP35 and Its Double Mutant. *J. Chem. Theory Comput.* **2013**, *9* (12), 5257–5266.

(46) Best, R. B.; Hummer, G.; Eaton, W. A. Native Contacts Determine Protein Folding Mechanisms in Atomistic Simulations. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110* (44), 17874–17879.

(47) Sittel, F.; Jain, A.; Stock, G. Principal Component Analysis of Molecular Dynamics: On the Use of Cartesian vs. Internal Coordinates. *J. Chem. Phys.* **2014**, *141* (1).

(48) Jain, A.; Stock, G. Hierarchical Folding Free Energy Landscape of HP35 Revealed by Most Probable Path Clustering. *J. Phys. Chem. B* **2014**, *118* (28), 7750–7760.

(49) Ernst, M.; Sittel, F.; Stock, G. Contact- and Distance-Based Principal Component Analysis of Protein Dynamics. *J. Chem. Phys.* **2015**, *143* (24).

(50) Mori, T.; Saito, S. Molecular Mechanism behind the Fast Folding/Unfolding Transitions of Villin Headpiece Subdomain: Hierarchy and Heterogeneity. *J. Phys. Chem. B* **2016**, *120* (45), 11683–11691.

(51) Nagel, D.; Weber, A.; Lickert, B.; Stock, G. Dynamical Coring of Markov State Models. *J. Chem. Phys.* **2019**, *150* (9), 094111.

(52) Reiner, A.; Henklein, P.; Kiefhaber, T. An Unlocking/Relocking Barrier in Conformational Fluctuations of Villin Headpiece Subdomain. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107* (11), 4955–4960.

(53) Beauchamp, K. A.; McGibbon, R.; Lin, Y. S.; Pande, V. S. Simple Few-State Models Reveal Hidden Complexity in Protein Folding. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109* (44), 17807–17813.

(54) Serrano, A. L.; Bilsel, O.; Gai, F. Native State Conformational Heterogeneity of HP35 Revealed by Time-Resolved FRET. *J. Phys. Chem. B* **2012**, *116* (35), 10631–10638.

(55) Scherer, M. K.; Trendelkamp-Schroer, B.; Paul, F.; Pérez-Hernández, G.; Hoffmann, M.; Plattner, N.; Wehmeyer, C.; Prinz, J.-H.; Noé, F. PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. *J. Chem. Theor. Comput.* **2015**, *11* (11), 5525–5542.

(56) Wang, E.; Tao, P.; Wang, J.; Xiao, Y. A Novel Folding Pathway of the Villin Headpiece Subdomain HP35. *Phys. Chem. Chem. Phys.* **2019**, *21* (33), 18219–18226.

(57) Nagarajan, S.; Xiao, S.; Raleigh, D. P.; Dyer, R. B. Heterogeneity in the Folding of Villin Headpiece Subdomain HP36. *J. Phys. Chem. B* **2018**, *122* (49), 11640–11648.

(58) Piana, S.; Lindorff-Larsen, K.; Shaw, D. E. How Robust Are Protein Folding Simulations with Respect to Force Field Parameterization? *Biophys. J.* **2011**, *100* (9), L47–L49.

(59) Hu, K. N.; Yau, W. M.; Tycko, R. Detection of a Transient Intermediate in a Rapid Protein Folding Process by Solid-State Nuclear Magnetic Resonance. *J. of the Am. Chem. Soc.* **2010**, *132* (1), 24–25.

(60) Do, C. B.; Batzoglou, S. What Is the Expectation Maximization Algorithm? *Nat Biotechnol* **2008**, *26* (8), 897–899.

(61) Sargsyan, K.; Grauffel, C.; Lim, C. How Molecular Size Impacts RMSD Applications in Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **2017**, *13* (4), 1518–1524.

(62) Plimpton, S. Fast Parallel Algorithms for Short-Range Molecular Dynamics. *J. Comp. Phys.* **1995**, *117* (1), 1–19.

CHAPTER 4. AUTOMATED ENZYME QM-MODEL GENERATION WITH PYTHON WORKFLOWS

4.1 Introduction

Greater understanding of enzymatic mechanisms aids the discovery of new targets for biologics, the development of biocatalytic transformations, and *de novo* enzyme design. Methods using quantum mechanical (QM) potentials, such as Density Functional Theory (DFT), have enabled complex multistep enzymatic mechanisms to be studied, often in quantitative detail. Nevertheless, there are two fundamental areas of improvement in this field that limit widespread application and predictive power, especially in the absence of experimental data. The first regards the standardization and reproducibility of multi-scale methods applied in computational biocatalysis. The second pertains to the differences between computational modeling and experimental observables; the former provides a microscopic description, while the latter is inherently macroscopic.

Computational biocatalysis demonstrates significant barriers to reproducibility, as made evident in a recent study that shows simply reporting QM model set-up is not sufficient for reproducibility.[1] Even default algorithm parameters must be reported as differences across software and versions can quantitatively alter results. The growing list of seemingly minor procedural details required for reproducibility compounds the issue further by increasing the chances of misreporting at the time of publication and introducing modeling inconsistencies or errors into the investigation. As noted by the National Academies of Sciences, Engineering, and Medicine in a 2019 consensus report, *Reproducibility and Replicability in Science,* computational workflows are exceedingly valuable to facilitate reproducibility.

I have established an open-source, MIT licensed, Python code called enzyASM that takes fundamental steps towards improving reproducibility and aiding high-throughput investigation in QM-based enzyme modeling. To the best of my knowledge, enzyASM is the first publicly available module that automates truncated enzyme active site model (ASM) generation. The current functionalities are rooted in the fundamental aspects of developing an input structure for QM geometry optimization and described in more detail in Section 4.2.1. Expansion of the enzyASM module and its integration with existing toolkits to facilitate a broad range of applications is discussed in Section 4.2.2. The code is freely available via the GitHub repository (https://github.com/hklem/enzyASM), that also contains a Jupyter notebook tutorial example that generates an initial Gaussian input file for geometry optimization from the PDB structure 3zr4[2] of IGPS.

4.2 The enzyASM Framework

The only non-native dependency of the enzyASM code is the RDKit package, a widely used open-source cheminformatics software written in C++.[3] RDKit enables efficient processing and manipulation of atomic- and residue-level properties and attributes that come from Protein Data Bank[4] (PDB) data. Open Babel,[5] another common, open-source chemistry toolbox, is not required, however, in examples is used to generate protonated structures. The Open Babel protonate function is particularly useful because it will protonate structures according to a specified pH. The last not-required, but highly recommended package to use in conjunction with enzyASM is the Automated Quantum Mechanical Environment (AQME).[1] AQME has various modules that facilitate automation and reproducibility in QM workflows. Although it was originally developed with small molecule applications in mind, collaborations are ongoing to supplement

macromolecular support. Integration of enzyASM with AQME requires Python version 3.6 or newer.

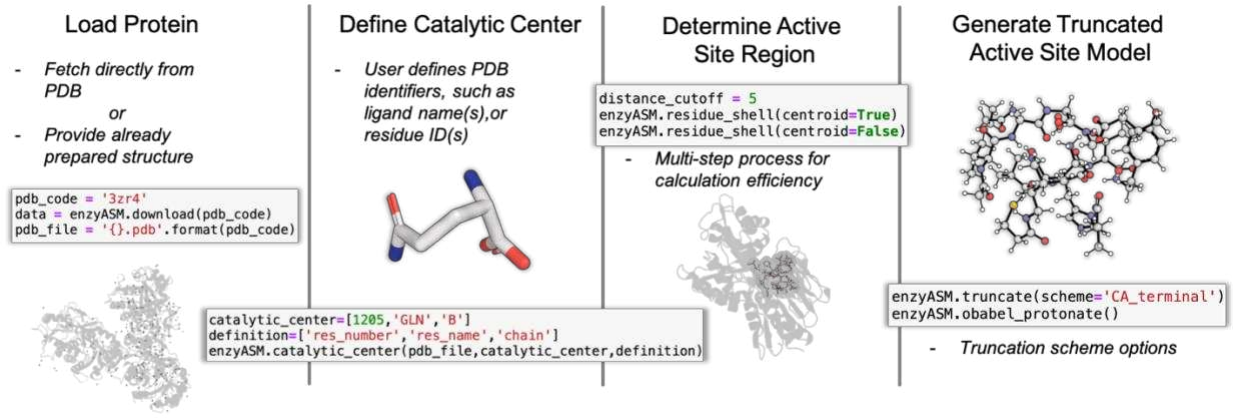


Figure 4.1 Process of enzyASM to generate truncated active site models. Images shown are direct snapshots from the PyMOL visualization script available in the enzyASM GitHub. Code snippets are taken from the Jupyter notebook example, also on the enzyASM GitHub.

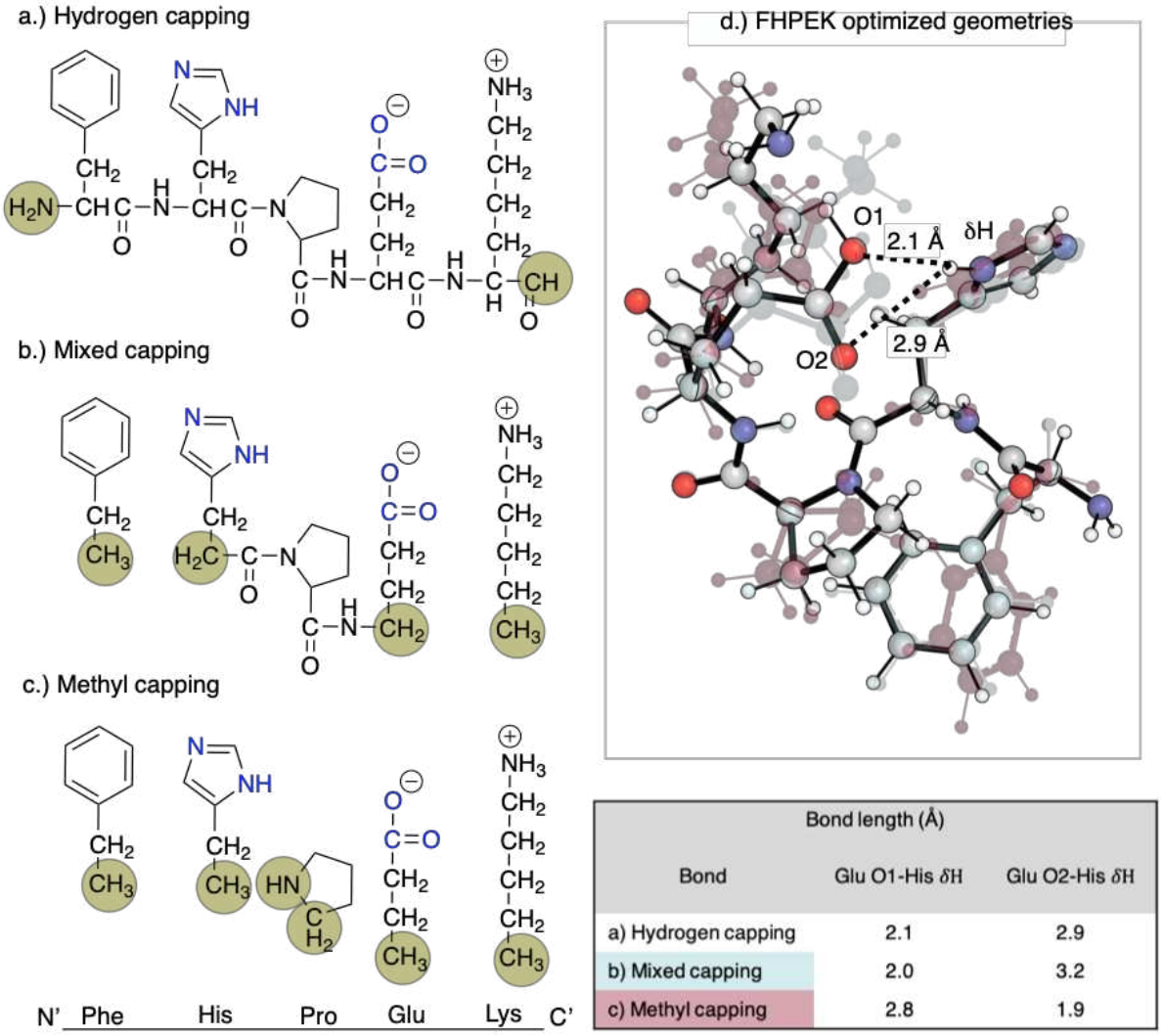
Code initialization. The starting point in the enzyASM framework can vary depending on its application. In the most rudimentary scenario, it begins with the 4-letter PDB code assigned to the system of interest. Given this, enzyASM will download the ".pdb" file directly from the Protein Data Bank server and parse it for relevant information such as if hydrogen atoms or ligands are present and number of amino acid residues (Figure 4.1). In other applications, the user may opt to supply a prepared ".pdb" file, for example one generated during MD simulation. Other than the PDB code or ".pdb" file, the user is only required to define a catalytic center and cutoff distance.

Active site reduction. The catalytic center describes the core of the ASM, and its definition can impact the quality of the generated model. The reacting substrate is often a reasonable choice unless it has not been crystallized in a catalytically competent pose. In that case, it may be farther than expected from the residues that are functionally relevant. The user is cautioned to select an appropriate input structure to avoid such concerns. Once the catalytic center is defined, the active site region is selected via an *ad hoc* distance cutoff (Figure 4.1). All residues, including water

molecules, with at least one atom within a user-defined distance from an atom of the catalytic center are kept in active site model. This requires the calculation of pairwise-atomic distances between catalytic center atoms and all other atoms, which can become costly. This calculation is expedited by performing an initial approximate reduction of atoms, in which distances are measured only from the catalytic center centroid. The distance cutoff is buffered by taking the maximum distance between the catalytic center centroid and its atoms, in doing so it is ensured that all residues that would fall within the distance cutoff are maintained. Once the first pass is complete, the function runs a second time, during which the distances are measured from all catalytic center atoms. In the case of the 3zr4 IGPS structure (used in the example provided in the enzyASM GitHub repository), with the catalytic center defined by the 10 heavy atoms of the Gln ligand, this two-stage approach requires 95,012 less distance measurements.

Truncation and capping schemes. For modeling via the QM-cluster approach, different truncation schemes can, and should be considered. To highlight the influence different truncation schemes can have, I will model the *h*Phe177—*h*His178—*h*Pro179—*h*Glu180—*h*Lys181 (FHPEK) strand in the IGPS active site (Figure 4.2). Gas-phase geometry optimizations were performed at the ωB97XD/6-31+G(d) level of theory, and all Cα atoms and capping hydrogens were frozen. Truncating each residue in FHPEK at the Cα position (methyl capping) removes an interaction between the Glu180 sidechain atom, O2, and its amide backbone that is maintained in the hydrogen capping and mixed capping examples. As a result, Glu180 has more freedom and forms a stronger hydrogen bond between its O2 and the His178 δH (1.9 Å) versus the hydrogen capping (2.9 Å) and mixed schemes (3.2 Å). These structural differences do not alter any conclusions from CHAPTER 2 because the study focused on how structural differences between active site geometries altered the reaction energetics, and the interaction between Glu180 and

His178 was not distinctly different in those evaluated models. This exception, however, would not apply to investigations that aim to study the detailed reaction mechanism resulting from a single model. In such situations, the influence of different truncation schemes should be examined. The enzyASM code facilitates this by providing the functionality to generate multiple models in accordance with different truncation schemes (Figure 4.1).



Bond length (Å)		
Bond	Glu O1-His δH	Glu O2-His δH
a) Hydrogen capping	2.1	2.9
b) Mixed capping	2.0	3.2
c) Methyl capping	2.8	1.9

Figure 4.2. Truncation schemes applied to peptide FHPEK from the IGPS HisH subunit. a) Hydrogen capping: the terminal ends are cut, and a single hydrogen is added to each end. b) Mixed capping: the peptide backbone atoms are removed from the terminal residues and the Cα are capped, creating methyl groups, His and Glu are hydrogen capped at the N- and C-termini, respectively. c) Methyl capping: all peptide backbone atoms are removed from each residue and the alpha carbons methyl capped. d) Optimized geometries of each truncation scheme. The translucent red structure is the methyl capping scheme, the translucent blue is the mixed capping scheme, and the solid atom-colored model is the hydrogen capping scheme. The table shows O1 and O2 distances from the His δH resulting from different truncation schemes.

4.3 Ongoing Developments

Validation of multiscale enzyme modeling is nontrivial due to a lack in rigorous guidelines for best practices. Accuracy of results depends critically on QM-region design aspects, such as which atoms to include and what geometrical configuration(s) to evaluate. A variety of residue selection schemes exist; however, their general advantages, limitations and applicability are vastly underexplored. The enzyASM code will be extended to accommodate QM-region selection schemes other than the *ad hoc* distance cutoff. Furthermore, additional integration with AQME will continue to be developed.

QM region validation. Chemical intuition and prior studies are often used to define the QM region; however, schematic approaches are increasingly popular because they can be implemented in automated workflows that avoid bias. A distance-based scheme, although straightforward and easy to implement, often results in very large QM regions (500–1000 atoms) to converge important properties like partial charges.[6] Charge shift analysis (CSA) was developed as an atom-economical approach that selects residues based on the magnitude of per-residue partial charge redistribution in the presence versus absence of the substrate.[7] CSA still requires large (~1000 atoms) single-point calculations (no geometry optimization) at the outset to inform the economized QM region. Fukui shift analysis (FSA) was developed as a more efficient alternative.[8] The condensed Fukui function is a well-established conceptual DFT reactivity descriptor that summarizes local reactivity change in response to adding or removing an electron.[9,10] In FSA, the condensed Fukui functions are first calculated for the minimal, chemically essential components. Then, these values are recomputed as surrounding residues are added one at a time in systematic, parallelizable single-point calculations. The root-sum-squared (RSS) difference from the minimal model is used to measure the impact of each added residue. enzyASM will be developed to

automate the generation of models following a variety of selection schemes, and calculation of subsequent quantities of interest ranging from geometric (e.g., hydrogen bond distances, dihedral angles) to energetic/electronic descriptors (e.g., activation and reaction energies, pKa gradients, charges, Fukui indices, bond dipoles, electrostatic field). The results will be corroborated by experimental data when available and compared with previous models published in the literature. The broad scope of QoI used for assessment will provide valuable insight into the sensitivity and robustness of criteria to QM-model design aspects.

Integration with existing workflows. Currently, the AQME QPREP module is incorporated in the enzyASM Jupyter notebook example to automate the generation of QM input files. AQME was developed for small-molecule applications, therefore calculation details common in enzyme modeling (e.g., boundary conditions and system constraints) are not accounted for. Upcoming versions will of AQME and enzyASM will accommodate additional AQME modules such as CSEARCH, to perform conformational searching, and QCORR to assess and record output calculation data and metadata (Figure 4.3).

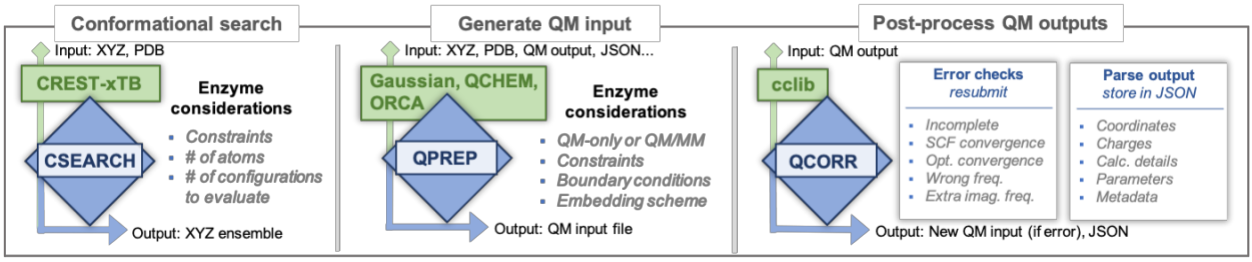


Figure 4.3 AQME modules that will be used in automating enzyASM.

Conformationally heterogenous active site modeling. The field of enzymology has recently evolved from structure-function to an ensemble-function paradigm due to state-of-the-art applications of molecular dynamics simulations and experiments, room-temperature (RT) and multi-state X-ray crystallography,[11–13] and cryogenic electron microscopy (cryoEM)[14–17] that

probe the conformational ensemble. From a statistical mechanics perspective, experimentally measured catalytic properties are macroscopic observables averaged over an ensemble of accessible microscopic states. However, computational predictions are often based on either a single value from a static structure or averaged-values from snapshots in short molecular dynamics (MD) trajectories.[18] Alternatively, the combination of extensive MD sampling and subsequent structural clustering has been proposed to identify and quantify relevant states required to compute accurate properties of population weighted QM enzyme models.[19] This approach is supported by recent advances in biomolecule clustering algorithms[20] (CHAPTER 3) and discussed in more detail in CHAPTER 5.

The small-molecule computational community routinely performs conformational sampling with semiempirical QM (SQM) methods like GFN2-xTB, a density functional tight binding potential parametrized for all biologically relevant elements.[21] GFN2-xTB optimized protein structures provide similar, and in some cases better, results than force fields specialized for protein simulation.[22] The Conformer-Rotamer Ensemble Sampling Tool (CREST), implemented by the free quantum chemistry package xTB,[23,24] can be used with GFN2-xTB to efficiently explore low-energy conformational space. CREST employs the iMTD-GC algorithm to perform iterative Metadynamics simulations to explore conformational space by adding bias potentials to the structure and final structure sorting via a genetic-crossing procedure.[24] CREST has been successfully applied to the prototypical enzyme catechol O-methyltransferase to identify lower energy transition state geometries.[23]

The application of the CREST-xTB framework in a conformational sampling workflow for enzyme models may yield active site configurations that enable more realistic, ensemble-informed modeling, and result in a generalizable procedure for the automated workflow. Although this

approach does not provide information regarding conformational transition timescales, the configurations can be used to inform biasing potentials for enhanced sampling MD to provide such information, supplementing an active area of research.[25] Furthermore, this approach may be particularly promising for predictive-structure applications.[26] For example, AlphaFold2 structures are averaged over ligand-free and ligand-bound states and found to structurally distort the active site to some extent,[27] necessitating model flexibility to accommodate for the structural uncertainties. There is concern that the SQM conformational sampling approach might not yield realistic configurations in the absence of the greater protein environment. Although I expect proper parameter tuning to resolve such issues, this would not be seen as a failure. A contingency plan will use these configurations to inform *de novo* design principles, inspired by the concept of theozyme modeling, to address questions like *what is an optimal arrangement of residues to favor pathway A over pathway B?*

REFERENCES

(1) Giudetti, G.; Polyakov, I.; Grigorenko, B. L.; Faraji, S.; Nemukhin, A. V.; Krylov, A. I. How Reproducible Are QM/MM Simulations? Lessons from Computational Studies of the Covalent Inhibition of the SARS-CoV-2 Main Protease by Carmofur. *J. Chem. Theory Comput.* **2022**, *18* (8), 5056–5067.

(2) List, F.; Vega, M. C.; Razeto, A.; Häger, M. C.; Sterner, R.; Wilmanns, M. Catalysis Uncoupling in a Glutamine Amidotransferase Bienzyme by Unblocking the Glutaminase Active Site. *Chem. Biol.* **2012**, *19* (12), 1589–1599.

(3) Landrum, G. RDKit: Open-Source Cheminformatics.

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

(5) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An Open Chemical Toolbox. *J. Cheminformatics* **2011**, *3* (1), 33.

(6) Kulik, H. J.; Zhang, J.; Klinman, J. P.; Martínez, T. J. How Large Should the QM Region Be in QM/MM Calculations? The Case of Catechol *O* -Methyltransferase. *J. Phys. Chem. B* **2016**, *120* (44), 11381–11394.

(7) Karelina, M.; Kulik, H. J. Systematic Quantum Mechanical Region Determination in QM/MM Simulation. *J. Chem. Theory Comput.* **2017**, *13* (2), 563–576.

(8) Qi, H. W.; Karelina, M.; Kulik, H. J. Quantifying Electronic Effects in QM and QM/MM Biomolecular Modeling with the Fukui Function. *Acta Phys. Chim. Sin.* **2018**, *34* (1), 81–91.

(9) Yang, Weitao.; Mortier, W. J. The Use of Global and Local Molecular Parameters for the Analysis of the Gas-Phase Basicity of Amines. *J. Am. Chem. Soc.* **1986**, *108* (19), 5708–5711.

(10) Faver, J.; Merz, K. M. Utility of the Hard/Soft Acid−Base Principle via the Fukui Function in Biological Systems. *J. Chem. Theory Comput.* **2010**, *6* (2), 548–559.

(11) Fraser, J. S.; van den Bedem, H.; Samelson, A. J.; Lang, P. T.; Holton, J. M.; Echols, N.; Alber, T. Accessing Protein Conformational Ensembles Using Room-Temperature X-Ray Crystallography. *Proc. Natl. Acad. Sci.* **2011**, *108* (39), 16247–16252.

(12) Fenwick, R. B.; van den Bedem, H.; Fraser, J. S.; Wright, P. E. Integrated Description of Protein Dynamics from Room-Temperature X-Ray Crystallography and NMR. *Proc. Natl. Acad. Sci.* **2014**, *111* (4), E445–E454.

(13) Yabukarski, F.; Doukov, T.; Pinney, M. M.; Biel, J. T.; Fraser, J. S.; Herschlag, D. Ensemble-Function Relationships to Dissect Mechanisms of Enzyme Catalysis. *Sci. Adv.* **2022**, *8* (41), eabn7738.

(14) Bai, X.; McMullan, G.; Scheres, S. H. W. How Cryo-EM Is Revolutionizing Structural Biology. *Trends Biochem. Sci.* **2015**, *40* (1), 49–57.

(15) Glaeser, R. M. How Good Can Cryo-EM Become? *Nat. Methods* **2016**, *13* (1), 28–32.

(16) Nogales, E. The Development of Cryo-EM into a Mainstream Structural Biology Technique. *Nat. Methods* **2016**, *13* (1), 24–27.

(17) Bonomi, M.; Vendruscolo, M. Determination of Protein Structural Ensembles Using Cryo-Electron Microscopy. *Curr. Opin. Struct. Biol.* **2019**, *56*, 37–45.

(18) Ryde, U. How Many Conformations Need To Be Sampled To Obtain Converged QM/MM Energies? The Curse of Exponential Averaging. *J. Chem. Theory Comput.* **2017**, *13* (11), 5745–5752.

(19) Klem, H.; McCullagh, M.; Paton, R. S. Modeling Catalysis in Allosteric Enzymes: Capturing Conformational Consequences. *Top. Catal.* **2022**, *65* (1–4), 165–186.

(20) Klem, H.; Hocky, G. M.; McCullagh, M. Size-and-Shape Space Gaussian Mixture Models for Structural Clustering of Molecular Dynamics Trajectories. *J. Chem. Theory Comput.* **2022**, *18* (5), 3218–3230.

(21) Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *J. Chem. Theory Comput.* **2019**, *15* (3), 1652–1671.

(22) Schmitz, S.; Seibert, J.; Ostermeir, K.; Hansen, A.; Göller, A. H.; Grimme, S. Quantum Chemical Calculation of Molecular and Periodic Peptide and Protein Structures. *J. Phys. Chem. B* **2020**, *124* (18), 3636–3646.

(23) Pracht, P.; Bohle, F.; Grimme, S. Automated Exploration of the Low-Energy Chemical Space with Fast Quantum Chemical Methods. *Phys. Chem. Chem. Phys.* **2020**, *22* (14), 7169–7192.

(24) Grimme, S. Exploration of Chemical Compound, Conformer, and Reaction Space with Meta-Dynamics Simulations Based on Tight-Binding Quantum Chemical Calculations. *J. Chem. Theory Comput.* **2019**, *15* (5), 2847–2862.

(25) Sasmal, S.; McCullagh, M.; Hocky, G. M. Learning Reaction Coordinates for Enhanced Sampling Directly Using Atomic Positions. *Biophys. J.* **2023**, *122* (3), 335a.

(26) Pereira, J.; Simpkin, A. J.; Hartmann, M. D.; Rigden, D. J.; Keegan, R. M.; Lupas, A. N. High-accuracy Protein Structure Prediction in CASP14. *Proteins Struct. Funct. Bioinforma.* **2021**, *89* (12), 1687–1699.

(27) Scardino, V.; Di Filippo, J. I.; Cavasotto, C. N. How Good Are AlphaFold Models for Docking-Based Virtual Screening? *iScience* **2023**, *26* (1), 105920.

CHAPTER 5. THE FUTURE OF COMPUTATIONAL BIOCATALYSIS[§]

"Science will never be a finished book. Any development over time reveals more and deeper difficulties." –Albert Einstein

The dynamic interconversion of enzyme conformations between active and inactive states, often involves length- and timescales inaccessible to QM treatments. This presents an ongoing challenge for the development of computational models of enzymes with conformationally distinct catalytic states. The previous chapters lay important groundwork that motivates solutions to these current challenges. The DFT model of the glutamine hydrolysis reaction performed by IGPS revealed there are multiple structural solutions to lowering the activation energy barrier (CHAPTER 2). The accessibility of active site conformations has not been studied in depth; it is unknown if a single, or multiple structures contribute to the macroscopic behavior by existing as interconverting points along the complex, multidimensional potential energy surface of IGPS. As science is a never-ending effort, extension of the QM-based IGPS research to incorporate the structural clustering developments of CHAPTER 3 are discussed in this chapter. Lastly, a broadly applicable, rigorous workflow that combines MD, structural clustering, and QM is described as a solution that could facilitate computational biocatalysis in the 21st.

5.1 Continued Work

In CHAPTER 3, The shapeGMM clustering approach that was demonstrated to yield physically meaningful states along the protein folding process of HP35 has applicability to a broad range of biomolecules, including globular proteins. Ongoing efforts in the McCullagh research

[§]This chapter is composed of selected content and reproduced with permission from Klem, H., McCullagh, M., & Paton, R. S. Modeling Catalysis in Allosteric Enzymes: Capturing Conformational Consequences. *Topics in Catalysis* (2022) **65**:165–186.

group aim to evaluate the allosteric mechanism in the benchmark K-type allosteric enzyme liver pyruvate kinase (L-PYK) via shapeGMM clustering and a Hessian based allosteric pathway analysis we developed earlier in my graduate work.[1] IGPS is another viable system to illustrate the versatility of shapeGMM. Evaluation of the conformational space of the glutamine hydrolysis active site in the various IGPS substrate states can further elucidate the relationship between structure and catalytic function, and potentially validate conformational changes proposed by the minimal QM models designed in CHAPTER 2.

Preliminary Results. Extensive all atom conventional MD data of IGPS in its four substrate states were graciously provided by Yao and Hamelberg at Georgia State University. The original simulation data used for analysis in their allosteric pathways study[2] contains 10 μs trajectories of each complex with a frame rate of 1 ps, totaling 40 μs of simulation. To expedite the transfer process, the simulation data was reduced to contain only heavy backbone atoms every 5 ps, resulting in approximately 200 GB of data (as opposed to 3.5 TB of the original data). All simulation details can be found in their research publication.[2] The simulations reproduce only some structural aspects of the proposed active conformation in PDB 7ac8[3] chains E/F. Probability distributions of the backbone dihedrals ψ of *h*Gly50, φ and ψ of *h*Val51 were measured to identify oxyanion strand structural differences between the substrate state. Interestingly, the *h*Val51 dihedral flip associated with the presumed active conformation is observed only in the simulation with neither Gln nor PrFAR bound (Apo). The Gln-bound, PrFAR-bound, and Ternary (both Gln- and PrFAR-bound) simulations were found to stabilize the presumed inactive *h*Val51 φ angle. The Apo and PrFAR-bound states were found to sample multiple *h*Val51 ψ dihedral states in a region (–50° to +50°) distinct from the angles identified in crystallographic structures (approx. +80° for the active and approx. +120° for the inactive conformation). All simulations sampled the inactive

*h*Val51 ψ dihedral and none precisely sampled the active dihedral value. The angle between residues *h*Gly52, *h*W123 and *f*F120 is used to define the extent of interfacial opening. A hinge breathing motion has been previously connected to allosteric activation in IGPS, resulting in a reduced angle and increased frequency.[3–6] Although the extent of interfacial closure observed in 7ac8 chains E/F is not achieved in any simulation, the smallest angle is sampled most in the Ternary simulation, and the largest in Apo, following a trend consistent with previous studies.

Under the assumption that the relationship between structure and glutamine hydrolysis function can be adequately captured by local conformational changes in the glutaminase active site, the heavy backbone atoms of all residues used in the QM models presented in CHAPTER 2 are selected as features for clustering. This selection results in 60 atoms. With a frame rate of 5 ps, the trajectories from each IGPS simulation yield a combined 8 million frames. To make trajectory processing more efficient, every 5 frames were used for analysis. With all four simulations this results in an input data set of size (1,600,000 x 60 x 3) that contains backbone atom positions of active site residues every 10 ps. The trajectories were amalgamated in the following order: Apo, Gln-bound, PrFAR-bound, and Ternary.

To inform an appropriate number of clusters, the log likelihood per frame of weighted shapeGMM (wSGMM) trained as a function of cluster size was evaluated. An improved code of shapeGMM (available to pip install as shapeGMMTorch) that speeds up the singular value decomposition calculation was used for this work. Cluster sizes of 2, 4, 6, 8, 10, 15, 20, and 25 were investigated. Each wSGMM was trained on 40,000 frames and cross-validated on the remaining data. Results from five training-test splits are shown in Figure 5.1. The heuristic elbow method is not highly informative with these results, as there is not a single obvious reduction in the gain of log likelihood with increasing cluster size (as was the case in the beaded helix transition

example in CHAPTER 3, Figure 3.3.). This is not surprising, as it can be expected that identification of an optimal cluster size increases in difficulty with increasing simulation length and feature size, as was the case in the HP35 cluster scan (APPENDIX B. Figure B1). Nevertheless, the change in log likelihood decreases as a function of cluster size, particularly at a size of 6 and 10, although a selection of 10 might be misleading due to sizes 11, 12, 13 and 14 not being explicitly evaluated. Another important aspect of this IGPS cluster scan is the deviation between the training and cross-validation data, signifying model over-fitting. The extent of over-fitting is likely exacerbated by the ratio of train to validation data (1:40), however, a larger training size increases the calculation time significantly. Alternatively, modification of the input features may yield clearer results.

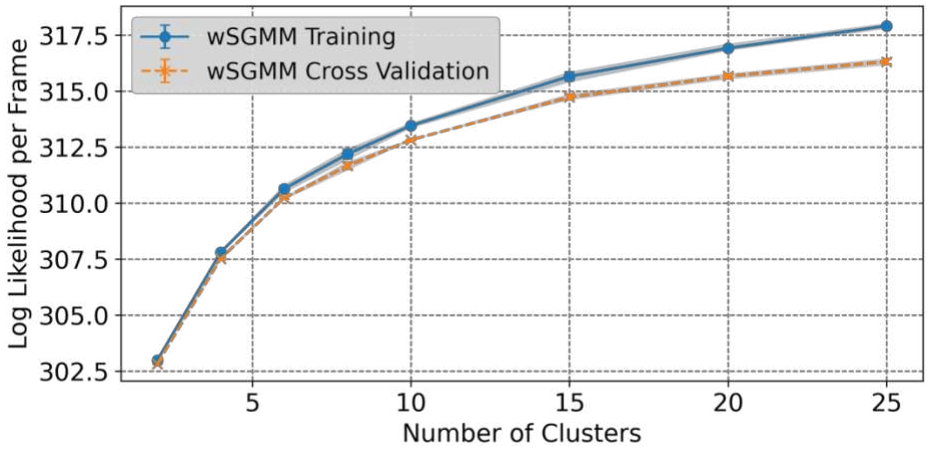


Figure 5.1. Log likelihood of weighted shapeGMM (wSGMM) as a function of number of clusters for the amalgamated IGPS substrate state trajectories. Each wSGMM was trained on 40,000 frames (blue solid) in a 5-fold cross validation scheme using the remainder of frames for validation (orange dashed). The shaded region indicated the 90% confidence interval.

A 6-state wSGMM was trained on 160,000 frames for further analysis, and the remaining frames were predicted from the resulting model. The state classification of each frame in trajectory order provides a simple overview of the differences between IGPS simulations (Figure 5.2). For example, in roughly the first half of the Apo trajectory, States 1, 2, 3, and 5 are sampled, while

States 4 and 6 are sampled only in the second half, and States 1, 2, 3 and 5 again near the end of the Apo simulation. While the first three states and State 5 are sampled in all trajectories (though to different extents) States 4 and 6 appear primarily in the Apo and PrFAR-bound simulations.

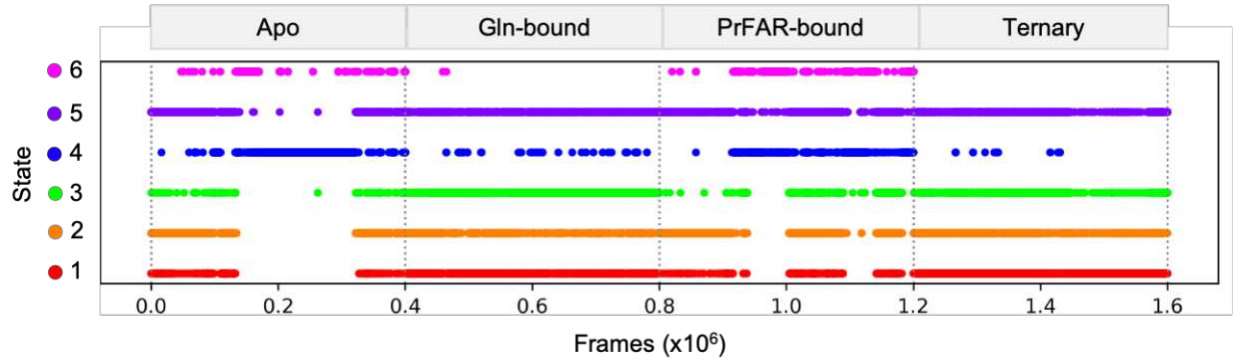


Figure 5.2. State discretization of the 6-state wSGMM across amalgamated trajectory frames of IGPS simulations.

The interfacial HisF residue *f*Gln123 and HisH oxyanion strand residues *h*Gly50, *h*Val51 and *h*Gly52 contribute to the classification of distinct states. Figure 5.3 displays the variance in position of each residue backbone atom according to the state it is ascribed. This graph is similar to Figure 3.6 of the 4-state HP35 wSGMM results in CHAPTER 2. The difference is that no estimate of sampling error has been performed yet. Although in the case of HP35 the standard deviation in the cluster feature variance was minimal (displayed error bars were 10 standard deviations from averages), the difference in variance that separated different states was larger (y-axis values ranged from near 0 to 70 Å^2). The range in IGPS active site residue backbone position variance is much smaller (near 0 to 1.2 Å^2). Sampling error may have a larger impact on conclusions drawn from these results, although another trend observed in the HP35 data is that features with lower variance had smaller deviations across different trajectory samples.

The current single-sampled results indicate the backbone in *f*Gln123 varies more than any other active site residue in every state, and that it fluctuates most in State 1, and least in State 3. All three oxyanion strand residues fluctuate most in State 4, and least in State 1. State 5 is

distinguished by higher variations in the backbone atomic displacements primarily of *h*Val51 and *h*Gly52, but interestingly not *h*Gly50 of the oxyanion strand. The backbone O atom of *h*Val51 in State 5 shows the largest spike in variance and is the outstanding feature of that state. The variance of features in shapeGMM, although useful to classify feature differences between and within states, it alone is not particularly informative on what those structural distinctions are.

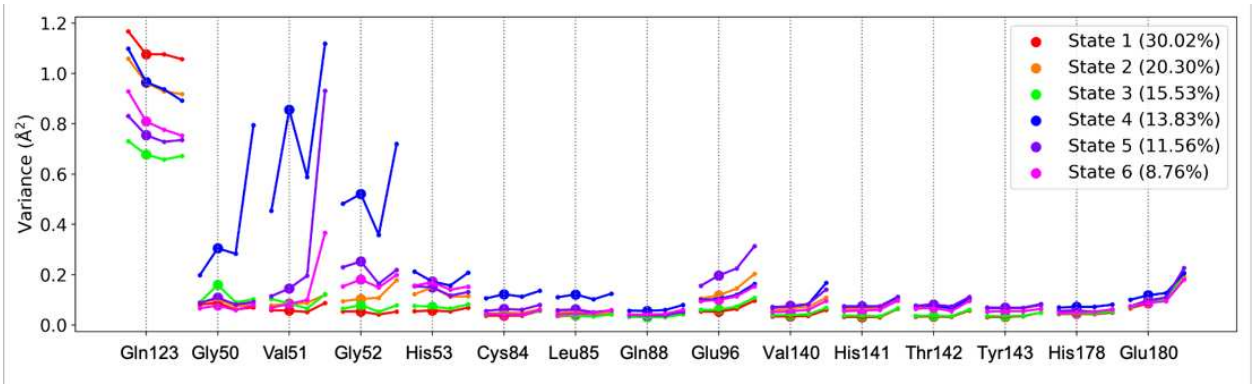


Figure 5.3. Variation in the backbone atomic displacement for each state in the 6-state wSGMM results for IGPS simulations. Enlarged points represent the Cα atoms of each residue. Atoms belonging to the same residue are connected by a line, whereas no line connects two different residues, even if they are neighbors in the IGPS sequence.

To gain a structural perspective on the changes that distinguish each state, ramacolor plots of each independent simulation are shown in Figure 5.4. The interfacial *f*Gln123 residue is expected to represent differences in the extent of interfacial opening, which is shown in these simulations to adopt a more closed conformation in all ligand bound simulations. The variety in the *f*Gln123 backbone atom position variance ascribed to each of the 6 states in Figure 5.3 is presumably due to differences in the proximity of *f*Gln123 to the HisH active site. The consistent green ramacolor assigned to *f*Gln123 across all states in each simulation confirms that its backbone dihedral is not distinguishing between the states.

Alternatively, the oxyanion strand residues are expected to sample dihedral conformations differently in the simulations. It should be noted that certain atoms necessary for a complete φ dihedral description were not included as features for clustering. For example, the backbone C

atom of *h*Pro49 required to determine the φ dihedral of *h*Gly50, was not included as a feature. This may have an impact on the clustering conclusions, especially since the *h*Gly50 φ dihedral has been previously proposed as a collective variable to describe active to inactive conformational changes in IGPS, in combination with the *h*Val51 φ dihedral.[7] Nevertheless, there are some commonalities in the averaged *h*Gly50 (φ,ψ) angles of certain states.

In particular, *h*Gly50 dihedrals are distinctly different in State 4 of each simulation. The probability distributions of the *h*Gly50 ψ dihedral reported by Yao and Hamelberg show two states at approximately 160° and 300° are significantly sampled in the Apo simulation. The value of 300° is consistent with the ramacolor difference of State 4 which is primarily and similarly populated by frames of the Apo and PrFAR-bound trajectories (apprx. 14% each). The 300° *h*Gly50 ψ dihedral was not observed in the probability distributions of the PrFAR simulation, however. This indicates another feature contributes to State 4 besides the *h*Gly50 backbone atoms. Figure 5.3 suggests the oxyanion strand residues collectively distinguish State 4 from the other states. Qualitatively, the ramacolors associated with residues *h*Gly50 and *h*Gly52 are consistent between the Apo and PrFAR-bound simulations, however *h*Val51 has a distinctly different ramacolor in the Apo State 4, compared to all other simulations, again indicating that additional structural factors are contributing to the classification of State 4. One possibility is that State 4 is defined by differences in the relative positioning of the oxyanion strand residues.

State 4 is visited particularly later in the trajectories of Apo and PrFAR-bound, as observed in Figure 5.3. It appears that States 4 and 5 are visited within the same time in PrFAR-bound simulations, whereas there is an obvious absence of State 5 in the Apo simulations around the same time State 4 is populated. The differences in the State 5 populations between Apo (11.39%) and PrFAR-bound (19.61%) reveal there is a possible population shift to State 5 in the presence of

PrFAR. This same behavior is also found in State 6, with an even larger deviation in population. Alternatively, States 1 and 3 are more populated in Apo than in the PrFAR-bound simulation. State 1 stands out in Figure 5.4 due to the high variance in the *f*Gln123 backbone atom positions, therefore it is expected that State 1 is distinguished by the dimer interface.

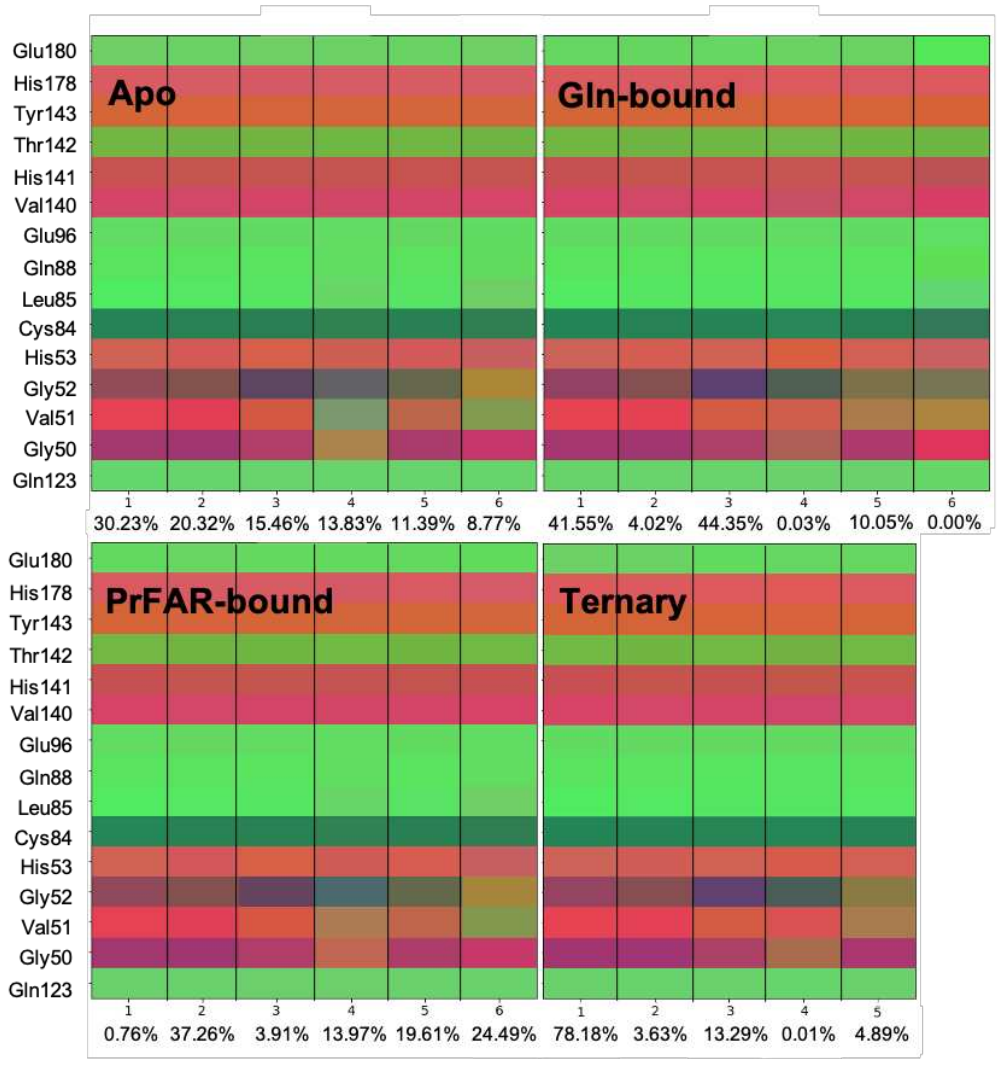


Figure 5.4. Ramacolor plots for the IGPS 6-state wSGMM divided into their corresponding simulations. The state populations are simulation-specific, and do not correspond to the overall state populations listed in Figure 5.3. Each simulation trajectory contains 400,000 frames with a frame step corresponding to 10 ps. State 6 is very minimally populated (<1,600 frames) in the Gln-bound simulation and not populated at all in the Ternary simulation.

The presented analyses are by no means complete; however, the preliminary results show promise in the application of shapeGMM to characterize active site structural change. The

similarities in how State 4 is sampled in Apo and PrFAR-bound simulations, in addition to the observed differences in the *h*Val51 dihedral pair are interesting and require a more comprehensive investigation of structural differences, such as comparison of the average structures that define each state. These commonalities in the Apo and PrFAR-bound simulations identifies in the shapeGMM cluster and reported by Yao and Hamelberg[2] are surprising and will be investigated further in future analyses. One of the more impactful outcomes of this preliminary clustering work is the emphasis on the oxyanion strand residues *h*Gly50, *h*Val51 and *h*Gly52 and the interfacial residue *f*Gln123, that all have evidence of involvement in the allosteric activation of IGPS. The notable differences in the ramacolors of *h*Gly52 across states, as well as distinctions of *h*Gly52 atom variations between each state may have important implications consistent with conclusions drawn from the IGPS QM models in CHAPTER 2; namely that *h*Gly52 can competently form the oxyanion hole in the absence of a *h*Val51 dihedral flip, but only if the oxyanion strand is optimally repositioned. Future directions include substructure matching of the constructed Inactive Val51 model of CHAPTER 2 to states identified from shapeGMM clustering. The feature selection may also be modified to fully describe the atoms involved in the *h*Gly50 φ.

5.2 Conclusions

A recurring theme in this Dissertation is that evaluation of conformational states is essential to model protein function, whether that is a folding process of a small protein like HP35, or the reaction of an allosterically regulated enzyme like IGPS. In general, biological activity is influenced by the populations and interconversions of conformational states and, specifically in the case of biocatalysis, the intrinsic reactivity (i.e., activation barrier) of these states. As discussed in Section 1.2, a combination of multi-scale computational methods is well-suited for this challenge. Therefore, a multistep computational workflow can be envisaged for rigorous modeling

of catalysis in enzymes with multiple functionally relevant conformational states such as IGPS (Figure 4.5). First, explicitly solvated aaMD simulation of the four different substrate states (E, E+X, E+X, E+X+S) is required. Statistically meaningful sampling of the dynamic equilibrium between inactive and active IGPS states is nontrivial and will undoubtedly require extensive simulation lengths and/or enhanced sampling techniques. Second, structural clustering of trajectories is performed to yield geometric representations of each state, along with populations. Third, multistep reaction Gibbs energy profiles for glutamine hydrolysis are generated for each conformational cluster from MD. Fourth, individual k_{cat} values are weighted according to the conformational populations determined from clustering to predict overall turnover frequency according to Equation 1.5 in Section 1.5.

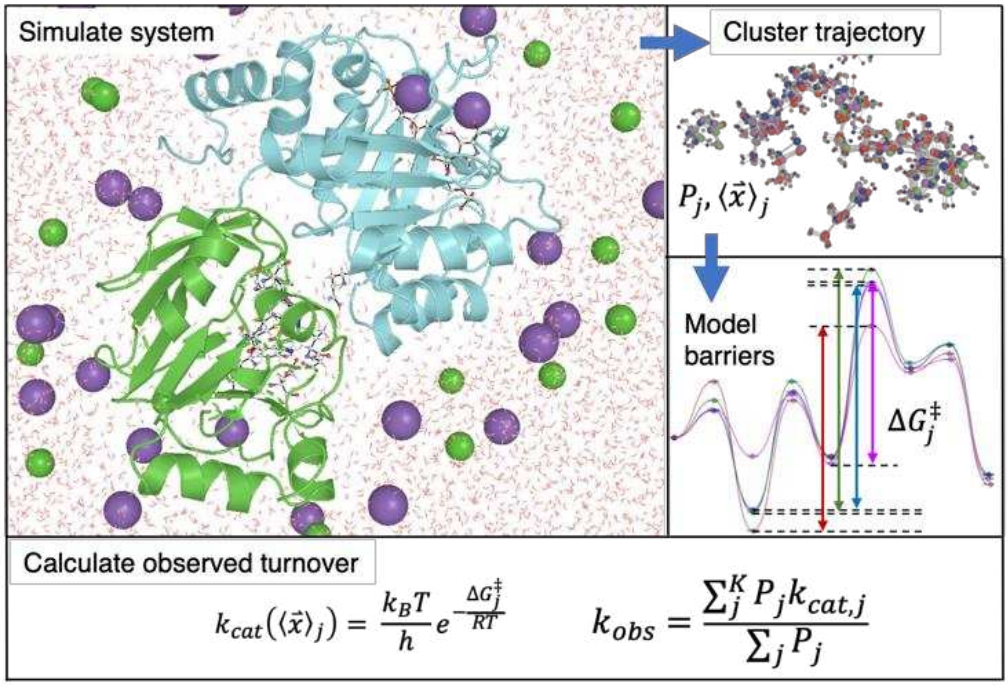


Figure 5.5 A stepwise workflow combining MD simulation, conformational clustering, and QM calculations to estimate the macroscopic catalytic turnover number. The first step is to simulate the system with explicit solvent all-atom MD. Next, structural clustering of the trajectory yields populations, P_j and cluster average structure coordinates, $\langle \vec{x} \rangle_j$. These coordinates are used to generate models for QM calculations of barrier heights of each cluster, $\Delta G_j^{\ddagger}$. The macroscopically observed turnover frequency, k_{obs} can be calculated as an expected value of $k_{cat,j}$ averaged over the conformational ensemble.

5.3 Final Remarks

The combination of MD sampling of the protein conformational ensemble and QM modeling of the enzymatic reaction provides a practical framework for modeling enzymatic mechanisms. Catalytically inactive X-ray structures, long timescales associated with protein motions, and conformational landscapes influenced by allosteric ligands present challenges for QM-only approaches that require the intervention of protein sampling. In this final chapter, I emphasize that a bridge between MD simulations and QM cluster models comes in the form of conformational clustering and analysis since this generates a tractable number of states that could then be used in QM cluster models. The enzyASM code discussed in CHAPTER 4 supplements this workflow by automating QM model generation, which may become tedious in scenarios with many conformational states to evaluate. This integrated approach is also suitable for use with QM/MM generated barriers, and future developments of enzyASM aim to accommodate for the various QM-involved modeling strategies. The choice of approach to calculate the reaction barriers may be based on user preference as well as attributes of the system of study. As the combination of MD sampling, conformational clustering, and subsequent energy barrier calculations are increasingly pursued, computational chemists will be able to access new mechanistic information about complex biological systems, such as allosteric and directed evolved enzymes. Although the computational framework is discussed here in the context of a V-type allosteric enzyme, IGPS, it can also be applied more generally to study the enzyme conformational ensemble and its importance for enzyme catalysis. While this approach is computationally demanding, advances in GPU hardware, enhanced sampling strategies, QM approaches, and calculation automation makes this tractable for numerous enzymes.

REFERENCES

(1) Lake, P. T.; Davidson, R. B.; Klem, H.; Hocky, G. M.; McCullagh, M. Residue-Level Allostery Propagates through the Effective Coarse-Grained Hessian. *J. Chem. Theory Comput.* **2020**, *16* (5), 3385–3395.

(2) Yao, X.-Q.; Hamelberg, D. Residue–Residue Contact Changes during Functional Processes Define Allosteric Communication Pathways. *J. Chem. Theory Comput.* **2022**, *18* (2), 1173–1187.

(3) Wurm, J. P.; Sung, S.; Kneuttinger, A. C.; Hupfeld, E.; Sterner, R.; Wilmanns, M.; Sprangers, R. Molecular Basis for the Allosteric Activation Mechanism of the Heterodimeric Imidazole Glycerol Phosphate Synthase Complex. *Nat Commun* **2021**, *12* (1), 2748.

(4) Amaro, R. E.; Sethi, A.; Myers, R. S.; Davisson, V. J.; Luthey-Schulten, Z. A. A Network of Conserved Interactions Regulates the Allosteric Signal in a Glutamine Amidotransferase. *Biochemistry* **2007**, *46* (8), 2156–2173.

(5) Kneuttinger, A. C.; Rajendran, C.; Simeth, N. A.; Bruckmann, A.; König, B.; Sterner, R. Significance of the Protein Interface Configuration for Allostery in Imidazole Glycerol Phosphate Synthase. *Biochemistry* **2020**, *59* (29), 2729–2742.

(6) Rivalta, I.; Lisi, G. P.; Snoeberger, N.-S.; Manley, G.; Loria, J. P.; Batista, V. S. Allosteric Communication Disrupted by a Small Molecule Binding to the Imidazole Glycerol Phosphate Synthase Protein–Protein Interface. *Biochemistry* **2016**, *55* (47), 6484–6494.

(7) Calvó-Tusell, C.; Maria-Solano, M. A.; Osuna, S.; Feixas, F. Time Evolution of the Millisecond Allosteric Activation of Imidazole Glycerol Phosphate Synthase. *J. Am. Chem. Soc.* **2022**, *144* (16), 7146–7159.

APPENDIX A. THE CATALYTIC EFFECTS OF ACTIVE SITE CONFORMATIONAL CHANGE IN THE ALLOSTERIC ACTIVATION OF IMIDAZOLE GLYCEROL PHOSPHATE SYNTHASE SUPPORTING INFORMATION

A.1 IGPS active site model components and protonation states

Table A1. IGPS active site models from PDB 7ac8 chains C/D (Inactive model) and E/F (Active model)

Number	Name	Chain	Charge	Side chain	Main chain	Atom constraints
121	Gly	E (C)	0	Kept	Cut at N-terminus	C-alpha
122	Ser	E (C)	0	Removed	Kept	None
123	Gln	E (C)	0	Kept	Kept	None
124	Ala	E (C)	0	Removed	Cut at C-terminus	C-alpha
50	Gly	F (D)	0	Kept	Cut at N-terminus	C-alpha
51	Val	F (D)	0	Kept	Kept	None
52	Gly	F (D)	0	Kept	Kept	None
53	His	F (D)	0	Removed	Cut at C-terminus	C-alpha
A84C	Cys	F (D)	0	Kept	Cut at N-terminus	C-alpha
85	Leu	F (D)	0	Kept	Cut at C-terminus	C-alpha
88	Gln	F (D)	0	Kept	Removed	C-alpha
96	Glu	F (D)	-1	Kept	Removed	C-alpha
140	Val	F (D)	0	Kept	Cut at N-terminus	C-alpha
141	His	F (D)	0	Removed	Kept	None
142	Thr	F (D)	0	Kept	Kept	None
143	Tyr	F (D)	0	Kept	Cut at C-terminus	C-alpha
178	His	F (D)	0	Kept	Removed	C-alpha
180	Glu	F (D)	-1	Kept	Removed	C-alpha
425 (418)	HOH	F (D)	0	–	–	–
426 (420)	HOH	F (D)	0	–	–	–
301	L-Gln	Substrate	0	–	–	–

Components of HisF residues 121,122 and 124 were included to expand the main chain around *f*Gln123 and allow the terminal C-alpha atoms of this molecular fragment to be frozen. This is done to give the *f*Gln123 residue slight flexibility in geometry optimizations. The same reasoning applies to the oxyanion strand molecular fragment, where only the terminal C-alpha

atoms are frozen. Atoms were constrained using the "–1" optimization flag syntax in Gaussian. This enforces the gradient of these atoms to be 0 during the optimization updates.

The initial protonation states are assumed since hydrogen atoms are not resolved in the X-ray crystal structure. The His178 Nδ protonation state is supported given the position relative to Glu180 in numerous IGPS crystal structures and the expected general acid/base behavior of histidine in enzyme mechanisms involving cysteine catalytic triads.[1] The protonation state of Cys84 is interesting to consider, because it is not necessarily required to reform in the catalytic cycle. Therefore, it is possible Cys84 protonation is obviated during multiple turnover events. A hybrid QM/MM study of a cysteine protease acylation mechanism found the ionic state of the Cys–His dyad (Cys^{-}, $HisH^{+}$) to be more stable in the free enzyme, whereas binding of the substrate shifted the equilibrium such that the neutral dyad became more stable.[2]

Glu180 hydrogen bonds with His178 in all structures and remains in the carboxylate form for the results discussed here. **Int1** geometries with a Glu180 carboxylic acid were evaluated and the difference in protonation was not found to alter any mechanistic conclusions. Furthermore, site-directed mutagenesis experiments support a mechanism where Glu180 serves a structural role to position His178, as opposed to directly effecting catalysis.[3]

A.2 Construction of the Inactive Val52 and Inactive *f*Gln123 models

To probe the individual structural differences between Active and Inactive, two additional models, Inactive *f*Gln123 and Inactive Val51, were constructed using the active geometry as the foundation. To construct Inactive Val51, the Active model was altered by manually flipping the Val51 amide such that the N–H is directed away from the binding site (Figure A2, top), mimicking this aspect of the Inactive conformation. Inactive *f*Gln123 was built by aligning the coordinates of the Cys84 S and reactive Gln C, N, and Oε atoms of the optimized Active- and Inactive-**TS3**

structures. This was done to acquire the position of *f*Gln123 in Inactive relative to the reactive center. After alignment, the atomic coordinates corresponding to the molecular fragment containing *f*Gln123 in Active are replaced with the corresponding Inactive coordinates (Figure A2, bottom).

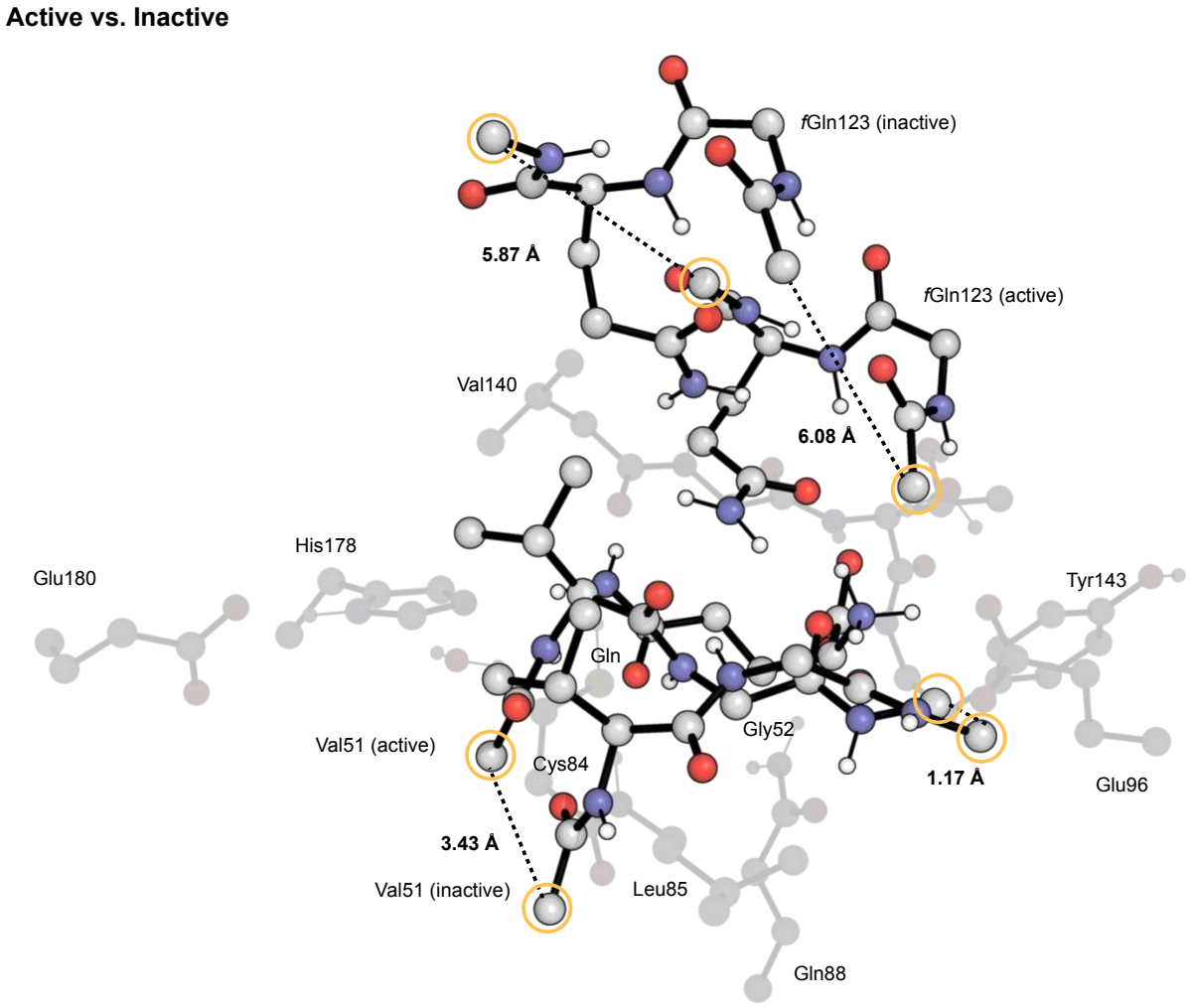


Figure A1. Geometric comparisons of the oxyanion strand and fGln123 position between Active and Inactive models. For clarity, transparent residues and the Gln substrate coordinates are only shown for Active. Coordinates from each model are taken from their optimized **ES** structures.

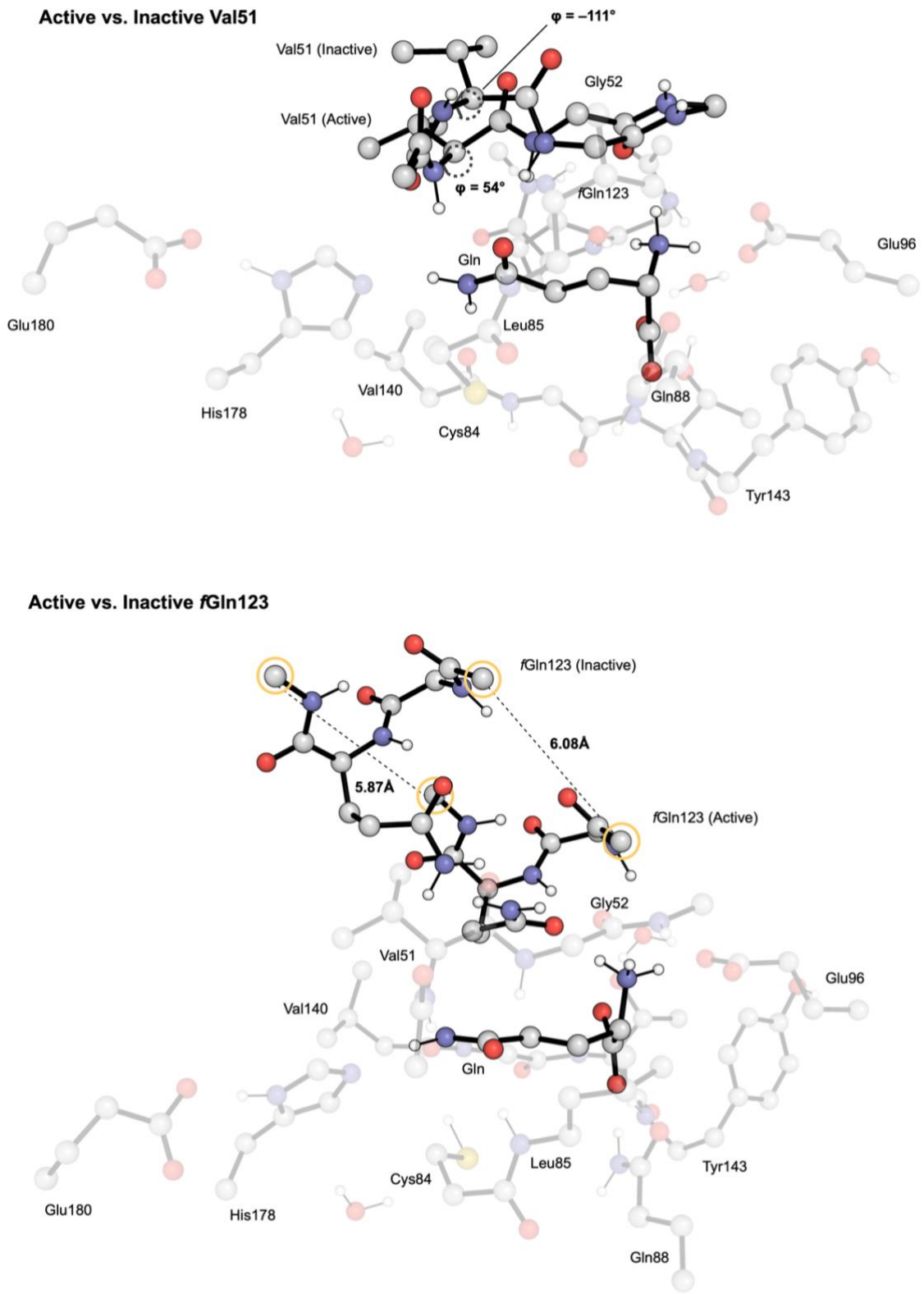


Figure A2. Geometric comparisons of Active and the Inactive Val51 constructed model (top) and the Inactive fGln123 constructed model (bottom). For clarity, transparent residues and the Gln substrate coordinates are only shown for Active. Coordinates from each model are taken from their optimized **ES** structures.

A.3 NBO calculation details and data

The **ES**, **TS2**, **TS3** and **TS4** stationary points were optimized in Inactive *f*Gln123 and Inactive Val51. Energy barriers were calculated for each TS relative to the **ES** structure of the corresponding model. The stabilizing oxyanion hole strength was evaluated in each TS by estimating the hydrogen bond interaction energy of Leu85, Gly52 and Val51 with the Gln Oε lone pair electrons via Natural Bonding Orbital (NBO) analysis version 7.0.5.[4]

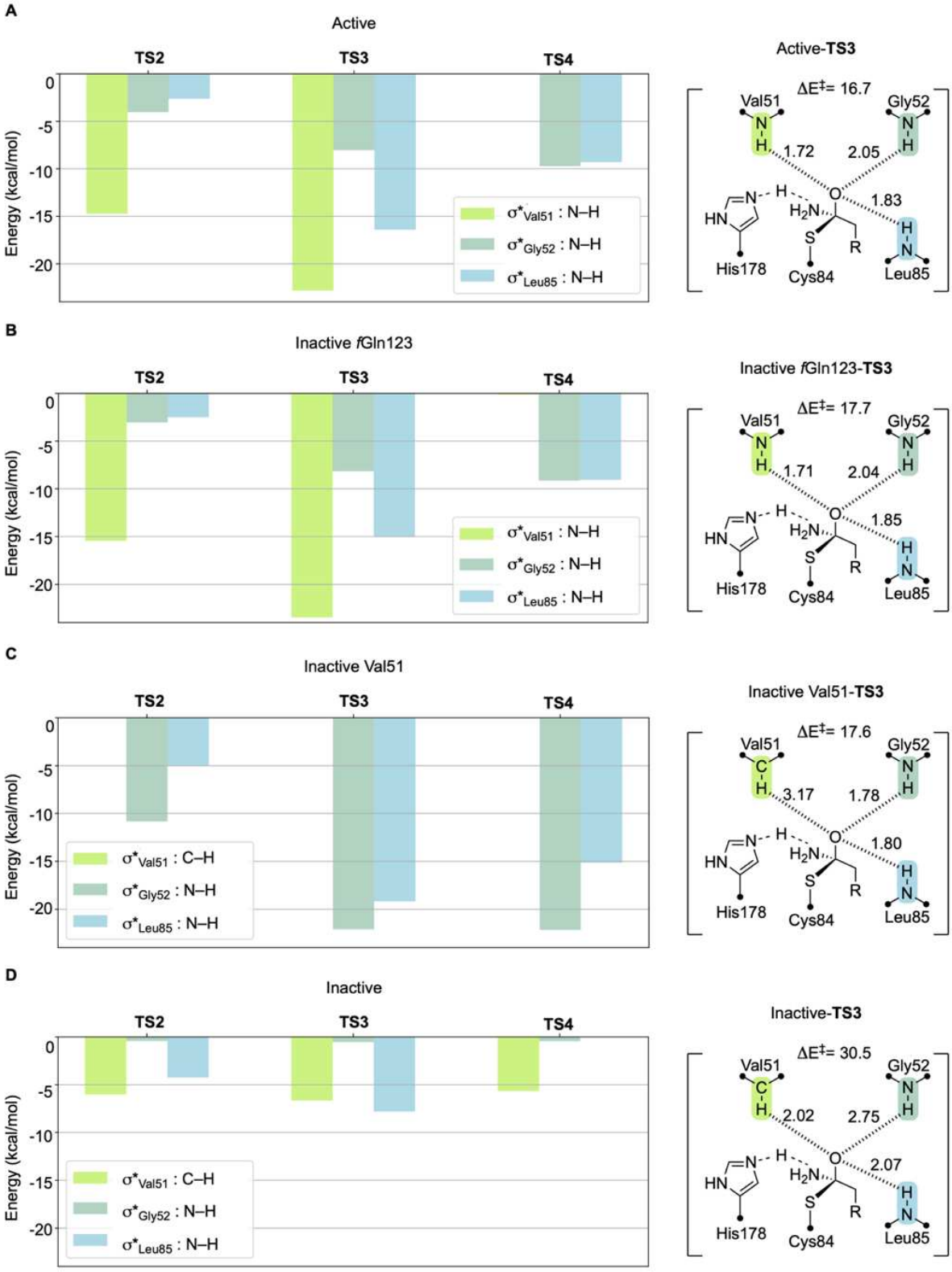


Figure A3. Left: Oxyanion hole contributions of **TS2**, **TS3** and **TS4**. Right: **TS3** oxyanion hole hydrogen bond distances for A) Active, B) Inactive fGln123, C) Inactive Val51, and D) Inactive models with barriers relative to each corresponding ES.

Table A2. Interactions between oxyanion hole components and Gln Oε le pair electrons. A value of 0 arises if there is either no interaction or the strength of that interaction is below the calculation cutoff (0.05 kcal/mol). "n/a" indicates the corresponding lone pair electrons were not detected in the structure.

Model	TS	Residue	Orbital	Interaction with LP 1 (kcal/mol)	Interaction with LP 2 (kcal/mol)	Interaction with LP 3 (kcal/mol)	Total
Active	TS1	Val51	BD*(1) N 22- C 23	4.25	10.49	n/a	14.74
Active	TS1	Gly52	BD*(1) N 29- H224	3.80	0.24	n/a	4.04
Active	TS1	Leu85	BD*(1) N 39- H154	2.63	0	n/a	2.63
Active	TS2	Val51	BD*(1) N 98- H232	7.19	11.31	4.31	22.81
Active	TS2	Gly52	BD*(1) N105- H228	3.43	1.49	3.13	8.05
Active	TS2	Leu85	BD*(1) N 5- H 53	7.71	0	8.70	16.41
Active	TS3	Val51	BD*(1) N202- H236	0	0	n/a	0
Active	TS3	Gly52	BD*(1) N203- H232	3.88	5.84	n/a	9.72
Active	TS3	Leu85	BD*(1) N 44- H 68	5.09	4.21	n/a	9.30
Inactive fGln123	TS1	Val51	BD*(1) N118- H152	4.79	10.64	n/a	15.43
Inactive fGln123	TS1	Gly52	BD*(1) N119- H148	2.74	0.30	n/a	3.04
Inactive fGln123	TS1	Leu85	BD*(1) N190- H205	2.51	0	n/a	2.51
Inactive fGln123	TS2	Val51	BD*(1) N161- H195	7.02	11.62	4.83	23.47
Inactive fGln123	TS2	Gly52	BD*(1) N162- H191	3.56	1.46	3.15	8.17
Inactive fGln123	TS2	Leu85	BD*(1) N 5- H 29	7.09	0	7.96	15.05
Inactive fGln123	TS3	Val51	BD*(1) N166- H200	0.06	0.05	n/a	0.11
Inactive fGln123	TS3	Gly52	BD*(1) N167- H196	3.70	5.44	n/a	9.14
Inactive fGln123	TS3	Leu85	BD*(1) N 8- H 32	4.89	4.17	n/a	9.06
Inactive Val51	TS1	Gly50	BD*(2) C232- O233	0.10	0.07	n/a	0.17
Inactive Val51	TS1	Val51	BD*(1) C 44- H 64	0	0	n/a	0
Inactive Val51	TS1	Gly52	BD*(1) N 41- H 65	9.16	1.66	n/a	10.82
Inactive Val51	TS1	Leu85	BD*(1) N 69- H 84	4.49	0.48	n/a	4.97
Inactive Val51	TS2	Gly50	BD*(2) C232- O233	0.23	0.11	n/a	0.34
Inactive Val51	TS2	Val51	BD*(1) C212- H140	0	0	0	0
Inactive Val51	TS2	Gly52	BD*(1) N117- H141	7.31	4.97	9.82	22.1
Inactive Val51	TS2	Leu85	BD*(1) N 5- H 30	7.20	2.34	9.63	19.17
Inactive Val51	TS3	Gly50	BD*(2) C232- O233	0.15	0.09	0	0.24
Inactive Val51	TS3	Val51	BD*(1) C205- H225	0	0	0	0
Inactive Val51	TS3	Gly52	BD*(1) N202- H226	7.71	3.36	11.09	22.16
Inactive Val51	TS3	Leu85	BD*(1) N 44- H 68	6.76	0	8.37	15.13
Inactive	TS1	Val51	BD*(1) C 23- H167	5.86	0.15	n/a	6.01
Inactive	TS1	Gly52	BD*(1) N 29- H175	0.08	0.34	n/a	0.42
Inactive	TS1	Leu85	BD*(1) N 39- H156	3.72	0.53	n/a	4.25
Inactive	TS2	Val51	BD*(1) C 23- H167	5.49	0.24	0.92	6.65
Inactive	TS2	Gly52	BD*(1) N 29- H175	0	0.21	0.33	0.54
Inactive	TS2	Leu85	BD*(1) N 39- H156	2.88	0.81	4.11	7.80
Inactive	TS3	Val51	BD*(1) C 23- H167	4.92	0.74	n/a	5.66
Inactive	TS3	Gly52	BD*(1) N 29- H175	0	0.43	n/a	0.43
Inactive	TS3	Leu85	BD*(1) N 39- H156	0	0	n/a	0

A.4 ALMO-EDA calculation details and data

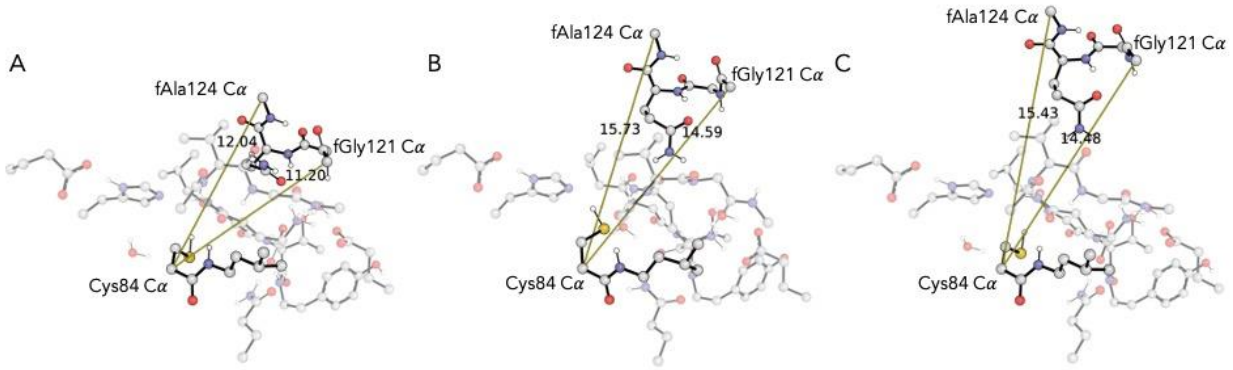


Figure A4. Distances from the frozen Cys84 Cα atom to the frozen *f*Gly121 and *f*Ala124 Cα atoms in A) Active, B) Inactive and C) Inactive *f*Gln123. These distances cannot change during geometry optimizations.

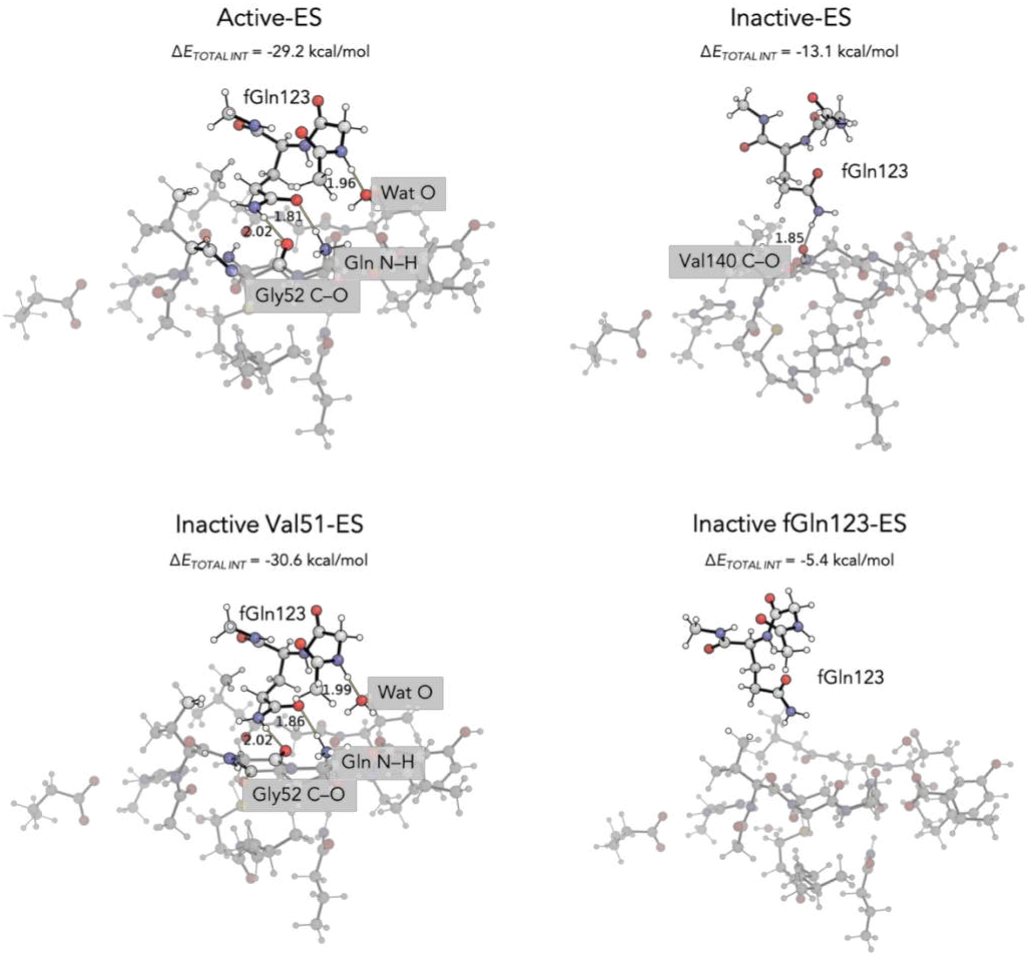


Figure A5. Differences of fGln123 interactions in the **ES** structure of each active site model that effect the fGln123 total interaction energy. Any atom that is not transparent is within 3 Å of an atom in the fGln123 molecular fragment, and it labeled.

The hydrogen bond to Val140 is not present in crystal structures (*f*Gln123:N—*h*Val140:O distance of 5.3 Å in 7ac8 chains C/D) because of a change in the *f*Gln123 rotamer state during geometry optimization. The optimized rotamer state in the Inactive model is structurally feasible due to the open dimer interface in the inactive conformation crystal structure, however it cannot be ruled out that it may be a consequence of the truncated model used here. In the Inactive model, *f*Gln123 forms a single hydrogen bond (1.85 Å) with its NH_2 group and the backbone carbonyl oxygen of Val140. In contrast, *f*Gln123 in the Inactive *f*Gln123 model does not form hydrogen bonds with other active site residues in any of the optimized structures. The presence of this hydrogen bond is likely the reason why the *f*Gln123 interaction energy is more favorable in the Inactive model than the Inactive *f*Gln123 construct, as opposed to the difference in oxyanion strand conformation. The models with *f*Gln123 in close proximity (Active and Inactive Val51) are highly consistent across all interaction energy terms, indicating the conformation of the oxyanion strand (and subsequently the orientation of the oxyanion hole) does not effect interactions with *f*Gln123. The absence of that hydrogen bond in the Inactive *f*Gln123 construct results from differences in the relative positioning of *f*Gln123, which was based on the Active model rather than the Inactive model.

Table S3. *f*Gln123 interaction energies for **ES** structures measured with ALMO-EDA.

Component	Active	Inactive Val51	Inactive *f*Gln123	Inactive
ES				
ΔELEC	-88.37	-89.54	-11.11	-35.39
ΔSOLV	16.34	15.17	5.64	12.85
ΔPAULI	92.68	94.75	5.91	23.19
ΔDISP	-31.94	-33.94	-4.89	-8.56
ΔPOL	-4.78	-4.36	-0.16	-1.67
ΔCT	-13.08	-12.73	-0.73	-3.47
ΔTotal INT	-29.15	-30.64	-5.35	-13.05
TS2				
ΔELEC	-86.6	-73	-9.9	-31
ΔSOLV	15.1	14.9	5.1	11.3
ΔPAULI	91.4	77.1	4.9	20.7
ΔDISP	-31.6	-31.7	-4.6	-8
ΔPOL	-4.5	-3	-0.1	-1.2
ΔCT	-12.8	-9.6	-0.7	-3
ΔTotal INT	-29	-25.3	-5.3	-11.2
TS3				
ΔELEC	-86.3	-72.1	-9.5	-30.3
ΔSOLV	15.3	14.7	4.5	10.8
ΔPAULI	90.6	75.8	5.6	20.5
ΔDISP	-31.4	-31.8	-4.7	-7.9
ΔPOL	-4.4	-2.9	-0.1	-1.2
ΔCT	-12.6	-9.6	-0.7	-3
ΔTotal INT	-28.8	-25.8	-5	-11.1
TS4				
ΔELEC	-87.2	-76.4	-9.3	-31.9
ΔSOLV	15.3	14.5	4.3	11.3
ΔPAULI	92.1	81	5.7	21.5
ΔDISP	-32.2	-33.5	-4.9	-7.5
ΔPOL	-4.5	-3	-0.2	-1.5
ΔCT	-12.7	-10.1	-0.7	-3.2
ΔTotal INT	-29.1	-27.5	-5.1	-11.4

A.5 Absolute energies from geometry optimizations and single point corrections for all evaluated structures

Table S4. Electronic energies at the geometry optimization, ωB97X-D/6-31G*(C,H,N);6-31+G*(S,O), and single point correction, B3LYP-D3(BJ)/6-311+G(2d,2p), levels of theory of each model evaluated.

Model	Structure	Geometry Optimization (kcal/mol)	Single Point Correction (kcal/mol)	Relative to ES Structure (kcal/mol)
Inactive	ES	-3580983.85	-3583478.03	0.00
	TS1	-3580981.35	-3583474.89	3.17
	Int1	-3580988.24	-3583480.54	-2.87
	TS2	-3580960.65	-3583454.82	23.22
	Int2	-3580961.90	-3583452.32	25.35
	TS3	-3580956.26	-3583447.30	30.51
	Int3	-3580961.90	-3583453.57	24.58
	TS4	-3580955.00	-3583451.06	26.60
	Int4	-3580959.39	-3583456.71	21.28
Active	ES	-3581009.57	-3583496.22	0.00
	TS1	-3580993.26	-3583481.79	14.52
	Int1	-3581008.31	-3583495.59	0.88
	TS2	-3580993.89	-3583481.79	14.49
	Int2	-3581002.67	-3583484.93	11.42
	TS3	-3580995.77	-3583479.28	16.70
	Int3	-3580997.65	-3583482.42	13.75
	TS4	-3580989.50	-3583477.40	18.67
	Int4	-3581003.30	-3583491.20	4.81
Inactive *f*Gln123	ES	-3580977.58	-3583472.39	0.00
	TS2	-3580963.79	-3583459.21	12.56
	TS3	-3580962.53	-3583454.20	17.64
	TS4	-3580957.51	-3583454.20	17.76
Inactive Val51	ES	-3581001.97	-3583490.92	0.00
	TS2	-3580991.34	-3583481.10	9.82
	TS3	-3580988.46	-3583474.45	16.47
	TS4	-3580977.34	-3583474.03	16.89

Absolutely Localized Molecular Orbitals based Energy Decomposition Analysis (ALMO–EDA) implemented in Q–Chem 5.4 was performed at the B3LYP-D3(BJ)/6-311+G(2d,2p) level of theory and CPCM with ε=4.24, to be consistent with the single-point energies used to produce

the reaction profiles.[5,6] Within this framework, the interaction energy is defined as $\Delta E_{INT} = \Delta E_{FRZ} + \Delta E_{POL} + \Delta E_{CT}$. E_{POL} and E_{CT} refer to the polarization and charge transfer energy components, respectively. The E_{FRZ} term combines the interaction energy contributions of the unrelaxed fragment densities, namely attractive dispersion (E_{DISP}), repulsive Pauli (E_{PAULI}) and permanent electrostatics (E_{ELEC}): $\Delta E_{FRZ} = \Delta E_{DISP} + \Delta E_{PAULI} + \Delta E_{ELEC}$.

A.6 Geometries and relevant atomic distances of each evaluated transition state

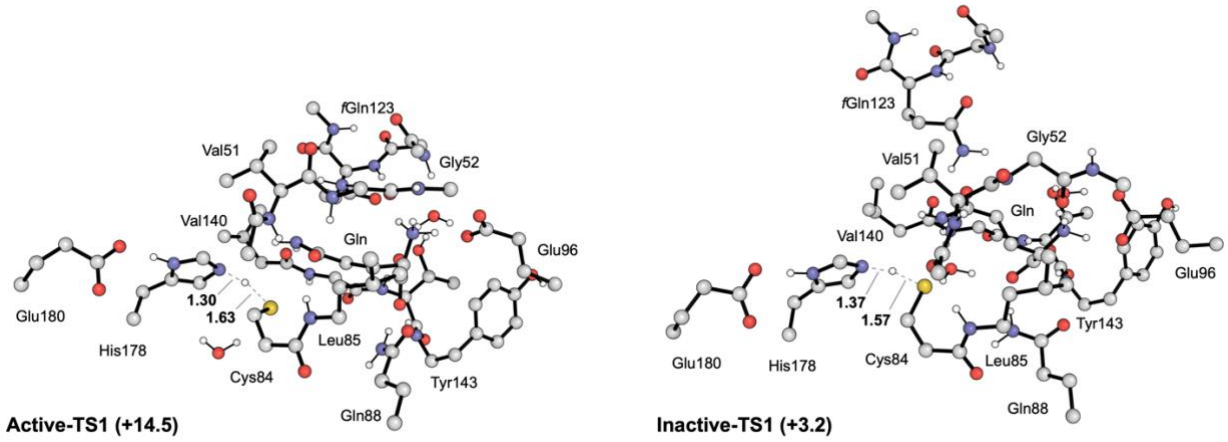


Figure A6. Optimized geometries of **TS1**.

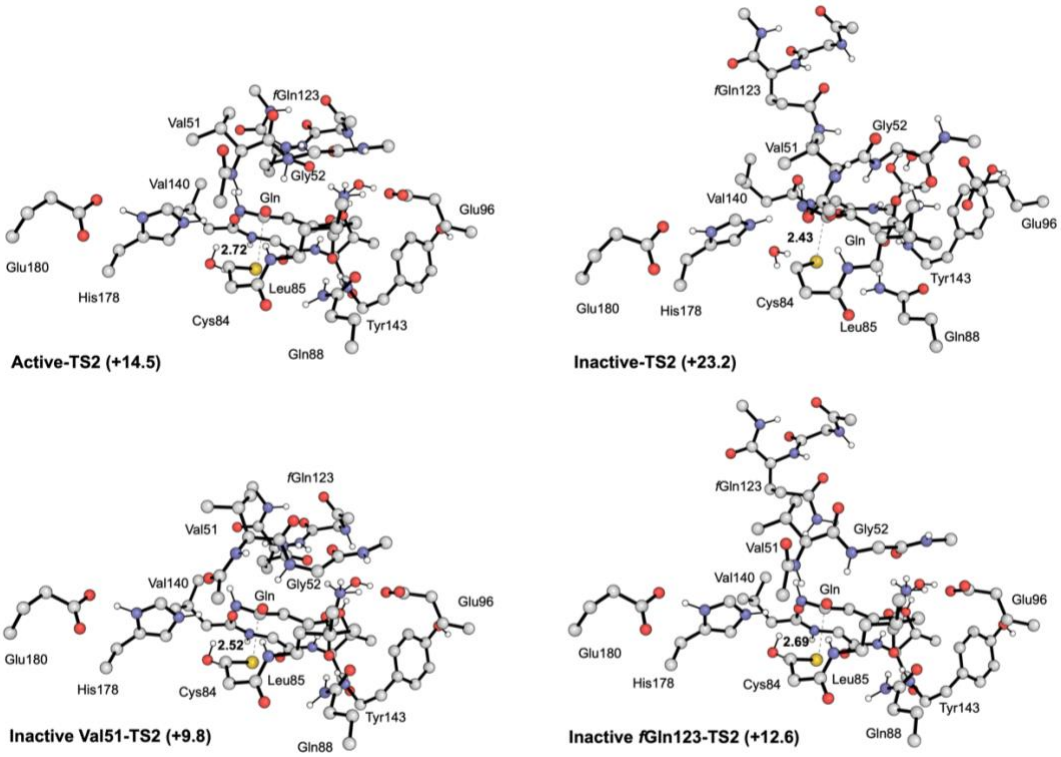


Figure A7. Optimized geometries of **TS2**.

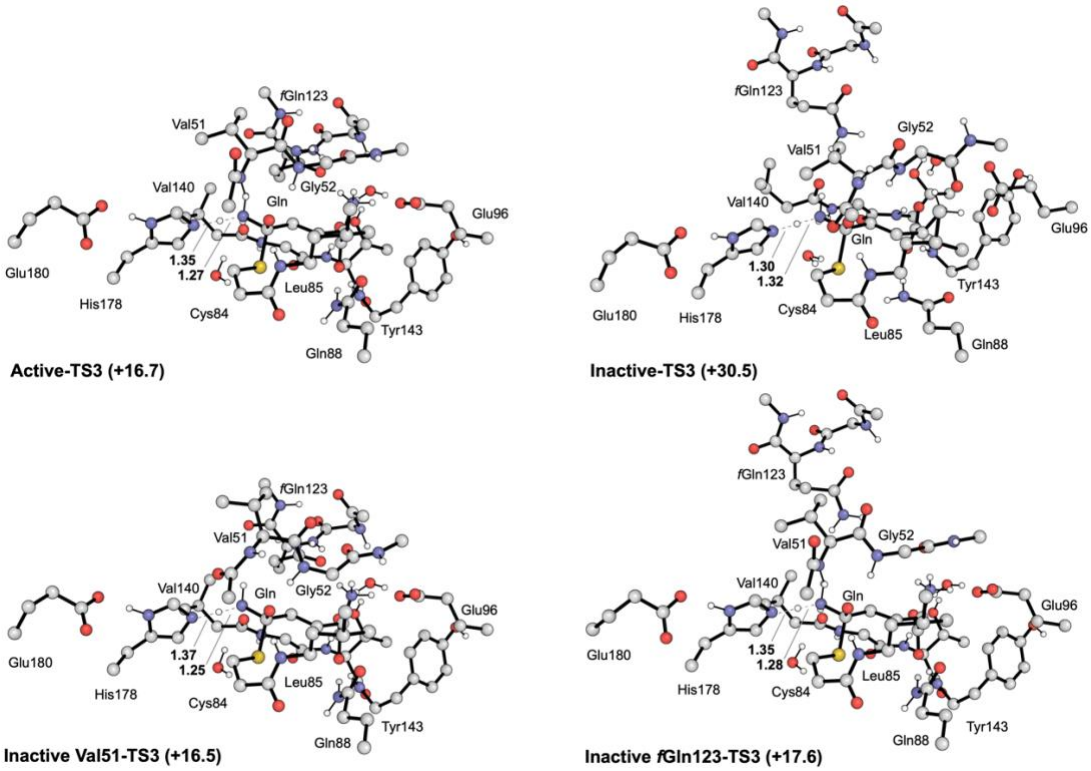


Figure A8. Optimized geometries of **TS3**.

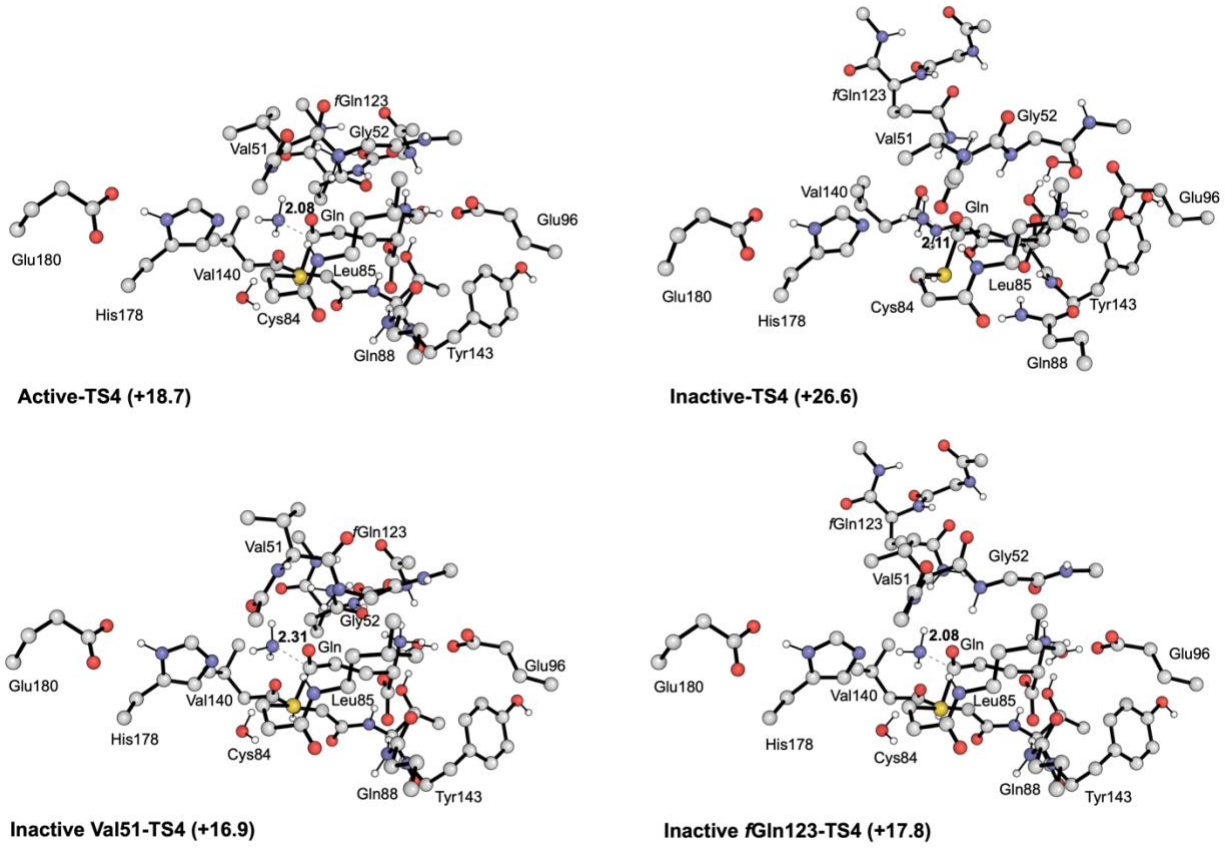


Figure A9. Optimized geometries of **TS4**.

A.7 Structural comparison with PLP synthase and CPS

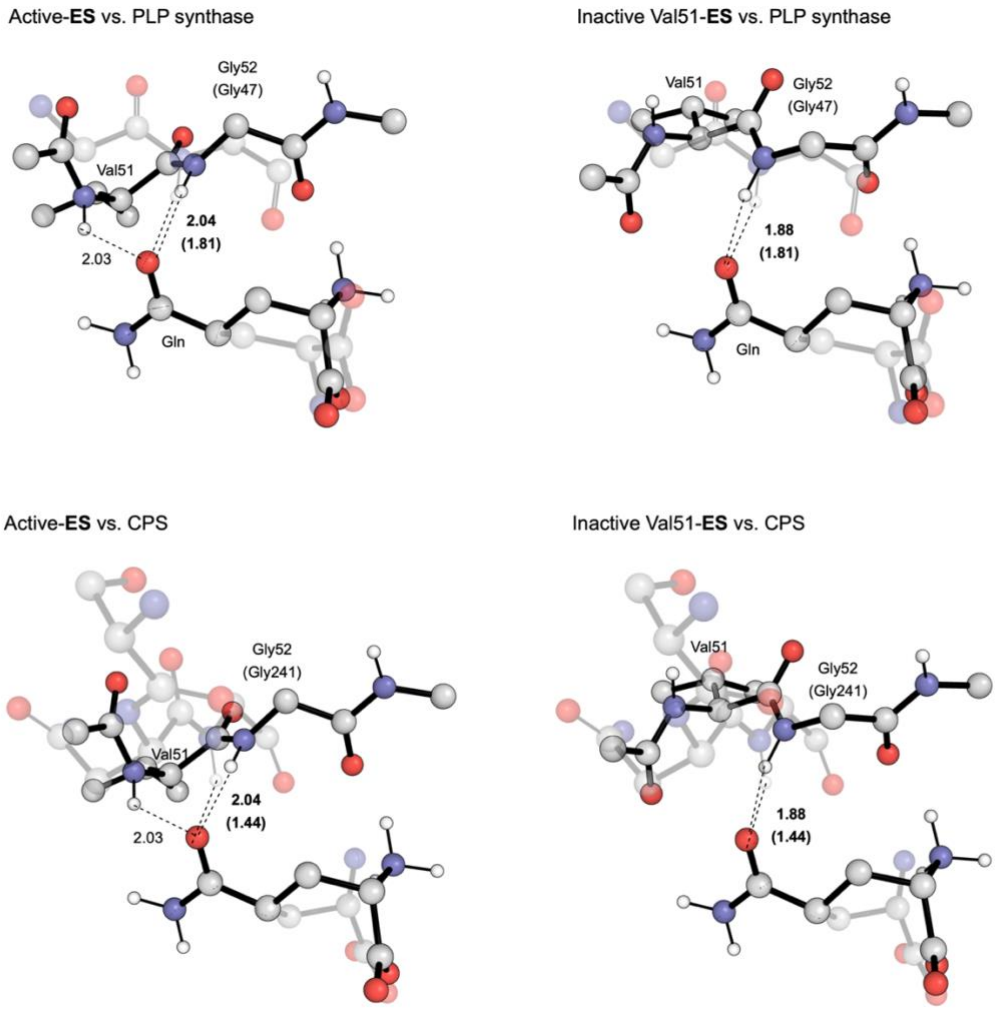


Figure A6. Optimized IGPS geometries of Active-**ES** (left) and Inactive Val51-**ES** (right) models aligned to crystallographic models of PLP synthase (PDB: 2NV2,[7] transparent) and CPS (PDB: 1C3O,[8] transparent) via Gln substrate carboxamide atoms. The preformed oxyanion holes from PLP and CPS are more structurally consistent with the IGPS Gly52 than with Val51, especially in the manually constructed Inactive Val51 model. Interestingly, the backbone dihedral of Gly47 in PLP synthase rotates upon dimer complexation and Gln binding, similar to what is proposed for Val51 in IGPS upon PrFAR binding. The H atoms of PLP synthase Gly47 and CPS Gly241 were added with the PyMol add_h function since they are not resolved in the crystal structures. Hydrogen bond distances are shown in Å and are in parentheses for the non-optimized structures.

REFERENCES

(1) Czapinska, H.; Bochtler, M. The Nϵ-Rule for Serine, but Not Cysteine Catalytic Triads. *Angew. Chem. Int. Ed.* **2022**, *61* (42), e202206945.

(2) Arafet, K.; Świderek, K.; Moliner, V. Computational Study of the Michaelis Complex Formation and the Effect on the Reaction Mechanism of Cruzain Cysteine Protease. *ACS Omega* **2018**, *3* (12), 18613–18622.

(3) Roux, B.; Walsh, C. T. P-Aminobenzoate Synthesis in Escherichia Coli: Mutational Analysis of Three Conserved Amino Acid Residues of the Amidotransferase PabA. *Biochemistry* **1993**, *32* (14), 3763–3768.

(4) Glendening, E. D.; Landis, C. R.; Weinhold, F. *NBO 7.0* : New Vistas in Localized and Delocalized Chemical Bonding Theory. *J. Comput. Chem.* **2019**, jcc.25873.

(5) Horn, P. R.; Mao, Y.; Head-Gordon, M. Probing Non-Covalent Interactions with a Second Generation Energy Decomposition Analysis Using Absolutely Localized Molecular Orbitals. *Phys. Chem. Chem. Phys.* **2016**, *18* (33), 23067–23079.

(6) Epifanovsky, E.; Gilbert, A. T. B.; Feng, X. Software for the Frontiers of Quantum Chemistry: An Overview of Developments in the Q-Chem 5 Package. *J. Chem.* **2021**.

(7) Strohmeier, M.; Raschle, T.; Mazurkiewicz, J.; Rippe, K.; Sinning, I.; Fitzpatrick, T. B.; Tews, I. Structure of a Bacterial Pyridoxal 5′-Phosphate Synthase Complex. *Proc. Natl. Acad. Sci.* **2006**, *103* (51), 19284–19289.

(8) Thoden, J. B.; Huang, X.; Raushel, F. M.; Holden, H. M. The Small Subunit of Carbamoyl Phosphate Synthetase: Snapshots along the Reaction Pathway. *Biochemistry* **1999**, *38* (49), 16158–16166.

APPENDIX B. SIZE-AND-SHAPE SPACE GAUSSIAN MIXTURE MODELS FOR STRUCTURAL CLUSTERING OF MOLECULAR DYNAMICS TRAJECTORIES SUPPORTING INFORMATION[§]

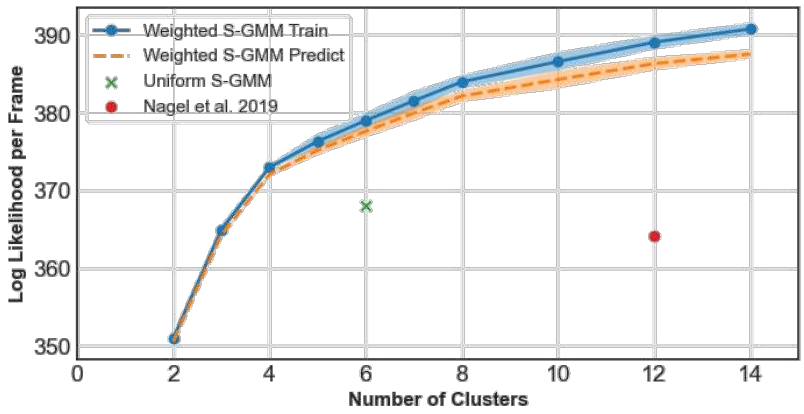


Figure B1: Log likelihood of weighted shape-GMM (S-GMM) as a function of number of clusters for HP35 trained on approximately $25K$ frames (blue line) in a 10-fold cross validation (orange dashed) scheme. The shaded region indicates the 90% confidence interval. The features are C, CA and N backbone atoms of residues 2-34, the C atom of residue 1 and N atom of residue 35 (101 atoms total). The log likelihood value from a uniform 6-state shape-GMM is shown as a green x marker. The red circle marker indicates the log likelihood that results from the HP35 discretization of Nagel *et al*. clustering procedure resulting in 12-states.[1]

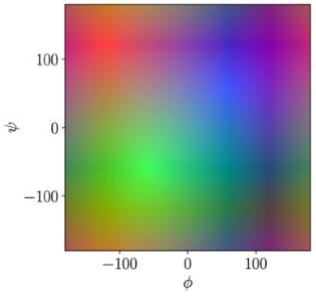


Figure B2: RGB values corresponding to phi-psi angle pairs used to assignment a unique color value to a dihedral conformation. The major dihedral states are indicated with more vibrant colors: red, green, and blue correspond respectively to β-strand, α_R-helical, and α_L-helical dihedral character. The code used to produce this mapping can be found at https://github.com/moldyn/ramacolor.[1,2]

[§] Reproduced with permission from **Klem, H**.; Hocky, G. M.; McCullagh, M. Size-and-Shape Space Gaussian Mixture Models for Structural Clustering of Molecular Dynamics Trajectories. *J. Chem. Theory Comput.* **2022**, *18* (5), 3218–3230.

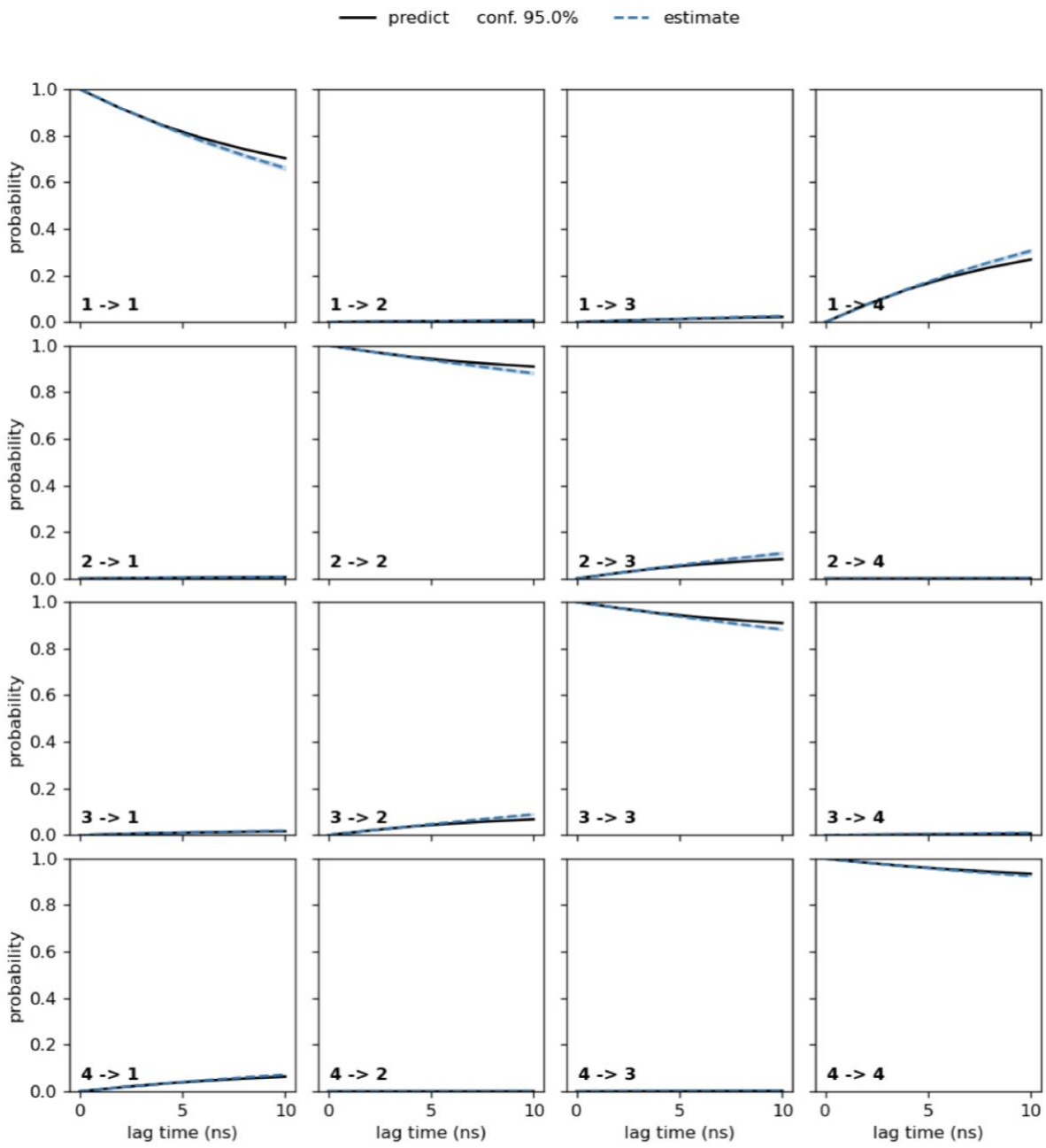


Figure B3: Chapman-Kolmogorov test for 4-state MSM resulting from weighted shape-GMM clustering on HP35. Computed using PyEMMA software.[3]

Table B1: Scores for a variety of clusterings of the HP35 trajectory. VAMP-2 scores were computed after construction of MSMs using a 10-fold cross validation on five trajectory segments using PyEMMA.[3] These were computed using either the first four ($k = 4$) or the first 10 ($k = 10$) eigenvalues of the transition matrix. Log likelihoods of a weighted shape-GMM (wSGMM) model were also computed on 10 randomly selected trajectory segments. A dynamic coring (here denoted dCore) procedure was applied to each clustering.[1] Error as estimated by 10-fold trajectory chunking are reported in parentheses.

Clustering Model	VAMP-2 Score (K=4)	VAMP-2 Score (K=10)	SGMM Log Likelihood Per Frame
wSGMM 4-state	3.3(2)	n.a.	372.4(1)
wSGMM 4-state (dCore)	3.87(3)	n.a.	372.0(1)
wSGMM 6-state	3.4(1)	n.a.	379.5(1)
wSGMM 6-state (dCore)	3.89(1)	n.a.	378.6(1)
uSGMM 6-state	2.9(1)	n.a.	368.0(1)
uSGMM 6-state (dCore)	3.93(1)	n.a.	367.66(9)
wSGMM 12-state	3.7(1)	7.2(6)	388.4(1)
wSGMM 12-state (dCore)*	3.91(1)	9.39(5)	383.16(9)
Stock 12-state (dCore)	3.91(1)	9.30(4)	364.2(3)
Stock 12-state (state specific dCore)*	3.93(1)	9.46(3)	363.9(2)

*Dynamic coring of wSGMM 12-state yielded a 10-state model.

REFERENCES

(1) Nagel, D.; Weber, A.; Lickert, B.; Stock, G. Dynamical Coring of Markov State Models. *J. Chem. Phys.* **2019**, *150* (9), 094111.

(2) Sittel, F.; Stock, G. Robust Density-Based Clustering to Identify Metastable Conformational States of Proteins. *J. Chem. Theory Comput.* **2016**, *12* (5), 2426–2435.

(3) Scherer, M. K.; Trendelkamp-Schroer, B.; Paul, F.; Pérez-Hernández, G.; Hoffmann, M.; Plattner, N.; Wehmeyer, C.; Prinz, J.-H.; Noé, F. PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. *J. Chem. Theor. Comput.* **2015**, *11* (11), 5525–5542.