ORIGINAL PAPER

Modeling Catalysis in Allosteric Enzymes: Capturing Conformational Consequences

Heidi Klem[1] · Martin McCullagh[2] · Robert S. Paton[1]

Abstract

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, have enabled complex multistep enzymatic mechanisms to be studied, often in quantitative detail. Nevertheless, the dynamic interconversion of enzyme conformations between active and inactive catalytic forms, involving length- and timescales inaccessible to QM treatments, presents a formidable challenge for the development of computational models for allosterically modulated enzymes. We present an overview of the key concepts underlying multistate models of enzyme catalysis, enzyme allostery, and the challenge that large-scale conformational changes pose for methods using QM, QM/MM, and MM potentials. Structural clustering is highlighted as a valuable approach to bridge molecular dynamics conformational sampling of MM potentials and quantum chemical cluster models of catalysis. Particularly relevant to this discussion is structural allostery, which serves as the exemplar of conformational consequences. Here, a well-characterized allosteric enzyme, imidazole glycerol phosphate synthase, is used to showcase the importance of multiple conformations and guide a new direction for qualitative understanding and quantitative modeling in enzyme catalysis.

Keywords Biocatalysis · Allostery · Conformational ensembles · Molecular modeling · Molecular dynamics · Computational chemistry · Reaction mechanisms

1 Introduction

Enzymes carry out the homogeneous catalysis of many essential biological transformations and industrial processes, yielding kinetic rate enhancements of up to 26 orders of magnitude [1]. The quantitative understanding of enzymes has expanded drastically, propelled by biophysical, biochemical, and theoretical studies. Fischer's "lock and key" model became the foundation for understanding enzyme–substrate mechanics in 1894 [2]. Koshland modified this model in 1958, accounting for enzyme flexibility in developing the induced fit model [3]. The field of enzymatic catalysis has since evolved to account for critical contributions of conformational substates [4]. Additional current topics of interest are enzyme preorganization, reorganization, near attack conformations, desolvation effects, transition state stabilization (TSS), reactant state destabilization (RSD), and dynamic contributions [5–9]. The reader is directed towards reference [10] and references therein for a historical perspective on modeling enzyme catalysis. In this perspective, we focus on enzymes for which conformational changes are critical to (a) understanding the catalytic mechanism and (b) computing quantitatively accurate barriers and energy changes.

Three classes of potentials are commonly employed in computational studies of enzymes: quantum mechanics (QM) only, hybrid QM and molecular mechanics (QM/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 (Sect. 4.1), notably the cluster approach, are now routinely used to study

✉ Martin McCullagh
martin.mccullagh@okstate.edu

✉ Robert S. Paton
robert.paton@colostate.edu

[1] Department of Chemistry, Colorado State University, Fort Collins, CO 80523, USA

[2] Department of Chemistry, Oklahoma State University, Stillwater, OK 74074, USA

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 [11]. 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 [12–14]. Cluster models have similarly been used to study how enzyme active sites control regioselectivity, for example, in facilitating intrinsically disfavored epoxide-opening pathways [15]. A related QM-only method, the *theozyme* approach, has been used to model a theoretical enzyme active site as a selection of functional groups directly involved in catalysis [16]. Houk and co-workers have used theozyme models to illustrate how side-chain motions are minimized in the multistep serine esterase catalytic cycle [7] and in the computational design of enzymes for abiological reactions [17].

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 (Sect. 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) [18], provide near-quantitative results for the activation enthalpies and free energies of the reactions catalyzed by chorismate mutase and *para*-hydroxybenzoate hydroxylase [19]. 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 Sect. 5. As with the selection of a QM cluster model, the choice of QM/MM boundary and how MM partial point charges at the boundary are described should be handled with care and validated [20].

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 are essential to the catalytic mechanism and for which 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 [21–23] and, most recently, in directed evolution [24–26]. MD has proven to be especially apt at sampling "local" conformational diversity in enzymes such as the different rotamer states adopted by side-chains. However, some conformational changes necessary to achieve a catalytically competent active site, such as those involving loop and helix motions, occur on the millisecond timescale [27, 28]. In such cases, enhanced sampling MD methods can be used.

Of significance to enzyme catalysis is the ability of MD simulation to provide statistical details for catalytically relevant conformations. In general, this is aided by clustering the MD snapshots into structurally similar groups to yield populations of significant enzyme conformations (Sect. 5). While classical MD simulations excel in conformational sampling, conventional MM potentials do not describe the breaking and forming of bonds along a reaction pathway. In contrast, while low-cost QM potentials are gaining traction, high-accuracy QM approaches are ill-suited to large-scale conformational analyses. Therefore, the combination of QM and classical MD approaches to study enzyme catalysis in the context of the conformational ensemble is of great interest to the community [29, 30].

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 [31]. The validity of these assumptions is difficult to test unless *apriori* knowledge or hypotheses regarding the catalytic mechanism exist. Thorough investigations of catalysis will evaluate mechanistic possibilities indicated in the literature and preliminary results. In Sect. 3, we discuss myriad enzymatic reactions now known for which these simplifying assumptions do not hold. These systems present challenges for computational chemistry 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. Section 6 describes how computational workflows based on macromolecular conformational sampling with MD, structural clustering to obtain ensemble populations, and QM calculations are poised to accelerate the study of enzyme catalysis further. The explicit consideration of multiple enzyme conformations to support mechanistic conclusions and calculate barrier heights has now been used in multiple studies [28, 32–35].

The main goal of perspective is to illustrate the role of conformational effects upon enzymatic catalysis. We provide a brief account of the main computational approaches, procedures, and limitations and discuss their relevance

in modeling protein motions that occur across different timescales. We introduce fundamental concepts of protein allostery and discuss how these systems present a unique challenge for computational chemistry and conformational sampling. We suggest that a combination of cutting-edge techniques in MD sampling and QM modeling provides a particularly appealing approach to study enzyme catalysis. Clustering methods are discussed as a means to bridge the results of classical simulations with QM-cluster models. The final section outlines the case study of glutamine hydrolysis performed by the allosterically-regulated glutamine amidotransferase (GAT) imidazole glycerol phosphate synthase (IGPS). This system beautifully illustrates the importance of conformational changes in catalysis, the existence of inactive and active forms in the solid-state and solution, and the importance of computational sampling of the enzyme's conformational space. The rate of glutamine hydrolysis in IGPS is critically dependent on the presence of the allosteric ligand over 25 Å away from the glutamine binding site [36]. Until recently, the atomic changes that yield the hydrolysis rate enhancement were not realized due to the absence of an observed catalytically active conformation.
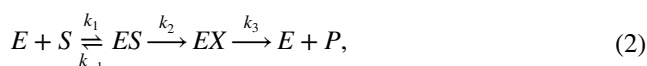
2 Enzyme Catalysis: The Reaction Coordinate and Protein Dynamics

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:

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

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 Eq. 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:

$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} ES \overset{k_2}{\longrightarrow} EX \overset{k_3}{\longrightarrow} E + P, \tag{2}$$

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 Eq. 2, the steady-state parameters are defined as $k_{cat} = k_2 k_3 / k_2 + k_3$ 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 Eqs. 1 and 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 Eq. 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 Eq. 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 where product inhibition occurs. However, we limit the scope of this perspective to examples for which the rate-determining step is associated with bond formation or cleavage.

One critical enzyme attribute missing from Eqs. 1 and 2 is conformational heterogeneity. Much of the focus of enzyme catalysis in the twenty-first century has been on the role of conformational changes involving both experimental and computational expertise [4, 10, 27, 30, 37]. 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 (Fig. 1) [38].

Equations 1 and 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 [38]. Consequently, quantitative agreement of
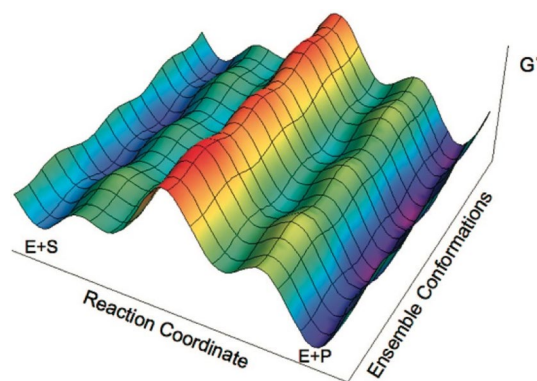

Fig. 1 Enzyme free energy landscape projected onto the reaction and conformational coordinates. Adapted with permission from [38]. Copyright 2021 American Chemical Society

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 Fig. 1, strong coupling can arise where even qualitative agreement with experiment requires consideration of the ensemble [27, 28]. Construction of the free energy landscape for large systems is a laborious task. As discussed above, such studies require computational methods in classical simulation techniques, while the reaction coordinate may be more amenable to QM approaches. As a result, a combination of distinct computational methodologies and expertise is often required.

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 Fig. 2. Multiple factors can influence the quality of the Gibbs energy profile and its mechanistic interpretation [39]. 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 [40]. 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 [41].

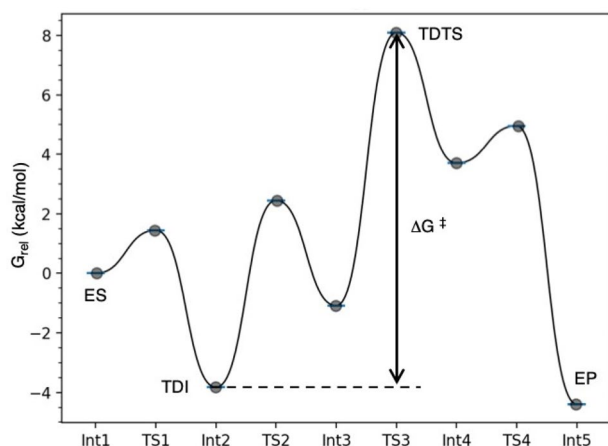In comparing the computed catalytic (Gibbs) energy with experiment, the concepts and language introduced by Kozuch and Shaik are both illuminating and influential [43]. In contrast to using rate constants to define a catalytic cycle (the k-representation), computational studies generate state energies (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) (Fig. 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:

$$k_{cat} = \frac{kk_BT}{h} \exp\left(\frac{\Delta G^{\ddagger}}{RT}\right) \tag{3}$$

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 $\kappa = 1$, but more advanced techniques have been developed to evaluate non-equilibrium effects and the contributions of recrossing and tunneling to enzymatic rate constants [44, 45]. 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 [46]. As cautioned by Kozuch and Shaik, care must be taken when calculating $\Delta G^{\ddagger}$ to be used in Eq. 3. In some situations the internal energy can be a fair approximation for this value, however 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 [47] and theozyme [16] approaches, and (2) mixed QM/MM studies. These approaches are surveyed in Sect. 4. More detailed technical summaries of these approaches can be found elsewhere, e.g., in reference [31]. Herein, we 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.

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

Fig. 2 Example reaction coordinate, created with GoodVibes [42], 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, $\Delta G^{\ddagger}$

function. Proteins are known to populate multiple metastable structures (the conformational ensemble) under typical physiological conditions [48]. Mounting evidence suggests that the ensemble nature of proteins is intricately tied to their function [49, 50]. 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 (Fig. 3).

Experimental techniques used to probe protein ensembles include X-ray crystallography, nuclear magnetic resonance (NMR) [50], small-angle X-ray scattering (SAXS), atomic force microscopy (AFM) [51–53], and more recently, cryogenic electron microscopy (cryoEM) [54–57]. 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 [58] and recognition that a given crystal likely has multiple structures in the unit cell [59, 60]. 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 [61] and dipolar couplings [62], are ensemble averages over multiple metastable states. NMR has successfully been coupled to SAXS [63, 64], spectroscopic techniques, and molecular simulations [65–67] 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 [54–57]. 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 [68]. Other than reparametrization of standard functional forms, new directions in force field development include polarizable [69] and machine-learned force fields [70]. 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
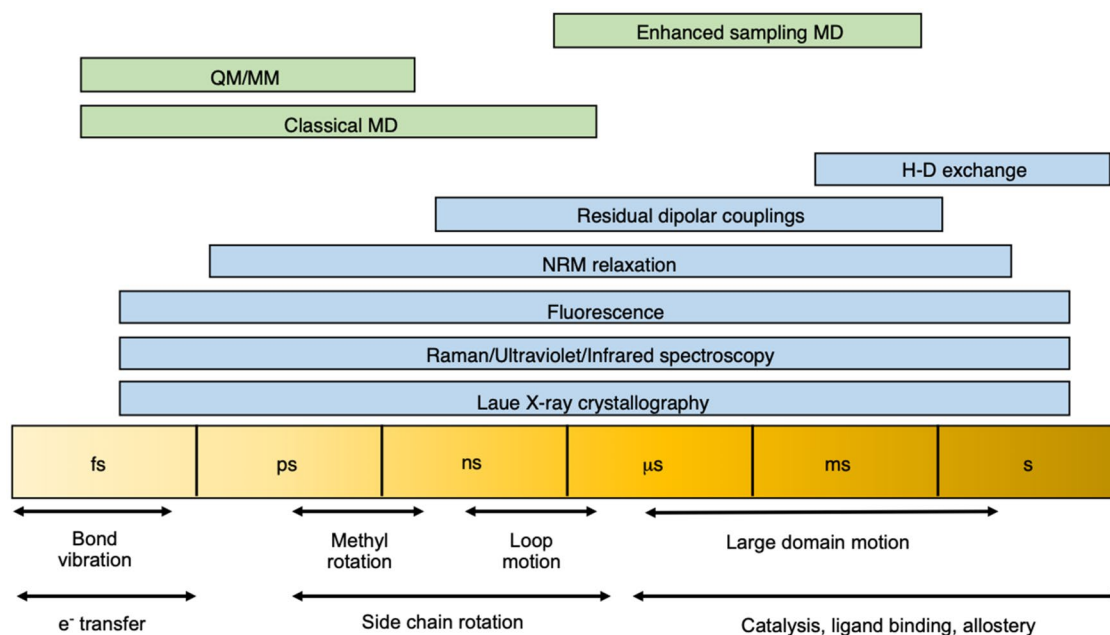

Fig. 3 Methods for studying dynamic processes in proteins ranging from femtosecond to second timescales. Experimental and computational techniques are shown in blue and green, respectively

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 [71] or enhanced sampling techniques. The latter is a more approachable solution and includes techniques such as replica exchange [72, 73], metadynamics [74], and adaptive sampling [75–77]. 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 [78].

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 [24, 25, 79–81], and a direct correlation between reaction rate and active conformation population influenced by different allosteric ligands and enzyme mutations has been quantified [28].

Allosteric enzymes exemplify the importance of conformational ensembles [3]. Allosteric regulation 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 (Fig. 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,
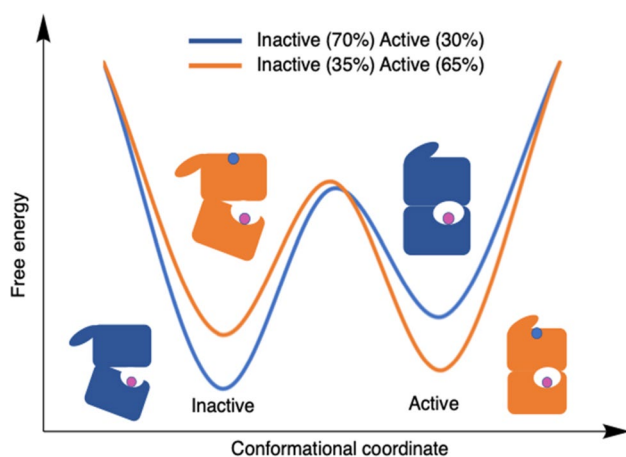

Fig. 4 A free energy profile projected onto a two-state conformational coordinate. The blue line and toy enzyme model illustrate a system without the allosteric effector bound. Orange corresponds to a system with the allosteric effector bound. The legend shows relative populations of the active and inactive states that result from different ligand states

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 is altered, and the resulting inactive and active conformation populations are altered. This conformational shift phenomenon is often used in discussions of structural allostery but is not limited to such enzymes.

3 Allostery in Enzyme Catalysis

The long-ranged coupling of sites in an allosteric system can be explained through short-range interactions linking the distant sites [82]. 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 [83]. Biological systems harness allostery to respond to changes in their environment. This is recognized in various biological processes such as signal transduction [84, 85], transcriptional regulation [86, 87] and metabolism [88]. 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 [85]. In 2011, Huang and co-workers created the Allosteric Database (ASD) to provide a comprehensive collection of allosteric data [89]. Presently, the ASD contains 1,949 allosteric proteins and 82,070 allosteric modulators [90, 91]

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 Fig. 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 Fig. 5, has been developed to quantify the allosteric effect of a K-type system [83]. 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 relatively unexplored in the context of allostery.
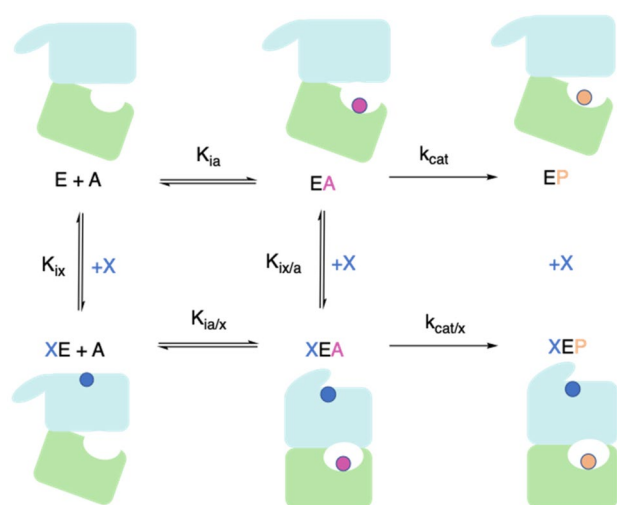
Fig. 5 Complete thermodynamic cycle of an allosteric enzyme, E. The top row represents the enzyme in the presence of only substrate, A. The bottom row represents the enzyme in the presence of the substrate and allosteric effector, X. Binding of X modifies the activity of E by either altering the binding affinity of E for A, as in K-type allostery, or influencing the rate-determining step, as in V-type allostery. Nomenclature from Ref. [83] is used here for consistency

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 [92]. 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 [93]. It was observed that in apohemoglobin (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 [94]. This model favors an "induced fit" mechanism, whereby the apoprotein 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

[95]. 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 [96]. 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 [23, 85, 88, 97, 98]. Allostery research, which traditionally emphasized static comparisons, is currently faced with the challenge of leveraging the ensemble nature of allostery [99]. 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 [25].

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 [28, 95]. The whole-scale enzymatic influence on the chemical rate in V-type allosteric systems is particularly intriguing in this regard.

Enzymatic catalytic activity is typically compared against the background reaction rate in (aqueous) solution to investigate catalytic origins [8]. Alternatively, one could compare the same reaction in enzymes with different activities. Such an approach can be adopted in evaluating V-type allosteric enzymes. 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 enzyme conformational ensemble modeling.

4 Quantum Mechanics for Enzyme Catalysis

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 [47]. This approach has been pioneered in the study of biologically relevant metal centers by Siegbahn and Blomberg,

with model sizes around 60–70 atoms [100], and by Himo, who has used cluster models that can be much larger, even surpassing 300 atoms [11]. For more focused reviews on the quantum chemical cluster approach, we urge the interested reader to explore several excellent reviews [11, 47, 101]. Here, we summarize the steps taken in building a quantum chemical cluster model.

4.1.1 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 [102]. If crystallographic waters or ions are present in the active site, they may need to be explicitly included in the model as well [103]. It is typical to test various sizes of clusters, ranging from 100 to 300 atoms [11]. Convergence studies suggest that QM-cluster models give reliable energetics when the model size is large enough [47, 104–108]. 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 [109], however, it has been proposed that informed residue selection should take priority over a simple distance cutoff [110].

4.1.2 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-C_α and C_α-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-C_α bond or C-terminus (C'), C_α-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. The effects of different truncation schemes are evaluated in Sect. 6.2.2.

4.1.3 Coordinate Locking

In most cluster models, it is necessary to lock or freeze certain atoms to preserve the active site geometry, maintain side-chain 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 [111]. The effects of coordinate locking have been explored in Ref. [112], 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 reduce 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 [11] or not locking residues that directly interact with the substrate [113] 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.

4.1.4 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 and Becke-Johnson damped (D3 and D3(BJ), respectively) corrections are recommended with typical GGA (generalized gradient approximation) or hybrid-GGA density functionals such as B3LYP [114]. 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 (c.f. 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 [47, 105]. 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.

4.2 Application of the Quantum Cluster Approach: Benzoylformate Decarboxylase

Benzoylformate decarboxylase (BFDC) is a thiamine diphosphate (ThDP)-dependent enzyme that catalyzes the decarboxylation of benzoylformate into benzaldehyde and carbon dioxide in a critical step of the mandelic acid degradation pathway [115, 116]. Additionally, BFDC can enantioselectively catalyze carboligation reactions, depending on the substrate. X-ray structures of enzymes in the ThDP-dependent decarboxylase family show common features such as the presence of ThDP, an almost invariant glutamate, two ionizable acidic residues, and two proximal histidine residues on an ordered loop termed the HH-motif. However, the first X-ray structure of BFDC contains serine (Ser26) as the only ionizable acidic residue in the active site, and the two proximal histidines (His70, His281) belong to separate monomers rather than an ordered loop.

Himo and co-workers used the cluster approach to study BFDC-catalyzed decarboxylation. Their model consisted of 307 atoms and an overall − 1 charge. Geometry optimizations were performed at the B3LYP-D3(BJ)/6-31G(d,p) level of theory, with 6–311 + G(2d,2p) single-point corrections using the SMD solvation model with a dielectric of $\varepsilon = 4$. While entropy changes along the reaction coordinate are often neglected in studies using cluster models, the release of CO_2 gas involves a significant increase in translational entropy. This was calculated to be 11.3 kcal/mol, and so this

value was included in each step after CO_2 formation, in line with previous estimation methods [13, 117]. The key roles of active site residues were identified, and the authors also identified a kinetically relevant off-cycle species produced by intramolecular cyclization of the cofactor [118].

In a subsequent study, Himo and co-workers focused on enantioselective catalysis by BFDC [119]. Following benzoylformate decarboxylation, the (Breslow) enamine intermediate can participate in a benzoin condensation with benzaldehyde or acetaldehyde electrophiles. Interestingly, these transformations proceed with an opposite sense of enantioselectivity. The computed energy difference between competing TSs in the enantioselectivity determining step for benzaldehyde addition (TS3, Fig. 6) is 9.3 kcal/mol, consistent with complete stereocontrol observed experimentally. Steric interactions between substrate and multiple residues contribute to the higher energy of the disfavored pathway. For acetaldehyde, the energy difference (0.3 kcal/mol) favors the (S)-product, consistent with the contrast in selectivity observed for the two substrates. This, and other studies by Himo, illustrate the importance of a reasonably large cluster to capture the multiple steric and other noncovalent interactions influencing enantioselectivity in enzyme catalysis.

4.3 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 [29, 120–128].

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 [129]. Further, the role of protein reorganization along the reaction coordinate, which can be consequential, may also be captured and quantified [130]. Readers are referred to detailed reviews on the status of QM/MM approaches in modeling enzymatic catalysis [120, 127, 131, 132]. 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
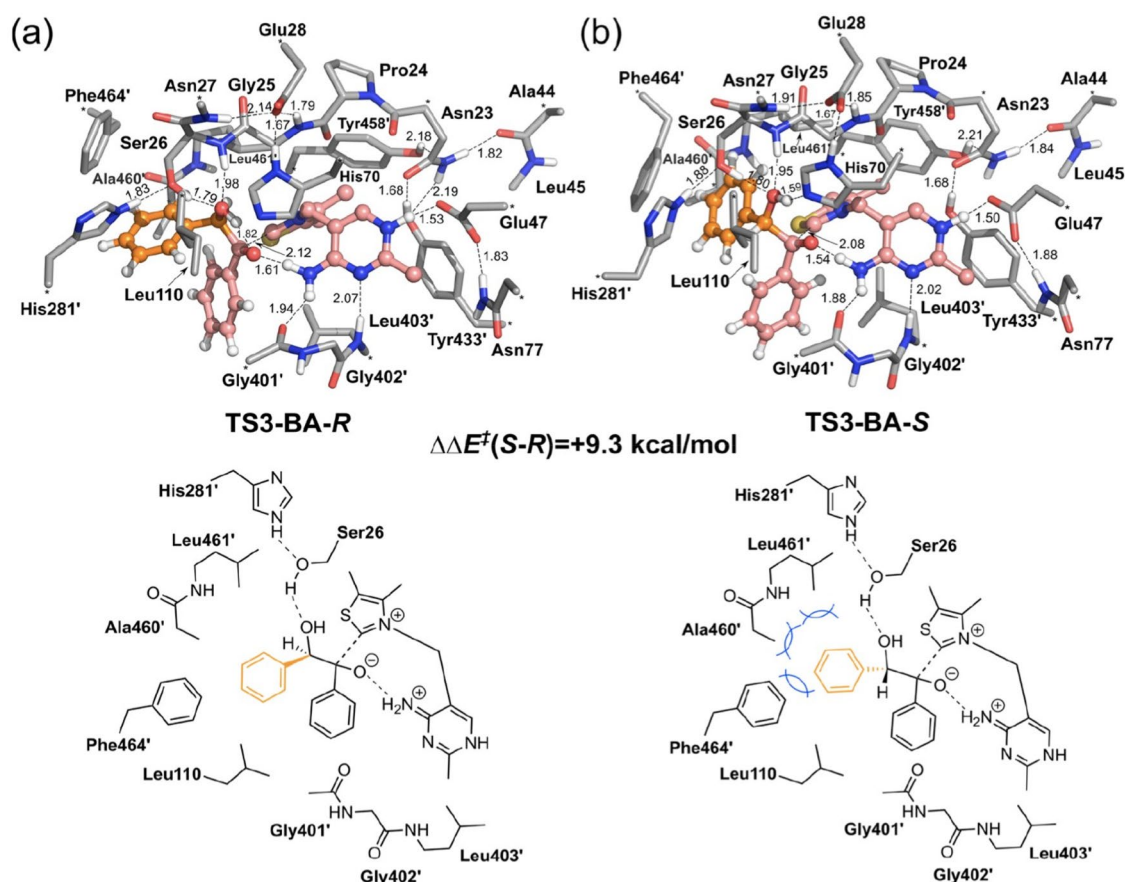
Fig. 6 Optimized geometries of selectivity-determining transition state structures for **a** (*R*)-benzoin and **b** (*S*)-benzoin in the BFDC carboligation reaction. The benzaldehyde acceptor and enamine donor are shown in orange and pink, respectively. The blue lines in the lower panel of (b) indicate the steric clashes from Ala460, Leu461, Phe464 that drive stereoselectivity. Asterisks designate the atoms frozen during calculations. All non-polar hydrogens are hidden for simplicity. Figure adapted from [112]

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 [120, 133, 134]. The computational demands of QM/MM simulations can be influenced by choice of the QM region's 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 [135–138], and attention to the region definition is also required [139–142].

Although, compared to QM-only models, Warshel has demonstrated a relative lack of sensitivity to including more distal groups in the QM region [143]. The use of SQM methods generally requires careful benchmarking or parameterization against more expensive QM results [144]. 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 [145–147]. 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 [148]. Classical MD simulations over microsecond timescales, inaccessible while employing QM/MM potentials, have been necessary to observe binding site formation, for example, in bromodomains [149].

4.4 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 [150]. 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 [151]; however, questions continue to be raised, such as what is required to converge energies in this approach [104] and how to account for conformations with differing reactivities [30, 152, 153]. Alternatively, the Boltzmann ensemble can be evaluated with a combination of extended MD and structural clustering. 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. The details of this approach are outlined in the next section.

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, $\mathbf{X}$, 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,

$$\langle M \rangle = \frac{\sum_i^N M(\vec{x}_i) P_i}{\sum_i P_i} \tag{4}$$

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, Eq. 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 Eq. 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 [71] or combined trajectories [154]. Given a clustering of a conformational ensemble into $K < N$ clusters with well-defined average (or mediod) structures, observable can be estimated as

$$\langle M \rangle = \frac{\sum_j^K M\left(\langle \vec{x} \rangle_j \right) P_j}{\sum_j P_j} \tag{5}$$

where $\langle \vec{x} \rangle_j$ is the average structure and P_j is the probability of cluster j. The accuracy of Eq. 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. Here, we describe five steps that most clustering protocols follow: (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. We note that while this is a typical order of the steps, it is not unique. It is possible, for example, to cluster first and then perform dimensionality reduction (swapping steps 3 and 4).

5.1 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 a protein, there are ~ 100 K total atoms and thus 300 K 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, CA 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 CA 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 CA atoms of each residue. The choice of all CA 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.

5.2 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 CA atoms in the top-right panel of Fig. 7. With 20 residues (and thus
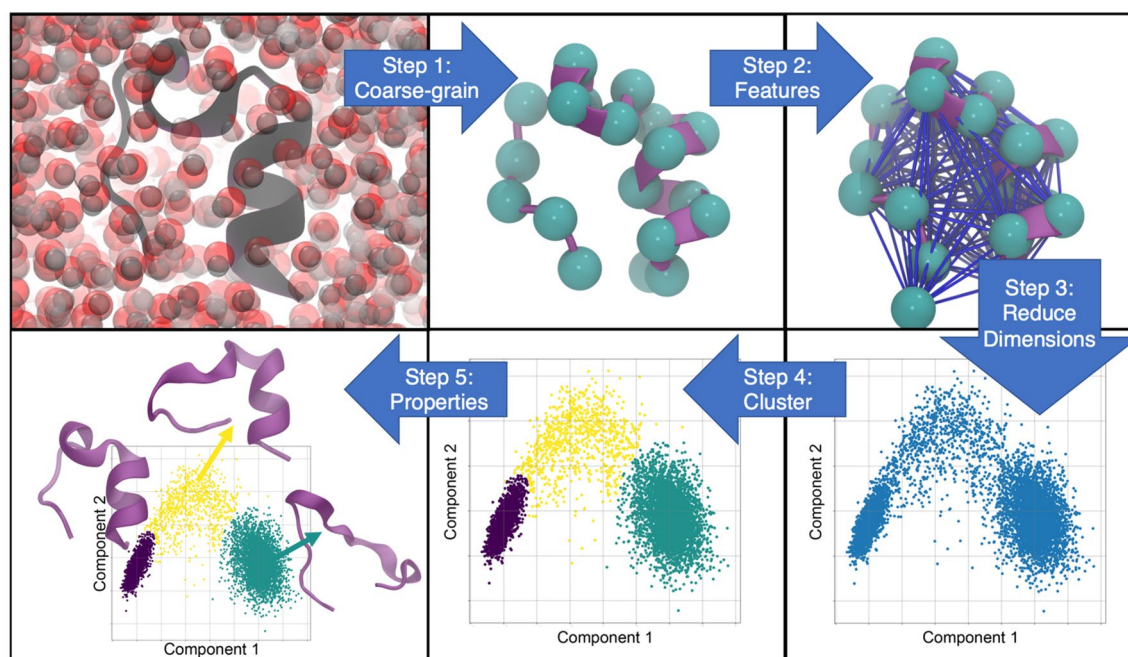

Fig. 7 Five-step schematic for structural clustering from all-atom molecular dynamics trajectories. Starting from the top left, the first step is to choose a particle coarse-grained mapping and apply it to an explicit solvent all-atom trajectory. From there, a set of features to describe the system, 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)

CA 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 [155].

5.3 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.

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 time-scale of collective motions [156]. tICA is often coupled with Markov State Modeling (MSM) to determine rates of transitions between metastable states [157]. Other dimensionality reduction methods have been employed, such as Sketch Map [158] and UMAP [159], 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.

5.4 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 the most commonly employed for MD trajectory data. The most common example is k-means clustering and its variants. k-means determines the means and trajectory partitionings 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 [160]. 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 [160]. 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.

5.5 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. This assumption can be tested to assess the validity of the clustering.

6 Unexplored Land in Computational Catalysis: A Case Study on IGPS

In this section, we focus on the mechanism of glutamine hydrolysis by the allosterically regulated enzyme IGPS. This system illustrates particular challenges for computation, such as catalytically-inactive crystal structures, the presence of active and inactive enzyme conformations, and allosteric rate enhancement.

6.1 Imidazole Glycerol Phosphate Synthase (IGPS)

Efficient transportation of a reaction product from one location to be used as a reactant in the next is a critical step in biological pathways. In some cases, the efficiency depends on whether or not that product must travel through bulk solvent to arrive at its next stop. The glutamine amidotransferase (GAT) family of enzymes produce ammonia through hydrolysis of glutamine and utilize the ammonia in a subsequent reaction, harnessing several traits to ensure their effectiveness in biochemical pathways. The bienzyme imidazole glycerol phosphate synthase (IGPS) is a GAT composed of glutaminase and cyclase subunits, nominally known as HisH and HisF in *T. maritima* (Fig. 8). It operates in histidine and purine biosynthesis in plants, fungi, bacteria, and archaea. In the glutaminase subunit, IGPS performs the hydrolysis of glutamine to form glutamate and ammonia, the latter of which goes on to react with the allosteric ligand (phosphoribulosyl-formimino-AICAR-phosphate, PrFAR) over 25 Å
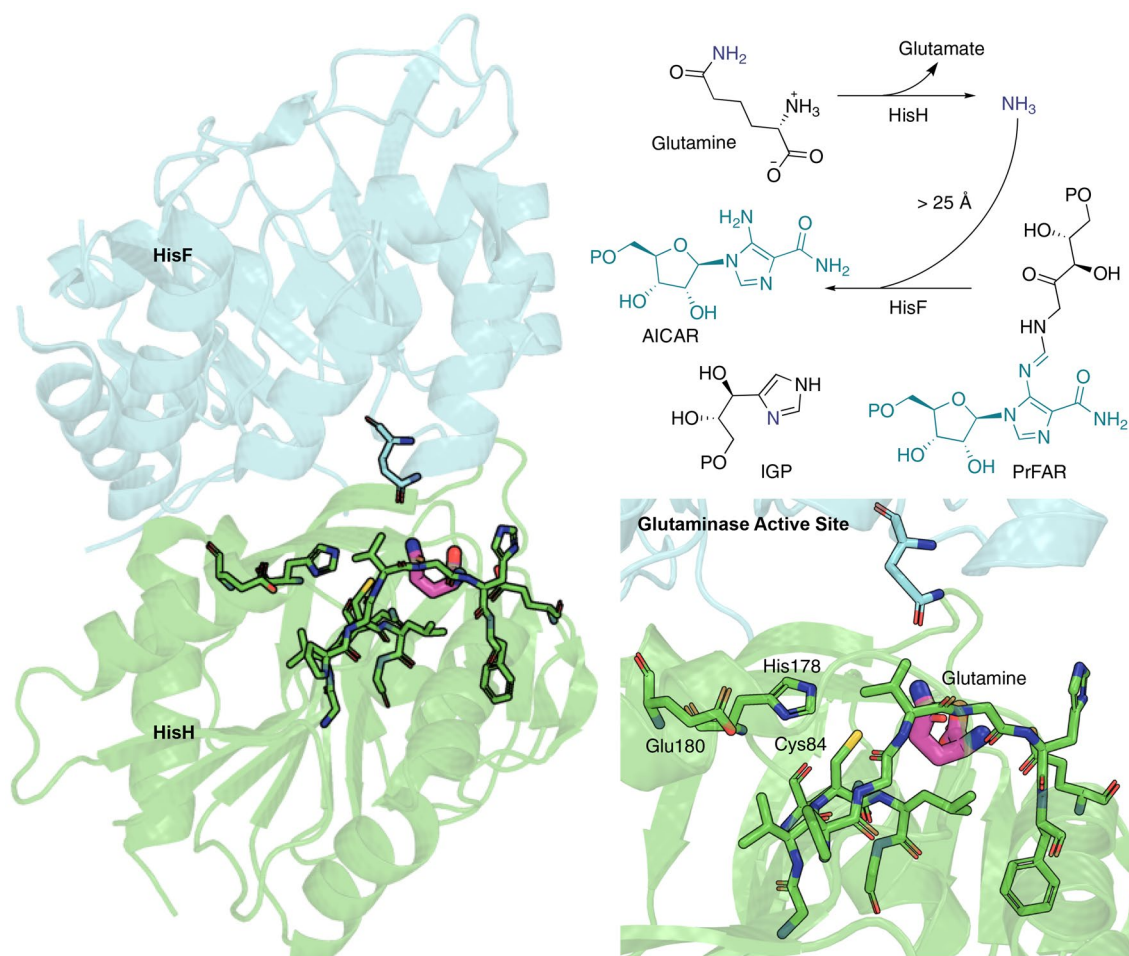

Fig. 8 IGPS from *T. maritima* and overview of glutaminase and cyclase coupled reactions. HisH is shown in green, and HisF is shown in blue. In HisH, glutamine reacts with water to form glutamate and ammonia, which is shuttled over 25 Å to the cyclase active site in HisF, where it reacts with PrFAR to form AICAR and ImGP

away in the cyclase active site. The binding of PrFAR results in a 4500-fold increase in glutamine hydrolysis, making IGPS a V-type allosteric system [36]. The mechanism of allosteric regulation in IGPS and how the two active sites are coupled through residue pathways has received intense interest from experimentalists and theoreticians alike [67, 82, 161, 162].

Until 2021, it was unclear how conformational ensembles impact allosteric regulation of IGPS and if changes to the conformational equilibrium occur when PrFAR binds. Furthermore, it was unknown whether PrFAR binding induces specific changes to catalytic residues involved in glutaminase activity. Comparison of IGPS to similar GATs suggested the conformation of a conserved oxyanion strand (P49 G50 V51 G52) to be critical to the rate enhancement. The formation of an oxyanion hole, a catalytic motif that stabilizes a substrate's developing negative charge through hydrogen-bonding interactions, has been a leading theory to account for rate enhancement in GAT mechanisms. It has been proposed that for many GAT enzymes (e.g., pyridoxal 5'-phosphate synthase and IGPS from *T. maritima*) [65, 163], binding of the effector induces a backbone flip in the oxyanion strand, forming an emergent oxyanion hole that stabilizes an oxyanion tetrahedral intermediate. Visual inspection of the apo and ternary IGPS crystal structures, 3zr4 [161] and 1ox5 [164], respectively, shows the Val51 backbone has not flipped in either case, suggesting either the flip in the Val backbone is not induced by effector binding or the crystal structure with PrFAR bound is not in the catalytically enhanced state, as proposed from solution NMR and MD studies [65].

In 2021, an IGPS crystal structure from Sprangers and co-workers showed the oxyanion hole fully formed, with the Val51 amide flipped (Fig. 9) [28]. Furthermore, the authors observed that both glutamine and PrFAR ligands must be present for the catalytically competent state to be reached. Numerous structural changes occur, breaking the hydrogen bond between P10 and V51 in HisH, a critical interaction stabilizing the backbone conformation of the oxyanion strand. Consequently, the V51 backbone is free to flip, accessing a conformation in which the backbone N–H is favorably oriented towards the carbonyl group of the substrate. The dynamic equilibrium between inactive and active IGPS conformations was found to occur on the millisecond timescale, explaining the difficulty of observing all conformational transitions in previous computational studies. Of particular importance, this study shows a direct correlation between the equilibrium of the active versus.

inactive states and glutamine turnover, just as the ensemble model of allostery predicts. The population of the active conformation can be manipulated experimentally by distant mutations. In these studies, a linear relationship is observed between destabilization of the active conformation and
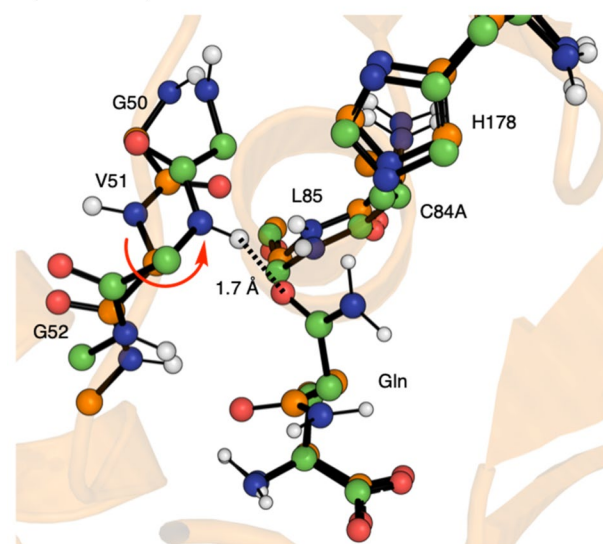
Fig. 9 Aligned crystal structure conformations of IGPS in ball-and-stick representation. All oxygen, nitrogen, and polar hydrogen atoms are colored red, blue, and white. Orange carbons correspond to the catalytically inactive conformation, chains AB from crystal structure 7ac8. Green carbons correspond to the catalytically active conformation, chains EF from crystal structure 7ac8. The V51 backbone flip is highlighted with a red arrow. This dihedral flip leads to a strong hydrogen bond (1.7 Å) between the V51 amide and the oxygen atom of the bound Gln substrate

$ln(k_{cat})$, as predicted by the ensemble allostery model [28]. Furthermore, k_{cat} is directly influenced by the population of the active conformation.

6.2 Glutamine Hydrolysis in IGPS – Key Considerations for Future Computational Studies

Glutamine hydrolysis in IGPS occurs via a multistep mechanism; however, no previous quantum chemical or QM/MM studies have been conducted. Before 2021, the active enzyme conformation had not been captured crystallographically, and as outlined above, the millisecond timescale of conformational transitions between inactive and active states is challenging to sample with cMD simulations. In the following sections, we discuss current prospects for modeling glutamine hydrolysis in IGPS and identify the challenges ahead.

In IGPS, a catalytic triad (Glu180, His178, and Cys84) performs glutamine hydrolysis. These residues would be the minimal ingredients of a QM region. However, as outlined above, catalytic activity results from large-scale conformational changes leading to a backbone flip, and the substrate is in contact with several additional residues in the active site. For this reason, a physically realistic active site model

of IGPS involves over 200 atoms. We suggest QM treatment of the following residues are necessary: f98D, f123Q, h10P, h11G, h12N, h50P, h51G, h52V, h53G, h84C, h85L, h88Q, h96E, h141H, h142T, h143Y, h176H, h178E, along with the glutamine substrate. Additionally, inactive (Chains A and B) and active conformations (Chains E and F) can be found from the 7ac8 crystal deposition. The overall charge of this 223-atom active site model is -3. For modeling via the QM-cluster approach, different truncation schemes (Sect. 4.1) can be considered for IGPS (Fig. 10), of which some illustrative examples are discussed below.

ωB97XD/6–31 + G(d) optimizations of different truncated models of the FHPEK stand in the IGPS active site, in which all C_α atoms and capping hydrogens were frozen, reveal the importance of decisions made during this stage of the modeling process. Terminating the side-chains by methyl groups leads to much greater atomic displacement relative to a model in which the polypeptide backbone is preserved (Fig. 10a–c). This disrupts an H-bond between His178 and Glu180, which is preserved in the more conservative capping scheme. While the impact of these structural differences upon the reaction mechanism has not been studied,
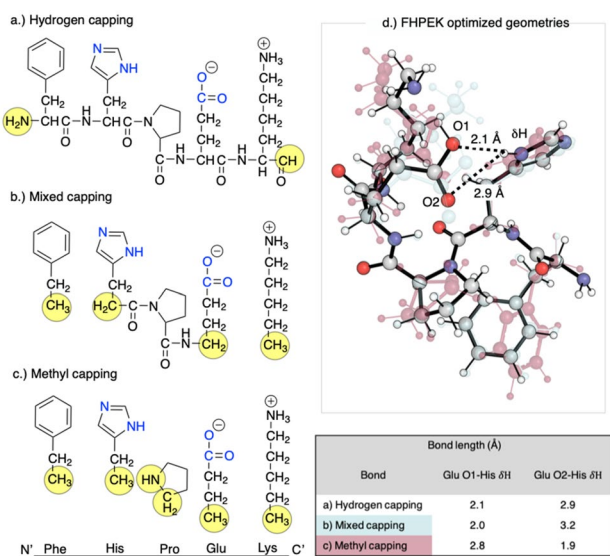
various schemes may need to be examined in models of catalysis in the active site of IGPS.

6.3 Connecting Macroscopic Turnover Frequency to the Microscopic Conformational Ensemble

IGPS and other enzymes explore multiple conformations with different levels of catalytic activity. One such example is the acyltransferase, LovD, which is part of the lovastatin biosynthetic pathway. Exploring these conformational dynamics is essential to model reactivity. For example, the directed evolution of LovD produces a favorably mutated version (LovD9), whose catalytic performance can only be captured by microsecond MD simulations [32]. The overall catalytic activity is influenced by the populations of different conformational states and their interconversion, and the intrinsic reactivity (i.e., catalytic activation barrier) of these states. Therefore, a multistep computational workflow can be envisaged for modeling catalysis in allosteric enzymes such as IGPS (Fig. 11). 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 enhanced sampling techniques. Second, structural clustering of trajectories will be performed to yield representatives of each cluster, along with populations. Third, multistep reaction Gibbs energy profiles for glutamine hydrolysis will be generated for each conformational cluster from MD. Fourth, individual k_{cat} values are weighted

Fig. 10 Truncation schemes applied to peptide sequence 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 alpha carbons are capped, creating methyl groups, His and Glu are hydrogen capped at the N and C-terminus, 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
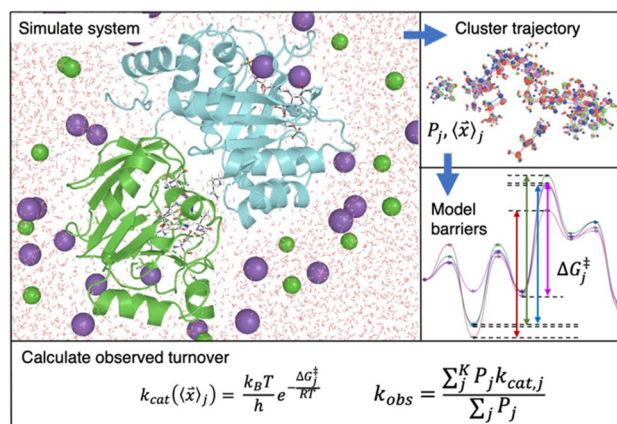

Fig. 11 A stepwise workflow combining MD simulation, conformational clustering, and QM calculations to calculate catalytic turnover. The first step is to simulate the system with explicit solvent all-atom MD. Next, structural clustering of the trajectory yields populations, P_j and cluster average structure coordinates, $\langle \bar{x} \rangle_j$. 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

according to conformational populations to predict overall turnover frequency according to Eq. 5.

7 Conclusions and Future Directions

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. In particular, 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 perspective, we have emphasized 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. This integrated approach is also suitable for use with QM/MM generated barriers. 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 allosteric enzymes. Although the computational framework discussed herein applies to V-type allosteric systems, 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, the advent of GPU accelerated MD and QM approaches makes this tractable for numerous enzymes.

The empirical correlation between active conformation population and catalytic activity in IGPS shows the value in using MD and structural clustering to account for the influence of conformational heterogeneity on enzyme catalysis. This behavior is expected to be consistent among other enzymes that follow the ensemble model of allostery. Furthermore, the concept of free energy redistribution is a more general concept that numerous enzymes have exhibited as a way to regulate catalytic activity. Therefore, the workflow illustrated in Fig. 11 may be applicable across various enzymatic systems, although there may be cases where the results are consistent with single-conformation studies, particularly in more rigid systems.

Using a combined MD/QM conformational ensemble approach, computational studies of enzyme catalysis will address critical open questions. These include:

1. What are the relative rates of conformational sampling and enzymatic reactions for specific enzymes?

2. Can a single "active" conformation capture observed catalytic properties, or do enzymatic reactions proceed from various enzyme–substrate conformations?

3. Are there general mechanisms that can be applied to classes of enzymes? For example, do V-type allosteric enzymes all conform to the ensemble model of allostery?

Additionally, several assumptions underlying such approaches require further interrogation, such as (i) whether a single structure from a conformational cluster can be used to compute the energy barrier for a reaction, as in Eq. 5; and (ii) whether different enzymatic reactions progress within the same conformational state, or whether in the limit of relatively fast conformational transitions, the Curtin-Hammett principle of chemical reactivity can be applied to enzyme catalysis. Further exploration of these questions will provide insightful results that propel the field of computational enzyme catalysis towards a land of opportunities.

Declarations

Conflict of interest The authors do not have conflicts of interest to declare.

References

1. Edwards DR, Lohman DC, Wolfenden R (2012) Catalytic proficiency: the extreme case of S-O cleaving sulfatases. J Am Chem Soc 134:525–531. https://doi.org/10.1021/ja208827q
2. Fischer E (1894) Einfluss der configuration auf die wirkung der enzyme. Ber Dtsch Chem Ges 27:2985–2993. https://doi.org/10.1002/cber.18940270364
3. Koshland DE (1958) Application of a theory of enzyme specificity to protein synthesis. Proc Natl Acad Sci 44:98–104. https://doi.org/10.1073/pnas.44.2.98
4. Ramanathan A, Savol A, Burger V et al (2014) Protein conformational populations and functionally relevant substates. Acc Chem Res 47:149–156. https://doi.org/10.1021/ar400084s
5. Warshel A (1978) Energetics of enzyme catalysis. Proc Natl Acad Sci USA 75:5250–5254
6. Marcus RA (1956) On the theory of oxidation-reduction reactions involving electron transfer. I. J Chem Phys 24:966–978. https://doi.org/10.1063/1.1742723
7. Smith AJT, Müller R, Toscano MD et al (2008) Structural reorganization and preorganization in enzyme active sites: comparisons of experimental and theoretically ideal active site geometries in the multistep serine esterase reaction cycle. J Am Chem Soc 130:15361–15373. https://doi.org/10.1021/ja803213p
8. Warshel A, Sharma PK, Kato M et al (2006) Electrostatic basis for enzyme catalysis. Chem Rev 106:3210–3235. https://doi.org/10.1021/cr0503106

9. Lodola A, Sirirak J, Fey N et al (2010) Structural fluctuations in enzyme-catalyzed reactions: determinants of reactivity in fatty acid amide hydrolase from multivariate statistical analysis of quantum mechanics/molecular mechanics paths. J Chem Theory Comput 6:2948–2960. https://doi.org/10.1021/ct100264j

10. Benkovic SJ (2003) A perspective on enzyme catalysis. Science 301:1196–1202. https://doi.org/10.1126/science.1085515

11. Himo F (2017) Recent trends in quantum chemical modeling of enzymatic reactions. J Am Chem Soc 139:6780–6786. https://doi.org/10.1021/jacs.7b02671

12. Lind MES, Himo F (2013) Quantum chemistry as a tool in asymmetric biocatalysis: limonene epoxide hydrolase test case. Angew Chem Int Ed 52:4563–4567. https://doi.org/10.1002/anie.201300594

13. Lind MES, Himo F (2014) Theoretical study of reaction mechanism and stereoselectivity of arylmalonate decarboxylase. ACS Catal 4:4153–4160. https://doi.org/10.1021/cs5009738

14. Lind MES, Himo F (2016) Quantum chemical modeling of enantioconvergency in soluble epoxide hydrolase. ACS Catal 6:8145–8155. https://doi.org/10.1021/acscatal.6b01562

15. Hotta K, Chen X, Paton RS et al (2012) Enzymatic catalysis of anti-Baldwin ring closure in polyether biosynthesis. Nature 483:355–358. https://doi.org/10.1038/nature10865

16. Tantillo DJ, Jiangang C, Houk KN (1998) Theozymes and compuzymes: theoretical models for biological catalysis. Curr Opin Chem Biol 2:743–750. https://doi.org/10.1016/S1367-5931(98)80112-9

17. Kiss G, Çelebi-Ölçüm N, Moretti R et al (2013) Computational enzyme design. Angew Chem Int Ed 52:5700–5725. https://doi.org/10.1002/anie.201204077

18. Schütz M (2000) Low-order scaling local electron correlation methods. III. Linear scaling local perturbative triples correction (T). J Chem Phys 113:9986–10001. https://doi.org/10.1063/1.1323265

19. Claeyssens F, Harvey JN, Manby FR et al (2006) High-accuracy computation of reaction barriers in enzymes. Angew Chem 118:7010–7013. https://doi.org/10.1002/ange.200602711

20. Lin H, Truhlar DG (2005) Redistributed charge and dipole schemes for combined quantum mechanical and molecular mechanical calculations. J Phys Chem A 109:3991–4004. https://doi.org/10.1021/jp0446332

21. Yu EW, Koshland DE (2001) Propagating conformational changes over long (and short) distances in proteins. Proc Natl Acad Sci 98:9517–9520. https://doi.org/10.1073/pnas.161239298

22. Nussinov R, Tsai C-J (2014) Unraveling structural mechanisms of allosteric drug action. Trends Pharmacol Sci 35:256–264. https://doi.org/10.1016/j.tips.2014.03.006

23. Nussinov R (2016) Introduction to protein ensembles and allostery. Chem Rev 116:6263–6266. https://doi.org/10.1021/acs.chemrev.6b00283

24. Otten R, Liu L, Kenner LR et al (2018) Rescue of conformational dynamics in enzyme catalysis by directed evolution. Nat Commun 9:1314. https://doi.org/10.1038/s41467-018-03562-9

25. Otten R, Pádua RAP, Bunzel HA, et al (2020) How directed evolution reshapes the energy landscape in an enzyme to boost catalysis. Science eabd3623. https://doi.org/10.1126/science.abd3623

26. Broom A, Rakotoharisoa RV, Thompson MC et al (2020) Ensemble-based enzyme design can recapitulate the effects of laboratory directed evolution in silico. Nat Commun 11:4808. https://doi.org/10.1038/s41467-020-18619-x

27. Bhabha G, Lee J, Ekiert DC et al (2011) A dynamic knockout reveals that conformational fluctuations influence the chemical step of enzyme catalysis. Science 332:234–238. https://doi.org/10.1126/science.1198542

28. Wurm JP, Sung S, Kneuttinger AC et al (2021) Molecular basis for the allosteric activation mechanism of the heterodimeric imidazole glycerol phosphate synthase complex. Nat Commun 12:2748. https://doi.org/10.1038/s41467-021-22968-6

29. Ahmadi S, Herrera LB, Chehelamirani M et al (2018) Multi-scale modeling of enzymes: QM-cluster, QM/MM, and QM/MM/MD: a tutorial review. Int J Quantum Chem 118:e25558. https://doi.org/10.1002/qua.25558

30. Maria-Solano MA, Serrano-Hervás E, Romero-Rivera A et al (2018) Role of conformational dynamics in the evolution of novel enzyme function. Chem Commun 54:6622–6634. https://doi.org/10.1039/C8CC02426J

31. Lonsdale R, Harvey JN, Mulholland AJ (2012) A practical guide to modelling enzyme-catalysed reactions. Chem Soc Rev 41:3025. https://doi.org/10.1039/c2cs15297e

32. Jiménez-Osés G, Osuna S, Gao X et al (2014) The role of distant mutations and allosteric regulation on LovD active site dynamics. Nat Chem Biol 10:431–436. https://doi.org/10.1038/nchembio.1503

33. Maria-Solano MA, Romero-Rivera A, Osuna S (2017) Exploring the reversal of enantioselectivity on a zinc-dependent alcohol dehydrogenase. Org Biomol Chem 15:4122–4129. https://doi.org/10.1039/C7OB00482F

34. Maria-Solano MA, Iglesias-Fernández J, Osuna S (2019) Deciphering the allosterically driven conformational ensemble in tryptophan synthase evolution. J Am Chem Soc 141:13049–13056. https://doi.org/10.1021/jacs.9b03646

35. Nussinov R, Tsai C-J (2015) Allostery without a conformational change? Revisiting the paradigm. Curr Opin Struct Biol 30:17–24. https://doi.org/10.1016/j.sbi.2014.11.005

36. Myers RS, Jensen JR, Deras IL et al (2003) Substrate-induced changes in the ammonia channel for imidazole glycerol phosphate synthase. Biochemistry 42:7013–7022. https://doi.org/10.1021/bi034314l

37. Glowacki DR, Harvey JN, Mulholland AJ (2012) Taking Ockham's razor to enzyme dynamics and catalysis. Nat Chem 4:169–176. https://doi.org/10.1038/nchem.1244

38. Benkovic SJ, Hammes GG, Hammes-Schiffer S (2008) Free-energy landscape of enzyme catalysis. Biochemistry 47:3317–3321. https://doi.org/10.1021/bi800049z

39. Eisenstein O, Ujaque G, Lledós A (2020) What makes a good (computed) energy profile? In: Lledós A, Ujaque G (eds) New directions in the modeling of organometallic reactions. Springer, Cham, pp 1–38

40. Yoon T (2009) Commentary: reviewer comments a discussion of "can reaction mechanisms be proven?" Chemical Education Today.

41. Peng Q, Duarte F, Paton RS (2016) Computing organic stereoselectivity – from concepts to quantitative calculations and predictions. Chem Soc Rev 45:6093–6107. https://doi.org/10.1039/C6CS00573J

42. Luchini G, Alegre-Requena JV, Funes-Ardoiz I, Paton RS (2020) GoodVibes: automated thermochemistry for heterogeneous computational chemistry data. F1000Res 9:291. https://doi.org/10.12688/f1000research.22758.1

43. Kozuch S, Shaik S (2011) How to conceptualize catalytic cycles? The energetic span model. Acc Chem Res 44:101–110. https://doi.org/10.1021/ar1000956

44. Masgrau L, Truhlar DG (2015) The importance of ensemble averaging in enzyme kinetics. Acc Chem Res 48:431–438. https://doi.org/10.1021/ar500319e

45. Garcia-Viloca M, Gao J, Karplus M, Truhlar DG (2004) How enzymes work: analysis by modern rate theory and computer simulations. Science 303:186–195. https://doi.org/10.1126/science.1088172

46. Bistoni G, Polyak I, Sparta M et al (2018) Toward accurate QM/MM reaction barriers with large QM regions using domain based pair natural orbital coupled cluster theory. J Chem Theory Comput 14:3524–3531. https://doi.org/10.1021/acs.jctc.8b00348

47. Siegbahn PEM, Himo F (2011) The quantum chemical cluster approach for modeling enzyme reactions. WIREs Comput Mol Sci 1:323–336. https://doi.org/10.1002/wcms.13

48. Frauenfelder H, Sligar SG, Wolynes PG (1991) The energy landscapes and motions of proteins. Science 254:1598

49. Frauenfelder H, McMahon B (1998) Dynamics and function of proteins: the search for general concepts. Proc Natl Acad Sci 95:4795–4797. https://doi.org/10.1073/pnas.95.9.4795

50. Baldwin AJ, Kay LE (2009) NMR spectroscopy brings invisible protein states into focus. Nat Chem Biol 5:808–814. https://doi.org/10.1038/nchembio.238

51. Ando T, Kodera N, Takai E et al (2001) A high-speed atomic force microscope for studying biological macromolecules. Proc Natl Acad Sci 98:12468–12472

52. Junker JP, Ziegler F, Rief M (2009) Ligand-dependent equilibrium fluctuations of single calmodulin molecules. Science 323:633–637. https://doi.org/10.1126/science.1166191

53. Kodera N, Yamamoto D, Ishikawa R, Ando T (2010) Video imaging of walking myosin V by high-speed atomic force microscopy. Nature 468:72–76. https://doi.org/10.1038/nature09450

54. Bai X, McMullan G, Scheres SHW (2015) How cryo-EM is revolutionizing structural biology. Trends Biochem Sci 40:49–57. https://doi.org/10.1016/j.tibs.2014.10.005

55. Glaeser RM (2016) How good can cryo-EM become? Nat Methods 13:28–32. https://doi.org/10.1038/nmeth.3695

56. Nogales E (2016) The development of cryo-EM into a mainstream structural biology technique. Nat Methods 13:24–27. https://doi.org/10.1038/nmeth.3694

57. Bonomi M, Vendruscolo M (2019) Determination of protein structural ensembles using cryo-electron microscopy. Curr Opin Struct Biol 56:37–45. https://doi.org/10.1016/j.sbi.2018.10.006

58. Fraser JS, van den Bedem H, Samelson AJ et al (2011) Accessing protein conformational ensembles using room-temperature X-ray crystallography. Proc Natl Acad Sci 108:16247–16252. https://doi.org/10.1073/pnas.1111325108

59. Qin H, Lim L, Song J (2012) Protein dynamics at Eph receptor-ligand interfaces as revealed by crystallography NMR and MD simulations. BMC Biophys 5:2. https://doi.org/10.1186/2046-1682-5-2

60. Fenwick RB, van den Bedem H, Fraser JS, Wright PE (2014) Integrated description of protein dynamics from room-temperature X-ray crystallography and NMR. Proc Natl Acad Sci 111:E445–E454. https://doi.org/10.1073/pnas.1323440111

61. Mulder FAA, Filatov M (2010) NMR chemical shift data and ab initio shielding calculations: emerging tools for protein structure determination. Chem Soc Rev 39:578–590. https://doi.org/10.1039/B811366C

62. Tolman JR, Ruan K (2006) NMR residual dipolar couplings as probes of biomolecular dynamics. Chem Rev 106:1720–1736. https://doi.org/10.1021/cr040429z

63. Sibille N, Bernadó P (2012) Structural characterization of intrinsically disordered proteins by the combined use of NMR and SAXS. Biochem Soc Trans 40:955–962. https://doi.org/10.1042/BST20120149

64. Schwalbe M, Ozenne V, Bibow S et al (2014) Predictive atomic resolution descriptions of intrinsically disordered hTau40 and α-synuclein in solution from NMR and small angle scattering. Structure 22:238–249. https://doi.org/10.1016/j.str.2013.10.020

65. Rivalta I, Sultan MM, Lee N-S et al (2012) Allosteric pathways in imidazole glycerol phosphate synthase. Proc Natl Acad Sci 109:E1428–E1436. https://doi.org/10.1073/pnas.1120536109

66. Lisi GP, Loria JP (2017) Allostery in enzyme catalysis. Curr Opin Struct Biol 47:123–130. https://doi.org/10.1016/j.sbi.2017.08.002

67. Negre CFA, Morzan UN, Hendrickson HP et al (2018) Eigenvector centrality for characterization of protein allosteric pathways. Proc Natl Acad Sci USA 115:E12201–E12208. https://doi.org/10.1073/pnas.1810452115

68. Bonomi M, Heller GT, Camilloni C, Vendruscolo M (2017) Principles of protein structural ensemble determination. Curr Opin Struct Biol 42:106–116. https://doi.org/10.1016/j.sbi.2016.12.004

69. Jing Z, Liu C, Cheng SY et al (2019) Polarizable force fields for biomolecular simulations: recent advances and applications. Annu Rev Biophys 48:371–394. https://doi.org/10.1146/annurev-biophys-070317-033349

70. Unke OT, Chmiela S, Sauceda HE, et al (2021) Machine learning force fields. Chem Rev. https://doi.org/10.1021/acs.chemrev.0c01111

71. Shaw DE, Maragakis P, Lindorff-Larsen K et al (2010) Atomic-level characterization of the structural dynamics of proteins. Science 330:341–346. https://doi.org/10.1126/science.1187409

72. Hansmann UHE (1997) Parallel tempering algorithm for conformational studies of biological molecules. Chem Phys Lett 281:140–150. https://doi.org/10.1016/S0009-2614(97)01198-6

73. Sugita Y, Okamoto Y (1999) Replica-exchange molecular dynamics method for protein folding. Chem Phys Lett 314:141–151. https://doi.org/10.1016/S0009-2614(99)01123-9

74. Laio A, Parrinello M (2002) Escaping free-energy minima. Proc Natl Acad Sci 99:12562–12566. https://doi.org/10.1073/pnas.202427399

75. Zhou T, Caflisch A (2012) Free energy guided sampling. J Chem Theory Comput 8:2134–2140. https://doi.org/10.1021/ct300147t

76. Bacci M, Vitalis A, Caflisch A (2015) A molecular simulation protocol to avoid sampling redundancy and discover new states. Biochim Biophys Acta 1850:889–902. https://doi.org/10.1016/j.bbagen.2014.08.013

77. Zimmerman MI, Bowman GR (2015) FAST conformational searches by balancing exploration/exploitation trade-offs. J Chem Theory Comput 11:5747–5757. https://doi.org/10.1021/acs.jctc.5b00737

78. Shrestha UR, Smith JC, Petridis L (2021) Full structural ensembles of intrinsically disordered proteins from unbiased molecular dynamics simulations. Commun Biol 4:243. https://doi.org/10.1038/s42003-021-01759-1

79. Campbell E, Kaltenbach M, Correy GJ et al (2016) The role of protein dynamics in the evolution of new enzyme function. Nat Chem Biol 12:944–950. https://doi.org/10.1038/nchembio.2175

80. Hong N-S, Petrović D, Lee R et al (2018) The evolution of multiple active site configurations in a designed enzyme. Nat Commun 9:3900. https://doi.org/10.1038/s41467-018-06305-y

81. Romero-Rivera A, Garcia-Borràs M, Osuna S (2017) Role of conformational dynamics in the evolution of retro-aldolase activity. ACS Catal 7:8524–8532. https://doi.org/10.1021/acscatal.7b02954

82. Lake PT, Davidson RB, Klem H et al (2020) Residue-level allostery propagates through the effective coarse-grained hessian. J Chem Theory Comput 16:3385–3395. https://doi.org/10.1021/acs.jctc.9b01149

83. Fenton AW (2008) Allostery: an illustrated definition for the 'second secret of life.' Trends Biochem Sci 33:420–425. https://doi.org/10.1016/j.tibs.2008.05.009

84. Changeux J-P (2005) Allosteric mechanisms of signal transduction. Science 308:1424–1428. https://doi.org/10.1126/science.1108595

85. Gunasekaran K, Ma B, Nussinov R (2004) Is allostery an intrinsic property of all dynamic proteins? Proteins 57:433–443. https://doi.org/10.1002/prot.20232

86. Li J, White JT, Saavedra H, et al (2017) Genetically tunable frustration controls allostery in an intrinsically disordered transcription factor. eLife 6:e30688. https://doi.org/10.7554/eLife.30688

87. Davidson RB, Hendrix J, Geiss BJ, McCullagh M (2018) Allostery in the dengue virus NS3 helicase: insights into the NTPase cycle from molecular simulations. PLoS Comput Biol 14:e1006103. https://doi.org/10.1371/journal.pcbi.1006103

88. Wodak SJ, Paci E, Dokholyan NV et al (2019) Allostery in its many disguises: from theory to applications. Structure 27:566–578. https://doi.org/10.1016/j.str.2019.01.003

89. Huang Z, Zhu L, Cao Y et al (2011) ASD: a comprehensive database of allosteric proteins and modulators. Nucl Acids Res 39:D663–D669. https://doi.org/10.1093/nar/gkq1022

90. Huang Z, Mou L, Shen Q et al (2014) ASD v2.0: updated content and novel features focusing on allosteric regulation. Nucl Acids Res 42:D510–D516. https://doi.org/10.1093/nar/gkt1247

91. Liu X, Lu S, Song K, et al (2019) Unraveling allosteric landscapes of allosterome with ASD. Nucl Acids Res. https://doi.org/10.1093/nar/gkz958

92. Monod J, Wyman J, Changeux J-P (1965) On the nature of allosteric transitions: a plausible model. 31

93. Perutz MF, Rossmann MG, Cullis AF et al (1960) Structure of hæmoglobin: a three-dimensional fourier synthesis at 5.5-Å. Resolution, obtained by x-ray analysis. Nature 185:416–422. https://doi.org/10.1038/185416a0

94. Koshland DE, Némethy G, Filmer D (1966) Comparison of experimental binding data and theoretical models in proteins containing subunits*. Biochemistry 5:365–385. https://doi.org/10.1021/bi00865a047

95. Cooper A (1984) Protein fluctuations and the thermodynamic uncertainty principle. Prog Biophys Mol Biol 44:181–214. https://doi.org/10.1016/0079-6107(84)90008-7

96. Hilser VJ, Wrabl JO, Motlagh HN (2012) Structural and energetic basis of allostery. Annu Rev Biophys 41:585–609. https://doi.org/10.1146/annurev-biophys-050511-102319

97. Wei G, Xi W, Nussinov R, Ma B (2016) Protein ensembles: how does nature harness thermodynamic fluctuations for life? the diverse functional roles of conformational ensembles in the cell. Chem Rev 116:6516–6551. https://doi.org/10.1021/acs.chemrev.5b00562

98. Motlagh HN, Wrabl JO, Li J, Hilser VJ (2014) The ensemble nature of allostery. Nature 508:331–339. https://doi.org/10.1038/nature13001

99. Schueler-Furman O, Wodak SJ (2016) Computational approaches to investigating allostery. Curr Opin Struct Biol 41:159–171. https://doi.org/10.1016/j.sbi.2016.06.017

100. Siegbahn PEM, Blomberg MRA (1999) Density functional theory of biologically relevant metal centers. Annu Rev Phys Chem 50:221–249. https://doi.org/10.1146/annurev.physchem.50.1.221

101. Siegbahn PEM, Himo F (2009) Recent developments of the quantum chemical cluster approach for modeling enzyme reactions. J Biol Inorg Chem 14:643–651. https://doi.org/10.1007/s00775-009-0511-y

102. Liao R-Z, Yu J-G, Raushel FM, Himo F (2008) Theoretical investigation of the reaction mechanism of the dinuclear zinc enzyme dihydroorotase. Chem Eur J 14:4287–4292. https://doi.org/10.1002/chem.200701948

103. Kazemi M, Sheng X, Kroutil W, Himo F (2018) Computational study of *Mycobacterium smegmatis* acyl transferase reaction mechanism and specificity. ACS Catal 8:10698–10706. https://doi.org/10.1021/acscatal.8b03360

104. Ryde U (2017) How many conformations need to be sampled to obtain converged QM/MM energies? the curse of exponential averaging. J Chem Theory Comput 13:5745–5752. https://doi.org/10.1021/acs.jctc.7b00826

105. Siegbahn PEM, Li X (2017) Cluster size convergence for the energetics of the oxygen evolving complex in PSII. J Comput Chem 38:2157–2160. https://doi.org/10.1002/jcc.24863

106. Sumowski CV, Schmitt BBT, Schweizer S, Ochsenfeld C (2010) Quantum-chemical and combined quantum-chemical/molecular-mechanical studies on the stabilization of a twin arginine pair in adenovirus Ad11. Angew Chem Int Ed 49:9951–9955. https://doi.org/10.1002/anie.201004022

107. Hu L, Söderhjelm P, Ryde U (2013) Accurate reaction energies in proteins obtained by combining QM/MM and large QM calculations. J Chem Theory Comput 9:640–649. https://doi.org/10.1021/ct3005003

108. Liao R-Z, Thiel W (2013) Convergence in the QM-only and QM/MM modeling of enzymatic reactions: a case study for acetylene hydratase. J Comput Chem n/a-n/a. https://doi.org/10.1002/jcc.23403

109. Raugei S, Seefeldt LC, Hoffman BM (2018) Critical computational analysis illuminates the reductive-elimination mechanism that activates nitrogenase for N_2 reduction. Proc Natl Acad Sci USA 115:E10521–E10530. https://doi.org/10.1073/pnas.1810211115

110. Hu L, Eliasson J, Heimdal J, Ryde U (2009) Do quantum mechanical energies calculated for small models of protein-active sites converge? J Phys Chem A 113:11793–11800. https://doi.org/10.1021/jp9029024

111. Siegbahn PEM (2016) 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 138:10485–10495. https://doi.org/10.1021/jacs.6b03846

112. Chen S-L, Fang W-H, Himo F (2008) Technical aspects of quantum chemical modeling of enzymatic reactions: the case of phosphotriesterase. Theor Chem Account 120:515–522. https://doi.org/10.1007/s00214-008-0430-y

113. Blomberg MRA, Borowski T, Himo F et al (2014) Quantum chemical studies of mechanisms for metalloenzymes. Chem Rev 114:3601–3658. https://doi.org/10.1021/cr400388t

114. Becke AD, Johnson ER (2005) A density-functional model of the dispersion interaction. J Chem Phys 123:154101. https://doi.org/10.1063/1.2065267

115. Hegeman GD (1966) Synthesis of the enzymes of the mandelate pathway by *Pseudomonas putida* I. Synthesis of enzymes by the wild type. J Bacteriol 91:1140–1154. https://doi.org/10.1128/jb.91.3.1140-1154.1966

116. Hegeman GD (1970) Benzoylformate decarboxylase (*Pseudomonas putida*). In: Methods in enzymology. Elsevier, New York, pp 674–678

117. Blomberg MRA, Siegbahn PEM (2012) Mechanism for N_2O generation in bacterial nitric oxide reductase: a quantum chemical study. Biochemistry 51:5173–5186. https://doi.org/10.1021/bi300496e

118. Planas F, Sheng X, McLeish MJ, Himo F (2018) A theoretical study of the benzoylformate decarboxylase reaction mechanism. Front Chem 6:205. https://doi.org/10.3389/fchem.2018.00205

119. Planas F, McLeish MJ, Himo F (2019) Computational study of enantioselective carboligation catalyzed by benzoylformate decarboxylase. ACS Catal 9:5657–5667. https://doi.org/10.1021/acscatal.9b01084

120. Magalhães RP, Fernandes HS, Sousa SF (2020) Modelling enzymatic mechanisms with QM/MM approaches: current status and future challenges. Isr J Chem 60:655–666. https://doi.org/10.1002/ijch.202000014

121. Ramos MJ, Fernandes PA (2008) Computational enzymatic catalysis. Acc Chem Res 41:689–698. https://doi.org/10.1021/ar7001045

122. Monard G, Merz KM (1999) Combined quantum mechanical/molecular mechanical methodologies applied to biomolecular systems. Acc Chem Res 32:904–911. https://doi.org/10.1021/ar970218z

123. Gao J, Truhlar DG (2002) Quantum mechanical methods for enzyme kinetics. Annu Rev Phys Chem 53:467–505. https://doi.org/10.1146/annurev.physchem.53.091301.150114

124. Rosta E, Klähn M, Warshel A (2006) Towards accurate ab initio qm/mm calculations of free-energy profiles of enzymatic reactions. J Phys Chem B 110:2934–2941. https://doi.org/10.1021/jp057109j

125. Lin H, Truhlar DG (2007) QM/MM: what have we learned, where are we, and where do we go from here? Theor Chem Acc 117:185. https://doi.org/10.1007/s00214-006-0143-z

126. Warshel A, Levitt M (1976) Theoretical studies of enzymic reactions: dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. J Mol Biol 103:227–249. https://doi.org/10.1016/0022-2836(76)90311-9

127. Senn HM, Thiel W (2009) QM/MM methods for biomolecular systems. Angew Chem Int Ed 48:1198–1229. https://doi.org/10.1002/anie.200802019

128. Gao J, Ma S, Major DT et al (2006) Mechanisms and free energies of enzymatic reactions. Chem Rev 106:3188–3209. https://doi.org/10.1021/cr050293k

129. Cortopassi WA, Simion R, Honsby CE et al (2015) Dioxygen binding in the active site of histone demethylase JMJD2A and the role of the protein environment. Chem Eur J 21:18983–18992. https://doi.org/10.1002/chem.201502983

130. Walker RC, de Souza MM, Mercer IP et al (2002) Large and fast relaxations inside a protein: calculation and measurement of reorganization energies in alcohol dehydrogenase. J Phys Chem B 106:11658–11665. https://doi.org/10.1021/jp0261814

131. van der Kamp MW, Mulholland AJ (2013) Combined quantum mechanics/molecular mechanics (QM/MM) methods in computational enzymology. Biochemistry 52:2708–2728. https://doi.org/10.1021/bi400215w

132. Senn HM, Thiel W (2007) QM/MM studies of enzymes. Curr Opin Chem Biol 11:182–187. https://doi.org/10.1016/j.cbpa.2007.01.684

133. Sousa SF, Ribeiro AJM, Neves RPP et al (2017) Application of quantum mechanics/molecular mechanics methods in the study of enzymatic reaction mechanisms. WIREs Comput Mol Sci 7:e1281. https://doi.org/10.1002/wcms.1281

134. Klähn M, Braun-Sand S, Rosta E, Warshel A (2005) On possible pitfalls in ab initio quantum mechanics/molecular mechanics minimization approaches for studies of enzymatic reactions. J Phys Chem B 109:15645–15650. https://doi.org/10.1021/jp0521757

135. Sumowski CV, Ochsenfeld C (2009) A convergence study of QM/MM isomerization energies with the selected size of the QM region for peptidic systems. J Phys Chem A 113:11734–11741. https://doi.org/10.1021/jp902876n

136. Blank ID, Sadeghian K, Ochsenfeld C (2015) A base-independent repair mechanism for DNA glycosylase—no discrimination within the active site. Sci Rep 5:10369. https://doi.org/10.1038/srep10369

137. Roßbach S, Ochsenfeld C (2017) 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 13:1102–1107. https://doi.org/10.1021/acs.jctc.6b00727

138. Hu L, Söderhjelm P, Ryde U (2011) On the convergence of QM/MM energies. J Chem Theory Comput 7:761–777. https://doi.org/10.1021/ct100530r

139. Solt I, Kulhánek P, Simon I et al (2009) Evaluating boundary dependent errors in QM/MM simulations. J Phys Chem B 113:5728–5735. https://doi.org/10.1021/jp807277r

140. Kulik HJ, Zhang J, Klinman JP, Martínez TJ (2016) How large should the QM region be in QM/MM calculations? the case of catechol O -methyltransferase. J Phys Chem B 120:11381–11394. https://doi.org/10.1021/acs.jpcb.6b07814

141. Lyman E, Pfaendtner J, Voth GA (2008) Systematic multiscale parameterization of heterogeneous elastic network models of proteins. Biophys J 95:4183–4192. https://doi.org/10.1529/biophysj.108.139733

142. Das S, Nam K, Major DT (2018) 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 14:1695–1705. https://doi.org/10.1021/acs.jctc.7b00964

143. Jindal G, Warshel A (2016) Exploring the dependence of QM/MM calculations of enzyme catalysis on the size of the QM region. J Phys Chem B 120:9913–9921. https://doi.org/10.1021/acs.jpcb.6b07203

144. Govender K, Gao J, Naidoo KJ (2014) AM1/d-CB1: a semiempirical model for qm/mm simulations of chemical glycobiology systems. J Chem Theory Comput 10:4694–4707. https://doi.org/10.1021/ct500372s

145. McCullagh M, Saunders MG, Voth GA (2014) Unraveling the mystery of ATP hydrolysis in actin filaments. J Am Chem Soc 136:13053–13058. https://doi.org/10.1021/ja507169f

146. Yagi K, Ito S, Sugita Y (2021) Exploring the minimum-energy pathways and free-energy profiles of enzymatic reactions with QM/MM calculations. J Phys Chem B 125:4701–4713. https://doi.org/10.1021/acs.jpcb.1c01862

147. Sun R, Sode O, Dama JF, Voth GA (2017) Simulating protein mediated hydrolysis of ATP and other nucleoside triphosphates by combining qm/mm molecular dynamics with advances in metadynamics. J Chem Theory Comput 13:2332–2341. https://doi.org/10.1021/acs.jctc.7b00077

148. Quinn TR, Steussy CN, Haines BE et al (2021) Microsecond timescale MD simulations at the transition state of Pm HMGR predict remote allosteric residues. Chem Sci 12:6413–6418. https://doi.org/10.1039/D1SC00102G

149. Cortopassi WA, Kumar K, Duarte F et al (2016) Mechanisms of histone lysine-modifying enzymes: a computational perspective on the role of the protein environment. J Mol Graph Model 67:69–84. https://doi.org/10.1016/j.jmgm.2016.04.011

150. Quesne MG, Borowski T, de Visser SP (2016) Quantum mechanics/molecular mechanics modeling of enzymatic processes: caveats and breakthroughs. Chemistry 22:2562–2581

151. Cooper AM, Kästner J (2014) Averaging techniques for reaction barriers in QM/MM simulations. ChemPhysChem 15:3264–3269. https://doi.org/10.1002/cphc.201402382

152. Romero-Téllez S, Cruz A, Masgrau L et al (2021) 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 23:13042–13054. https://doi.org/10.1039/D1CP01338F

153. von der Esch B, Dietschreit JCB, Peters LDM, Ochsenfeld C (2019) Finding reactive configurations: a machine learning approach for estimating energy barriers applied to sirtuin 5. J Chem Theory Comput 15:6660–6667. https://doi.org/10.1021/acs.jctc.9b00876

154. Rhee YM, Sorin EJ, Jayachandran G et al (2004) Simulations of the role of water in the protein-folding mechanism. Proc Natl Acad Sci 101:6456–6461. https://doi.org/10.1073/pnas.0307898101

155. Sittel F, Stock G (2018) Perspective: identification of collective variables and metastable states of protein dynamics. J Chem Phys 149:150901. https://doi.org/10.1063/1.5049637

156. Pérez-Hernández G, Noé F (2016) Hierarchical time-lagged independent component analysis: computing slow modes and reaction coordinates for large molecular systems. J Chem Theory Comput 12:6118–6129. https://doi.org/10.1021/acs.jctc.6b00738

157. Pérez-Hernández G, Paul F, Giorgino T et al (2013) Identification of slow molecular order parameters for Markov model construction. J Chem Phys 139:015102. https://doi.org/10.1063/1.4811489

158. Ceriotti M, Tribello GA, Parrinello M (2011) Simplifying the representation of complex free-energy landscapes using sketch-map. Proc Natl Acad Sci USA 108:13023–13028. https://doi.org/10.1073/pnas.1108486108

159. Trozzi F, Wang X, Tao P (2021) UMAP as a dimensionality reduction tool for molecular dynamics simulations of biomacromolecules: a comparison study. J Phys Chem B 125:5022–5034. https://doi.org/10.1021/acs.jpcb.1c02081

160. Westerlund AM, Delemotte L (2019) InfleCS: clustering free energy landscapes with gaussian mixtures. J Chem Theory Comput 15:6752–6759. https://doi.org/10.1021/acs.jctc.9b00454

161. List F, Vega MC, Razeto A et al (2012) Catalysis uncoupling in a glutamine amidotransferase bienzyme by unblocking the glutaminase active site. Chem Biol 19:1589–1599. https://doi.org/10.1016/j.chembiol.2012.10.012

162. Amaro RE, Sethi A, Myers RS et al (2007) A network of conserved interactions regulates the allosteric signal in a glutamine amidotransferase. Biochemistry 46:2156–2173. https://doi.org/10.1021/bi061708e

163. Strohmeier M, Raschle T, Mazurkiewicz J et al (2006) Structure of a bacterial pyridoxal 5'-phosphate synthase complex. Proc Natl Acad Sci 103:19284–19289. https://doi.org/10.1073/pnas.0604950103

164. Chaudhuri BN, Lange SC, Myers RS et al (2003) Toward understanding the mechanism of the complex cyclization reaction catalyzed by imidazole glycerolphosphate synthase: crystal structures of a ternary complex and the free enzyme. Biochemistry 42:7003–7012. https://doi.org/10.1021/bi034320h