COMPUTATIONAL STRATEGIES FOR UNDERSTANDING THE MOLECULAR BASIS OF BIOCHEMICAL AND BIOCATALYTIC PROCESSES

Carla Calvó-Tusell

Universitat de Girona

DOCTORAL THESIS

COMPUTATIONAL STRATEGIES FOR UNDERSTANDING THE MOLECULAR BASIS OF BIOCHEMICAL AND BIOCATALYTIC PROCESSES

Carla Calvó-Tusell

2023

DOCTORAL THESIS

Computational Strategies for Understanding the Molecular Basis of Biochemical and Biocatalytic Processes

Carla Calvó-Tusell

2023

Doctoral Programme in Chemistry

Supervised by: Prof. Sílvia Osuna Oliveras and Dr. Marc Garcia-Borràs

Tutor: Prof. Marcel Swart

Presented to obtain the degree of PhD at the University of Girona

Universitat de Girona

Prof. Dr. Sílvia Osuna Oliveras and Dr. Marc Garcia-Borràs, of University of Girona,

WE DECLARE:

That the thesis entitled “Computational Strategies for Understanding the Molecular Basis of Biochemical and Biocatalytic Processes”, presented to obtain a doctoral degree, has been carried and completed by Carla Calvó-Tusell under our supervision and that meets the requirements to opt for an International Doctorate.

For all intents and purposes, we hereby sign this document.

Signature

Prof. Dr. Sílvia Osuna Oliveras

Dr. Marc Garcia-Borràs

Girona, April 6th, 2023

Acknowledgements

I would like to specially thank the supervisors, researchers, PhD students, and other coworkers that I collaborated with during the elaboration of this thesis. In particular, I would like to thank my supervisors Sílvia Osuna and Marc Garcia-Borràs; the IGPS team, Miguel Angel Maria-Solano and Ferran Feixas; the NH-insertion team, Zhen Liu, Andrew Zhou, Kai Chen, and Frances Arnold; and the FDH team, Liliana Calzadiaz-Ramirez and Carlos Acevedo-Rocha. Thank you for the opportunity to participate in these amazing projects and all the discussions generated.

Also, I would like to thank Rommie Amaro and the Amaro Lab for hosting me during my research stay at University of California, San Diego. Special thanks to the members and staff of the Institut de Química Computacional i Catàlisi (IQCC) and *Comissió de Doctorat de Química* for the support during this period.

The projects of this thesis have been performed thanks to the financial support granted by: the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (ERC-2015-StG-679001, PI: Sílvia Osuna), the the Spanish MICINN for the project PGC2018-102192-B-I00 (PI: Sílvia Osuna) and PID2019-111300GA-I00 (PI: Marc Garcia-Borràs).

Full List of Publications

The following published articles have been included as chapters in this Thesis:

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

2. Liu, Z.; **Calvó-Tusell, C.**; Zhou, A. Z.; Chen, K.; Garcia-Borràs, M.; Arnold, F. H.; *Dual-Function Enzyme Catalysis for Enantioselective Carbon–Nitrogen Bond Formation. Nat. Chem.,* **2021**, 13 (12), 1166-1172.

3. **Calvó-Tusell, C.**[‡]; Liu, Z[‡]; Chen, K.; Arnold, F. H.; Garcia-Borràs, M.; *Reversing the enantioselectiviy of enzymatic carbene N-H insertion through mechanism-guided protein engineering. ChemRxiv,* **2022**, https://doi.org/10.26434/chemrxiv-2022-f02xh
([‡]Equally contributed authors)

4. Calzadiaz-Ramírez, L[‡]; **Calvó-Tusell, C.**[‡]; Stoffel, G.[‡]; Lindner, S. N.; Osuna, S.; Erb, T. J.; Garcia-Borràs, M.; Bar-Even, A., Acevedo-Rocha, C. G*.; In vivo selection for formate dehydrogenases with high efficiency and specificity towards NADP+, ACS Catal.,* **2020**, 10 (14), 7512–7525.
([‡]Equally contributed authors)

Other published articles not included in this Thesis:

5. Codony, S.; **Calvó-Tusell, C.**; Valverde, E.; Osuna, S.; Morisseau, C.; Loza, M. I.; Brea, J. M.; Pérez, C.; Rodríguez-Franco, M. L.; Pizarro, J.; Corpas, R.; Griñán-Ferré, C.; Pallàs, M.; Sanfeliu,C.; Vázquez-Carrera, M.; Hammock, B. D.; Feixas, F.; Vázquez, S; *From the Design to the In Vivo Evaluation of Benzohomoadamantane-Derived Soluble Epoxide Hydrolase Inhibitors for the Treatment of Acute Pancreatitis, J. Med. Chem.,* **2021**, 64 (9), 5429-5446.

6. Codony, S.; Pont, C.; Griñán-Ferré, C.; Pede-Mattatelli, A. D.; **Calvó-Tusell, C.**; Feixas, F.; Osuna, S.; Jarné-Ferrer, J.; Naldi, M.; Bartolini, M.; Loza, M. I.; Brea, J.; Pérez, B.; Bartra, C.; Sanfeliu, C.; Juárez-Jiménez, J.; Morisseau, C.; Hammock, B. D.; Pallàs, M.; Vázquez, S.; Muñoz-Torrero. D.; *Discovery and in Vivo Proof-of-Concept of a Highly Potent Dual Inhibitor of Soluble Epoxide Hydrolase and Acetylcholinesterase for the Treatment of Alzheimer's Disease, J. Med. Chem.,* **2022**, 65 (6), 4909-4925.

7. Codony, S.; Entrena, J.M.; **Calvó-Tusell, C.**; Jora, B.; González-Cano, G.; Osuna, S.; Corpas, R.; Morisseau, C.; Pérez, B.; Barniol-Xicota, M.; Griñán-Ferré, C.; Pérez, C.; Rodríguez-Franco, M.I.; Martínez, A.L; Loza, M.I.; Pallàs, M.; Verhelst, S.H.L.; Sanfeliu, C.; Feixas, F.; Hammock, B.D.; Brea, J.; Cobos, E.J., Vázquez S.; *Synthesis, In Vitro Profiling, and In Vivo Evaluation of Benzohomoadamantane-Based Ureas for Visceral Pain: A New Indication for Soluble Epoxide Hydrolase Inhibitors. J. Med. Chem.,* **2022**, 65 (20), 13660-13680.

List of Abbreviations

Abbreviation	Description
AdK	Adenylate Kinase
AICAR	5-aminoimidazole-4-carboxamide ribotide
AMBER	Assisted Model Building with Energy Refinement
aMD	accelerated Molecular Dynamics
ATP	Adenosine triphosphate
CDPKs	Calcium-dependent protein kinases
CHARMM	Chemistry at HARvard Macromolecular Mechanics
cMD	conventional Molecular Dynamics
CNA	Community Network Analysis
CPU	Central Processing Unit
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
CSR-SALAD	Cofactor Specificity Reversal – Structural Analysis and Library Design
CV	Collective Variables
DE	Directed Evolution
DFT	Density Functional Theory
DNA	Deoxyribonucleic acid

ee	Enantiomeric excess
EDA	Acyclic Ethyl Diazoacetate
er	Enantiomeric ratio
EVB	Empirical Valence Bond
FAD	Flavinadenine Dinucleotide
FDH	Formate Dehydrogenases
FEL	Free Energy Landscape
FF	Force Field
GAFF	Generalized Amber Force Field
GaMD	Gauddian accelerated Molecular Dynamics
GATase	Glutamine amidotransferases
GPCRs	G protein-coupled receptors
GPU	Graphics Processing Unit
GROMOS	GROningen MOlecular Simulation
IGP	Imidazole Glycerol Phosphate
IGPS	Imidazole Glycerol Phosphate Synthase
ISM	Iterative Saturation Mutagenesis
ITC	Isothermal Titration Calorimetry
KNF	Koshland, Némethy and Filmer
L-Gln	L-Glutamine

LAC	Lactone-Carbene
MCPB	Metal Center Parameter Builder
MD	Molecular Dynamics
MetaD	Metadynamics
MM	Molecular Mechanics
MSM	Markov State Models
MWC	Monod, Wyman and Changeux
NAC	Near Attack Conformation
NAD	Nicotinamide Adenine Dinucleotide
NADP	Nicotinamide Adenine Dinucleotide phosphate
NMR	Nuclear Magnetic Resonance
NVT	Canonical ensemble
OxH	Oxyanion hole
PBC	Periodic Boundary Conditions
PC	Principal Components
PCA	Principal Components Analysis
PKA	Protein Kinases A
PseFDH	*Pseudomonas* Formate dehydrogenases
PT	Parallel tempering
PTC	Proton transfer catalyst

QM	Quantum Mechanics
QM/MM	Quantum Mechanics/Molecular Mechanics
RA	Retro Aldolases
RE	Replica Exchange
RESP	Restrained Electrostatic Potential
RMSD	Root-mean-square deviation
RMSF	Root-mean-square fluctuation
SDM	Site Directed Mutagenesis
SPM	Shortest Path Map
SSM	Site Saturated Mutagenesis
tICA	time-structure Independent Component Analysis
tRNA	transfer Ribonucleic acid
TS	Transition state
TTN	Turnover number
UEST	Unconstrained Enhanced Sampling Techniques
US	Umbrella Sampling
vdW	Van der Waals
WISP	Weighted Implementation of Suboptimal Paths

Contents

List of figures

Summary

Enzymes are biomolecules involved in a wide range of biological and chemical processes. Understanding enzyme function at molecular level is essential for deciphering the mechanisms of allosteric regulation, enzyme catalysis, and inhibition and harbors relevant information to design enzymes with specific functions. The ability of enzymes to develop their function - and to evolve towards new functions - in an accurate and specific way is conferred in part by their flexibility and dynamism. A detailed description of the structural and dynamic changes is key to provide a full picture of enzymatic mechanisms. However, the transient nature of allosteric processes or enzymatic reaction intermediates make them difficult to be captured with experimental techniques. Therefore, computational techniques can provide the required atomistic view to explain the molecular basis of biological processes.

The general goal of this thesis is to explore the molecular basis of biochemical and biocatalytic processes by means of computational methods and examine its relationship with enzymatic properties such as allostery, cofactor specificity, and catalytic activity. These computational protocols combine different techniques including molecular dynamics simulations, enhanced sampling techniques, dynamical networks, and quantum mechanics. By gaining insight into the molecular basis of these enzymatic processes, we rationalized the novel enzymatic functions of laboratory-evolved enzymes and used this information to rationally design new enzyme variants.

The results section is divided in three different chapters. In Chapter 4, we focus on understanding the molecular basis of allosteric regulation. In particular, we characterize the molecular details of the allosteric activation of imidazole glycerol phosphate synthase (IGPS) in the ternary complex and identified hidden states relevant for IGPS catalytic activity with a computational strategy tailored to explore millisecond timescale events. In Chapter 5, we design a computational protocol to unravel the molecular mechanism of the enantioselective N–H insertion in P411 variants. First, we explore the molecular basis of this enzymatic transformation and elucidate the role of key mutations in promoting asymmetric carbene N–H insertion. Second, we generate through a mechanistically-guided design strategy a biocatalytic platform for enantiodivergent C-N bond formation. In Chapter 6, we rationalize the molecular basis of cofactor specificity in engineered formate dehydrogenase variants that present three

properties: kinetic efficiency with the non-natural NADP$^+$ cofactor, specificity toward the non-natural NADP$^+$ cofactor and affinity toward the substrate formate.

Overall, these studies highlight the importance of understanding enzyme function at the molecular level to harness this information to design enzyme variants with specific functions.

Resum

Els enzims són biomolècules implicades en una àmplia gamma de processos biològics i químics. Entendre la funció enzimàtica a nivell molecular és essencial per desxifrar els mecanismes de regulació al·lostèrica, catàlisi enzimàtica i inhibició i també aporta informació rellevant per dissenyar enzims amb funcions específiques. La capacitat dels enzims per desenvolupar la seva funció -i evolucionar-los cap a noves funcions- d'una manera precisa i específica ve donada en part per la seva flexibilitat i dinamisme. Una descripció detallada dels canvis estructurals i dinàmics és clau per proporcionar una imatge completa dels mecanismes enzimàtics. Tanmateix, la naturalesa transitòria dels processos al·lostèrics o dels intermedis de la reacció enzimàtica dificulta la seva captura amb tècniques experimentals. Per tant, les tècniques computacionals poden proporcionar la visió atomística necessària per explicar la base molecular dels processos biològics.

L'objectiu general d'aquesta tesi és explorar la base molecular dels processos bioquímics i biocatalítics mitjançant mètodes computacionals i examinar la seva relació amb propietats enzimàtiques com l'al·losteria, l'especificitat del cofactor i l'activitat catalítica. Aquests protocols computacionals combinen diferents tècniques, com ara simulacions de dinàmica molecular, tècniques de mostreig millorat, xarxes dinàmiques i mecànica quàntica. En conèixer la base molecular d'aquests processos enzimàtics, hem racionalitzat les noves funcions enzimàtiques d'enzims desenvolupats al laboratori i utilitzat aquesta informació per dissenyar de manera racional noves variants enzimàtiques.

La secció de resultats està dividida en tres capítols diferents. Al capítol 4, ens centrem a entendre les bases moleculars de la regulació al·lostèrica. En particular, caracteritzem els detalls moleculars de l'activació al·lostèrica de l'imidazol glicerol fosfat sintasa (IGPS) al complex ternari i identifiquem estats rellevants per a l'activitat catalítica de l'IGPS amb una estratègia computacional adaptada per explorar esdeveniments a escala temporal de mil·lisegons. Al capítol 5, dissenyem un protocol computacional per descriure el mecanisme molecular de la inserció enantioselectiva de N–H en variants de P411. En primer lloc, explorem les bases moleculars d'aquesta transformació enzimàtica i dilucidem el paper de les mutacions clau en la promoció de la inserció N-H asimètrica al carbè. En segon lloc, generem mitjançant una estratègia de disseny guiada a partir de coneixements mecanístics una plataforma

biocatalítica per a la formació d'enllaços C-N enantiodivergents. Al capítol 6, racionalitzem la base molecular de l'especificitat del cofactor en variants de format deshidrogenasa dissenyades que presenten tres propietats: eficiència cinètica amb el cofactor $NADP^+$, especificitat cap al cofactor $NADP^+$ no natural i afinitat cap al format del substrat.

En general, aquests estudis destaquen la importància d'entendre la funció enzimàtica a nivell molecular per utilitzar aquesta informació per dissenyar variants enzimàtiques amb funcions específiques.

Resumen

Las enzimas son biomoléculas involucradas en una amplia gama de procesos biológicos y químicos. Comprender la función enzimática a nivel molecular es esencial para descifrar los mecanismos de regulación alostérica, catálisis enzimática e inhibición y alberga información relevante para diseñar enzimas con funciones específicas. La capacidad de las enzimas para desarrollar su función - y evolucionar hacia nuevas funciones - de forma precisa y específica viene dada en parte su flexibilidad y dinamismo. Una descripción detallada de los cambios estructurales y dinámicos es clave para proporcionar una imagen completa de los mecanismos enzimáticos. Sin embargo, la naturaleza transitoria de los procesos alostéricos o los intermedios de reacción enzimática dificultan su descripción con técnicas experimentales. Por lo tanto, las técnicas computacionales pueden proporcionar la visión atomística requerida para explicar la base molecular de los procesos biológicos.

El objetivo general de esta tesis es explorar las bases moleculares de los procesos bioquímicos y biocatalíticos mediante métodos computacionales y examinar su relación con propiedades enzimáticas tales como alosteria, especificidad de cofactor y actividad catalítica. Estos protocolos computacionales combinan diferentes técnicas que incluyen simulaciones de dinámica molecular, técnicas de muestreo mejoradas, redes dinámicas y mecánica cuántica. Al obtener información sobre la base molecular de estos procesos enzimáticos, racionalizamos las funciones enzimáticas novedosas de las enzimas desarrolladas en el laboratorio y utilizamos esta información para diseñar racionalmente nuevas variantes de enzimas.

La sección de resultados se divide en tres capítulos diferentes. En el Capítulo 4, está enfocado en comprender la base molecular de la regulación alostérica. En particular, caracterizamos los detalles moleculares de la activación alostérica de la enzima imidazol glicerol fosfato sintasa (IGPS) en el complejo ternario e identificamos estados relevantes para la actividad catalítica de IGPS con una estrategia computacional diseñada para explorar eventos de escala de tiempo de milisegundos. En el Capítulo 5, diseñamos un protocolo computacional para desentrañar el mecanismo molecular de la inserción enantioselectiva de N-H en las variantes de P411. Primero, exploramos la base molecular de esta transformación enzimática y dilucidamos el papel de las mutaciones clave en la promoción de la inserción N-H asimétrica en carbeno. En segundo lugar, generamos a través de una estrategia de diseño guiada mecánicamente una

plataforma biocatalítica para la formación de enlaces C-N enantiodivergentes. En el Capítulo 6, racionalizamos la base molecular de la especificidad del cofactor en variantes de formato deshidrogenasa diseñadas que presentan tres propiedades: eficiencia cinética con el cofactor $NADP^+$, especificidad hacia el cofactor $NADP^+$ no natural y afinidad hacia el sustrato de formato.

En general, estos estudios destacan la importancia de comprender la función de la enzima a nivel molecular para utilizar esta información para diseñar variantes de enzimas con funciones específicas.

CHAPTER 1. INTRODUCTION

1.1. Enzyme function and dynamics

Enzymes are essential for living organisms. Most enzymes are proteins whose role is to catalyze chemical reactions relevant for different cellular processes including metabolism, synthesis of cell components, and signaling.[1] Catalysts increase the rate of chemical reactions enabling them to proceed at milder conditions than uncatalyzed reactions. Enzymes evolved to be highly efficient, specific, and selective catalysts as a consequence of billions of years of optimization through evolution. They work by converting substrates into products in pre-organized active sites that provide the proper environment for accelerating chemical reactions. Besides the substrate, enzymes often require the interaction with other ions or molecules to become functional. Some enzymes incorporate organic or inorganic cofactors that directly participate in the chemical reaction.[2] Moreover, enzyme function can be controlled by environmental changes such as ligand binding or post-translational modifications through allosteric regulation. Understanding these properties harbors essential information to unravel how enzymes work. This information can be used to rationally design enzymes with specific functions.

In the following sections key basic concepts of enzyme function, structure, dynamics, and design will be introduced.

1.1.1. Enzyme structure

Enzymes present specific structural characteristics that enable them to catalyze the conversion of substrates to products. The catalytic activity of enzymes is encoded in the different degrees of protein structural organization (see Figure 1.1):[1,3]

- Primary structure: refers to the linear sequence of amino acids that compose the enzyme polypeptide chain. This sequence is determined by the genetic code and is unique for each enzyme. The primary structure of an enzyme harbors important information about the enzyme's function and stability.
- Secondary structure: refers to local structural arrangements of amino acids within the enzyme's polypeptide chain; the most common secondary structural elements are alpha helices, beta sheets and loops, which are then arranged to determine the enzyme's three-dimensional structure.

- Tertiary structure: refers to the three-dimensional (3D) structure of the enzyme's individual subunit itself, *i.e.,* monomer. It typically folds into a specific shape that determines the enzyme activity.
- Quaternary structure: Many enzymes are composed of multiple subunits that come together to form the functional enzyme. A wide number of enzymes exists as oligomers forming dimer, trimer, tetramer or more complex assemblies.

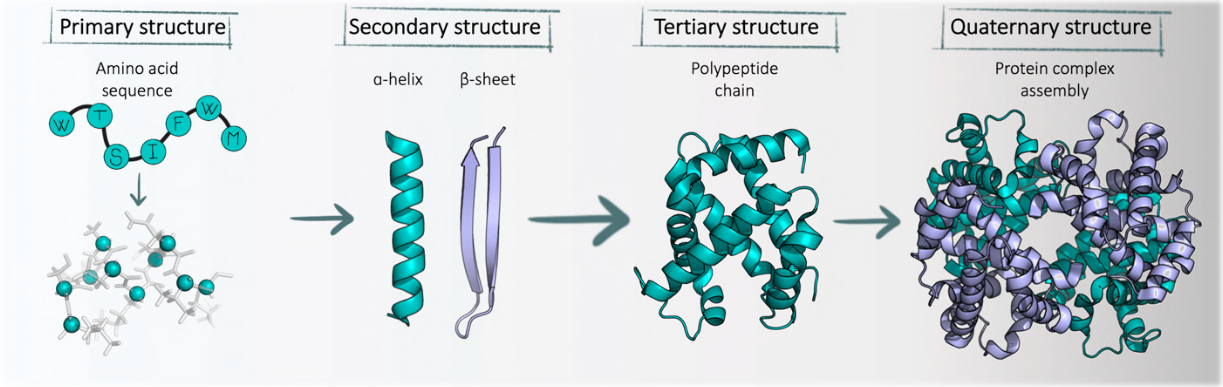


Figure 1.1. Protein structural organization. From left to right, the protein structure is organized by: primary structure, determined by the linear sequence of amino acids; secondary structure, refers to the local structural arrangements of amino acids that form the three-dimensional structure of the enzyme; tertiary structure, is the 3D structure of the enzyme's individual subunit; finally, some enzymes are composed of multiple subunits, which come together to form the functional enzyme, and this is called the quaternary structure.

These structural features work together to create an optimal environment for the chemical reaction to occur and are essential for the enzyme function. How amino acid residues are arranged in the 3D structure is key for the enzyme's activity, in particular, those that form the active site. However, residues beyond the active site such as the ones involved in substrate binding or residues that are found in distal positions from the catalytic site also play important roles in determining enzyme activity. The stability of an enzyme is also influenced by its structure, as well as its dynamics.[1] Understanding the structural features of natural enzymes is critical to unravel enzyme function and can aid in the design of new enzymes and in the optimization of existing ones.

Some enzymes require the interaction with additional chemical compounds to be active, called cofactors. Cofactors play an important role in enzymes by assisting in the catalytic activity and can be inorganic ions (*e.g.,* Zn^{2+}, Mg^{2}+, Cu^{+}, Cu^{2+}, Fe^{2}+, Fe^{3+}) or small organic molecules that bind to enzymes to help them perform their catalytic function, such as Nicotinamide Adenine Dinucleotide (NAD) or Flavin Adenine Dinucleotide (FAD) cofactors among many others.[1,2] Cofactors bind to specific sites on the enzyme and participate in enzymatic reactions by facilitating oxidation-reduction reactions and the transfer of electrons. Some enzymes require metal coordination complexes as cofactors, for example the heme cofactor (*i.e.,* iron ion coordinated to a porphyrin), which significantly enhances the chemical toolkit of enzymes. The study of the role of these cofactors in complex with the enzymes is one of the focuses of this thesis (see Chapters 5 and 6).

Generally, to refer to the complete active enzyme including the cofactor (also called coenzyme) and/or the metal ions, the term *holo*-enzyme is used, while to refer to the enzyme in the absence of cofactor or substrate, the term *apo*-enzyme is used.

Experimental methods such as *X-ray* crystallography, cryo-electron microscopy, or nuclear magnetic resonance (NMR) are commonly used to retrieve the 3D structure of enzymes.[2] However, sometimes these techniques do not have enough resolution to provide structural information of some segments such as flexible loops or side chains. Moreover, there are enzymes for which it is difficult to obtain the 3D structure. Computational protein structure-prediction techniques can be used to reconstruct the missing parts or suggest full 3D models using only the information of the primary structure.[4] Recently, artificial intelligence-based protein structure prediction methods have transformed the field providing accurate structures for a wide range of proteins.[5]

1.1.2. Enzyme catalysis

Enzymes are able to catalyze chemical reactions in biological environments by speeding up its reaction rate. The basic idea is that enzymes, as catalysts, do not modify the equilibrium of the reaction (*i.e.* the thermodynamics) but favor the kinetics of the reaction. Therefore, enzymes allow the equilibrium to be attained more rapidly without changing the structure and energetics of the reactants and products.[2]

To do so, enzymes form complexes with their substrates prior converting them into products. In simple terms, an enzyme-catalyzed reaction can be described in the following way:

$$E + S \rightarrow ES\ complex \rightarrow E + P$$

when the substrate (S) binds the enzyme (E), the ES complex is formed, making the reaction (formation of products (P)) more favorable. When the reaction is completed, the product dissociates from the enzyme, which is found again in its free form (E) being available to bind to another substrate molecule. Considering a single-step mechanism, the substrate is transformed to a transition state (TS) structure (transient form between substrate and product) and then converted into the product. Some enzymatic reactions are usually more complex, involving a series of TS and intermediates between substrates and products (see Figure 1.2).

The mechanism through which the enzyme increases the reaction rate is system dependent. However, the general idea is that upon the formation of the substrate-enzyme complex, the reaction becomes favorable because the activation energy of the reaction is lowered (see Figures 1.2 and 1.3).

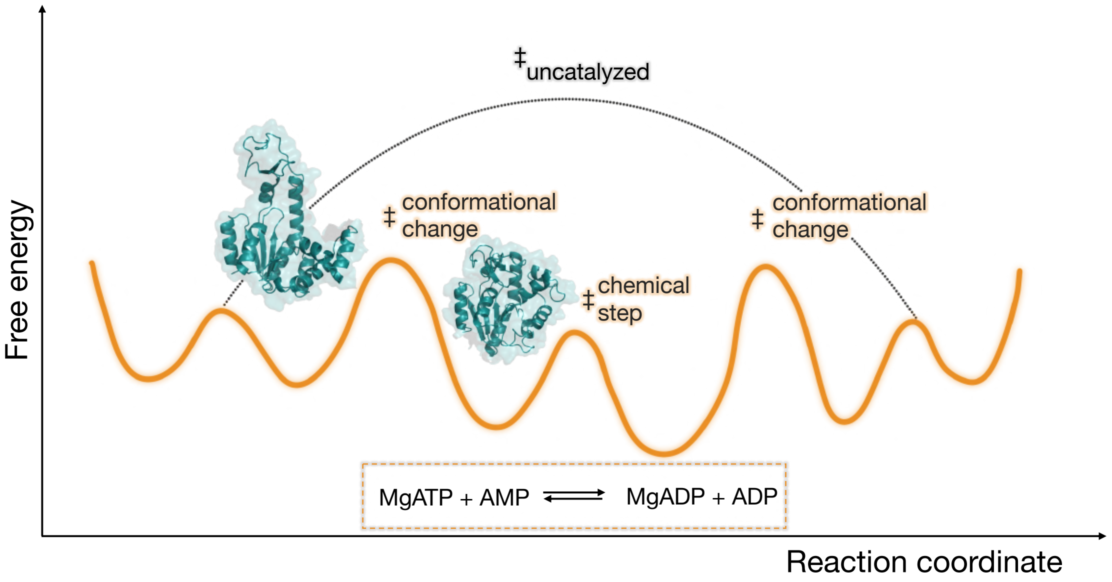


Figure 1.2. Reaction coordinate diagram representing the energy profile of a non-catalyzed reaction (gray) and an enzymatic reaction (orange). The free energy is shown as a function of the course of the chemical reaction for adenosine phosphate transfer. In orange, it is depicted the effect of the enzyme Adenylate Kinase (AdK) on reducing the activation energy in comparison to the

uncatalyzed reaction, shown in grey. The diagram is based on the catalytic energy landscape measured enzyme kinetics extracted from Kerns *et al.*[6]

Enzymes are proposed to lower the activation energy of a reaction by making the formation of the transition state energetically easier. This accelerates the rate at which the reaction occurs but does not alter the relative energy of the reactants and the products. To understand the role of the enzyme on the chemical reaction different strategies can be used including experimental and computational techniques (see Chapter 2. Computational Methods).

1.1.2.1. Transition state stabilization

At the molecular level, catalysis is performed on a specific region of enzyme 3D structure which usually is a pocket or cavity where substrate binds and the reaction takes place, called the active site. As mentioned before, enzymes act as a catalyst for the reaction of interest by decreasing the activation energy of such reactions (see Figure 1.3).[1,2]

The kinetics of the chemical reaction is determined by the energy required to reach the transition state (TS). The transition state is the transient configuration that exists between the reactants and the products where chemical bonds are being formed or cleaved. In case of multi-step reactions, the focus is put on the rate-limiting step.

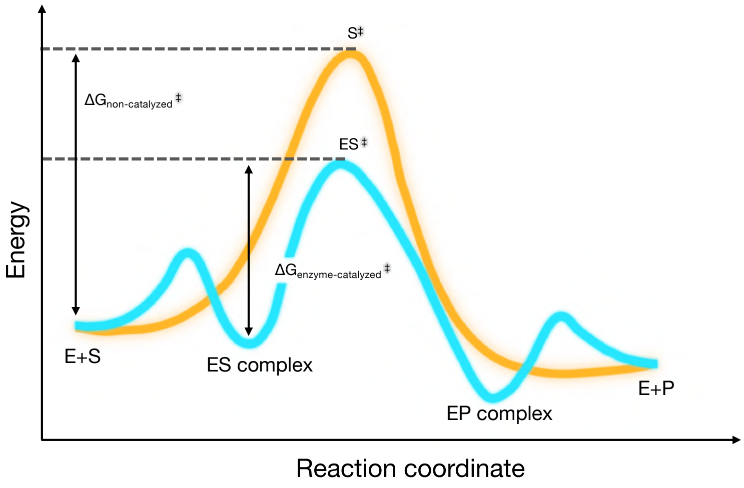


Figure 1.3. Transition state stabilization. The energy of the TS of the reactions represents the highest point of the blue profile. The sum of the favorable binding energy and the unfavorable activation energy leads to a lower activation energy for the reaction catalyzed by the enzyme.

Enzymes specifically bind substrates on their active sites to form enzyme-substrate (ES) complexes. By binding substrates, enzymes stabilize the energy of the transition state, which in turn stimulates the breakage of existing bonds or the formation of new bonds. The fundamental question on how enzymes decrease the activation energy relies on the binding energy and the reaction mechanism of the process.

One of the main characteristics of enzymes is their specificity. The specificity of enzymes refers to their ability to discriminate between different substrates and it arises mainly from two factors: the geometrical complementarity and the multiple non-covalent interactions established between the enzyme active site and its specific substrate.[7,8] The network of non-covalent interactions formed between the enzyme active site and the substrate provide the main source of free energy used by enzymes to decrease the activation energy.

To achieve the efficient decrease of the activation energy, it is proposed that the active site geometry and charge distribution is complementary to the transition state rather than the substrate.[2] The substrate is able to bind to enzyme active sites forming the ES complex assisted by some non-covalent interactions, but the total complementarity of the non-covalent interactions will be formed when the transition state is reached. In this way, the enzyme active site efficiently balances the substrate and TS stabilization by stabilizing substrate regions or conformations that can resemble the TS configuration.

It is important to note that if the enzyme active site is only complementary to the substrate, the ES complex would be highly stabilized, improving the binding of the substrate but increasing the energy barrier to reach the TS, which would have a negative impact on the rate of the chemical reaction. On the other hand, if both the substrate and TS configuration have been equally stabilized, the activation energy would remain the same as in the absence of the enzyme.

Additionally, the binding energy can also induce enzyme structural rearrangements to enhance the catalytic properties. These structural rearrangements can properly orient functional groups and provide additional non-covalent interactions to stabilize the TS. In Sections 1.2 and 1.3, the implications of conformational rearrangements for enzyme catalysis will be discussed.

Current experimental protocols still cannot capture the evolution of the chemical reaction inside enzymes at the molecular level. Using time-resolved *X-ray* crystallography it was possible to identify intermediate states in enzymatic reactions, but TS and reactive intermediates are still hidden for experimental techniques. To reconstruct the full reaction mechanism including TS, computational methods such as Quantum Mechanics (QM) and hybrid Quantum Mechanics/Molecular Mechanics (QM/MM) are commonly used. For example, Casalino *et al.* employed QM/MM to unravel the reaction mechanism of DNA cleavage in the CRISPR-Cas9 system.[9] Force-field based representations of biochemical reactions are used as a promising alternative to QM/MM methods. One example of these methods is Empirical Valence Bond (EVB), which was employed to explore the connection between loop dynamics and the phosphoryl transfer reaction catalyzed by protein tyrosine phosphatases.[10] With these techniques it is possible to calculate the energy profiles of enzymatic reactions and relate the findings with experimental information. More information on QM-based methods to model enzymatic reactions can be found in Chapter 2, Section 2.6.

1.1.2.2. Enzyme kinetics

Enzyme kinetics is the study of the factors that determine the speed of enzyme-catalyzed reactions. In experimental continuous enzyme assays, the rate of an enzyme-catalyzed reaction can be monitored by mixing the substrate with the enzyme and measuring the formation of the product over time.[1,2]

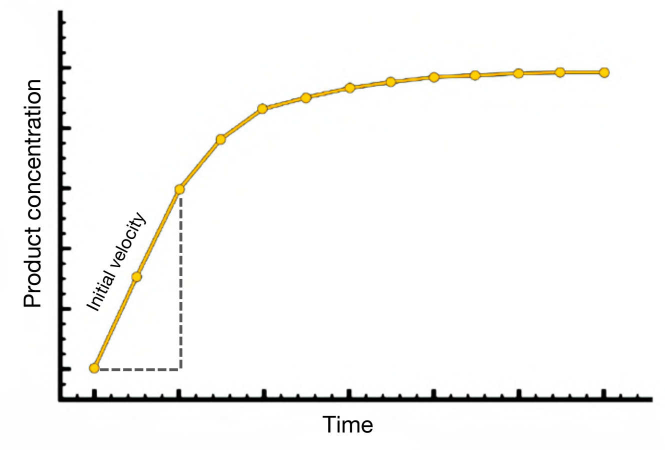


Figure 1.4 Formation of product in an enzyme-catalyzed reaction, plotted against time.

After a certain amount of time, the reaction starts to slow down (see Figure 1.4). This decrease in the speed of the reaction can be caused by the substrate being consumed and, thus, becoming limiting. Another factors can be that the enzyme is becoming unstable along the experiment (denaturation), or that the pH of the mixture is changing over time (some reactions release or consume protons). To avoid these limitations, the reaction rate is measured right after the enzyme has been added. This initial rapid rate is known as the *initial velocity* (v_0). Measuring the reaction rate in the initial stage of the reaction is straightforward, since the rate is effectively linear (see Figure 1.4).

To characterize the kinetic properties of an enzyme, we can perform similar enzyme assays to evaluate how the initial velocity changes upon changes in substrate, enzyme concentration, temperature or pH. For example, a simple linear relationship is observed between the initial velocity and the amount of enzyme (more active sites available to catalyze the reaction).

A more complex relationship emerges when the enzyme assays are performed using a fixed enzyme concentration and different substrate concentrations (see Figure 1.5). Initially, when the substrate concentration increases, the initial velocity increases significantly. Subsequently, as the substrate concentration keeps increasing, the effect on the initial velocity starts to decline, until a *plateau* is reached. At this point, increasing the substrate concentration does not affect the initial velocity.

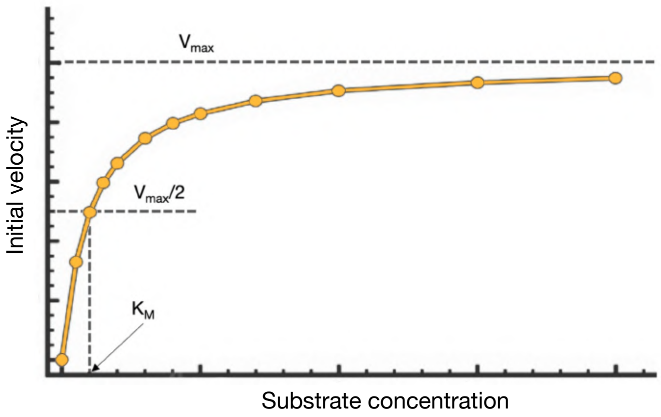


Figure 1.5. Relationship between substrate concentration and the initial velocity of an enzyme-catalyzed reaction.

In these conditions the enzyme is close to saturation with the substrate and operates at its *maximal velocity* (V_{max}). As depicted in Figure 1.5, V_{max} is a theoretical limit that, in practical terms, will not be achieved in any experiment.

The relationship between substrate concentration and the initial velocity shown in Figure 1.5 is represented as a rectangular hyperbola. To define this relation, the Michaelis-Menten equation can be applied:

$$Initial\ velocity\ (v_0) = V_{max} \times \frac{Substrate\ concentration}{Substrate\ concentration + K_M} \qquad \text{(Eq 1. 1)}$$

where K_M is the Michaelis constant, which is corresponds to the substrate concentration that provides half-maximal velocity (see below).

In 1913, Leonor Michaelis and Maud Menten mathematically derived Equation 1.1 from first principles, using simple assumptions regarding how the enzyme reacts with a substrate to form a product. The first assumption is that the reaction takes place through the formation of the ES complex which, at this point, can either productively dissociate to release the product or dissociate in the reverse direction to retrieve again the substrate (no product formed). The enzymatic reaction can, thus, be represented in the following way:

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

Where k_1, k_{-1}, and k_2 are the rate constants of the above-mentioned reaction steps.

Two additional assumptions are required to derive the Michaelis-Menten equation: (i) the initial velocity (v_0) of the reaction is obtained when the concentration of product is negligibly small (*i.e.,* [S]>>[P]), in this way the option of any product converting to substrate can be ignored, and (ii) the substrate concentration exceeds the enzyme concentration (*i.e.,* [S]>>[E]).

Considering these assumptions, the derivation of the Michaelis-Menten equation starts expressing the initial velocity (the rate of formation of the product) as the rate at which the ES complex dissociates to form the product. This depends on the rate constant k_2 and the ES complex concentration as:

$$v_0 = \frac{d[P]}{dt} = k_2\,[ES] \qquad \text{(Eq. 1. 2)}$$

Because ES represents an intermediate state, its concentration is unknown. At his point, an important assumption is that the reaction takes place in a steady-state condition, which means that the concentration of the intermediate involved in the reaction remains constant. This implies that the rate of formation of ES is the same as the rate of dissociation of ES. Thus, the rate of formation and breakdown of the ES complex must balance:

$$Rate\ of\ ES\ complex\ formation = \ k_1[E][S] \qquad \text{(Eq. 1. 3)}$$

and

$$Rate\ of\ ES\ complex\ breakdown = \ (k_{-1} + k_2)[ES] \qquad \text{(Eq. 1. 4)}$$

Hence, considering the steady-state the following important relationship emerges:

$$k_1[E][S] = k_{-1} + k_2[ES] \qquad \text{(Eq. 1. 5)}$$

Which can be rearranged to give the concentrations of ES (*i.e.* [ES]) as follows:

$$[ES] = \frac{k_1[E][S]}{k_{-1}+k_2} \qquad \text{(Eq. 1. 6)}$$

From this equation, the Michaelis constant (K_M) can be defined as:

$$K_M = \frac{k_{-1}+ k_2}{k_1} \qquad \text{(Eq. 1. 7)}$$

So, Equation 1.6 can be simplified as:

$$[ES] = \frac{[E][S]}{K_M} \qquad \text{(Eq. 1. 8)}$$

Since it is assumed that the substrate concentration exceeds the enzyme concentration (*i.e.,* [S]>>[E]), the concentration of free substrate [S] is almost equivalent to the total substrate concentration. On the other hand, the concentration of free enzyme [E] is equal to the total enzyme concentration $[E]_T$ minus the enzyme combined with substrate [ES], *i.e.* $[E]= [E]_T - [ES]$. Introducing these terms to Equation 1.8, rearranging for [ES], we obtain the following relation:

$$[ES] = \frac{[E]_T[S]}{[S]+K_M} \qquad \text{(Eq. 1. 9)}$$

Then, it can then be introduced in Equation 1.2 to give:

$$v_0 = k_2[E]_T\frac{[S]}{[S]+K_M} \qquad \text{(Eq. 1. 10)}$$

Where $k_2[E]_T$ corresponds to the maximal velocity, V_{max}. Using all this information, Michaelis and Menten were able to derive the final equation as follows:

$$v_0 = \frac{V_{max}[S]}{[S]+K_M} \qquad \text{(Eq. 1. 11)}$$

The *Michaelis constant* (K_M) have been experimentally determined for several enzymes and its values are generally in the lower millimolar range. The K_M depends on the enzyme, for example, enzymes from different organisms that catalyze the same reaction can present different K_M values. Moreover, an enzyme that accepts multiple substrates will probably have different K_M values for each substrate.

The value of K_M can tell several important things about a particular enzyme: low K_M values indicate that the enzyme requires only a small amount of substrate to become saturated. This implies that the maximum velocity is reached at relatively low substrate concentrations. On the other hand, high K_M values indicate that high substrate concentrations are required to achieve maximum velocity. Thus, K_M is typically used as a measure of the affinity of the enzyme for its substrate. The higher K_M the lower will be the affinity and *vice versa*.

The constant described as k_2 is also known as k_{cat}, the turnover number of the enzyme. In general terms, k_{cat} is a first-order rate constant that represents the maximum number of substrate molecules that are transformed into product molecules by a single active site in a given period of time, or the number of times that the enzyme "*turns over*" per unit time. To refer to enzyme specificity and catalytic efficiency, the relationship between k_{cat}/K_M is used.

In this thesis, k_{cat}, K_M, and k_{cat}/K_M values are used to discuss the catalytic performance of enzymes imidazole glycerol phosphate synthase and formate dehydrogenase with respect to computational simulation predictions performed in this thesis.

1.1.3. Enzyme dynamics

Enzymes are biomolecules involved in a wide range of biological and chemical processes. Besides its structural features, the ability of enzymes to develop their function – and to evolve towards new functions – in an accurate and specific way is conferred by their flexibility and dynamism.[11] Far from the traditional view of enzymes as a single, rigid, and static structure, they have to be considered as an ensemble of multiple thermally accessible conformations.[12] With conformation we refer to the specific shape or arrangement of the enzyme, similar conformations are grouped into a conformational state. The inherent dynamic properties of enzymes allow the interconversion between different conformational states, which are important for enzyme function. The conformational space that an enzyme can explore can be described in terms of a free energy landscape (FEL), see Figure 1.6. For a theoretical description of how the free energy landscape is reconstructed see Section 2.4.2.

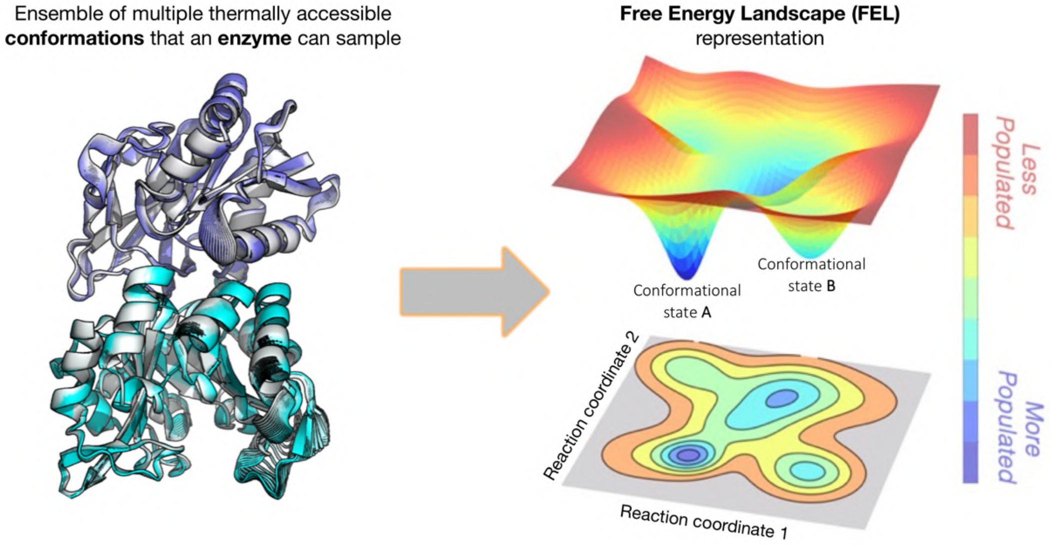

Ensemble of multiple thermally accessible **conformations** that an **enzyme** can sample
Free Energy Landscape (FEL) representation
Conformational state A
Conformational state B
Reaction coordinate 2
Reaction coordinate 1
Less Populated
More Populated

Figure 1.6. Enzyme conformations represented as a free energy landscape (FEL). Enzymes can sample an ensemble of multiple thermally accessible conformations that can be represented in a free energy landscape, where each axis corresponds to a different reaction coordinate. The minima

of the landscape correspond to stable conformational states and the color map represents the population of visited conformations, being blue the most populated and red the least populated.

1.1.3.1. Fundamentals of free energy landscapes

In the FEL representation of an enzyme conformational space, the different conformations sampled are populated based on their energies (thermodynamics) and the rates of interconversion (kinetics) between conformations depend on the height of the energy barriers that separate the different states (the higher the energy barrier, the longer will be the time required to reach one state from the other).[12] Despite the FEL is multidimensional, usually it is represented as a function of one or two representative dimensions (see Sections 2.4.1 and 2.4.2 for a more detailed and technical description of theses aspects).

The distribution and populations of conformational states in a FEL are not static and can be modulated.[13] A particular free energy landscape representation captures the distribution of a certain system in a particular set of conditions (*e.g.,* in a certain temperature, pressure or solvent conditions).[12,13] Therefore, environmental changes (*e.g.,* covalent or non-covalent binding), chemical modifications (*e.g.,* post-translational modifications), introduction of point mutations, molecule-molecule interactions or variations in the above-mentioned set of conditions can change the relative populations of the states (relative energy of each state) and the kinetics of interconversion between them (energy barriers), altering the FEL distribution.[13,14] The concept of population shift is key to understanding the intrinsic flexibility and dynamism of proteins and enzymes. In most cases, when a change or perturbation is introduced in the system, the relative stabilities of conformations change, redistributing the FEL populations.

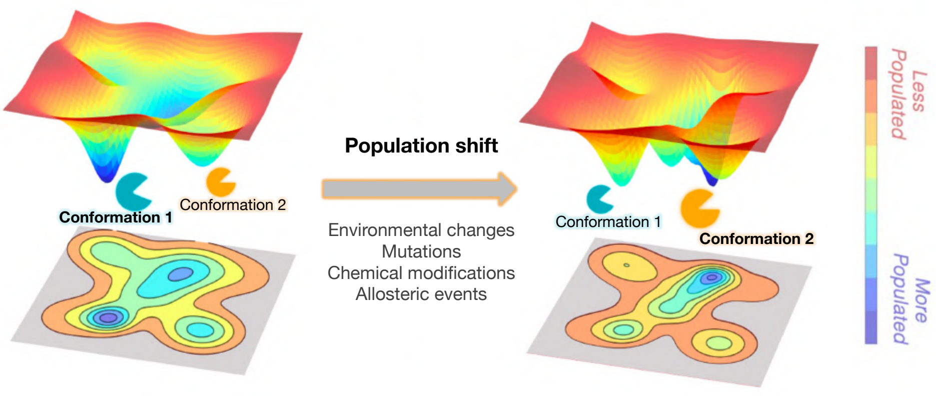


Figure 1.7. Representation of population shift in terms of a free energy landscape. The distribution of populations can change when perturbations (*e.g.,* environmental changes, mutations, chemical modifications, or allosteric events, among others) are introduced in the system.

As depicted in Figure1.7, perturbations and changes on the system conditions can promote population shifts. It can be illustrated with a particular example of a ligand binding event to an enzyme. An unbound conformation (*apo* state) might be unstable, however, the binding of the ligand (ligand-bound state) can stabilize it, resulting in an equilibrium shifting from the *apo* state (unbound protein) to the ligand-bound state, that may display different conformational features.[13,14] Therefore, when the ligand binds, the FEL of the enzymes changes, affecting the thermodynamic properties (relative populations of the conformational states) and also the dynamic properties, including the rates of conformational transitions and the amplitude of motions (different timescales).[15] Characterizing the thermodynamic, kinetic, and conformational features of biomolecules is one of the main challenges of experimental and computational approaches.[13]

1.1.3.2. Experimental and computational approaches to explore enzyme conformational dynamics

Recent advances in experimental and computational techniques have allowed more detailed descriptions of the role of conformational dynamics in enzyme function. To provide a full picture of enzymatic mechanisms, these advances require a tight coupling between experiments and computational work. The integration of both types of techniques can further improve our understanding, with experiments being used to refine simulations and simulations being used to interpret experimental data or *vice versa*.[4] In this section, a brief description of experimental and computational approaches for characterizing the role of conformational dynamics in enzymes and proteins is given.

Crystallographic methods have been crucial to establish the structural view of enzymes. *X-ray* crystallography can be used to determine the high-resolution structure of enzymes.[12] The same enzyme can be crystallized in different conformations, confirming the possibility of enzymes to adopt different shapes. With *X-ray* methods, one can measure the atomic vibrations within a crystal structure, known as B-factors. B-factors can provide information about the flexibility and mobility of a biomolecule. However, B-factors are not a direct measure of molecular dynamism, thus, *X-ray* crystallographic methods still cannot describe the rates of conformational transitions. Nuclear Magnetic Resonance (NMR) techniques are one of the most valuable experimental methods to characterize dynamical and structural changes in biomolecules. The main advantage of NMR relaxation experiments is that timescale transitions can be obtained together with atomistic level resolution. In NMR experiments, the dynamical information is extracted from the relaxation time of nuclei after excitation. Different NMR experiments can provide information of different timescales by assessing different types of nuclei (^{1}H, ^{2}H, ^{13}C and ^{15}N).[12,16] Other experimental methods such as fluorescence spectroscopy, cryo-electron microscopy, or mass spectrometry can also be used to study protein dynamics, but are out of the scope of this thesis. It is important to note that multiple techniques are often used in combination to provide a more complete understanding of protein dynamics. Nonetheless, with these experimental techniques, we can still not obtain both detailed structural description and dynamical changes at the same time.[17]

In the last decades, several computational methods have emerged to model the structure and dynamism of biomolecules. Molecular Dynamics (MD) based methods can be used as a tool for obtaining valuable information of structural changes that occur in solution and for determining the timescales of the transitions between the different conformational substates. MD simulations are a powerful tool to characterize protein conformational dynamics at atomic-level resolution on timescales from picoseconds to microseconds.[18] All-atom molecular dynamics simulation is the standard method used nowadays to study complex dynamics processes that occur in biological systems. In MD simulations, the positions and velocities of every atom in the system evolve according to the laws of classical physics. The forces that act on the atoms are calculated with a physical model, known as force-field (see Chapter 2. Computational Methods, Section 2.2).[4,18,19]

This all atom-based description of the system which evolves as a result of the forces acting on the atoms, generates a trajectory describing its evolution over the course of the simulation as a function of time. Thus, all-atom molecular dynamics simulations can capture conformational dynamism in different timescales and are highly complementary to the above-mentioned experimental techniques.[20,21]

Despite molecular dynamics simulations being an important tool to describe protein dynamics they currently present two challenges: the so-called "sampling problem" and the "force-field problem".[4] The sampling problem refers to the ability of MD simulations to sample all relevant states and events that occur on a biomolecular system enough times to be statistically significant and allow the comparison to experimental observables (*i.e.* the Ergodic hypothesis).[4,18] The force-field problem refers to the required improvement of physical models to provide an accurate description of molecular interactions (see Chapter 2, Section 2.2).

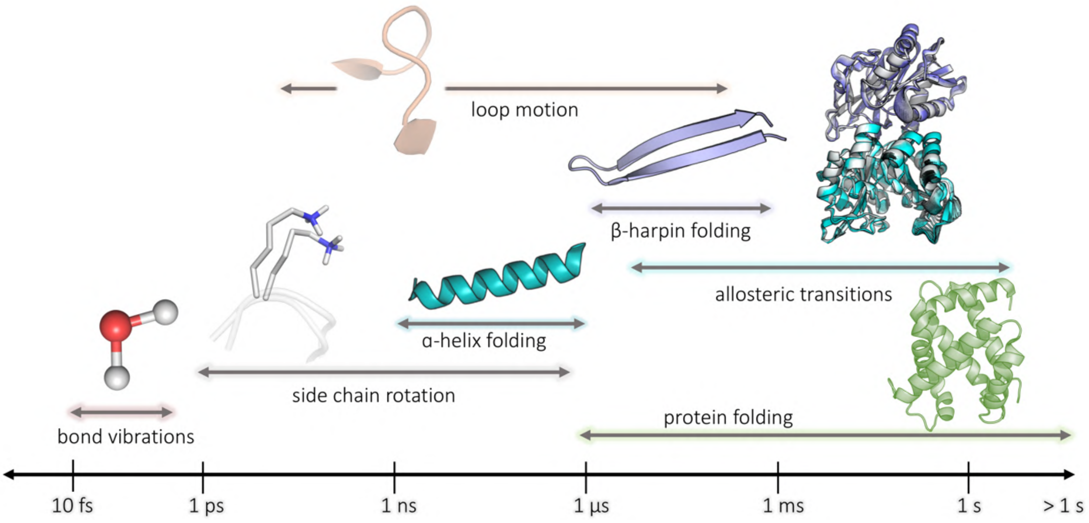


Figure 1.8. Representation of the timescale range for the different motions that enzymes can present.

Providing an accurate description of the system, MD simulations *a priori* offer the possibility of exploring a wide number of biological phenomena, as depicted in Figure 1.8.

Next, we summarize some of the representative motions that enzyme can display with their respective timescales. Bond vibration and side chain rotations can occur in timescales that range from femtoseconds (fs) to picoseconds (ps) and ps to microseconds (μs) respectively. Thus, these events take place on short timescales. On the other hand, loop motions and secondary structure formation can take place in timescales that range from nanoseconds (ns) to milliseconds (ms). Other protein motions such as protein folding or allosteric transitions occur in long timescales that range from microseconds (μs) to seconds (s).[18,21] The study of biomolecular events of interest – that usually take place in the scale of μs to milliseconds – require enough sampling of these events in order to be compared with experimental observables. Hence, to connect MD with experimental data, enough representative conformations should be computationally sampled to satisfy the ergodic hypothesis, confirming that the system has eventually passed through all possible states.[18,22] To achieve the sampling of all relevant events that occur in long timescales, an extensive amount of computational resources would be needed.

When relevant conformational changes occur in millisecond to second timescales, current conventional Molecular Dynamics (cMD) techniques do not properly reproduce with statistical significance the experimental data.[17,19] As introduced above, this is known as the sampling problem. More accurate descriptions of complex systems rely, thus, in alternative approaches to access long timescale events. Different strategies based on hardware and software advances have been developed to access long timescales events and sample rare events. These strategies are generally called *enhanced sampling techniques* and can be classified in unbiased and biased approaches.[18,19] In some cases, the definition of a reaction or a progress coordinate is required to drive or monitor the evolution of the system toward the conformational state of interest.

During the last decades, strategies based on using **unbiased** MD simulations to overcome the sampling problem have been proposed:

- *Multiple processor parallelization using CPUs.* The calculations of the forces that act on the atoms in MD simulations are split in multiple processors. This requires good communication between the processors to recover the force calculations in order to obtain the total energy of the system in each step.[23] This approach appeared when Duan and Kollman developed a parallelization algorithm in 1998 and the first μs biomolecule simulation in explicit solvent was performed with the AMBER software.[24,25]
- *Anton supercomputers*, developed by D. E. Shaw and coworkers.[23] Anton was designed to perform fast and accurate all-atom MD simulations of biological systems. This was achieved by designing special purpose parallel architectures in combination with specific parallelization algorithms. More recently, the third revision of Anton was released (Anton 3), which was optimized to perform long MD simulations (100 μs per day) of large biological systems, such as virus or ribosomes (systems with more than 1 million atoms and up to 50 million).[4]
- *Graphics processing units (GPU).* GPUs have the ability to perform a large number of calculations in parallel on a single GPU.[23] In the last decade, MD simulations of

medium-sized proteins (300-600 residues) are commonly performed in GPUs obtaining a total of 100-200 ns/day. This allows to routinely run microsecond MD simulations.

- *Distributed computing and Markov-State models (MSM)*. In 2000, Shirts and Pande started the project *folding@home*. With this approach, people around the world were able to run short MD simulations in their personal desktop machines. The idea of this project was to achieve massive parallelization of short MD simulations using multiple GPUs.[26] Analyzing all the data collected from hundreds of thousands of independent unbiased MD simulations became a challenge. An approach to analyze large MD data sets is to use a dimensionality reduction protocol (*e.g.,* Principal Component Analysis (PCA) or time-lagged Independent Component Analysis (tICA), see Section 2.4.1 of Chapter 2) in combination with the construction of a Markov State Model (MSM). MSM is a statistical mechanical algorithm to describe the conformational dynamics of the system. Several examples of works where large time scale processes have been studied using these unbiased methods have been recently reported. In 2015, Ingram and coworkers reported the activation mechanism of calcium-dependent protein kinases (CDPKs).[27] The experimental work carried out in this study in combination with extensive unbiased molecular dynamics (MD) simulations, showed the mechanism of inhibition and activation of protein kinases in response to changes in calcium concentrations.
- *Path sampling approaches.* Path sampling approaches, such as milestoning and weighted ensemble (WE), enhance the sampling of rare events by focusing on the transition between states.[28] The idea behind these methods is that since the transitions are infrequent but occur rather fast, they can be explored using ensembles of unbiased simulations along a progress coordinate. From the statistical treatment of these simulations, it is possible to obtain kinetic properties and ensembles of unbiased pathways connecting different states. For example, Sztain et al. used WE to characterize how a glycan gate controls the opening of the SARS-CoV-2 spike protein or Sohraby et al. reviewed its application to characterize ligand binding kinetics.[29,30]

On the other hand, **biased** enhanced sampling approaches are also used to overcome the sampling problem. These techniques are based on changing simulation parameters (*e.g.,* the introduction of artificial biased potentials) and can be classified as constrained and unconstrained methods. The **constrained** methods are based on the definition of a reaction coordinate or collective variable (CV) of the simulated biomolecule.[4,20,31] These methods include Umbrella Sampling and Metadynamics among others (see Figures 1.9 and 1.10). A basic requisite of both methods is that structural information of the system is required. When constrained methods are applied to study transitions that involve conformational changes (*i.e.,* going from an initial state A to a final state B), the method chosen will depend on the structural information available about the system. If there is structural information about the two states, A and B, but we do not have structural information of the intermediate states, metadynamics can be used to explore all possible transitions between the two states along a predefined collective variable (CV) that connect both states (see Sections 2.4.1 and 2.5.1 for a more detailed description). On the other hand, when detailed structural information is available (*i.e.,* both conformational states A and B are known in addition with intermediate states), Umbrella sampling allows the exploration of all desired conformational states defined in the transition path.

- *Umbrella sampling (US).* Umbrella sampling is based on performing MD simulations along a predefined transition path or reaction coordinate (see Figure 1.10). Using this method, each point of the transition is equally sampled, allowing to recover the probability distributions and the free energy along the transition path, thus, providing an estimation of the FEL along the selected coordinates. Detailed structural information of the initial, final, and intermediate conformational states of the desired transition path is required.[18,32]
- *Metadynamics.* Metadynamics is a method that is able to enhance sampling and reconstruct the FEL.[4,31] This method is based on the addition of time-dependent small Gaussian potentials to a selected set of CVs.[31,33] The addition of repulsive potentials discourages the system from visiting previous conformational states, forcing it to explore transitions along the predefined CV (see Figure 1.9 and Figure 1.10).[18,31,33] The selected collective variables should describe the process of interest in order to distinguish between initial, intermediate and final states at the same time that relevant events of the process of interest are sampled. For this reason, having some structural information (*i.e.,* two different conformational states) of the system is required to apply this methodology (see further description of this method in Section 2.5.1.1).

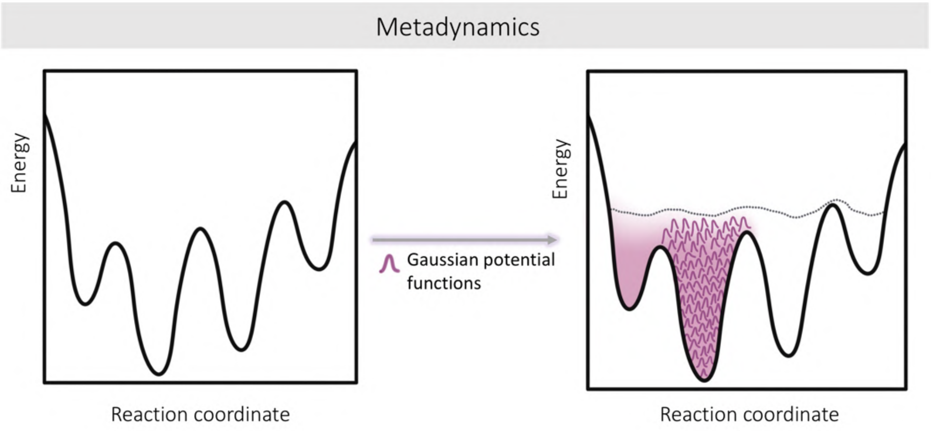


Figure 1.9. Schematic representation of metadynamics strategy used to reconstruct the conformational free energy landscape. In metadynamics simulations Gaussian potential functions are added to force the transitions between different states.

Choosing a set of reaction coordinates or CVs to define the system is preferred when structural information is available. Otherwise, if only one conformational state of the system is at hand or little information of the system is known, **unconstrained** biased methods can be used to explore biomolecular conformations without initial structural knowledge. Unconstrained biased techniques include methods such as Replica Exchange (RE) or Parallel Tempering (PT), accelerated Molecular Dynamics (aMD), and Gaussian accelerated Molecular Dynamics (GaMD).

- *Replica exchange (RE)* or *parallel tempering (PT)*. This method is based on running parallel replicas (*i.e.,* copies) of MD simulations at different temperatures. At certain time intervals, temperature of different replicas is exchanged to enhance sampling.[17,18] In this method, the modified simulation parameter is the temperature at which the system is simulated instead of a chosen CV. The main advantage of this approach is that the increasing of the temperature also increases the kinetic energy, lowering the transition barriers and allowing more sampling events.[4] However, the main limitation is that the number of simulations required to ensure temperature exchange is proportional to the number of atoms, thus, requiring expensive computational resources.[17,18]
- *Accelerated Molecular Dynamics (aMD) and Gaussian aMD.* In aMD and GaMD, a non-negative boost potential is added when the system potential is lower than a reference

energy (see Figure 1.10). By applying this boost potential, the magnitude of the energy barriers between states is decreased, accelerating the transition between conformational states.[25,34,35] aMD and GaMD are efficient approaches to enhance conformational space sampling without previous structural information of the studied system (see further description of these methods in Section 2.5.2.)

Different approaches to study long timescale events can be combined to provide a more reliable description of biomolecular processes. A good example of combining MD based methods was recently reported by Sultan and coworkers, where an initial exploration of the conformational landscape of a protein kinase is performed using aMD. Then, the conformational sampling achieved by aMD is used to run multiple conventional MD simulations to build a Markov State Model.[36]

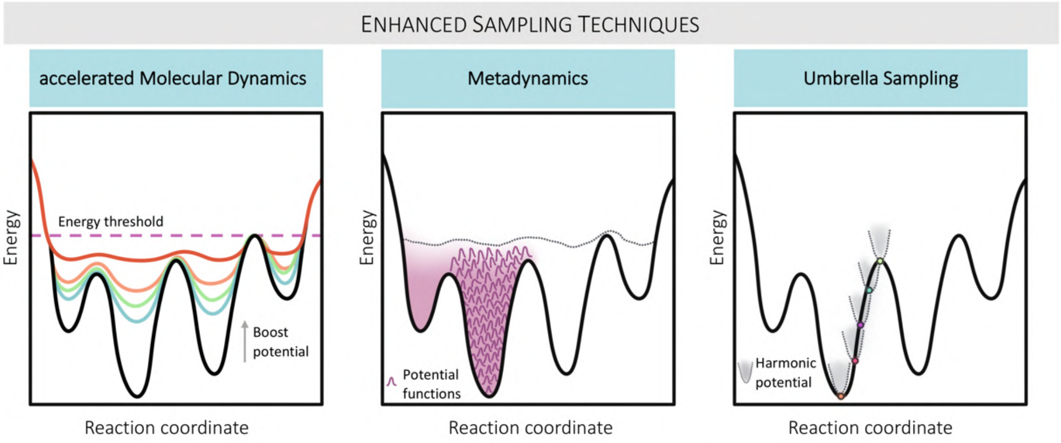


Figure 1.10. Representation of constrained and unconstrained approaches to reconstruct the conformational free energy landscape.

For a more detailed description of computational methods used in this thesis to characterize the dynamic properties and conformational free energy landscape of proteins including conventional MD, aMD, and Metadynamics, see Section 2.5.

1.1.4. Allostery

Allostery is an intrinsic property of proteins. This biological process refers to the mechanism or set of mechanisms that regulate protein function, activity, and conformational ensemble by distal positions within a protein.[37,38] To define the term allostery is not straight-forward. In general, it is described as the process by which the effect of binding at one site is transmitted to another, often distal, functional site, allowing for activity or function regulation.[38,39] However, the concept of allostery is broader and not limited to binding events. Allosteric regulation can be induced by the binding of a molecule (*e.g.,* small molecules, ions, nucleic acids, proteins), chemical modifications (*e.g.,* covalent post-translational modifications), or changes in external conditions (*e.g.,* light absorption).[15,39–41] Exploring the mechanisms of allosteric regulation is important not only to comprehend protein function at molecular level, but also to understand crucial biological processes such as enzyme activity, cell signaling or molecular basis of disease. For this reason, most of the studies on allosteric processes rely on unravelling the allosteric mechanisms in complex biological systems such as G-protein coupled receptors (GPCRs), protein kinases, ion channels or enzymes.[39,42]

As described in Section 1.1.4.4, recent advances in experimental and computational techniques have allowed more detailed descriptions of the role of dynamics in protein function and allostery, providing experimental and computational evidence to further understand allosteric mechanisms.[13] The purpose of the following section is to give an overview about the historical perspective of allostery, the allosteric concept, and its implications and role in enzyme design and catalysis.

1.1.4.1. Allostery: a historical perspective of the proposed models

Over the past century, the concept of allostery has evolved. Many models have been proposed to explain allosteric mechanisms and regulation. In this section, an overview of studies, concepts, and models to characterize allosteric regulation is given, through the lens of history.

In 1904, Christian Bohr described the biological relationship for protein haemoglobin where the presence of one molecule, carbon dioxide, directly affects the binding affinity of another molecule, oxygen.[43] The cooperative behavior driven by the binding of small molecules to distinct protein sites, was named the “Bohr effect”.[43] Another pioneer description of the cooperative effect, before the term allostery was coined, was done by Pauling in 1935, when he proposed a model for the intramolecular control in haemoglobin to explain the cooperativity effect in oxygen binding.[44]

Nowadays, the "Bohr effect" is known as the "allosteric effect". The term *allostery* – whose etymology accounts for the Greek terms "*allo-*" (other) and "*-steric*" (three-dimensional or arrangement of atoms in the space)[3] – was introduced for the first time by Jacques Monod and François Jacob in 1961, to describe the regulatory mechanism by which the enzyme function was inhibited by a *non-steric analogue of the substrate*.[45]

In 1960, Perutz reported the first *X-ray* structure of haemoglobin.[46] The elucidation of the *X-ray* structure revealed that oxygen was bound to spatially separated, distinct sites of haemoglobin.[47] With the aim to rationalize the experimental evidence, two different models were proposed to provide a description of allostery in haemoglobin: the *symmetric* or *concerted* model of Monod, Wyman and Changeux (MWC model) and the *sequential* model of Koshland, Némethy and Filmer (KNF model).[48]

In 1965, Monod, Wyman and Changeux proposed a "plausible model" to describe allosteric transitions.[47] Inspired by haemoglobin structure and the two major conformational states reported by Perutz, MWC model stated that:

I) Structural assumption: allosteric proteins are symmetric oligomers with identical proteomers, creating a cooperative assembly of subunits.[49]

II) Conformational transition assumption: at least two structural states (T for tense and R for relaxed in the case of haemoglobin) exist for each proteomer with different affinities for the ligand.[49] The basic assumption of the model is that interconversion between the two states (T to R) occurs in a concerted manner, leaving no option for the existence of a stable intermediate state (*e.g.,* TR state). In this model, the allosteric mechanism is thermodynamically driven, the binding of the ligand (oxygen) shifts the equilibrium between the R state (major conformation before the allosteric event) towards the T state (major conformation after the allosteric event).[44]

A year later, Koshland, Némethy and Filmer, recovered the Pauling model (1935) and challenged the MWC model by proposing the sequential KNF model.[44] This model stated that the interconversion between the two states T to R occurs at the same time for each subunit.[48] This assumption allows the conception of an existing stable intermediate (TR). In this model, the binding of a ligand follows an induced-fit pattern (induced-fit model was proposed by Koshland in 1959),[50] where the binding of the ligand (oxygen) drives the protein toward a new conformation (from R to T) that is more complementary to its ligand.

The remarkable observations of the models proposed were not only important for further advances in allostery, some of the statements also contributed to the field of molecular recognition, laying the foundations of induced-fit (Koshland, 1959) and conformational selection models that are still used and debated.[47,51] Both MWC and KNF models were focused on the importance of conformational change between two well-defined structural states of haemoglobin: T and R states. Although these models were able to successfully describe allostery, the mechanism of allosteric transition between conformational states was lacking.[39,40] In the last decades, the concept of allostery has evolved to provide a more accurate description that overcomes the two main limitations of the initially proposed models (MWC and KNF models):

1) Considering static structures. Proteins are an ensemble of different conformations; thus, they should be described statistically and not statically.
2) Considering only thermodynamic properties. Kinetic properties are relevant to describe transition pathways between the different conformational states.[15]

This definition of allostery is purely structural, the allosteric effector (*i.e.,* the molecule that binds in a distal position from the functional site) induces a population-shift that involves long-range conformational changes that affect the activity or function in a distal region of the protein (*e.g.,* conformational change from an inactive to an active conformation in the active site pocket). This allosteric event redistributes the conformational ensemble, affecting the pre-existing relative stabilities between active and inactive states and their rate of interconversion, meaning that the allosteric event selectively stabilizes a given conformational state over the others (see Figure 1.11). To refer to this event, the term conformational-based allostery is used and has been used to describe several protein and enzymatic processes such as protein kinases activation.[27] In 1984, Cooper and Dryden introduced the concept of dynamic-based allostery.[52] In contrast with conformational-based allostery, this concept relies on the changes of fluctuations in the protein conformational space that is limited to the basin of the free energy landscape (*i.e.,* absence of population shifts associated with long-range conformational changes). The nature of dynamic allostery relies on entropic effects in distant regions of the protein. The conformational entropy contributions of the allosteric event can alter motions that can range from the micro- to millisecond time scale (*e.g.,* slow backbone motions) to the sub-nanosecond time scale (*e.g.,* fast backbone or side chain motions).

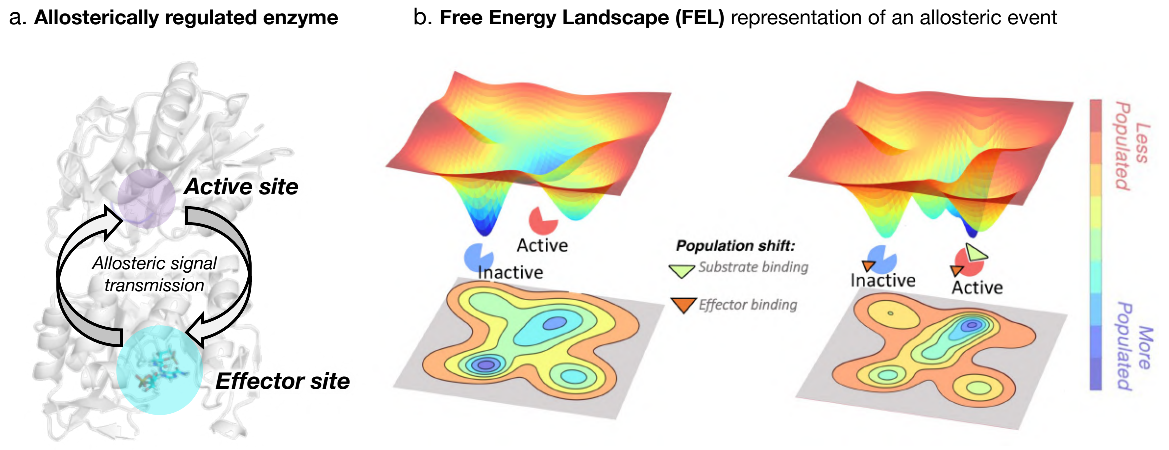

a. Allosterically regulated enzyme
b. Free Energy Landscape (FEL) representation of an allosteric event
Active site
Allosteric signal transmission
Effector site
Active
Inactive
Population shift:
Substrate binding
Effector binding
Inactive
Active
Less Populated
More Populated

Figure 1.11. Representation of a population shift induced by an allosteric event. a) Scheme of an allosterically regulated enzyme, where allosteric signal is transmitted between distal sites. **b)** Population shift of an allosteric event. Allosteric events can redistribute the free energy landscape, favoring the sampling of active conformations.

In a recent study, Taylor and co-workers compared the allosteric dynamic-based propagation with the vibration produced when a violin is played (the "violin" model).[53] The comparison relies on the redistribution of the vibration pattern on the string and the violin body induced by playing a particular note, like when an allosteric effector induces a redistribution of the dynamic fluctuations through the protein. Additionally, it is possible for both proposed allosteric mechanisms (conformational-based and dynamical-based) to act simultaneously.

1.1.4.2. The role of allostery in biological receptors and drug discovery

In the last decades, allostery has emerged as a critical feature to understand important biological processes. Allosteric regulation in biomolecules is a fundamental requirement for cell function, thus, for cell life. Nowadays, understanding allosteric mechanisms is key to fully describe biological systems function.

One of the most exploited applications of allostery is in the field of drug discovery. Identifying molecules that control biomolecular function through allosteric regulation can contribute to developing new drugs.[43] Allosteric drugs are more selective and less toxic than traditional therapeutics. For this reason, methods developed for rational discovery of allosteric drugs have been applied to complex receptors such as GPCRs and ion channels that are common drug targets.[54] Other studies of allosteric regulation are focused on the role of allostery in enzyme catalysis, which will be further explained as it is of relevance for this thesis.

1.1.4.3. Allostery in enzyme catalysis

Allostery in enzyme catalysis can be described as the mechanism by which chemical information is transferred between different sites of the enzyme.[16,55] With chemical information we refer to changes in the networks of molecular interactions that connect the distal sites.

For the sake of clarity, there are shared features in allosteric enzymes that allow to provide a simpler definition of the term:[56]

- The allosteric effector is chemically distinct to the main ligand or substrate.
- The effector causes a change in a functional property of the enzyme.
- The effector binds to a distinct, distal site from the functional site of the enzyme (no overlapping occurs).

Following this definition, the binding of an effector can induce an allosteric event that can play an important role in enzyme mechanisms facilitating, for example, the formation of enzyme-substrate complexes or the catalytic activity.[55]

In enzymology, allostery can regulate activation or deactivation of enzymatic function, therefore, modulating the catalytic activity by modifying the catalytic constants of the reaction (K_M, k_{cat} or both). Depending on the regulatory feature of the allosteric mechanism, allosteric enzymes are described as *K*-type or *V*-type enzymes.[47,56]

- *K-type enzymes:* the binding of an allosteric effector alters the affinity for the substrate, thus altering the K_M constant.
- *V-type enzymes:* those where the maximum activity, k_{cat}, is altered. V-type enzymes are usually inactive, and their activity and catalytic rate is critically dependent on ligand binding activation (both the allosteric effector and the substrate).

To understand how allosteric regulation can affect catalytic activity, the kinetic parameters of the enzyme should be determined. Kinetic parameters in combination with experimental and computational techniques such as *X-ray*, NMR, or MD simulations can provide new molecular insights to characterize the allosteric pathways.[55] One of the best examples of how allosteric regulation affects enzyme catalysis is the case of the enzyme Imidazole Glycerol Phosphate Synthase (IGPS). Chapter 4 is focused on studying the molecular basis of the allosteric mechanism of IGPS and its effect on catalytic activity.

The aim to enhance catalytic activities of enzymes and to perform chemical reactions of industrially relevant interest, relies on designing new enzyme variants. Exploring the characteristics of allosterically regulated enzymes can be a powerful tool to improve their catalytic activity.

1.1.4.4. Experimental and computational approaches to explore allosteric events

There are several techniques that can be used to explore allosteric events, as reviewed in Section 1.1.3 *X-ray* crystallography and NMR can be used to study the dynamics of enzymes and can provide information about the conformational changes that occur during allosteric regulation. In addition, for studying allosteric processes another important experimental technique commonly used is Isothermal Titration Calorimetry (ITC).[57] ITC allows quantifying the free energy of the binding process providing enthalpic and entropic components as global parameters, thus, can be used to study the thermodynamics of allosteric interactions between enzymes and their ligands.

Computational MD-based methods have been used to provide a mechanistic understanding about important allosteric processes that occur in long-timescales. An interesting study of allostery of protein kinases using distributed computing techniques was reported by Malmstrom and coworkers, where long timescale all-atom MD simulations carried out on GPUs were used to build a Markov State Model (MSM).[53] The thermodynamics and kinetics properties obtained from MSM analyses provided the elucidation of allosteric mechanisms of protein kinases A (PKA). Allosteric transition studies can be also performed using constrained biased approaches. Formoso and coworkers recently reported a nice example of studying large conformation motions that regulate Adenylate Kinase (AK) function by means of metadynamics simulations.[58] With the simulation protocol applied in this work, the authors identified the most relevant enzyme states and the thermodynamics and structural properties of AK function transition.

To facilitate the analysis of the vast data generated by MD methods and identify relevant residues for allosteric communication, a valuable option is to convert the complex conformational dynamics into a simple network that represents protein graph motions.[54] In a network representation of proteins, each protein amino acid is treated as a node (*e.g.,* the *Cα* or residue center of mass) and pairs of nodes are connected by edges. To convert MD-based simulation information into this graph representation, the relation between residues has to be measured along the MD trajectories. Different types of methods to analyze these data are based on residue-pair correlation. When correlation methods are used to construct the graphical network representation, edge distance between nodes is inversely proportional to the degree of correlation between two nodes (*i.e.,* large correlation, shorter edge distances).[54,59] Once the graphical network representation has been constructed, different network-analysis techniques can be applied to identify the most important communication pathways of residues between distal sites that may contribute to allostery:

- *Dynamical Network analysis:* The correlation of individual residue-residue contacts along the protein is evaluated through MD based simulation data. To identify the shortest optimal path that connects residues involved in the communication between two allosteric paths, different algorithms can be used (*e.g.,* Floyd Warshall algorithm, Dijkstra's algorithm, etc). Relevant examples of correlation methods developed to describe protein dynamics and to correlate protein dynamics with allosteric pathways have been recently reported. In 2014, VanWart and coworkers developed the Weighted Implementation of Suboptimal Path (WISP) program.[60] With this program, not only the shortest optimal path is computed, but also the suboptimal paths are identified to assess the effect of allosteric signaling through multiple near-optimal pathways besides the shortest one. In 2017, La Sala and coworkers developed an interesting tool that monitors pocket cross-talk to detect potential allosteric mechanisms.[61] This technique is able to detect allosteric pockets and to evaluate the allosteric communication between them, conferring a good approach to discover and characterize unknown allosteric networks in proteins. In 2017, Romero-Rivera and coworkers reported the

Shortest Path Map (SPM) tool.[59] Interestingly, this tool based on correlation methods, initially used to investigate allosteric regulation, was successfully applied to evaluate the effect of distal and active site mutations in a computationally designed retro-aldolase family of enzymes. SPM tool was able to predict the important residues for the enhanced retro-aldolase activity in the laboratory evolved variants and has been applied to other systems that present allosteric regulation such as tryptophan synthase.[62] Therefore, these results indicated that correlation tools are promising methods to improve the computational protocols for design of enzymes.

- *Community analysis:* Further analysis of individual residue contacts can be done by considering the correlation between communities of residues. Communities are defined with highly interconnected residue nodes. Analysis of the communication between these substructures provides an additional insight into the optimal shortest path analysis. The nodes that belong to the same community are more correlated between them and have weaker connections between nodes that are found in other communities.[54,63]

  A great example that combines both correlation methods was reported in 2009, by Sethi and coworkers where both dynamical and community network analysis were used to study the allosteric mechanisms of tRNA protein complexes.[63]

The computational methods shown in Sections 1.3 and 1.4 can be combined to achieve reliable computational descriptions of protein and enzyme dynamism and allostery. Recently, Vu and coworkers have reported the allosteric mechanism of enzymatic catalysis in human cyclophilin.[64] In this study, microsecond molecular dynamics simulations were used to construct a dynamical network analysis of three different variants of human cyclophilin isoforms, identifying two different allosteric pathways that were consistent with NMR experimental data.

1.2. Enzyme engineering and design

Over the last decades, the use of biocatalysts applied to industrial processes has expanded significantly. Enzymes as biocatalysts have been employed in pharmaceutical, food and biofuel industrial processes. These biocatalysts have emerged as a potential alternative with respect to traditional catalysts for many reasons. Enzymes can catalyze specific chemical reactions with great catalytic power and high specificity and stereoselectivity, being able to perform chemical reactions discriminating among different substrates with similar chemical structures and, when acting on *pro*-chiral substrates, precisely yielding a single pure stereoisomer.[65,66] Additionally, enzymes can be used to replace or reduce the use of harmful chemicals or solvents in industrial processes, making them more sustainable and environmentally friendly, as enzymes work in aqueous solutions and mild conditions of temperature, pressure, and pH. This can lead to increased efficiency, reduced costs, and a decrease in the use of harsh chemicals. Enzymes can also be used to create new products or improve existing ones.[67]

One of the main drawbacks of using enzymes for industrial purposes is that there are no natural enzymes that either efficiently accelerate the targeted reactions or that the substrate scope of the reaction does not meet the industrial requirements. Also, enzymes usually exhibit stability problems and low adaptability to non-aqueous solvents.

1.2.1. Experimental enzyme engineering approaches

Enzyme engineering can be used to address the limitations of enzymes in industrial applications.[67] This approach involves various strategies to improve the properties of enzymes, such as activity, stability, and specificity. Some of the earliest strategies involved modifying the reaction conditions (*i.e.,* temperature or pH) together with kinetic studies to optimize wild type enzymes for the production of natural compounds. More recent approaches use rational methods that involve mutagenesis techniques to make specific changes to the DNA sequence that encodes a particular of enzyme. A common mutagenesis technique is site-directed mutagenesis (SDM). This is a rational technique that involves making specific changes to the DNA sequence encoding an enzyme in order to alter its activity or stability by changing the amino acids of certain positions.[67] This is done by introducing mutations at specific positions in the enzyme's gene and the amino acid substitution is based on prior structural or functional knowledge.

More recently, Directed Evolution (DE) techniques, pioneered by Frances Arnold, Pim Stemmer, and Manfred Reetz among others, have greatly accelerated the optimization of biocatalysts, although at a high economic cost. DE is a protocol that introduces Darwin's theory of natural selection to improve the properties of enzymes. This approach mimics the process of natural evolution by generating a large number of variants of an enzyme, and then selecting the variants that improve the desired properties. To select the variants with improved properties, a screening of a large number of enzyme variants needs to be performed. The best hits (selected improved variants) are then subjected to a new DE round and the process is repeated multiple times until an enzyme with the desired property is achieved. Mutations are randomly distributed along the entire protein chain, including active site and distal positions. This random mutagenesis approach has led to the development of more efficient enzymes for industrial and biotechnological applications.

Other successful strategies are semi-rational versions of DE, that combine random methods (*e.g.,* DE) with elements of rational design (*e.g.,* previous or structural knowledge).[68] Two examples of semi-rational approaches are site-saturated mutagenesis (SSM) and iterative saturation mutagenesis (ISM). Site-saturation mutagenesis is a technique where all possible amino acid substitutions are systematically introduced at a specific position of the enzyme's protein sequence. This approach allows the evaluation of the effect of each possible amino acid change at a given position on the enzyme properties. On the other hand, ISM, developed by Manfred Reetz and coworkers, is a variation of SSM, in which multiple rounds of mutagenesis are performed at different sites within an enzyme's protein sequence.[69,70] This approach allows for the systematic exploration of multiple sites within an enzyme to identify the specific amino acid changes that are responsible for improving the enzyme's properties. ISM allows for a more comprehensive exploration of the enzyme's sequence and can lead to the identification of multiple key residues that control for the enzyme's properties. This information can be then used to design improved enzymes with specific properties using rational design approaches.

These laboratory engineering strategies have been used to successfully improve the properties of enzymes, such as increasing their stability, activity, or specificity, which has led to the development of more efficient enzymes for industrial and biotechnological applications.[71]

1.2.1.1. Important aspects for engineering stereoselectivity: the case of P450s

Enzymes possess a remarkable ability to selectively produce a single stereoisomer from *pro*-chiral substrates, a complex feat in biochemistry.[72] The significance of this process, known as stereoselectivity, lies in the fact that while one stereoisomer may possess biological activity, the other may be nor active or toxic. Through the process of evolution, natural enzymes have developed a high degree of stereoselectivity. Achieving high stereoselective enzymes towards one stereoisomer or reversing the stereoselectivity of the reaction is of great importance for chemical synthesis, biosynthesis and biocatalysis. Reversing the stereoselectivity of a natural reaction or introducing stereoselectivity from scratch targeting non-natural substrates requires introducing mutations in the enzyme's sequence to reshape the active site pocket and alter the conformational ensemble, thus promoting the formation of the desired stereoisomer.[73]

Heme-containing P450 oxygenases are versatile biocatalysts for a broad range of transformations (*e.g.,* hydroxylation, dealkylation or C-C bond cleavage). These processes rely on the formation of a high-valent iron(IV)-oxo cation radical species (*i.e.,* Compound I).[74] These enzymes are able to accommodate highly reactive intermediates in active sites that induce reactants to adopt desired conformations while avoiding enzymatic destruction or inactivation.[73] P450 enzymes present significant catalytic promiscuity that can be harnessed to evolve these proteins to catalyze new-to-nature reactions.

A great example of new-to-nature enzymatic transformations for constructing complex molecules are hemoprotein-catalyzed carbene and nitrene transformations. These laboratory-repurposed enzymes reported by Arnold and coworkers exploit the ability of proteins to achieve high activity and stereoselectivity.[73] Metal-carbene and metal-nitrene intermediates are of highly interest in catalysis due to their versatility for building complex molecules via carbene insertion reactions.[75] Synthetic methods used for building these complexes usually require noble metal catalysts and harsh reaction conditions. In addition, limited control over reactive carbene and nitrene species leads to low selectivity in these processes. Biocatalysts that can perform these transformations can be a good alternative, however, no enzymes able to perform these transformations were known in Nature. In 2013, carbene and nitrene transferase activities were reported as side reactions of hemoproteins due to their high promiscuity. Biocatalysts with low activity were identified from an existing library of enzymes. These 'hits' were later improved by DE.[76–78]

1.2.2. Computational enzyme design

Rational approaches for enzyme optimization based on computational methods like structural modelling and thermodynamic calculations to predict and evaluate mutations have also emerged as a powerful strategy and several examples and applications will be reviewed in this thesis.

Initial computational approaches for enzyme engineering towards non-natural reactions were focused on redesigning the active site of existing natural protein scaffolds (*e.g.,* using *Rosetta*[79] or *ORBIT*[80] software). The strategy consisted in selecting a set of active site residues for mutagenesis, while treating the rest of the protein as a rigid body.[80] Despite the initial success, low catalytic activities were achieved with the computationally designed enzymes, needing experimental techniques (*e.g.,* DE) to further achieve the desired enzymatic properties (*e.g.,* improved catalytic activities). Low catalytic activities obtained with computational designs have been attributed to a restricted definition of active site residues, difficulties in properly mimicking the active site arrangement for optimal TS stabilization and focusing mainly on chemical steps without considering crucial dynamic conformational changes for substrate binding and product release.[81]

A more ambitious approach is the *de novo* computational enzyme design, where the active site of an enzyme is generated from scratch within a protein scaffold that does not have any inherent enzymatic activity for the reaction of interest. More recently, also *de novo* protein scaffolds generated using artificial intelligence have been employed to install catalytic active sites.[82] The idea of *de novo* enzyme design provides the opportunity to generate any desired reaction of interest, however, is a more challenging task than redesigning an enzyme with initial low activities. One of the most successful approaches in *de novo* enzyme design is the inside-out protocol, that combines computational design software (*e.g., Rosetta* software) with the *theozyme* concept.[83] In this protocol, it is first performed an initial geometry optimization of the transition state, including truncated enzyme functional groups (*i.e.,* residues involved in binding and catalytic steps) to mimic the minimal active site. This model is then optimized by QM calculations that provide the idealized three-dimensional model, called *theozyme*, *i.e.* theoretical enzyme. This TS geometry is then placed into a protein scaffold. Further computational refinement can be performed (*e.g.,* using *RosettaDesign* software)[84] to optimize the packing between the TS, the functional side chains and the nearby residues.[85] The resulting designed enzymes can be experimentally tested and further improved by means of computational or experimental tools.

These protocols have been used to design *de novo* enzymes based on different natural scaffolds, achieving initial activities for non-natural reactions such as the Kemp elimination,[86,87] retro-aldol,[88] Diels-Alder[89] or ester hydrolysis[90] among others. However, the catalytic activities obtained by *de novo* enzyme design are still orders of magnitude lower than those of natural enzymes,[65,91] thus often requiring the employment of DE techniques to improve the catalytic activities by several orders of magnitude.

Advances on computational techniques have contributed to a deeper understanding on the importance of conformational dynamics and its role on enzymatic function. An interesting example of the importance of conformational dynamics and the concept of population shift in enzyme design for engineering new function was reported for retro-aldolase (RA) enzymes.[88] First, the inside-out protocol was applied to obtain initial RA enzymes that catalyze the cleavage of methodol substrate by a multistep reaction that involves the formation of a Schiff base intermediate involving the catalytic lysine and the substrate. Then, Hilvert and coworkers applied DE on the computationally designed enzymes to improve their catalytic activity towards methodol cleavage.[92] In the second round of DE, the mutations introduced completely remodeled the active site, which provided a better positioning of the Schiff base intermediate for catalysis. After multiple rounds of DE, a highly active RA variant was obtained, showing the power of DE in converting the original computational design into highly proficient enzymes that can reach similar activities than natural enzymes.[93] Interestingly, along the retro-aldolase DE evolved lineage of enzymes, the rate limiting step of the reaction shifted, as observed experimentally[94] and characterized by QM/MM calculations.[95]

In order to elucidate the conformational dynamics of the different RAs variants generated along the evolutionary pathway, Romero-Rivera, Garcia-Borràs and Osuna performed a computational analysis by means of microsecond timescale MD simulation.[59] The analysis of the conformational landscape of the different variants showed that the conformational heterogeneity of the first computational design and the least evolved variants was progressively tuned to stabilize the catalytically active conformational substates due to mutations, which become the major populated states in the most evolved variants. Therefore, characterizing the role of conformational dynamics it is possible to use this information to enhance the catalytic activity of enzymes. Maria-Solano *et al.*, combined the information of conformational dynamics of tryptophan synthase with ancestral sequence reconstruction and SPM to enhance the stand-alone catalytic activity of the wild-type enzyme.[62]

1.2.2.1. Combining different strategies to speed up enzyme design

Nowadays various technologies are available to facilitate the generation of libraries of modified protein catalysts, which can be screened for enhanced activities through DE campaigns. The key for developing an enzyme for a new transformation lies in finding a protein with some level of desired promiscuous activity that can then be optimized. However, screening collections of proteins for often-low levels of activity can be time-consuming and labor-intensive, and designing enzymes for new activities is still highly challenging.[73]

Combining computational and experimental techniques (semi-rational enzyme design approach) has shown to be a great strategy in enzyme design for many challenging transformations. In Chapter 5, we report a biocatalytic platform for creating new carbon-nitrogen bonds in an enantiodivergent and efficient way, generated through DE of a P411 enzyme (P450 with the axial cysteine ligand coordinated to iron is replaced by serine).[96] In this study, extensive molecular dynamics (MD) simulations and density functional theory (DFT) calculations in selected P411 variants were performed to gain insight into the mechanism of this enzymatic transformation. Then, the computational modelling generated was employed to guide engineering efforts to generate an enzyme with the reversed enantioselectivity, but maintaining the high catalytic activity previously obtained. By combining MD simulations and protein engineering, a reversed P411 enzyme variant was obtained in a single round of semi-rational DE.

Another strategy to reduce the time and cost of enzyme engineering consists of expanding the biocatalytic toolbox with metabolic engineering allowing to create specific cells that facilitate the screening process in DE campaigns, see Chapter 6.

Chapter 2. Computational Methods for Studying Enzyme Function

For understanding enzyme function at molecular level, it is essential to decipher the mechanisms of allosteric regulation, catalysis, and inhibition. To properly describe these processes, an atomistic view of the system is required.[17] In the field of structural biology, current methods to study protein structure and dynamics at atomic resolution are diverse. Experimental techniques such as *X-ray* crystallography mostly provide frozen pictures of enzymes. NMR spectroscopy provides information on protein dynamics on different time scales, however the transformation of gathered data to a 3D structure is not trivial. Therefore, experimental methods can provide information on rearrangement of atoms in stable structures, kinetic parameters of the chemical reaction, and binding affinities, but cannot offer a complete description of enzyme conformational dynamics and reaction mechanisms. The transient nature of allosteric processes or enzymatic intermediates make them difficult to be captured in the required atomistic detail, and this remains as one of the main challenges for structural biology. Therefore, computational techniques that provide an atomistic description have been used to explain the molecular basis of biological processes.

2.1. Computational Microscope for Enzyme Function

Enzymes are complex biomolecules consisting of thousands of atoms that display a variety of motions in a wide range of timescales (from fs to s). Chemical reactions catalyzed at the enzyme active site involve bond-breaking and bond-forming events driven by the electrostatic environment of the enzyme.[12] The interaction between an enzyme and its substrate or a conformational change of the enzyme can imply rearrangements of atomic charges and changes in protonation states. These enzymatic processes involve changes at the electronic structure level. However, modeling the static and dynamic properties of the whole enzyme in its environment (*e.g.,* solvent, ions, other molecules) with accurate quantum mechanics methods is currently unattainable.[17] Therefore, computational strategies with different levels of resolution are used to study the different aspects of enzyme function. For example, to describe the dynamic behavior of enzymes, Molecular Dynamics (MD) simulations based on a molecular mechanics description of the system can provide relevant information about the conformational dynamics at atomic resolution. To study reaction mechanisms of enzymatic reactions, quantum mechanical calculations based on Density Functional Theory (DFT) are commonly used to reconstruct the energy profile. Understanding the advantages and limitations of the different computational techniques is key for their correct applicability.[97]

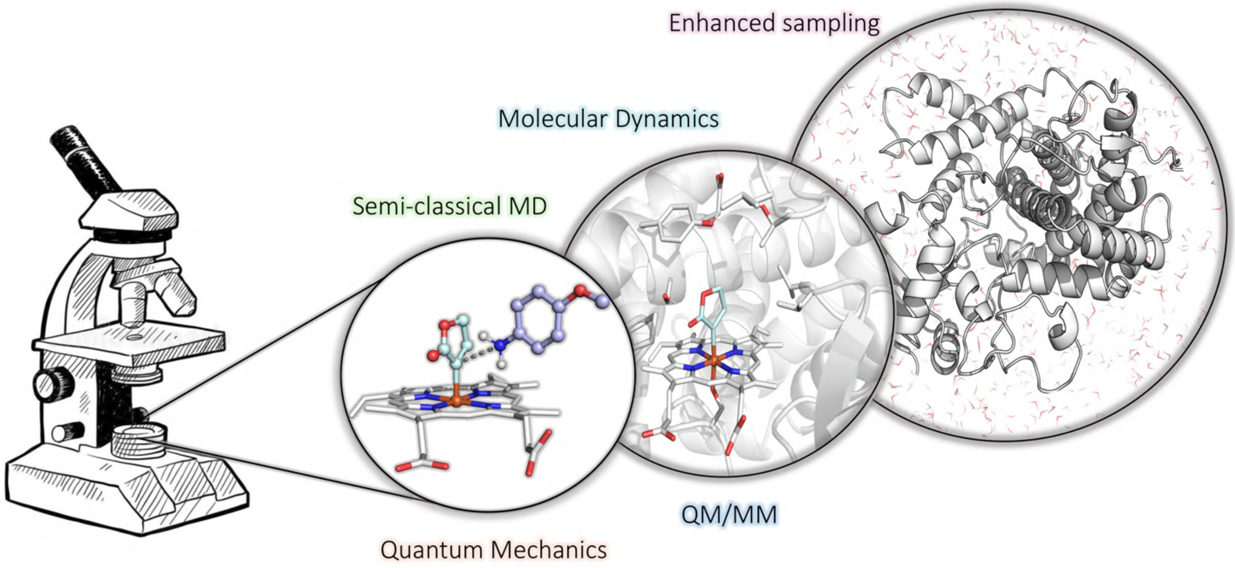


Figure 2.1. Computational Microscope. Representation of different layers of accuracy to study enzyme reactivity and conformational dynamics by means of computational techniques.

When Quantum Mechanics (QM) is used, electrons are explicitly represented in the model and the energy of the system is calculated by describing the electronic structure of molecules in the system with no (or few) empirical parameters. Therefore, with QM methods it is possible to describe bond rearrangements and describe the mechanism of enzymatic reactions. However, QM calculations are currently limited to a few hundreds of atoms due to computational requirements. This implies that only a portion of the enzyme region of interest can be described at the QM level. In addition, due to the high computational cost associated with each energy calculation, using QM to study global enzyme dynamics is still limited. Thus, exploring enzyme dynamics relies on reducing the level of resolution of QM methods from the electronic to the atomic structure.[17,97]

A strategy to reduce the computational cost of QM calculations is to use a molecular mechanics (MM) description of the system. In MM, molecules are composed of particles representing atoms (atomistic models) or groups of atoms (coarse-grained models), where each atom (or group of atoms) has assigned a fixed partial charge. This simplification allows us to define simple potential energy functions (force fields) based on classical physical models with a large number of empirical parameters (fitted to experimental, QM or other data) that can be used to calculate the energy of the system. This continuous energy function approximates the quantum mechanical description and allows one to perform tens of thousands of calculations in the time required for a single quantum mechanical evaluation of structure and energy. Therefore, MM methods reduce resolution and computational cost with respect to QM. However, in MM, bond-breaking and forming is generally not allowed (except for reactive force fields), so the topology or chemistry of the system remains constant as a function of time. This implies several approximations with respect to real conditions, like using fixed charges (unless polarizable force-fields are used) and protonation states for all components of the system. Therefore, MM methods allow us to study equilibrium geometries and relative energies between conformers of the same molecule. Using MM it is also possible to describe how different molecules interact through non-covalent interactions, which is of interest to understand for example the interaction between enzymes and substrates.[97,98]

Molecular mechanics force fields can be coupled with molecular dynamics (MD) algorithms to describe the time-dependent conformational dynamics of biomolecules. By coupling simple potential energy functions with Newton's laws of motion it is possible to simulate molecular evolution over time. MD simulations are widely used to study the time evolution of complex biological systems in its environment. However, to obtain accurate results, MD simulations should advance using short time steps. This limits the time scales of conventional MD simulations to a few microseconds, which sometimes are not enough to capture the motions of interest and to obtain meaningful results. To overcome such limitations, enhanced sampling techniques that provide the means to accelerate MD simulations can be used. Moreover, a huge amount of data is generated in each MD trajectory that needs to be carefully analyzed to extract the relevant information.[98]

These methods can be combined to elucidate the different aspects of enzyme function in detail. The computational multiscale view of enzymatic catalysis and function allows understanding the molecular basis of both conformational dynamics and mechanistic features, which can help in the process of rational enzyme design.

In the next sections, the theoretical basis and practical guidelines of the methods used in this thesis to study enzyme function of different systems will be briefly described.

2.2. Molecular Mechanics: Classical Physics Force Fields

Force fields (FF) based on molecular mechanics use simple potential energy functions to estimate the energy of a system (*e.g.,* enzyme surrounded by solvent) as a function of its conformation (coordinates, $\mathbf{r}_N$, where N is the number of particles) at low computational cost.[97,98] Physics-based FFs approximate the interactions between particles with simple classical physics expressions that all together account for the energy and forces of the system. In all-atom FFs the complexity of the systems is further reduced by considering each atom as a sphere with a fixed atomic charge linked to neighboring atoms through bonds describes as springs. Most common biomolecular force fields used in MD simulations are all-atom physics-based force fields with fixed atomic charges. Knowledge-based force fields and coarse-grained force fields are also widely used to model biological systems, however, these force fields are out of the scope of the present thesis and will not be discussed.

Physics-based force fields used for characterizing molecular systems are constructed by adding different components to describe the intra- and intermolecular forces within the simulated system. Different classes of FF exist depending on the level of complexity of the functional form. Here, we are going to describe the general terms of Class I FF, which are the ones used in the present thesis. Class I FF present simple functional forms that typically include harmonic potentials to model stretching and bending and the Coulomb term and Lennard-Jones potential to model electrostatic and van der Waals interactions. The potential energy ($V(\mathbf{r}_N)$) of a Class I force field is given by the following general potential energy function:

$$V(\boldsymbol{r}_N) = V_{bonded} + V_{non-bonded} \qquad \text{(Eq. 2.1)}$$

Which can be further split into the following components:

$$V(\boldsymbol{r}_N) = V_{str} + V_{bend} + V_{torsion} + V_{el} + V_{vdW} \qquad \text{(Eq. 2.2)}$$

The first general term of the force field (V_{bonded} in Equation 2.1) is used to model intramolecular interactions between bonded atoms (Equation 2.2). This includes bond stretching (V_{str}), angle bending (V_{bend}), and torsional ($V_{torsion}$) potentials of connected atoms. The second general term is used to describe non-bonded interactions between any pair of atoms (both intra- and intermolecular). This includes the electrostatic interaction between charged atoms and the van der Waals (vdW) interaction between pairs of atoms. Although more refined FFs contain additional terms, these components are generally maintained in most biomolecular force fields (see Figure 2.2).[98]

$$V_{total} = V_{str} + V_{bend} + V_{torsion} + V_{vdW} + V_{el}$$

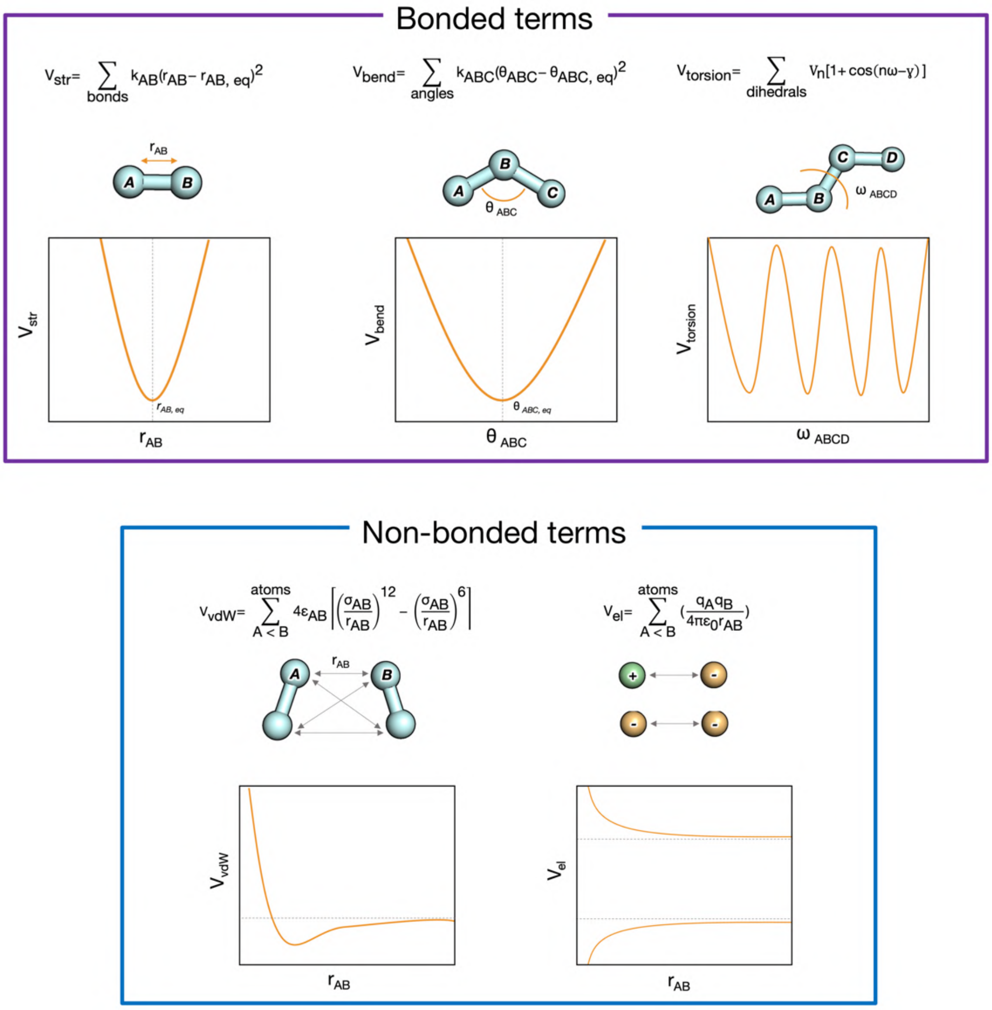

Bonded terms
$V_{str}= \sum_{bonds} k_{AB}(r_{AB}- r_{AB,\ eq})^2$
$V_{bend}= \sum_{angles} k_{ABC}(\theta_{ABC}- \theta_{ABC,\ eq})^2$
$V_{torsion}= \sum_{dihedrals} V_n[1+\cos(n\omega-\gamma)]$
r_{AB}
A
B
θ_{ABC}
C
D
ω_{ABCD}
V_{str}
$r_{AB,\ eq}$
r_{AB}
V_{bend}
$\theta_{ABC,\ eq}$
θ_{ABC}
$V_{torsion}$
ω_{ABCD}
Non-bonded terms
$V_{vdW}= \sum_{A<B}^{atoms} 4\varepsilon_{AB}\left[\left(\frac{\sigma_{AB}}{r_{AB}}\right)^{12} - \left(\frac{\sigma_{AB}}{r_{AB}}\right)^{6}\right]$
$V_{el}= \sum_{A<B}^{atoms} \left(\frac{q_A q_B}{4\pi\varepsilon_0 r_{AB}}\right)$
+
-
V_{vdW}
r_{AB}
V_{el}
r_{AB}

Figure 2.2. Force Field terms. Mathematical and graphical representation of force-field terms described in Section 2.2.1.

To calculate the energy of a molecular configuration using equation. 2.2, a set of parameters are required. Force field parameters for all types of atoms of the system are required *a priori*. These parameters are generally derived from experiments or QM data and differ among force fields. To select the required parameters for a potential energy calculation it is necessary to first assign an atom type to each atom of the system. The atom type contains information about the hybridization and connectivity of each atom. For example, sp^3 carbon atoms adopt tetrahedral geometry, while sp^2 carbon atoms are trigonal and sp carbons, linear. Additionally, bonds between different types of carbons will present different equilibrium distances. Using a limited number of atom types (for example using one atom type for all carbon atoms) can lead to inaccurate results. On the other hand, the larger the number of atom types defined, the larger is the number of parameters required. A compromise between the number of atom types, parameters, the simplicity of the functional form and accuracy should be found to define a force field. Therefore, it is the combination of the potential energy function with the set of atom types and parameters that compose a force field.[97]

2.2.1. Force Field Terms

Hereafter, the different force field terms are described in more detail.

- Bond stretching: The bond stretching term models the interaction between all pairs of bonded atoms (A-B). In biomolecular force fields, the functional form of the bonding energy is generally defined by a simple harmonic potential (Hooke's law) where the energy increases as the bond length deviates from a reference value ($r_{AB,eq}$). The degree of increase of energy along distance depends on the value of the force constant (k_{AB}) assigned to the bond.

Each force field has defined parameters for reference bond lengths and force constants for selected bonds (bonds expected to be stronger have larger force constants, for example there is a contrast between C-C, C=C, or C≡C). To avoid confusion with the term "equilibrium", we consider that the reference bond length is the value that the bond adopts when all other terms in the force field are set to zero, while the equilibrium bond length is the value that is adopted in a minimum energy structure where all the other terms of the force field also contribute. Therefore, in a molecular system a particular bond may deviate slightly to the reference value to compensate for other contributions to the energy. Since harmonic potentials are used in Class I FF, bonds cannot be formed nor broken. However, other expressions such as Morse potentials can be used to properly model the dissociation of two atoms within a MM framework.

- *Angle Bending:* The angle bending term is defined as the summation over all valence angles in the molecule using a harmonic potential as in the bond stretching (see Figure 2.2). A valence angle is the angle formed between three atoms, A-B-C, in which both A and C are bonded to B. As in bond stretching, the contribution of each angle to the potential energy is characterized by a force constant (k_{ABC}) and the deviation from a reference angle value ($\theta_{ABC,eq}$).

In both bond stretching and angle bending, substantial energies are required to cause significant deformations from the reference value. At room temperature, most of the variation in biomolecular structure and relative energies of the conformational landscape is due to torsional and non-bonded contributions.

- *Torsional terms:* The torsional term models energy variations associated with bond rotations. Most force fields use explicit torsional potentials with a contribution from all the dihedral angles formed between each bonded quartet of atoms A-B-C-D in the system. Properly describing rotational barriers of chemical bonds is fundamental to understand the structural properties of molecules and its conformational analysis. In particular, the torsion parameters of amino acid residues play a key role for the success of biomolecular FFs. Torsional potentials are generally expressed as cosine series expansions that depend on barrier height associated with the rotation (V_n) and a phase factor (γ) that indicates where the dihedral angle passes through its minimum energy value.

Additionally, improper torsion terms and out-of-plane bending motions can be also incorporated to force fields to properly reproduce certain geometrical aspects that are difficult to describe with just torsional terms.

Non-bonded interactions: The nonbonded term is calculated between all pairs of atoms that are separated by at least three bonds (intramolecular) and between all pairs of atoms of different molecules (intermolecular, including solvent molecules). Molecules interact through non-bonded forces that play an important role in determining the structure of individual molecular species. Thus, non-bonded parameters are usually adjusted to have a good balance between solute-solvent and solvent-solvent interactions. Considering that the non-bonded interactions vanish with distance, a distance cutoff is usually applied to prevent the calculation of all possible non-bonding interactions, which significantly reduces its computational cost. The non-bonded terms in physics-based force fields are usually considered in two groups, the electrostatic interactions, and the van der Waals interactions.

- Electrostatic term: Corresponds to the energy associated with the electrostatic interaction between point charges which is commonly modeled with a Coulomb potential (see Figure 2.2). To calculate the electrostatic contribution of a given pair of atoms the partial atomic charge (q_A and q_B) of each atom and the interatomic distance (r_{AB}) are required. Biomolecular force fields contain libraries with all the atomic charges for the standard amino acids and nucleobases, however, for non-standard molecules (*i.e.,* substrate or cofactor) the user should provide the atomic charges prior carrying out the calculation (see Section 2.2.3.1). The interatomic distance is obtained from the molecular structure. As mentioned before, these charges remain fixed during MD simulations, which is one of the main limitations of Class I force fields.
- *Van der Waals (vdW) term*: the interaction energy between atoms varies with distance irrespective of whether they are charged or not. The tendency is to have a balance between attractive (long range) and repulsive forces (short distances). Attractive forces are due to dispersion forces (instantaneous dipoles that arise during fluctuations in the electron clouds: an instantaneous dipole in a molecule can induce dipole in neighboring atoms). Repulsive forces, often referred as exchange forces or overlap forces, are due to interactions between electrons with the same spin (short range repulsive forces). The effect of exchange is to reduce the electrostatic repulsion between pairs of electrons that cannot occupy the same region of space. Dispersive and exchange-repulsive interactions can be modeled together using a simple empirical expression that can be rapidly calculated (as a consequence of the large number of vdW interactions that must be determined). The common vdW potential function used in biomolecular force fields is the Lennard-Jones 12-6 function controlled by two adjustable parameters, the distance at which attractive and repulsive forces balance (σ_{AB}) and the well-depth (ε_{AB}). In biomolecular FF, the σ_{AB} and ε_{AB} terms are generally obtained from the arithmetic and geometric mean of individual atom type parameters respectively.

General considerations: A key aspect for the efficiency of classical force fields is to select expressions for the different terms that are easy to calculate with a computer, as the ones shown in Figure 2.2. When performing an energy minimization or a MD simulation it is required to compute the first and second derivatives of the energy with respect to the atomic coordinates. Therefore, it is important to select FF expressions that can be easily differentiable at a computational level. As described in the introduction, when coupled to MD simulations, classical force fields are adequate for representing biomolecular systems at room temperature.

Biomolecular Force Fields: Several biomolecular force fields exist that can be applied to model proteins, DNA, and RNA. Among them, CHARMM[99], AMBER[100], and GROMOS[101] FFs are found among the most widely used to simulate biomolecules. The FF choice depends on the system studied (*e.g.,* protein in solution, protein embedded in a membrane). It is always recommended to perform preliminary simulations and analyze the stability of the structural elements of the protein. In this thesis, we have used AMBER ff14SB[102] to perform all simulations.

2.2.2. Water models

Biological and most chemical reactions take place in the presence of solvent. To appropriately describe biochemical processes, MD simulations are commonly carried out in the presence of explicit solvent molecules (*e.g.,* water). Thus, it is important to have a balanced description in terms of accuracy and computational efficiency of solute-solvent and solvent-solvent interactions in force field calculations[98]. Several water models have been suggested that account for different levels of complexity.

A common strategy to simulate explicit water in biomolecular simulations is using simple rigid water models such as TIP3P.[103] This is a three-point charge model that has been developed to reduce the degrees of freedom while accounting for water contribution to the system's energy (and to describe liquid phase properties). Therefore, TIP3P is a simple interaction-site model where each water molecule is maintained in a rigid geometry. The TIP3P model uses three sites for the electrostatic interactions where the partial positive charges on the hydrogen atoms are balanced by a negative charge located on the oxygen atom (see Figure 2.3). The van der Waals interactions between two water molecules are computed using a Lenard-Jones potential only applied to the oxygen atom, thus, no van der Waals interactions involving the hydrogen atoms are calculated. This reduces the computational cost when working with systems solvated with a large number of water molecules.

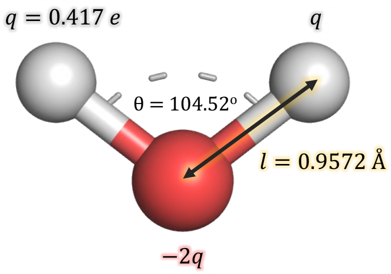


Figure 2.3. TIP3P water model parameters. The TIP3P water model assumes that each water molecule is composed of three-point charges: a negative charge on the oxygen atom and two positive charges on the hydrogen atoms.

Other models have been developed to better describe the solvent effects. For example, flexible models have been developed to permit internal changes in the molecule conformation and polarizable water models can explicitly include polarization and many-body effects, allowing for a more accurate representation of water behavior.[98] To significantly reduce the computational cost, implicit models of water can also be used, however, within this framework the role of explicit water molecules in the active site of enzymes cannot be described. In the present thesis we have used the TIP3P model, which provided accurate results for the studied systems.

2.2.3. Parametrization of molecules not included in biomolecular force fields

To study biochemical processes such as enzymatic catalysis, we require the simulation of biomolecules interacting with small molecules or ligands (*e.g.,* cofactors, substrates, inhibitors…). Biomolecular force fields are generally limited to only describe standard residues in proteins, DNA and RNA, and solvent. To include non-standard molecules in the simulation, additional parameters of the molecule of interest (reference bond, angles, dihedrals, and non-bonded (including atomic charges)) should be provided.

2.2.3.1. Ligand parameterization

To obtain the parameters for small organic molecules or ligands, generic force fields as the Generalized AMBER Force Field (GAFF) have been developed.[104] GAFF contains a set of parameters for most common atom types found in organic molecules. In contrast to biomolecular force fields, the atomic charges are not present in GAFF and need to be provided by the user. Therefore, to obtain the GAFF parameters for the ligand of interest the molecular structure and the atomic charges of the molecule are required. The general protocol used in this thesis to parametrize small organic molecules consists of the following steps. First, a QM optimization of the ligand is carried out. In general, we perform a DFT optimization using B3LYP with a small basis set (6-31G*). Second, a single point energy calculation of the previously optimized geometry is carried out to estimate the atom charges with the r*estrained electrostatic potential* (RESP)[105] method, which generates atom-centered point charges based on a charge fitting procedure to an electrostatic potential generated at the HF/6-31G* level of theory. Since the atomic charges of aminoacidic residues in the AMBER biomolecular force fields were generated at the HF/6-31G* level, this level of theory is commonly maintained for consistency. Finally, the generated RESP charges and the atom connectivity retrieved from the optimized molecular structure are used to assign atom types available in the GAFF force field. In this thesis, we have used GAFF to parametrize all the substrates and organic cofactors used in MD simulations. It is recommended to manually check the assignment of atomic charges, atom types, and connectivity generated by automatized protocols.

However, GAFF or other generic FFs applicable to biological systems do not include parameters to describe transition metal centers or metal-based cofactors present in metalloenzymes.

2.2.3.2. Metal Ion and Metal Complexes Parametrization Strategies

Metal ions are commonly found in biological systems either forming part of the enzyme structure, *i.e.*, metalloenzymes, or as part of the water solution, *e.g.*, sodium or chloride ions. Modeling metal ions with molecular mechanics presents several challenges due to the variety of oxidation and spin states, complex chemical bonding, and multiple coordination numbers.[106] There are basically three strategies to account for metals in classical molecular dynamics simulations depending on their structural environment and coordination mode: 1) the non-bonded model, 2) the cationic dummy model, and 3) the bonded model. Each of them requires defining appropriate atom types and force field parameters.

Within the framework of the **non-bonded model**, each metal ion is commonly described as a ball with a fixed integer charge, which generally corresponds to the oxidation state (*e.g.,* +1 for Na^+), and the corresponding van der Waals parameters (*e.g.,* $\boldsymbol{\sigma}$ and $\boldsymbol{\varepsilon}$ in Figure 2.2). The latter are parameterized to reproduce certain properties such as metal-ligand distances or hydration free-energies. In some cases, non-bonded models fail to properly reproduce the coordination sphere of transition metals along the MD simulations. To improve their performance modifications that incorporate additional terms in the Lennard Jones potential have been suggested.[107] In this thesis, the non-bonded model is used to model sodium cations and chloride anions required to neutralize the system for MD simulations.

In the **cationic dummy model**, a number of dummy atoms are distributed around the metal center to reproduce the metal coordination sphere.[108] Commonly an octahedral distribution of dummy atoms is used for most alkaline-earth and transition metals, although tetrahedral or other rearrangements have been employed to reproduce specific coordinations. Each dummy atom accumulates a certain partial positive charge and non-bonded vdW parameters are assigned to both central and dummy atoms. This model allows certain flexibility in terms of the coordination number, which can change along the MD simulation to adapt to the surrounding residues. As demonstrated by Duarte *et al.,* the use of the dummy model can be useful to describe the behavior of bound metal ions in metalloenzymes.[109]

In the **bonded model**, the metal ion is considered to be covalently bound to the surrounding ligands, which can be amino acids, water molecules, substrates, or organic ligands as part of a metal complex cofactor. This implies that both the bonded and non-bonded force field parameters between the metal and the directly linked residues have to be obtained. To facilitate the parametrization process, Li and Merz developed the Metal Center Parameter Builder program (MCPB.py) that allows for the parametrization of metal ions within the bonded model framework.[106] The general process consists of several steps that are summarized as follows. First, the bonded force field parameters are obtained from quantum mechanical calculations, usually at the DFT level. In particular, equilibrium bonds and angles are obtained from a geometry optimization while force constants are retrieved using the Seminario method from the cartesian Hessian matrix calculated at the QM level of theory. These calculations are performed on a simplified truncated model that contains the metal ion and the surrounding atoms. Second, partial atomic charges of the metal center and surrounding residues are calculated using quantum mechanical calculations (commonly DFT calculations). In general, RESP atomic charges are computed. Finally, the atom types and non-bonded parameters of the region of interest are assigned based on the geometric (molecular structure) and electrostatic (atomic charges) criteria. In this thesis, we have used the bonded model to perform MD simulations of a number of laboratory evolution haem-iron carbene transfer enzymes (see Chapter 5).

2.3. Molecular Dynamics Simulations

Molecular dynamics simulations are based on generating successive configurations (**r_N**) of the molecular system that evolve along time by integrating Newton's equations of motion. The result is a trajectory of the biomolecule that specifies how the positions and velocities of the atoms in the system vary with respect to time.[97,98]

The trajectory is obtained by solving the differential equations described by Newton's second law:

$$F_i = ma_i \qquad \text{(Eq. 2.3)}$$

$$m\frac{\partial r_{i2}}{\partial t^2} = \frac{\partial V(r_N)}{\partial r_i} \qquad \text{(Eq. 2.4)}$$

These equations describe the motion of a particle of mass *m* along a coordinate (*i*) with *Fi*, being the force exerted on the particle in that direction. The force is the derivative of the potential energy with respect to the position of the atoms.

In MD simulations particles follow Newton laws of motion: (i) a particle will continue to move at its current speed and direction unless a force acts upon it and (ii) the force equals the rate of change of momentum (*i.e.,* acceleration). In a molecular assembly, formed by many particles, the force that acts on each particle depends on its position with respect to the other particles, thus, a many-body problem that cannot be solved analytically, arises.

The total force that acts on a particle at a given time is the sum of its interactions with other particles. The forces are obtained from continuous potential models (force fields, Section 2.2), that are assumed to be pairwise additive. To solve the many-body problem, the equations of motion have to be integrated using numerical finite difference methods. The essential idea is breaking the integration of Equation 2.4 into many small steps, each separated by a fixed time (δt), assuming that the speed and acceleration are constant over these time step intervals.

Next, the parameters important for MD simulations will be described:

- *Time Step:* Selecting the time step is a key parameter for MD simulations. If the time step is too large it might result in instability due to high energy overlap between particles. If the time step is too small, it might lead to trajectories covering a limited phase space while increasing the computational cost of the MD simulation.
  Usually, the time step set for MD simulation is determined by the fastest frequency of motion. The highest-frequency vibrations are due to bond stretches, especially the O-H vibration (*ca.* 10 fs). Thus, accurately treating water molecules using a classical description would require solving the equations of motion using a small timestep (1 fs), which would dramatically increase the computational cost of MD simulations. To overcome this, bonds involving hydrogen atoms (including water molecules) are often treated as a rigid body (imposing a minimal set of fixed bond lengths and angles) to allow the use of a larger timestep (often double the length, 2 fs). This is done by using different methods like SHAKE or LINCS algorithms.

- *Initial velocities:* Before starting the MD simulation, initial positions and velocities of each atom have to be assigned. The selected initial positions (initial conformation or configuration) ideally come from experimental data (*X-ray* or NMR) and the initial velocities can be randomly generated by applying a Maxwell-Boltzmann distribution at a certain temperature.
  The Maxwell-Boltzmann equation provides the probability of having an initial velocity of v_{ix} in the x direction at a given temperature T of an atom i with mass m_i. With this, a Gaussian distribution with random velocity values is obtained. Given this distribution, multiple separate independent MD simulations starting from the same conformation (*i.e.,* replicas) will start with different initial velocities. Therefore, running multiple replicas starting from the same initial configuration is a commonly used sampling strategy: starting from different initial velocities allows a different conformational exploration over time for each MD replica, speeding up the conformational sampling and collecting better

statistics. However, the sampling strategies used will depend on the characteristics of the system (*i.e.,* number of atoms) and the process under study *(i.e.,* time scale of the given process).

- *Periodic boundary conditions:* Macroscopic, lab-scale or bulk systems consist of multiple moles of atoms or molecules, thus, from a simulation perspective can be considered infinite systems. In computational biochemistry, we attempt to simulate biomolecules by simulating finite small systems (*i.e.,* biomolecules in a solvent environment). An approximation used to overcome the finite size effects in simulations is to set periodic boundary conditions (PBC).

  PBC allow that the simulated system in a periodic box interacts with periodic images of the same system, mitigating the finite size effect. This can be a good approximation to the behavior of a small subsystem in a larger bulk phase. However, it is not desirable that a single particle interacts with the same particle multiple times. To prevent this, a cutoff distance is applied to the non-bonded interactions between all pairs of atoms (*i.e.,* non-bonded interactions are set to 0 when atoms are further apart than the defined cutoff distance). Thus, the employment of the cutoff decreases the computational cost of the simulation.

  The cutoff is usually set to be less than half the length of the simulation box in any dimension. In biomolecular systems, a cutoff between 8 and 12 Å is generally recommended.[97]

2.3.1. Setting up and Molecular Dynamics simulation production run

Setting up the system to properly run MD simulations includes certain steps. In this thesis, to prepare the different systems we followed the next steps: (i) System preparation; (ii) Minimization; (iii) Heating; (iv) Equilibration and (v) Production run.

System preparation is one of the most important steps to set up the system of interest. Since in MD simulations, bond-breaking and forming is generally not allowed, the topology or chemical description of the system will remain constant as a function of time. This implies that important features of the system like the selected initial configuration, the protonation states of the protein constituent, the addition of solvent and counter ions and the proper selection of the FF used will be critical to describe the molecular system.[97] Once the system has been carefully prepared, the energy minimization step proceeds. In this step, the purpose is to find a local minimum in the potential energy surface to identify a relatively stable structure that will avoid instabilities when running the MD simulation. After minimizing the system, it is found in a state where the numerical integration of the equations of motion can begin without major displacements of the atoms. However, to begin a simulation, not only the position of the atoms is required, but also the assignment of initial velocities of each particle of the system (see previous Section 2.3. for more details on the assignment of initial velocities of atoms). To attain the desired temperature of the system, a heating process is performed, where the temperature is gently increased in temperature spans of 50 K. Then, a short MD run is performed to equilibrate the system. The aim of the equilibration step is to bring the system to the target state point and to ensure that the simulation will be run in a particular thermodynamic ensemble (*e.g.,* microcanonical ensemble, defined by a fixed total energy, volume, and number of particles, NVE, or canonical ensemble, defined by a fixed temperature, volume, and number of particles, NVT) to collect data for analysis in the appropriate conditions. Generally, the equilibration step is performed in a constant NPT ensemble (*i.e.,* isothermal-isobaric ensemble), which is characterized by a fixed number of atoms (N), pressure (P) and temperature (T). In our case, to evaluate the proper equilibration of the system, we monitor the density of the solvent and the evolution of the solvation box volume over time. If at the end of the equilibration these two parameters remain constant, it can be considered that the system is properly equilibrated.

Once the equilibration is complete, the production run can start. In this thesis, in contrast to the equilibration step, the production run has been performed in a constant NVT ensemble (characterized by a fixed number of atoms (N), volume (V) and temperature (T)). From the output of the production run, data for the subsequent analysis will be collected.

The MD simulations are usually carried out with a thermostat that adds or removes heat from the system to maintain a target temperature of the system. In our case, the Langevin thermostat algorithm, which also includes frictional force, is used. Many other thermostat algorithms have been developed,[97] but are out of the scope of this thesis.

2.4. Analysis of MD simulations

2.4.1. Dimensionality reduction and identification of relevant states

MD simulations of an enzyme in explicit solvent describe the complex motions of a high number of atoms over time. With the increase in computational power, nowadays it is possible to run MD simulations beyond microsecond time scales and perform multiple replicas at the same time. This causes the system to explore regions of the conformational space that are far from the starting point accumulating a significant amount of data to be interpreted. As a consequence of the large data generated in MD simulation trajectories, it is challenging to identify the essential details that may be relevant to understand enzyme function and connect the observations with experimental data. A potential solution to this problem is to reduce the number of dimensions associated with the atomic coordinates of our system to a few degrees of freedom making the interpretation of MD simulation more simple.[18,97] The general procedure consists of selecting molecular observables, **s**, that describe the process of interest and then, performing a structural analysis (probability distribution P(**s**) or mean) or dynamical analysis (time evolution). Considering that the starting point is a high-dimensional set of input coordinates $\mathbf{r_N}$ = (r_1, r_2, …, r_N), the goal is to construct a low-dimensional observable **s** = (s_1, s_2, …, s_d), where *d* is the number of dimensions, that describes and captures the essential dynamics of the system. Each s_i represents a degree of freedom or a collective variable (CV) that we then used to project our MD trajectories. CVs are also used by some enhanced sampling methods to bias the MD simulation (see Section 2.5.1).

Choosing appropriate degrees of freedom/collective variables. Choosing the appropriate degrees of freedom or CVs to represent our data is crucial because otherwise relevant information can be omitted. This is particularly important to reconstruct the free-energy landscape of the process of interest (see Section 2.4.2 and 2.5.1). For example, it is possible that the selected degrees of freedom to reconstruct the FEL clearly differentiate between relevant stable states but do not capture the presence of intermediate states that are key for understanding the kinetics of the process.[18] Hereafter, a summary of the different strategies available to select the CVs is provided.

- *Collective variables based on (bio)chemical knowledge.* The simplest strategy to select relevant CVs is to visually inspect MD simulations and select key geometrical parameters that we consider relevant for the problem of interest, which can be based or not on previous information of the system. These parameters can be internal coordinates such as distances, angles, or dihedral angles calculated between any atoms in our system. A common practice is to first remove overall translation and rotation of the protein prior the analysis, which can be achieved via common analysis and visualization MD packages. For example, in this thesis, we have monitored the substrate binding into the active site of enzymes using only the distance between one atom of the substrate and one atom of a key residue in the active site. This one-dimensional description is a good option to evaluate the number of binding events and to characterize the molecular basis of the process but may be limited to reconstruct the thermodynamics and kinetics of the binding process because relevant intermediate states controlled by the enzyme conformational dynamics may be found at similar distance values.

- *Linear dimensionality reduction techniques.* Additional strategies to automatically reduce the dimensionality of the MD simulations also exists. Principal Component Analysis (PCA) can be used to reduce the dimensionality of MD simulations (either cartesian coordinates or selected distances, dihedrals, etc.) by maximizing the variance of a data set as much as possible.[110] The principal components (PC) resulting from the low dimensional PCA space will contain the correlated conformational motions of higher amplitude in our system. In this thesis, we have used PCA to study the global dynamic changes of an allosterically regulated enzyme. A different strategy consists of maximizing the time scales of the components instead of the variance such as in PCA. This can be achieved by means of the time-lagged Independent Component Analysis (tICA), which is commonly used to build Markov State Models.[111,112] However, biomolecular processes may be determined by small structural changes (difficult to be described by PCA) or depend on coupled processes that occur on various time scales (difficult to be described by tICA). Therefore, analyzing the outcome of these methods is always recommended to understand in detail the CVs generated. A description of advantages and limitations of these techniques is provided by Sittel and Stock[113] and is out of the scope of the present thesis.

Identification of metastable states. Another strategy to interpret complex MD data is to clusterize MD trajectories based on a certain metric. Geometrical clustering methods allow grouping conformational states based on structural information. Clustering can be performed considering all atoms or a subset (*i.e.,* active site residues or protein backbone atoms) of the atoms of the system.[113] In practice, the outcome of the clustering process will be the separation of MD frames into clusters providing the total population of each cluster. Then, representative structures of each cluster can be structured and compared with others to identify differences between metastable states. A typical geometric clustering technique is *k*-means, which partitions the MD data into *k* subsets (clusters) that minimize the sum of square distances between the objects and the corresponding clustering centroid. In this thesis, we have used *k*-means clustering to clusterize two-dimensional conformational landscapes. Additionally, hierarchical agglomerative clustering as implemented in *cpptraj* has been used to directly clusterize MD trajectories and extract the most relevant conformations. In this type of clustering, MD frames start as single clusters and are hierarchically merged up until a certain criterion (number of clusters) is met. In this thesis, we have focused on the conformational analysis of clusters, therefore, both k-means and hierarchical provide a relevant separation between conformational states. However, if clusters are used to retrieve thermodynamic and kinetic data, then, more appropriate algorithms such as density-based clustering algorithms should be used.

2.4.2. Free Energy Landscape reconstruction

In this Section, we will describe how we can use statistical mechanics concepts to retrieve the conformational free-energy landscape of an enzyme (see Section 1.3 for a general description).

The energy landscape of a molecule visually represents the shape of the potential energy function, V($\mathbf{r_N}$), as a function of the molecule configurations (each configuration (microstate) represents a set of atomic coordinates $\mathbf{r_N}$) at a certain temperature, T. The high-dimensional energy landscape can be simplified if we represent it in a single dimension **s** (see Section 2.4.1.).[22] Through the Boltzmann factor it is possible to connect energy with relative probabilities and obtain the probability of a certain configuration of the energy landscape. The probability density of a configuration can be retrieved from the normalized Boltzmann factor as:[114]

$$\rho(\boldsymbol{s}) = \frac{\exp\left[-V(\boldsymbol{s})/\boldsymbol{k_B T}\right]}{\int_V \exp\left[-V(\boldsymbol{s})/\boldsymbol{k_B T}\right]} \qquad \text{(Eq. 2.5)}$$

Where k_B is the Boltzmann constant, *T* the absolute temperature, and *V* is the total number of configurations. Therefore, the relative probability is lower for high energy configurations. Moreover, the relative probabilities become more equal when temperature increases.

Configurations that belong to the same energy basin can be grouped into a state. However, in practical terms, the word state is also used to group similar energy basins. Irrespective of the state definition, since it contains multiple configurations, it is not possible to speak about the potential energy of state A, but it will be possible to speak about the average potential energy of state A. Using the Boltzmann factor, we can calculate the probability of a certain state A by integrating over the volume of state A.

$$\rho_A(\boldsymbol{s}) = \frac{\int_{V_A} \exp\left[-V(\boldsymbol{s})/\boldsymbol{k_B T}\right]}{\int_{V} \exp\left[-V(\boldsymbol{s})/\boldsymbol{k_B T}\right]} \qquad \text{(Eq. 2.6)}$$

where V_A is the region that contains all configurations assigned to state A.

From the probabilities of each state, it is possible to calculate the free energy F_A(**s**) of a particular state following the expression:

$$F_A(s) = k_B T ln(p_A(s)) \qquad \text{(Eq. 2.7)}$$

These equations of statistical mechanics can be used to describe the conformational free energy landscape of biomolecules. Enzymes are described as an ensemble of conformational states populating an energy landscape. Transitions between conformational states are related to enzyme function. Conformational states are collections of configurations that belong to the same energy basin in the selected dimensions. Considering the typical example of protein folding, a particular enzyme can be represented using a two-state description: the folded and unfolded states. By grouping the configurations that form each state and using the equations described above, we can then retrieve the relative free energies of the different (un)folded states of the protein. This can be applied to other processes such as ligand binding or conformational transitions between inactive and active conformational states of an enzyme.

In this thesis, we have used enhanced sampling techniques to reconstruct the FEL of an allosterically regulated enzyme and identify its functionally relevant states (see Chapter 4).

2.4.3. Network representation of proteins

Another strategy to facilitate the analysis of the vast data generated by MD simulations is to convert the complex conformational dynamics of proteins into a simple network that represents protein motions. In a network representation of proteins, each protein constituent (*e.g.,* amino acid residue) is treated as a node (*i.e.,* using the *Cα* or residue center of mass) and pairs of nodes are connected by edges. In this thesis we have used two different network representation tools: the Shortest Path Map (SPM) and Community Network Analysis (CNA).

2.4.3.1. Shortest Path Map

The Shortest Path Map (SPM) is a dynamical network analysis tool that has been used in this thesis to evaluate the allosteric communication.[59] The first step of the Shortest Path Map (SPM) calculation relies on the construction of a graph based on the computed mean distances and residue-pair correlation values observed along the MD simulations (see Figure 2.4). For each residue pair of the protein a node is created and centered on the *Cα* atom of each residue if both residues display a mean distance of less than 6 Å in 3D space along the simulation time. The length of the line connecting both residues is drawn according to their correlation value (d_{ij}=-*log* $|C_{ij}|$, where C_{ij} is the correlation matrix). Larger correlation values (closer to 1 or -1) will have shorter edge distances, whereas less correlated residue pairs (values closer to 0) will have edges with long distances. At this point, we make use of the Dijkstra algorithm to identify the shortest path lengths and generate the SPM graph (see Figure 2.4). The algorithm goes through all nodes of the graph and identifies which is the shortest path to go from the first until the last protein residue. The method therefore identifies which are the edges of the graph that are shorter, *i.e.,* more correlated, and that are more frequently used for going through all residues of the protein, *i.e.*, they are more central for the communication pathway. These edges have a higher contribution (*i.e.,* represented with thicker lines in the final SPM graph). More details about applications of the SPM tool can be found in recent publications.[59,62,115]

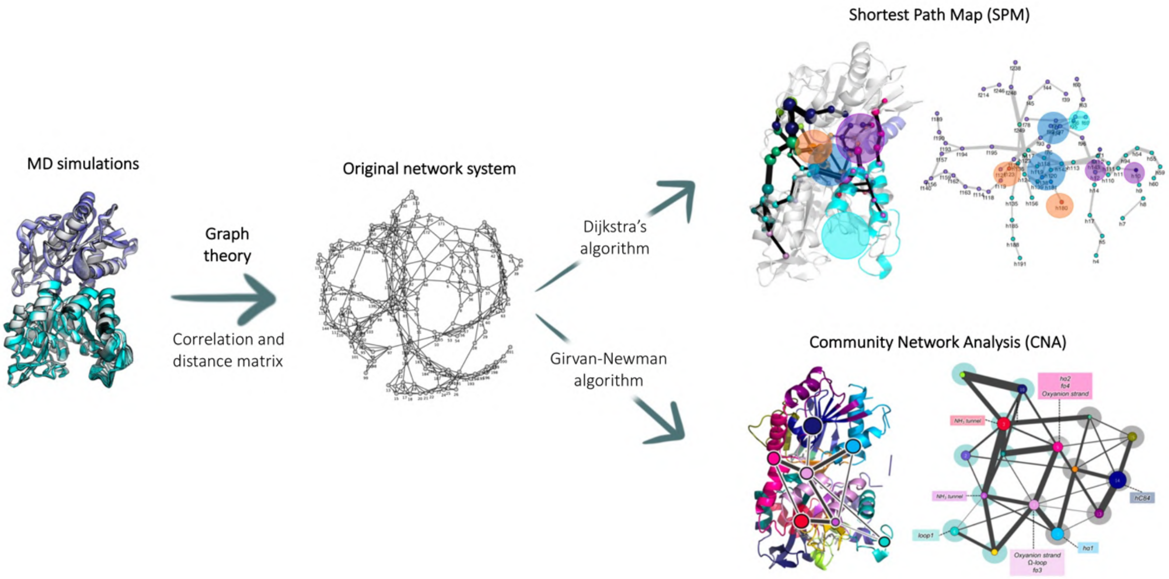


Figure 2.4. Schematic view of different network representations of proteins. Scheme of Shortest Path Map (SPM) and Community Network Analysis (CNA) construction from the MD simulation dataset. In both cases, the spheres represent the graph nodes while the lines that connect them, the edges.

2.4.3.2. Community Network Analysis

A similar analysis can be performed to evaluate the community networks from a graph constructed from MD simulations data. The communities are identified from the original graph generated based on the mean distances and correlation values from the MD simulations using the Girvan-Newman algorithm using a similar protocol as the one described by Sethi *et al.*[63]

2.5 Enhanced Sampling Methods

The study of biomolecular events of interest, that usually take place in the scale of microseconds to milliseconds, requires obtaining enough sampling of these processes to allow the comparison with experimental observables. Hence, to connect MD with experimental data, enough representative conformations should be computationally sampled to satisfy the ergodic hypothesis, confirming that the system has eventually passed through all possible states.[22] To achieve the sampling of all relevant events that occur in long timescales, an extensive amount of computational resources would be needed.[12] When relevant conformational changes occur in millisecond to second timescales, conventional MD simulations usually remain stuck around one energy minimum and, thus, do not properly reproduce the experimental data with statistical significance. More accurate descriptions of complex systems rely, thus, in alternative approaches to access long timescale events.

In recent years, a lot of efforts have been put into developing simulation techniques that speed up and improve the efficiency of conformational sampling while keeping the atomistic description of the system. Ideally, the aim of these methods is to provide strategies for visiting all relevant metastable states for the problem of interest in an affordable simulation time. In this line, enhanced sampling techniques represent an attractive alternative strategy to long unbiased simulations. Enhanced sampling methods can be divided into different groups whether they are based on biased and unbiased MD simulations or if they depend on the definition of a reaction coordinate (see Section 1.1.3.2).[18]. In this thesis, we used two different types of enhanced sampling methods: the ones that bias a set of collective variables (CVs) and biased unconstrained methods.

2.5.1. Collective variables-based methods

The CV-based methods, such as umbrella sampling and metadynamics, require an *a priori* definition of a reaction coordinate: either a transition pathway between known initial and final states or a set of CVs that need to be defined to drive the course of the simulation (see further description in Section 1.1.3.2.).[31,32] These methods significantly improve convergence of free-energy calculations. Therefore, the free energy landscape of the process of interest can be retrieved given a set of CVs.

Relevant aspects for defining proper collective variables for biased MD simulations. As mentioned in 2.4.1, CVs (**s**) are expressed as a function of the coordinates of the system and offer a low-dimensional representation of the molecular system.[33] When used for biasing an MD simulation, the selected CVs should: 1) clearly differentiate the relevant metastable states and transition states of the process of interest; 2) include all the slow modes of the system; and 3) be limited in number. CVs are commonly defined based on chemical/biological information or intuition. Typical CVs include distances, angles, or dihedral angles. However, more complex CVs can be defined such as paths of collective variables. Recently, automated selection of CVs based on machine-learning algorithms has also been used.[116] Once the CVs are selected, the probability distribution (P($\mathbf{r_N}$)) can be represented as a function of the CV as P(**s**). Therefore, the free energy can be written as a function of the probability distribution associated with the selected CV, $F(\boldsymbol{s})$=-Tln$(P(\boldsymbol{s}))$ as described in 2.4.2. If the CV is properly defined, the F(**s**) will retain the essential metastable states of the process of interest.

2.5.1.1. Metadynamics Simulations

Metadynamics is a constrained biased approach based on introducing an artificial bias potential to a set of CVs.[31,33] This allows for enhancing conformational sampling and reconstructing the FEL as a function of the selected set of CVs. This method is based on applying small repulsive Gaussian potentials at regular time intervals (see Figure 1.9). Repulsive potentials discourage the system from visiting previous points. For a metadynamics simulation using classical energy potentials to calculate the energy of the system we have:

$$V_t^*(r_N) = V_{FF}(r_N) + V_t^G(\boldsymbol{s}) \qquad \text{(Eq. 2.8)}$$

where $V_t^*(\mathbf{r_N})$ is the total energy of the system in the modified potential (at a certain time t) and $V_{FF}(\mathbf{r_N})$ is the force field energy of the system. $V_t^G(\mathbf{s})$ is a history-dependent bias potential which is a function of the CVs added during metadynamics simulation. Generally, the $V_t^G(\mathbf{s})$ at certain time *t* is expressed as a sum of Gaussian functions deposited in previous times (*t'*) centered at the corresponding values of s, as:

$$V_t^G(\boldsymbol{s}) = w\sum_{t'<t} exp\left(\frac{(s-s_{t'})^2}{2\sigma^2}\right) \qquad \text{(Eq. 2.9)}$$

where $s_{t'}$ determines where the Gaussian is centered in each time *t'*, w is an energy rate (w=W/τ_G, where W is the Gaussian height and τ_G the deposition stride), and σ is the width of the Gaussian. Both w and σ should be specified by the user prior to the simulation. After enough simulation time, the bias potential added during metadynamics ($V_t^G(\mathbf{s})$) provides an estimate of the underlying free energy:

$$F(\boldsymbol{s}) = -V_t^G(\boldsymbol{s}) \qquad \text{(Eq. 2.10)}$$

Therefore, metadynamics can be used to both escape from free energy minima and to reconstruct the FEL in a CV space. Ideally, the metadynamics should be stopped when all the relevant minima are filled with Gaussian and when the motion of CVs becomes diffusive.

However, using Equation 2.9. the potential $V_t^G(\mathbf{s})$ does not converge to the free energy, oscillating around its value. Therefore, it is not trivial to decide when to stop the simulation. *Well-tempered* metadynamics solves this problem by decreasing the Guassian height based on the time spent at a given point using a parameter that controls how quickly the height of the Gaussian decreases. In this thesis, we have used well-tempered metadynamics to reconstruct the FEL of the allosteric activation of an allosterically regulated enzyme (see Chapter 4).

General limitations of CV-based enhanced sampling methods: The definition of CVs sometimes represents a challenge and can be non-trivial. In particular, when the final state of the complex is unknown (*e.g.,* active state of a protein or location of binding site). The inappropriate definition of CVs can affect the number of metastable states visited in the simulation and the barriers that separate them. This will have direct consequences on the estimations of binding energies and rates. The more complex the process is, the more difficult it is to define the proper CVs.

2.5.2. Unconstrained Enhanced Sampling Techniques

An interesting strategy to overcome the above-mentioned limitations is the use of Unconstrained Enhanced Sampling Techniques (UEST). These methods do not rely on the *a priori* definition of a set of CVs and provide unconstrained conformational sampling to freely explore the biomolecular conformational space and transition pathways (*e.g.,* conformational changes of biomolecules and protein folding). These methods are used to identify possible unknown intermediate and metastable states of biomolecules. The list of UEST includes methods such as replica-exchange MD and accelerated molecular dynamics (aMD) among many others (see Section 1.1.3.2.). Therefore, the simulation will take place without the need of defining a reaction coordinate, making the exploration of the full conformational space less complex and more efficient.

2.5.2.1. Accelerated molecular dynamics simulations

Accelerated molecular dynamics (aMD) is a versatile enhanced sampling technique that speeds up molecular dynamics and does not rely on the *a priori* definition of reaction coordinates.[25] In the last years, aMD has been used to provide a dynamic atomistic view of relevant biomedical challenges. According to Miao *et al.* “hundreds of nanoseconds aMD simulations are able to capture millisecond-timescale events in both globular and membrane proteins”.[117]

The theoretical foundations of aMD are the following: aMD enhances sampling through modification of the system’s Hamiltonian in a relatively simple way (only two parameters are required). In addition, it does not rely on the definition of a reaction coordinate or a set of CVs and it conserves the essential details of the FEL. aMD typically modifies the underlying potential energy surface by applying a boost potential at each point of the MD trajectory according to Equations 2.11 and 2.12. The value depends on the difference between a reference, ‘threshold’, or ‘boost’ energy, E, and the actual potential energy V($\mathbf{r}_\mathrm{N}$).

$$\begin{aligned} V^*(r) &= V(r), & V^*(r) &\geq E \\ V^*(r) &= V(r) + \Delta V(r), & V^*(r) &< E \end{aligned} \qquad \text{(Eq. 2.11)}$$

where V($\mathbf{r}_\mathrm{N}$) is the original potential energy, E is the reference energy, and V*($\mathbf{r}_\mathrm{N}$) is the modified potential. If the potential energy V($\mathbf{r}_\mathrm{N}$) has a lower value than the reference energy E, the potential energy is modified V*($\mathbf{r}_\mathrm{N}$) by adding a non-negative boost potential to the original energy (Equations 2.11 and 2.12 and Figure 2.5). In aMD, the boost potential (ΔV($\mathbf{r}_\mathrm{N}$)) is defined by:

$$\Delta V(r) = \frac{(E-V(r))^2}{\alpha+E-V(r)} \qquad \text{(Eq. 2.12)}$$

The α parameter regulates the level of acceleration, and their optimal values are system specific (depend on the number of residues and total number of atoms in the simulation). The larger the difference between V and E, the greater the modification of the potential energy surface becomes, pushing up low-energy valleys and decreasing the magnitude of energy barriers, facilitating the transition between low-energy states. Therefore, aMD works remarkably well to efficiently explore the conformational changes of biomolecules (*e.g.,* open-closed transitions in proteins).

aMD and GaMD (see below) are particularly designed to study slow conformational changes within a folded protein. These transitions are primarily induced by torsional changes on protein backbone and side chains. Therefore, in standard aMD, the acceleration (boost potential) is mainly applied to the dihedral term of the force field (‘dihedral-boost’). Originally, it was found the ‘dihedral-boost’ alone was insufficient for conformational sampling of many biomolecules.

Thus, another boost can be added to the system total potential energy ('total-boost', that include all force field terms), which accelerates diffusion of the solvent molecules and provides further enhanced sampling of biomolecules. This led to the development of 'dual-boost' aMD, which is the version of aMD implemented in most common simulation packages. Usually, a higher acceleration, in relative terms, is applied to the dihedral term ('dihedral-boost') than to the total energy term ('total-boost'), see below for an example.

The acceleration parameters for the total and dihedral boost are usually calculated using the following expressions:

$$E_{dihed} = V_{dihed\ avg} + a_1 N_{res} \qquad \text{(Eq. 2.13)}$$

$$a_{dihed} = a_2 \frac{N_{res}}{5} \qquad \text{(Eq. 2.14)}$$

$$E_{total} = V_{total\ avg} + b_1 N_{atoms} \qquad \text{(Eq. 2.15)}$$

$$a_{total} = b_2 N_{atoms} \qquad \text{(Eq. 2.16)}$$

where N_{res} is the number of protein residues, N_{atoms} is the total number of atoms, and V_{dihed_avg} and V_{total_avg} are the average dihedral and total potential energies calculated from 100 ns cMD simulations, respectively. In this thesis, two different levels of acceleration have been used. The parameters used for aMD simulations are high acceleration (a_1=3.5, a_2=3.5; b_1=0.175, b_2=0.175) and moderate acceleration (a_1=2, a_2=3.5; b_1=0.16, b_2=0.16).

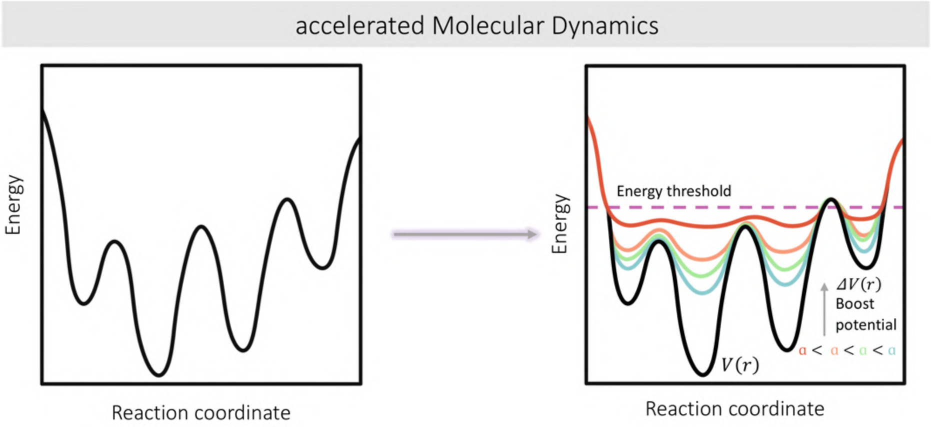


Figure 2.5. Accelerated Molecular Dynamics (aMD). Representation of the conformational landscape as a function of the reference energy (E) and the parameter a.

General limitations of Accelerated Molecular Dynamics: However, a long-standing problem of aMD is the accurate reweighting of aMD simulations and the recovery of the original free energy landscapes, especially for large biomolecules. Therefore, aMD works remarkably well to efficiently explore the conformational space of biomolecules but suffers to estimate the free energy barriers for transitions between different conformational states.

2.5.2.2. Gaussian Accelerated Molecular Dynamics

Recently, Miao *et al.* proposed an elegant way to recover the FEL with a method called Gaussian Accelerated Molecular Dynamics (GaMD).[35] This method allows for both accurate unconstrained enhanced sampling and accurate energetic reweighting. GaMD is an unconstrained enhanced sampling technique that offers accurate reweighting of the free energy surface in comparison to aMD simulations. GaMD enhances conformational sampling by applying a harmonic boost potential to flatten the energy landscape. As in aMD simulations, the harmonic boost potential is only added to the system when the system potential is lower than a reference energy:

$$\Delta V(r) = \frac{1}{2}k(E - V(r))^2, \qquad V(r) < E \qquad \text{(Eq. 2.17)}$$

$$\Delta V(r)\ 0, \qquad V(r) \geq E \qquad \text{(Eq. 2.18)}$$

where ΔV($\mathbf{r_N}$) is the harmonic boost potential, E is the reference energy, and k is the harmonic force constant. E and k are determined following the criteria:

$$V_{max} \leq E \leq V_{min} + \frac{1}{k} \qquad \text{(Eq. 2.19)}$$

where V_{min} and V_{max} correspond to the minimum and maximum potential energies, respectively. In our GaMD simulations the threshold energy E was set to the upper bound E = V_{min} + 1/k. By using harmonic boost potential it is possible to obtain a Gaussian distribution of boost potentials that yields a more accurate reweighting of the FEL.[118]

2.6. Quantum Chemical Calculations of Enzyme Mechanisms

Quantum mechanics (QM) based methods are widely used to model the reaction mechanism of (metallo)enzymes.[17] These methods are based on solving the Schrödinger Equation to obtain the energy and molecular properties of a molecular system. Because it does not exist an exact way of solving the Schrödinger equation for polyatomic systems, different approximations have been proposed. Among the different QM methods, DFT is still the preferred choice for modeling enzymatic reactions because of the compromise between accuracy and computational efficiency. DFT is based on using the electron density to describe molecular properties. A wide variety of DFT functionals exists and their choice depends on the features of the studied system and the properties we aim to predict. A detailed description of the theoretical foundations of DFT and DFT functionals to study chemical problems can be found for example in Cohen *et al.,*[119] and will not be discussed here.

From the practical point of view, DFT calculations in general allow for the accurate prediction of equilibrium and transition state geometries along a reaction mechanism. A proper choice of the DFT functional and basis set is critical to accurately predict reaction energies and energy barriers. The hybrid B3LYP is one of the most popular functionals for describing the thermodynamic and kinetic properties of mechanisms of organic reactions and also reactions involving transition metal complexes (as the ones found in metalloenzymes). In particular, B3LYP has been widely used and validated to study reaction mechanisms involving haem-iron cofactors as the ones studied in the present thesis. Since the high computational cost associated with including transition metals in DFT calculations (they have a large number of electrons), an efficient strategy to calculate energy profiles of chemical reactions is to: 1) Carry out the geometry optimizations and frequency calculations with DFT using a lower level of theory with small basis set; 2) use the optimized geometry to carry out single point calculations with a larger basis set to refine the energetics and reconstruct the energy profile (see for example Chapter 5). One of the issues with most DFT functionals is the lack of a description of long-range dispersion interactions (such van der Waals), which may play an important role in enzymatic catalysis. This can be partially solved by the inclusion of dispersion corrections as an empirically determined additional term, that is added to the total energy estimated from DFT calculations. In the DFT calculations performed in this thesis, D3-BJ dispersion correction has been only applied to the single point calculations with large basis sets.

A limitation of DFT methods is that only a limited number of atoms of the enzyme system can be treated at this level of theory because the computational cost increases exponentially with the size of the system (as described in Section 2.1).[17] This requires deciding which atoms will be included in the QM calculation while still considering (part) of the enzyme environment. Next, different strategies to model enzymatic reactions with DFT will be briefly described.

The first strategy is to carry out a DFT calculation with a truncated model of the enzyme active site. A limited number of atoms should be selected using the knowledge of the reaction mechanism or by closeness to the substrate. Two protocols are commonly used to understand enzymatic mechanisms using truncated models: the *theozyme* and the cluster model.

- Theozyme: Within the *theozyme* (short for *theoretical enzyme*) approach, a truncated portion of the enzyme is selected to include the essential elements to describe the transition-state of the reaction of interest *i.e.,* catalytically relevant active site residues, cofactors, and substrate.[83] This provides the ideal arrangement of the active site residues to stabilize the optimal TS of the targeted reaction. In Chapter 5, the reaction mechanism of haem-iron carbene transfer reaction is studied with the *theozyme* model that includes the porphyrin pyrrole core, the iron center, a methoxy group to mimic serine as iron-axial ligand, the lactone-carbene bound to the iron center, and the substrate.[96] In Chapter 6, a truncated model with a portion of the NADH/NADPH cofactor with formate (substrate) was used to calculate the TS optimal angle and distance for hydride transfer reaction.[120]

Since only a portion of the enzyme active site is considered, *theozyme* calculations are usually performed in implicit solvent with a dielectric constant that aims to mimic the polarization of the enzyme active site. For example, a dielectric constant $\varepsilon = 4$ has been proved to be a good and general model to account for electronic polarization and small backbone fluctuations in buried enzyme active sites.

The optimal geometrical parameters obtained from the *theozyme* calculations are often used as a reference to analyze the preorganization of enzyme active sites in MD simulations in the presence or absence of the substrate (see Chapters 5 and 6). Since the *theozyme* DFT calculations are commonly performed without using positional restraints, the characterized ideal orientation of residues around the substrate may change with respect to the one found in the enzyme active site (*i.e., X-ray* structure). This provides key information on how the enzyme controls the reaction in comparison to the truncated model. Afterwards, MD simulations can be used to understand the molecular basis of the enzymatic control in key reaction intermediates that determine the enantioselectivity of the product formed (see Chapter 5).

- *Cluster Model*: In general, cluster model approaches include a higher number of atoms than truncated *theozyme* models.[17] Within the cluster model not only the residues that play a key role in the enzymatic mechanism but also the residues that determine the shape of the active site are commonly included in the truncated enzyme model. Therefore, the cluster model also captures the steric effects imposed by the enzyme active site. Current cluster models can include more than 250 atoms in the calculation and positional restraints are commonly applied to selected atoms to keep the protein backbone conformation.[121]

Using a truncated enzyme model means that a larger part of the enzyme environment (protein scaffold and solvent) is not considered in the calculation, which may have implications on the accuracy of the calculations. In some cases, it is required to explicitly consider the effects of the enzyme environment (sterics and polarization) and solvent molecules and their influence on the calculation of the reaction profile. A strategy to include a more representative portion of the enzyme environment consists in combining QM with MM approaches.

- *Hybrid Quantum Mechanics/Molecular Mechanics (QM/MM).* In QM/MM calculations a small portion of the chemical system, usually the enzyme active site, is treated using QM, and the rest of the system is described using a computationally more efficient level of theory, such as MM (force fields).[17] A key aspect for the computational accuracy and efficiency of QM/MM calculations is how the coupling term that controls the interactions between the QM and MM parts is treated, *i.e.,* electrostatic or polarizable environment. More information about QM/MM methods and their applications to enzymatic catalysis can be found in the review by Sousa *et al.*[122]

The criteria to select the best strategy to calculate an enzyme reaction profile depends on the system, reaction of interest and the information one is looking for. All methods have their pros and cons and, thus, checking the available bibliography is always recommended to identify previous studies in similar systems.

In the present thesis, the reaction profiles will be calculated using a truncated *theozyme* model of the enzyme active site (see Chapters 5 and 6). The cluster model and QM/MM approaches are not used. More information on how QM calculations are technically performed can be found in the computational details of Chapter 5.

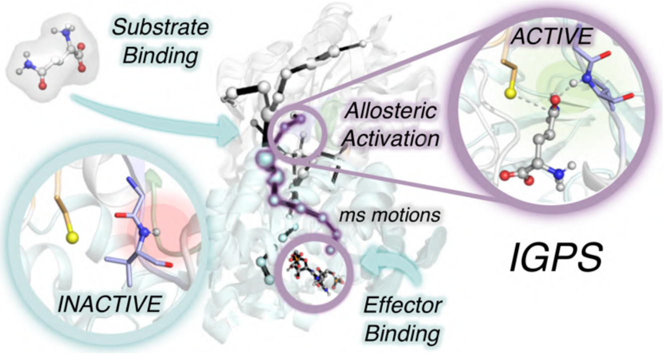
Chapter 3. Objectives

The general objective of this thesis is to explore the molecular basis of biochemical and biocatalytic processes by means of computational methods and examine its relationship with enzymatic properties such as allostery, cofactor specificity, and catalytic activity. The goal is to gain insights into the molecular basis of enzymatic processes and rationalize the novel enzymatic functions of laboratory-evolved enzymes through the application of computational protocols that combine different techniques: from molecular dynamics simulations to quantum mechanics. This information will then be used to rationally design new enzyme variants. The objectives of this thesis are divided into three main categories, each with specific goals related to the systems under study:

I. **Time Evolution of the Millisecond Allosteric Activation of Imidazole Glycerol Phosphate Synthase (Chapter 4):**

The main goal of this project is to characterize the molecular details of the allosteric activation of IGPS in the ternary complex and identify hidden states relevant for IGPS catalytic activity with a computational strategy tailored to explore millisecond timescale events. To accomplish this goal, the specific objectives for this project are the following:

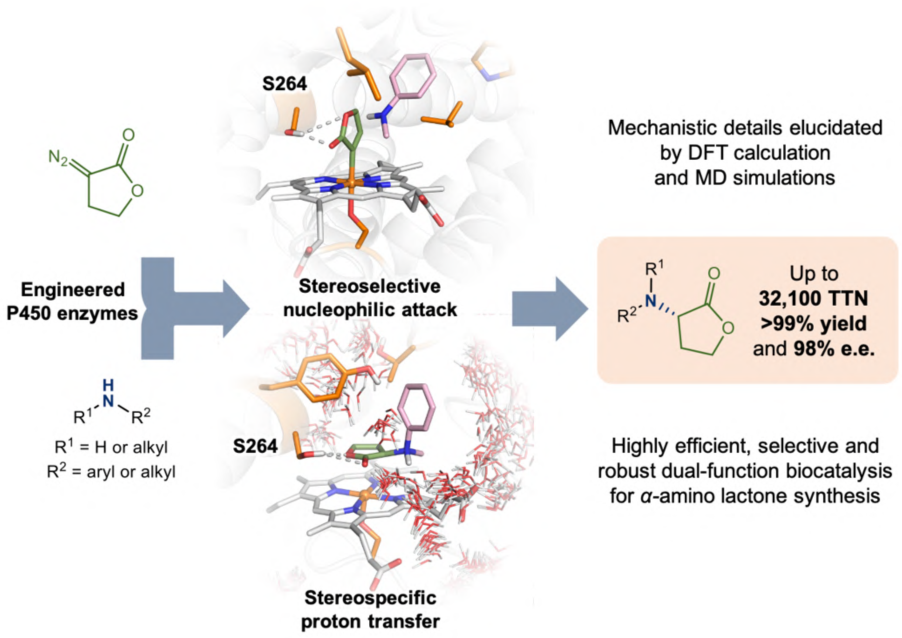
- To characterize the allosterically active state and relevant intermediate states for the allosteric regulation process of IGPS.
- To reconstruct the process of substrate binding and allosteric activation mechanism regulating IGPS activity and the sequence of events toward the formation of the active ternary complex.
- To design a general computational strategy to characterize allosteric transitions in the millisecond timescale and to decipher the molecular basis of allosteric mechanisms in related enzymes.



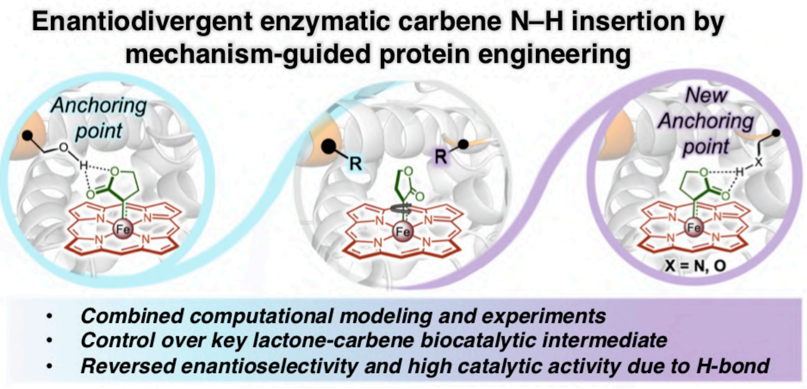
II. Molecular basis of efficient and enantioselective biocatalytic C-N bond formation: from understanding to design (Chapter 5)

The main goal of this project is to design a computational protocol tailored to unravel the molecular mechanism of the enantioselective carbene N–H insertion catalyzed by P411 enzyme variants. To accomplish this, two main specific objectives are described as follows:

- To understand the molecular basis of this enzymatic transformation and to elucidate the role of key mutations in promoting asymmetric carbene N–H insertion.


S264
Engineered P450 enzymes
R1 = H or alkyl
R2 = aryl or alkyl
Stereoselective nucleophilic attack
S264
Stereospecific proton transfer
Mechanistic details elucidated by DFT calculation and MD simulations
Up to 32,100 TTN >99% yield and 98% e.e.
Highly efficient, selective and robust dual-function biocatalysis for α-amino lactone synthesis

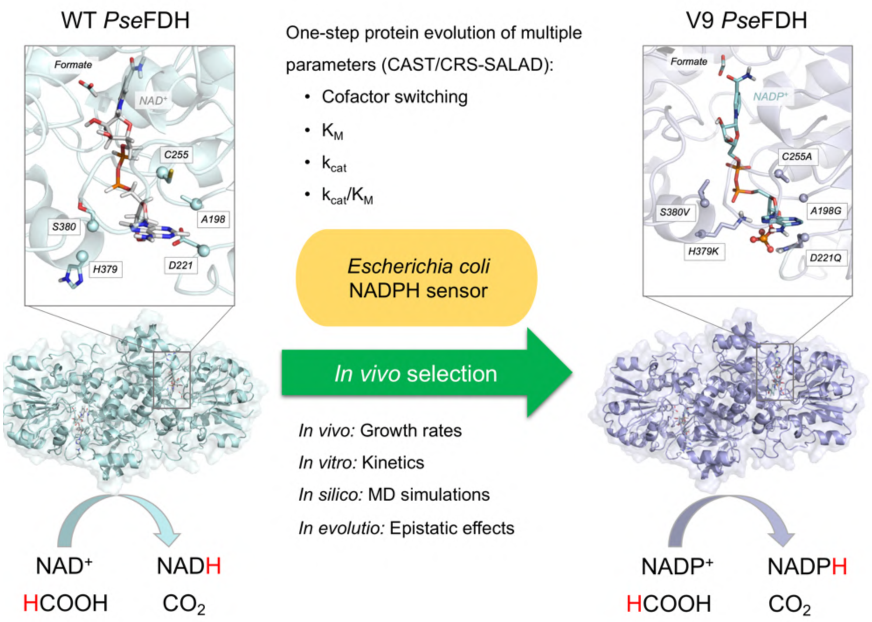
- To generate a mechanistically-guided design protocol for engineering a biocatalytic platform for enantiodivergent C-N bond formation based on the previous knowledge acquired from the analysis of enantioselective P411 variants.


Enantiodivergent enzymatic carbene N–H insertion by mechanism-guided protein engineering
Anchoring point
R
R
New Anchoring point
X = N, O
Combined computational modeling and experiments
Control over key lactone-carbene biocatalytic intermediate
Reversed enantioselectivity and high catalytic activity due to H-bond

III. Molecular basis for the Selection of Formate Dehydrogenases with High Efficiency and Specificity toward NADP+ (Chapter 6)

The main goal of this project is to rationalize the molecular basis of $NADP^+$ specific engineered Formate Dehydrogenase (FDH) variants that present three properties: kinetic efficiency with the non-natural $NADP^+$ cofactor, specificity toward the non-natural $NADP^+$ cofactor and affinity toward the substrate formate. The specific objectives of the computational modelling of FDH enzymes can be summarized as follows:

- To rationalize the molecular basis of cofactor specificity in the WT *Pse*FDH enzyme.
- To identify key residues involved in cofactor specificity.
- To unravel the molecular basis of the superior kinetics of $NADP^+$-dependent engineered FDH variants.



Chapter 4. Time Evolution of the Millisecond Allosteric Activation of Imidazole Glycerol Phosphate Synthase

The following Chapter describes the computational work to reconstruct the molecular details of the millisecond allosteric activation of imidazole glycerol phosphate synthase. The results described in this Chapter were reported in *J. Am. Chem. Soc.* **2022**, 144, 7146-159. We (Carla Calvó-Tusell, Miguel Angel Maria-Solano, Sílvia Osuna and Ferran Feixas) performed the complete computational study. Reprinted with permission from *J. Am. Chem. Soc.* **2022**, 144, 7146-159 (https://pubs.acs.org/doi/10.1021/jacs.1c12629). Copyright 2022 American Chemical Society. Further permissions to the material shown should be directed to the ACS.

4.1. State of the art

Proteins can modulate their function and dynamism in response to environmental changes (binding, point mutations, post-translational modifications) through allosteric regulation.[123] Allostery is defined as the process in which two (often distal) sites of a protein or protein complex are coupled in functional terms.[124] The prominent role of allostery in enzyme catalysis, signal transduction, and disease triggered many studies to unravel the underlying molecular mechanisms of allosteric communication and to harness its power for developing new therapeutics and engineering novel enzyme functions. Deciphering the molecular basis of allosteric mechanisms is challenging because it is system specific and involves a delicate balance between energy, structure, dynamics, and function tailored through networks of inter- and intramolecular interactions of binding partners.[125]

In allosterically regulated enzymes, the effect of binding at one site (effector site) is transmitted to a distal functional site (catalytic active site).[55] In particular, effector binding can modify the thermodynamic and/or kinetic parameters of the enzymatic reaction.[47,52] In K-type allosteric enzymes, the binding of the effector alters the affinity for the substrate (K_M) while in V-type the change is located in the k_{cat}, while keeping the affinity for the substrate unaffected. The chemical information transferred between the two coupled sites is regulated by structural and dynamical changes that activate the enzyme by pre-organizing the active site for the reaction.[126,127] To reach the enzyme catalytically active state, the binding of an allosteric effector finely tunes the conformational ensemble of the enzyme. This can alter both the population of the different conformational states and/or the dynamic properties, which include the rates of conformational transitions between states and the inner flexibility of the enzyme.[15] In the context of allostery, the population shift concept is often referred as the conformational-based allostery, while the changes in fluctuations (*i.e.,* activation/suppression of protein motions) are referred as dynamic-based allostery.[15]

The allosteric communication process initiated by effector binding is controlled by the interplay between enthalpic and entropically driven motions encoded in the protein ensemble that facilitates the population of the active state.[15] The communication between distal sites is of bidirectional nature and takes place when the ternary complex is formed. In the ternary complex, both the effector and the substrate are bound at their respective enzyme sites. The transfer of information between sites propagates through networks of interactions between amino acid residues. Therefore, to describe the time evolution of the allosteric activation process of an enzyme in their ternary complex involves characterizing the interplay of fast and slow conformational dynamics coupled to effector and substrate binding.[128,129] However, the transient nature of the allosteric transition in enzymes undergoing turnover complicates the molecular characterization of allosteric mechanisms and the identification of functionally relevant states.[51,130–134] Thus, the allosterically active state and relevant intermediate states for the allosteric regulation process remain often uncharacterized for most enzymes.

An archetypical model enzyme to study allosteric regulation is *Thermotoga maritima* Imidazole Glycerol Phosphate Synthase (IGPS). Allostery in IGPS has been extensively explored from both experimental and computational perspectives.[135–146] IGPS is a heterodimeric enzyme that belongs to class I glutamine amidotransferases (GATase) formed by two subunits HisH and HisF that are catalytically coupled (see Figure 4.1). The HisH monomer catalyzes the hydrolysis of glutamine to produce glutamate and ammonia. Then, ammonia migrates through an internal tunnel reaching the HisF active site. The HisF subunit presents cyclase activity and is responsible of coupling the HisF substrate N’-[(5’-phosphoribulosyl)formimino]-5-aminoimidazole-4- carboxamide ribonucleotide (PRFAR) with ammonia to produce imidazole glycerol phosphate (IGP) and 5-aminoimidazole-4-carboxamide ribotide (AICAR), (see Figure 4.1). PRFAR is the substrate of HisF but also is the allosteric effector for the glutaminase hydrolysis reaction in the HisH subunit. The HisH and HisF active sites are *ca.* 30 Å far away, indicating that a long-range allosteric communication is required. Based on experimental kinetic studies it was shown that the binding of PRFAR enhances 4500-fold the basal glutaminase activity of HisH in IGPS (from $4.6\text{x}10^{-1}$ to $2.08\text{x}10^{3}$ $M^{-1}s^{-1}$), while the substrate affinity is only moderately altered. This points out the existence of strong allosteric coupling between the two subunits. However, the mechanism through which IGPS is allosterically activated by PRFAR is still unknown. The most significant advances on the understanding of IGPS allosteric regulation from structural and dynamical perspective are summarized.

The initial hypothesis based on mechanistically similar enzymes is that the binding of PRFAR in the HisF subunit activates IGPS, thus, triggering the formation of an oxyanion hole at the HisH active site that is key for triggering glutamine hydrolysis. The oxyanion hole is suggested to be formed by the amide H^N backbone of residue *h*V51 (throughout the text the *h* and *f* labels are used to indicate HisH or HisF residues, respectively). Thus, the formation of *h*V51 oxyanion hole is believed to be responsible of pre-organizing the HisH active site for hydrolysing glutamine in an efficient manner.[147,148] *h*V51 is a HisH residue situated in the oxyanion strand (*h*49-PGVG), which is located near the IGPS catalytic triad consisting of *h*C84, *h*H178, and *h*E180 (Figure 4.1b). Taking into consideration mechanistic observations of similar GATases, the glutaminase reaction requires the formation of a transient tetrahedral intermediate with a negative charge. The formation of an oxyanion hole is a requisite to stabilize the negative charge formed during glutamine hydrolysis (Figure 4.1c). However, any of the available *X-ray* crystal structures corresponding to wild-type IGPS (*wt*IGPS) show the amide H^N backbone of *h*V51 oriented toward the HisH active site, indicating that the pre-organized HisH active site is not a major state in the conformational ensemble of *wt*IGPS.[149] Alternatively, it was proposed that the tetrahedral intermediate is stabilized by H^N of contiguous oxyanion strand residue *h*G52 (which is already pointing toward the active site in available *X-ray* structures) and that the effect PRFAR binding is to induce the productive closure of the HisF:HisH interface enhancing the glutaminase activity (see Figure 4.1c).[150] However, the rapid glutamine turnover observed upon *wt*IGPS allosteric activation hampered the structural characterization of the enzyme with the *h*V51 oxyanion hole formed or with the HisF:HisH interface closed. Thus, the available *X-ray* structures of *wt*IGPS present the oxyanion hole unformed and an open HisF:HisH interface, which based on all structural knowledge corresponds to the inactive enzyme.

IGPS has also been deeply studied from a dynamic perspective. Using NMR experiments it was suggested that the binding of PRFAR binding activates millisecond motions that trigger allosteric communications between HisF and HisH subunits resulting in increased IGPS conformational flexibility. Additionally, NMR measurements indicate that PRFAR-free, PRFAR-bound, and the ternary complex present different patterns of millisecond motions.[145] The formation of the ternary complex activates correlated millisecond motions of an allosteric network of IGPS residues that connect the HisF and HisH binding sites. Wurm and coworkers recently characterized the dynamics of *h*C84S mutant under turnover conditions with NMR showing that the inactive and active states of IGPS are in dynamic equilibrium.[151] This mutant presents significantly reduced glutaminase activity and allows the detection of active and inactive states. In the case of *wt*IGPS under turnover conditions, the active state was not detected because it was found below the detection limit of NMR experiments. Based on these results it was established that in the *h*C84S mutant conformational ensemble the active state is stabilized.

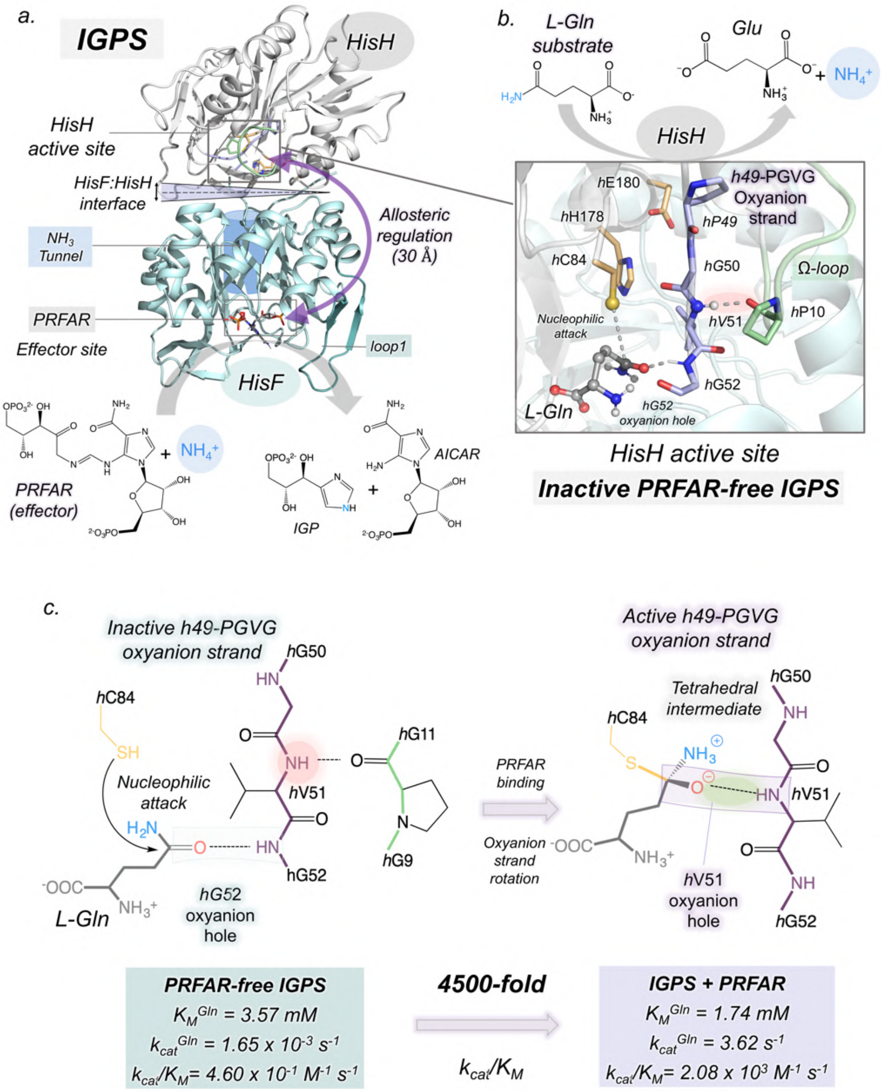


Figure 4.1. Overview of IGPS structure and global mechanism. a) IGPS is a heterodimeric enzyme formed by two subunits (PDB: 1GPW): HisH (shown in white) and HisF (shown in cyan). **b)**

HisH active site (PDB: 3ZR4) with substrate glutamine (*L-Gln*, gray) bound in inactive h49-PGVG oxyanion strand (purple) is depicted. The catalytic and Ω-loop residues are highlighted in orange and green, respectively. The NH backbone of *h*V51 is shown in spheres. **c)** Scheme of the hypothesis of *h*V51 oxyanion hole formation and kinetic parameters for glutamine hydrolysis in PRFAR-free and PRFAR-bound IGPS, extracted from Rivalta, *et al.*[152]

IGPS has also been the focus of extensive computational studies based on Molecular Dynamics (MD) simulations. MD simulations in the nanosecond time scale pointed out that the effect of PRFAR binding influences the dynamics of *f*loop1 (HisF subunit) and propagates through a network of salt bridges that interconnect HisF and HisH subunits. Rivalta and coworkers rationalized that these rearrangements alter the conformational dynamics of the HisF:HisH interface making the hydrogen bond between *h*P10 (located at the Ω-loop) and the amide H^N backbone of *h*V51 weaker (Figure 4.1b).[147] Dynamical network models were used to study the allosteric communication in IGPS. These works revealed that PRFAR improves HisF:HisH interdomain communication.[38,60,147,153–159] Since it is challenging to reach millisecond time-scale events with MD simulations, the rotation of the oxyanion strand to form the *h*V51 oxyanion hole characteristic of the allosteric active ternary complex was not explored. Despite all these relevant studies, the allosteric mechanism regulating IGPS activity and the sequence of events toward the formation of the active ternary complex of IGPS was still uncharacterized. However, unraveling the time-evolution of the millisecond allosteric activation of IGPS will provide relevant information for the understanding of enzyme function and engineering.

In this Chapter, we design a computational strategy to explore millisecond timescale events that is applied to describe the allosteric activation of *wt*IGPS at the molecular level and reveal hidden states key for IGPS functional activity (Figure 4.2 and A4.1). Our goal is to reconstruct the complete allosteric activation mechanism without using *a priori* information of the active state. This protocol consists of combining extensive conventional MD simulations, enhanced sampling techniques, and dynamical networks to unravel how the conformational ensemble evolves toward reaching the allosterically active state of IGPS. Our simulations reveal the formation of a pre-organized HisH active site that presents the *h*V51 oxyanion hole oriented to stabilize the glutamine substrate (*L-Gln*). Spontaneous substrate *L-Gln* binding MD simulations and the analysis of dynamical networks show that substrate binding activates correlated motions that control the allosteric activation of IGPS in the ternary complex. In addition, we identified that the productive closure of the HisF:HisH interface is required to attain the pre-organized HisH active site.

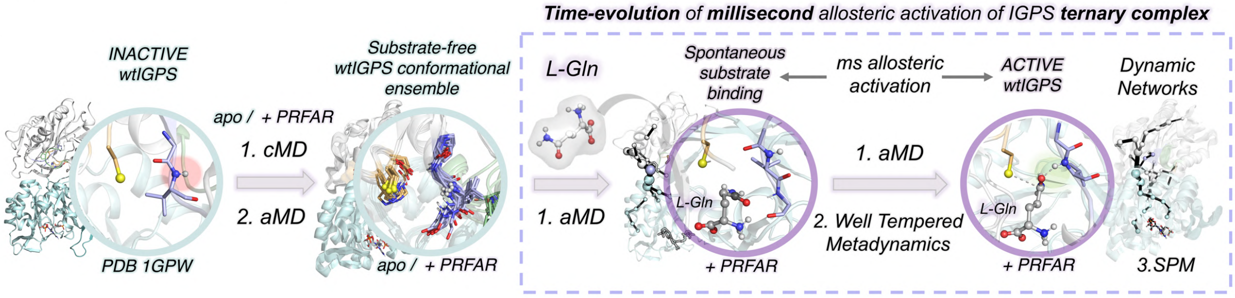

Time-evolution of millisecond allosteric activation of IGPS ternary complex
INACTIVE wtIGPS
PDB 1GPW
apo / + PRFAR
1. cMD
2. aMD
Substrate-free wtIGPS conformational ensemble
apo / + PRFAR
L-Gln
1. aMD
Spontaneous substrate binding
L-Gln
+ PRFAR
ms allosteric activation
1. aMD
2. Well Tempered Metadynamics
ACTIVE wtIGPS
L-Gln
+ PRFAR
Dynamic Networks
3.SPM

Figure 4.2. Scheme of the computational strategy used to characterize the molecular basis of the millisecond allosteric activation of wtIGPS. Initial conventional molecular dynamics (cMD) and accelerated molecular dynamics (aMD) simulations were performed starting from the *X-ray* structure of IGPS found in the inactive state (PDB:1GPW (chains A/B)). This was followed by aMD simulations to study the spontaneous *L-Gln* binding to the HisH active site, and to capture *wt*IGPS in the active state. Well-tempered metadynamics simulations were performed to estimate the free energy surface of the activation process, and, finally, dynamic network tools (SPM) were applied to identify the most relevant residues involved in the allosteric communication between subunits.

During the elaboration of this project, Wurm *et al.,* crystallized the allosterically activated conformation of *h*C84A IGPS mutant bound to *L-Gln* substrate and an allosteric effector which is a precursor of PRFAR.[151] The *h*C84A mutation catalytically inactivates IGPS. As also indicated by our simulations (see section 4.3.4), the *X-ray* *h*C84A IGPS structure shows a closed HisF:HisH interface with the HisH active site with the *h*V51 oxyanion hole formed. These experiments provide evidence to the allosteric activation that we computationally characterized independently and described in this Chapter. As we show in the last section 4.4, this computational protocol can be harnessed to unravel the impact of mutations on allosteric regulation, which harbors essential information for enzyme design and drug discovery.

4.2. Computational Details and Protocols

The complete details of the computational methods, system set up, and analysis tools used in this Chapter can be found in Appendix A. However, a summary of the most relevant information is provided in this section.

MD simulations of IGPS in the absence of glutamine were performed in two different states: without PRFAR (*apo* IGPS) and in the presence of the allosteric effector PRFAR (PRFAR-IGPS). Spontaneous substrate (glutamine) binding simulations were performed in both the absence (PRFAR-free, *L-Gln* bound) and presence of PRFAR (IGPS ternary complex). All simulations started from the inactive IGPS conformation corresponding to chains A and B of PDB 1GPW. In this PDB, the *h*49-PGVG oxyanion strand presents an inactive conformation where the *h*V51 oxyanion hole is not formed in the HisH active site and the HisF:HisH interface adopts a partially open state with an interface angle of *ca*. 25°.

From this starting structure, we used the following computational protocol to capture the allosteric activation of IGPS. First, we studied how the binding of PRFAR alters the oxyanion strand conformational dynamics using microsecond time scale conventional MD (cMD) simulations. These MD simulations were used to characterize microsecond time scale motions of IGPS in the *apo* and PRFAR-bound states providing the conformational ensemble of the oxyanion strand. In parallel, we performed accelerated molecular dynamics (aMD) simulations to explore local and global millisecond motions of IGPS in the absence of the substrate. Then, we clusterized cMD simulations to extract representative structures corresponding to the most relevant conformational states of the *h*49-PGVG oxyanion strand. These different conformations were employed as a starting point for spontaneous glutamine substrate (*L-Gln*) binding aMD simulations in both PRFAR-free and PRFAR-bound states. Our goal with these aMD simulations is to characterize the spontaneous formation of the active ternary complex which comprises the substrate-binding process and is followed by the complete millisecond allosteric activation that results in the pre-organized HisH active state. Subsequently, the most relevant structures sampled during these aMD simulations in the PRFAR-free and PRFAR-bound states were used as starting points for well-tempered metadynamics (WT-MetaD) simulations to recover the free-energy landscape of the oxyanion strand rotation associated with the allosteric activation. In total, we selected 10 representative conformations that present characteristic global and local features of allosterically inactive and active IGPS states for WT-

MetaD. Finally, to explore the role of correlated motions during the allosteric activation, we used the shortest path map (SPM) tool to analyze the changes of dynamical networks along the allosteric activation process. We dissected the aMD trajectory that captures the completed allosteric activation to calculate the SPM in time spans of 600 ns considering all Cα of the protein to build the dynamical network. The structures of most relevant functional states and MD trajectories were uploaded at https://github.com/ccalvotusell/igps.

4.3. Results

4.3.1. How PRFAR Binding Effects IGPS Conformational Dynamics: Structural Characterization of Transient hV51 Oxyanion Hole Formation in HisH

The hypothesis based on NMR experiments is that the binding of the effector PRFAR in HisF subunit alters the conformational dynamics of HisH *h*49-PGVG oxyanion strand motif, which eventually makes accessible the catalytically competent HisH active site with the characteristic *h*V51 oxyanion hole formed. To unravel the effect of PRFAR binding, we carried out microsecond conventional molecular dynamics (cMD) simulations of IGPS without the presence of the substrate in the *apo* state (PRFAR effector and *L-Gln* not present) and PRFAR-bound (*L-Gln* not present). All simulations were started from the inactive IGPS conformation (see Appendix A1 for further details). Since we aimed to understand how PRFAR (HisF) impacts the HisH active site, we use the following terminology throughout this Chapter: *L-Gln* will be considered as the substrate and PRFAR as the allosteric effector. Conventional MD simulations indicate that PRFAR significantly alters both the orientation and dynamism of the HisH *h*49-PGVG oxyanion strand, even when the substrate is not present (see sections 4.3.3-4.3.5).

We reconstructed the conformational landscape of the oxyanion strand conformational dynamics using ten replicas of 1.5 μs each of cMD simulations for both *apo* and PRFAR-IGPS. To illustrate the relevant *h*49-PGVG conformations, we used the ϕ dihedral angles of *h*V51 and *h*G50 oxyanion strand residues (Figures 4.3a, A2, and A3). In contrast to the rigidity observed in available *X-ray* structures of wild-type IGPS, cMD simulations indicate that the oxyanion strand is significantly flexible and can sample different orientations: three major conformations are observed when PRFAR is bound (inactive, active, and unblocked) while only two in the *apo* state (inactive and unblocked). In both cases, the most populated conformation corresponds to the inactive-OxH state (indicating that the HisH active site is not pre-organized), which is the *X-ray* like conformation. In the inactive-OxH state, the oxyanion hole is not formed because the amide H^N backbone of *h*V51 is pointing toward *h*P10 located at the Ω-loop. A stable hydrogen-bond is established between the carbonyl of *h*P10 and the amide H^N of *h*V51 that prevents the rotation of the *h*V51 backbone to form the oxyanion hole. Beside the Inactive-OxH state, both *apo* and PRFAR-bound states show another conformation not observed in any *X-ray* that presents the oxyanion strand unblocked (Figure 4.3b). In the unblocked-OxH state, ϕ-*h*G50 partially rotates providing some flexibility to the ϕ-*h*V51 dihedral. The oxyanion strand flexibility is enhanced because the *h*P10–*h*V51 hydrogen bond completely breaks separating the oxyanion strand from the Ω-loop (Figure A4). The observed interconversion between inactive-OxH and unblocked-OxH states in the microsecond cMD simulations, indicates that this event occurs in the microsecond time scale (Figure A2).

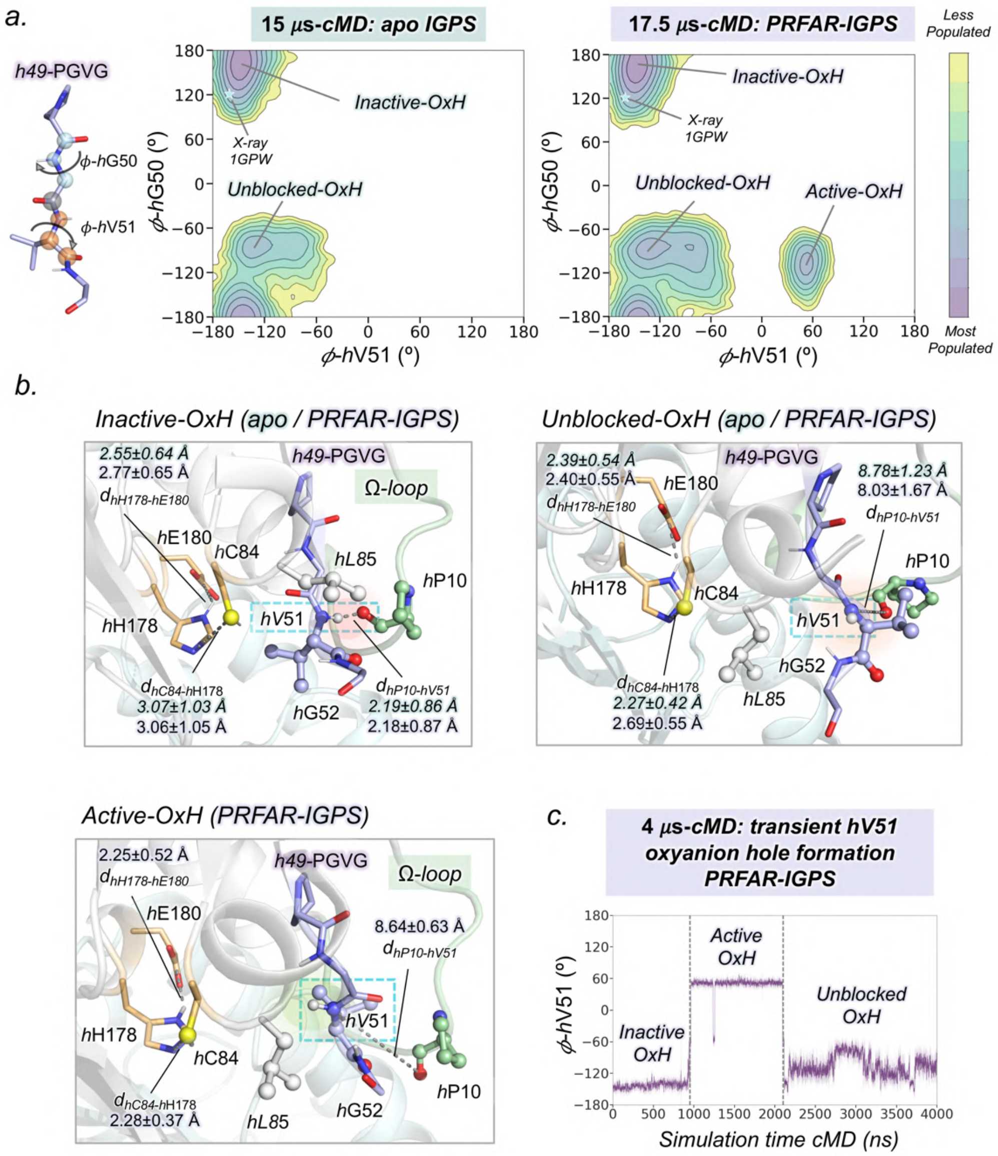


Figure 4.3. Conformational landscape of h49-PGVG oxyanion strand obtained from conventional molecular dynamics (cMD) simulations of substrate-free IGPS. a) Conformational landscape of substrate-free *apo* (in the absence of both PRFAR and *L-Gln*) and PRFAR-IGPS (in the absence of *L-Gln*) constructed using the *ϕ* dihedral angles of *h*V51 and *h*G50. The cyan star symbol indicates the *h*V51 and hG50 of *X-ray* IGPS structure (1GPW chain A/B) used as starting point in cMD simulations. **b)** Representative HisH active site structures of most populated states in the PRFAR-IGPS conformational landscape (2a). The NH backbone of hV51 is highlighted inside a cyan dashed box. Average distances (in Å) are depicted in green and purple for *apo* and PRFAR-bound states, respectively. **c)** Transient hV51 oxyanion hole formation observed in a cMD trajectory of 4 µs.

Interestingly, when PRFAR is bound, the conformational landscape indicates that an additional state is explored with respect to *apo* IGPS. We call this conformation the active-OxH state because it involves the complete rotation of ϕ-hV51 dihedral with respect to the inactive IGPS structure used as starting point (Figure 4.3a,b). More importantly, these results indicate that, even when the *L-Gln* substrate is not present, PRFAR-IGPS can access the conformation with the *h*V51 oxyanion hole formed. The active-OxH state of PRFAR-IGPS shows significant similarity with the recently reported *X-ray* structure of the substrate bound *h*C84A IGPS variant. Also, this state resembles the active site arrangements observed in other GATases that present an equivalent oxyanion hole formed (Figure A5).[149,160,161] In more detail, the active-OxH is characterized by the amide H^N backbone of *h*V51 oriented toward the catalytic *h*C84, thus, pre-organizing the active site for catalysis. Moreover, the *h*P10–*h*V51 hydrogen bond is clearly broken and the interactions between the residues forming the catalytic triad are strengthened with respect to the inactive-OxH and unblocked-OxH conformations (Figures 2b and A4). In both active-OxH and unblocked-OxH states, the side chain of *h*V51 leaves the active site region due to the oxyanion strand reorientation. However, this extra space in the HisH active site is occupied by the bulky side chain of *h*L85 that blocks the access to the catalytic *h*C84 (Figures 4.3b and A6). When the active-OxH is sampled, subtle HisF:HisH and HisF interface structural and dynamical changes are observed that may contain significant features of the allosteric activation (see Appendix A1 text and Figures A2–A10 for a general description of HisF and HisH conformational dynamics in *apo* and PRFAR-bound IGPS).

In PRFAR-bound cMD simulations, we observed the *h*V51 oxyanion hole formation occurs only in 1 out of 10 replicas, pointing out that the pre-organization of the HisH active is a rare event, at least in the microsecond time scale of the cMD simulations. To gain more knowledge on the *h*V51 oxyanion hole formation, we extended the cMD trajectory that captures the complete rotation of the oxyanion strand up to 4 µs (Figures 4.3c and A7). The analysis of the individual 4 µs trajectory revealed that a transient *h*V51 oxyanion hole formation (active-OxH) took place after 1 µs of simulation time, remained formed during 1 µs of simulation time, and subsequently interconverted to the unblocked-OxH state. Considering the results of all µs-cMD of PRFAR-IGPS, the active-OxH conformation is probably a high-energy state in the conformational landscape (see Section 4.3.4. for a quantitative description based on well-tempered metadynamics). Overall, the results obtained from µs-cMD simulations indicate that the hypothesized catalytically competent HisH active site pre-exists in solution for *wt*IGPS in the presence of PRFAR and in the absence of *L-Gln* substrate. These results do not discard that the pre-organized active site also exists in the *apo* IGPS state. However, with μs-cMD simulations we cannot completely describe the millisecond time scale events characteristic of IGPS allosteric communication.

4.3.2. IGPS can adopt a Closed HisF:HisH Interface in the presence of PRFAR

To explore the impact of PRFAR binding in slower time scales, we carried out accelerated molecular dynamics (aMD) simulations for both IGPS *apo* and PRFAR-bound states (in total 10 replicas of 1 μs, see Appendix for more details). As mentioned in the methodology chapter, aMD is an enhanced sampling technique that provides unconstrained sampling without using *a priori* information of the reaction coordinate. Based on previous works, with microsecond aMD simulations it is possible to access millisecond time scale events characteristic of allosteric transitions.[162,163] The aMD simulations in the presence of PRFAR captured multiple infrequent and transient rotations of the oxyanion strand to form the *h*V51 oxyanion hole, indicating that the active-OxH state of IGPS is explored. On the other hand, a significantly lower number of oxyanion hole formations are observed in the *apo* state (Figures A11 and A12). In line with previous cMD simulations, aMD simulations seem to indicate that the active-OxH state is a high energy state in comparison with the inactive-OxH state in the IGPS conformational ensemble without the presence of the substrate (see Section 4.3.4. for a more detailed description using metadynamics).

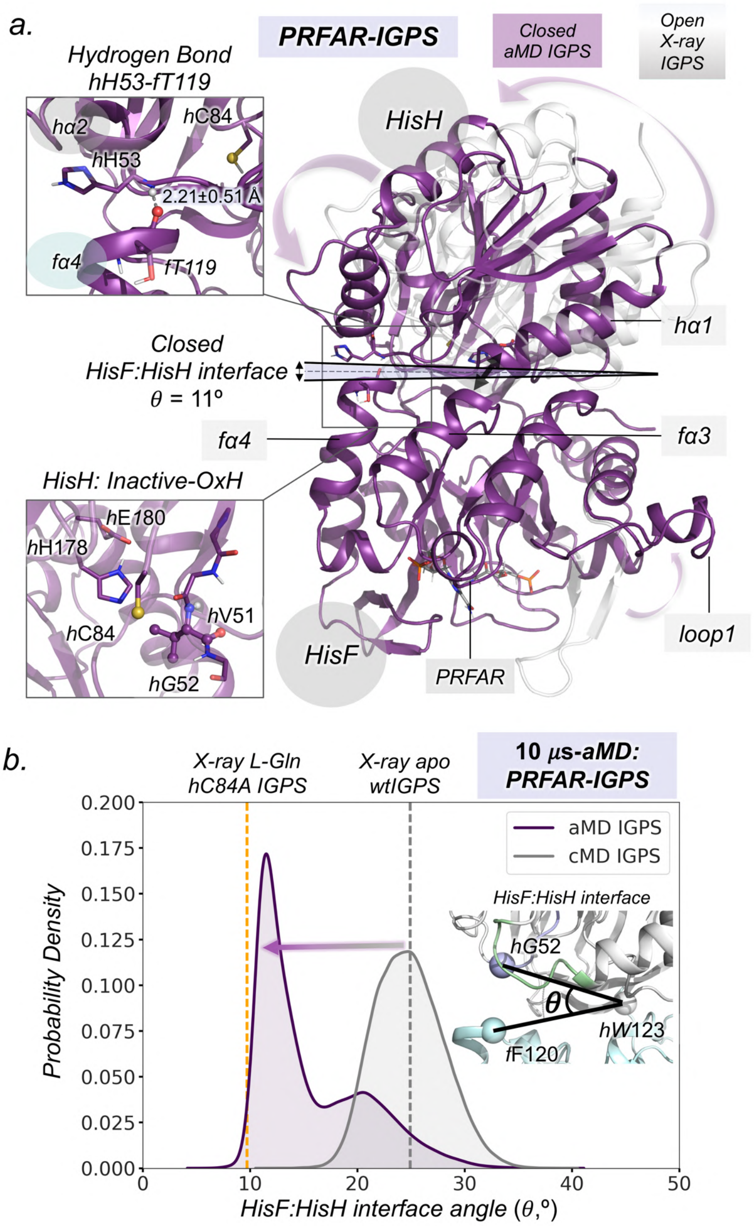


Figure 4.4. *L-Gln*-free PRFAR-IGPS Accelerated Molecular Dynamics (aMD) simulations: Identification of IGPS productive closure. a) Structural comparison of open (in gray, PDB 1GPW chains A/B) and closed HisF:HisH interfaces obtained from aMD simulations (in purple). The hydrogen bond between the backbones of *h*H53 and *f*T119 that stabilizes the closed HisF:HisH interface and the conformation of the HisH active site is depicted. **b)** Probability density distribution for HisF:HisH interface angle obtained in cMD and aMD simula-tions of PRFAR-IGPS. The angle (θ) of the HisF:HisH inter-face is calculated from the alpha-carbons of fF120, hW123 and hG52. The

vertical dashed orange and gray line corresponds to the to the hC84A IGPS (PDB: 7AC8 (chains E/F)) and *wt*IGPS (PDB:1GPW (chains A/B)) *X-ray* HisF:HisH interface angles, respectively.

Moreover, we explored global conformational changes in IGPS in these aMD simulations. In particular, we observed that in the simulations with PRFAR-bound the HisF:HisH interface unlocks triggering both the rotation and closure of the interdomain region (Figures 4.4 and A13). The conformational ensemble is displaced toward closed states of IGPS with respect to cMD simulations. In inactive IGPS crystal structures, the HisF:HisH interface angle (θ, defined between the Cα of *f*F120, *h*W123, and *h*G52) takes values around 25°. In aMD simulations, we explored a closed metastable state that shows an average HisF:HisH interface angle of about 11° (in the *X-ray* corresponding to the substrate-bound *h*C84A IGPS reported recently by Wurm and coworkers, the angle of the HisF:HisH interface is approximately 9.5°).[161] This closed state is particularly stabilized by a hydrogen bond that is formed between the backbones of one HisH residue, *h*H53, and one HisF residue, *f*T119, that is able to slow down the HisF:HisH opening–closing motion. Additionally, in the closed state the HisH Ω-loop moves close to the top of *f*α3 at the same time that the alignment of *h*α1 and *f*α3 takes place. These collective motions increase the communication between HisF and HisH subunits, which was previously suggested to be essential for glutamine hydrolysis. Another relevant aspect observed in substrate-free aMD simulations is that the closure of the HisF:HisH interdomain region is not correlated with the formation of the *h*V51 oxyanion hole, *i.e.* when the closed state is attained the oxyanion hole is not formed and vice-versa (Figure A14). We also sampled the infrequent formation of closed states in the IGPS *apo* state simulations. This indicates that the HisF:HisH interface closure is not an exclusive allosteric effect of PRFAR binding (Figure A13).

4.3.3. Exploring the Substrate Binding Process in IGPS: L-Glutamine Binds the Inactive-OxH State in Both PRFAR-Free and PRFAR-Bound IGPS

The question now is how the characteristic μs-ms conformational dynamics shown by IGPS in the absence of the substrate is coupled to the binding of *L-Gln* substrate to attain the allosterically active ternary complex (IGPS + PRFAR + *L-Gln*). As in *V*-type allosterically regulated enzymes, IGPS presents similar $K_M^{L\text{-}Gln}$ values in the mM range for both PRFAR-free and PRFAR-bound IGPS. This points out that binding of *L-Gln* to the HisH active site occurs in both states but that infrequent binding events are expected if low concentration of *L-Gln* is used. To explore the binding pathways of *L-Gln*, we performed spontaneous substrate binding aMD simulations. To couple the spontaneous binding with intrinsic IGPS conformational dynamics, we designed a strategy that includes the following steps: first, a single *L-Gln* molecule is positioned approximately 25-30 Å away from the catalytic *h*C84 active site residue; and, second, multiple replicas of aMD simulations are performed starting from different IGPS conformations sampled in the previous cMD simulations, including the active-OxH, inactive-OxH, and unblocked-OxH states (Figures 4.5a, A15, and A16). A total of 60 replicas of 600 ns of unconstrained aMD simulations were performed (gathering a total accumulated time of 36 μs) starting from *L-Gln* in the solvent. In both PRFAR-free (*apo* IGPS + *L-Gln*) and PRFAR-bound IGPS (IGPS + PRFAR + *L-Gln*), we observed the spontaneous binding of *L-Gln* to the HisH active site without applying any constraint.

To assess how the binding of *L-Gln* in the HisH active site depends on the conformational dynamics of the *h*49-PGVG oxyanion strand, we represented all spontaneous binding aMD simulations in a conformational landscape based on two simple coordinates: first, the nucleophilic attack distance (d_{nuc}) between the catalytic *h*C84 (*i.e.* sulfur atom of the thiol group) and *L-Gln* (*i.e.* reactive amide carbon); and, second, the ϕ-*h*V51 dihedral angle that is the one that best differentiates between the inactive-OxH and active-OxH states of the oxyanion strand (Figure 4.5b). The conformational landscape of the binding process indicates that the recognition of the substrate (d_{nuc} above 12 Å) at the entrance of the HisF:HisH interface represents the bottleneck of *L-Gln* binding in both PRFAR-free and PRFAR-bound simulations. The high polarity of *L-Gln* (high solubility in water) and the open-closed transitions of the HisF:HisH interface are responsible for slowing down the recognition of the substrate. Once recognized in the HisF:HisH interface entrance, the binding of *L-Gln* (d_{nuc} below 6 Å) depends on the orientation of the oxyanion strand. Remarkably, binding only takes place when the *h*49-PGVG oxyanion strand is found in the inactive-OxH conformation (ϕ-hV51 in the range of [−180°, −100°]). Of note is that this occurs both in PRFAR-free and PRFAR-bound simulations.

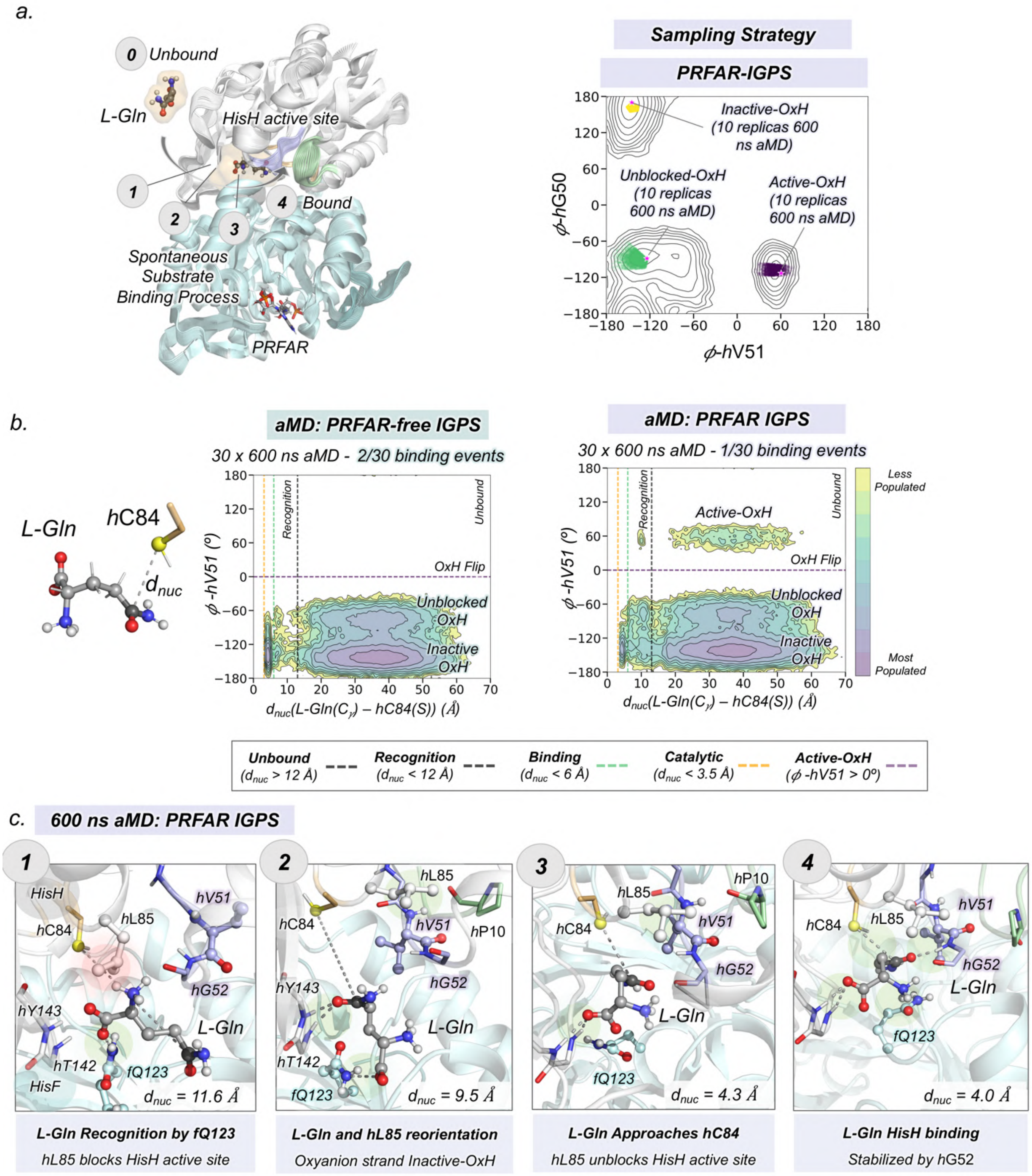


Figure 4.5. Molecular basis of *L-Gln* binding in IGPS. a) General scheme of spontaneous *L-Gln* substrate binding process in the PRFAR-free and PRFAR-IGPS states. The numbers indicate the most relevant steps of the binding process. **b)** Conformational landscape obtained from the nucleophilic attack distance (d_{nuc}) between the thiol group of catalytic *h*C84 (in yellow) and the amide carbon of *L-Gln* (in black), and the ϕ dihedral angle of *h*V51. The purple horizontal dashed line

indicates the oxyanion strand flip, and the gray, green and orange vertical dashed lines indicate the distance where recognition, binding, and catalysis take place. We consider that *L-Gln* is captured by HisH active site when d_{nuc} is below 6 Å, while catalytically productive distances will be only sampled when both the d_{nuc} is shorter than 3.5 Å and the active-OxH state (ϕ-hV51 *ca.* 60°) is attained. **c)** Molecular representation of selected key conformational states of the *L-Gln* binding pathway. The substrate is shown in gray, the oxyanion strand residues in purple, the catalytic residues in orange, the Ω-loop in green and the other HisH and HisF residues in white and cyan, respectively.

To identify the sequence of events that determine *L-Gln* binding, we analyzed the independent aMD trajectories that reconstructed the complete binding event (Movies A2 and A3, see Appendix A). By visually inspecting a representative PRFAR-bound IGPS aMD simulation, we determined the key steps of the *L-Gln* binding pathway as indicated in Figures 4.5c and A17–A19. In this particular simulation, the starting IGPS conformation presents an active-OxH oxyanion strand conformation and an open HisF:HisH interface. To recognize the *L-Gln* substrate, the HisF:HisH interface of IGPS significantly opens (interface angle up to 30°). Once captured in the HisF:HisH interface, the amide side chain of interface residue *f*Q123 interacts with the carboxylate group of *L-Gln* (step 1). Since the oxyanion strand remains in the active-OxH conformation, the substrate approach to the catalytic *h*C84 is prevented by the side chain of *h*L85. Since *h*L85 blocks the entrance to the HisH active site the nucleophilic attack distance between the amide carbon of *L-Gln* and the sulfur of catalytic *h*C84 is near 10 Å, which is too long for facilitating the nucleophilic attack. With the substrate still in the HisH active site, the oxyanion strand changes its conformation from the active-OxH to inactive-OxH state. As observed before in the cMD simulations, in the inactive-OxH the *h*L85 side chain is displaced from the HisH active site, which facilitates the reorientation of *L-Gln* within the pocket. In this new orientation, the backbones of *h*T142 and *h*Y143 residues interact with the carbonyl of *L-Gln* (step 2 in Figure 4.5c, see Figures A18 and A19). Since the oxyanion strand remains in the inactive-OxH state, the substrate rapidly repositions to move closer to the *h*C84 catalytic residue by extending the side chain along the HisH active site (steps 3 and 4).

At this point, we consider that *L-Gln* is bound in the inactive-OxH HisH active site (step 4). In this substrate-bound pose, the carbonyl of the *L-Gln* side chain establishes a hydrogen-bond with the H^N backbone of *h*G52 (2.47 ± 1.07 Å). However, in this pose the nucleophilic attack distance between the *h*C84 and the amide carbon of *L-Gln* is still too long (4.72 ± 0.67 Å) to promote glutamine hydrolysis. Additional interactions essential to keep in the substrate in the active site include the interactions between the side chains of *h*L85 and *f*Q123 and the side chain of *L-Gln*. That matches with experiments that show how the mutation of *f*Q123 residue affects the K_M values. Overall, glutamine hydrolysis cannot occur because the oxyanion strand is in the inactive-OxH conformation. Therefore, attaining the active-OxH seems a requirement to reach the short nucleophilic attack distances required for efficient catalysis (see next section).[164,165]

The *L-Gln* pose predicted and characterized from spontaneous binding aMD simulations perfectly matches available *X-ray* structures of IGPS crystallized without the presence of PRFAR and with the presence of the substrate (Figure A22). Indeed, we obtained the same substrate-bound pose in PRFAR-free simulations (Figures A20–A22). Interestingly, we did not observe unbinding of *L-Gln* upon the formation of the hydrogen bond between *L-Gln* and *h*G52. More importantly, when *L-Gln* binds the inactive-OxH state, the oxyanion strand does not reorient to form the active-OxH state within the 600 ns aMD simulation. This shows that the complete allosteric activation has still not occurred, which means that longer simulations will be required in line with the millisecond motions reported with NMR experiments.

Importantly, spontaneous binding aMD simulations show that the binding of PRFAR in the HisF subunit does not modify the initial steps toward the formation of the allosterically active IGPS ternary complex because similar steps of *L-Gln* binding are captured in both PRFAR-free and PRFAR-bound. Still, one question remains unanswered: once both *L-Gln* and PRFAR are bound in IGPS, which is the sequence of events that allosterically activate IGPS?

4.3.4. IGPS Caught in the Active State: Time Evolution toward the Millisecond Allosteric Activation of IGPS Ternary Complex

To reconstruct the complete allosteric activation of IGPS, we extended the spontaneous binding aMD simulations that captured the substrate-bound pose (with *L-Gln* bound in an inactive-OxH oxyanion strand conformation) in both PRFAR-free and PRFAR-bound systems. Based on NMR experiments, the allosteric activation of IGPS ternary complex is associated with the activation of correlated millisecond motions.[145] Therefore, our goal is also to understand how the formation of the IGPS ternary complex controls the oxyanion strand conformational dynamics and the IGPS conformational ensemble in comparison to the uncorrelated motions observed in the substrate-free simulations (see section 4.3.2 and Appendix A).

To capture and characterize the active state, we required five replicas of aMD simulations that accumulated a total of 30 μs (Figures 4.6, A23, and A24) starting from the substrate-bound pose. In these simulations, we evaluated the formation of the *h*V51 oxyanion hole in the presence of the substrate by monitoring the orientation of the *h*49-PGVG oxyanion strand along the aMD simulations. The aMD simulations show that the *h*V51 oxyanion hole formation is accessible, frequent, and long-lived when both the *L-Gln* and PRFAR are bound at their respective HisH and HisF active sites (Movie A4).

We analyzed the independent aMD trajectories to describe the sequence of events that occur upon *L-Gln* binding being able to determine the allosteric activation mechanism of IGPS in the ternary complex (Figures 4.6a and A25–A27). Remarkably, we observed that the formation of the *h*V51 oxyanion hole is always preceded by the productive closure of the HisF:HisH interface. After *L-Gln* binding in the inactive-OxH HisH active site, the HisF:HisH interface angle evolves from 20 to 15°. Despite this partial closure, the nucleophilic attack distance remains around 4.5 Å, indicating that the active site is still not properly pre-organized. After 1 μs of aMD simulation time, the HisF:HisH interface attains the productively closed state with HisF:HisH angles below 13° which decreases the nucleophilic attack distance to 4 Å. At this point, the oxyanion strand remains in the inactive-OxH state. Due to the productive closure, the side chain of *f*Q123 reorients to interact with *L-Gln*, which enhances HisF:HisH intersubunit communication through the *L-Gln* substrate. Importantly, the HisF:HisH interface closure significantly increases the intrinsic flexibility of the oxyanion strand triggering the rotation of the oxyanion strand and the formation of the ϕ-*h*V51 oxyanion hole. Upon the formation of the oxyanion hole, *L-Gln* moves closer to the catalytic *h*C84 and consequently the nucleophilic attack distance decreases below 3.6 Å (see Figure 4.6a). The HisF:HisH interface remains closed during the simulation time indicating a slow closed-to-open transition when both *L-Gln* and PRFAR are bound (Figure A28). Therefore, the coupled effect of the substrate and effector contribute to stabilize the closed conformation of IGPS, which might help to retain *L-Gln* in the active site for efficient catalysis and favor the transfer of NH_3 through the internal HisF tunnel for the subsequent cyclase activity in HisF.

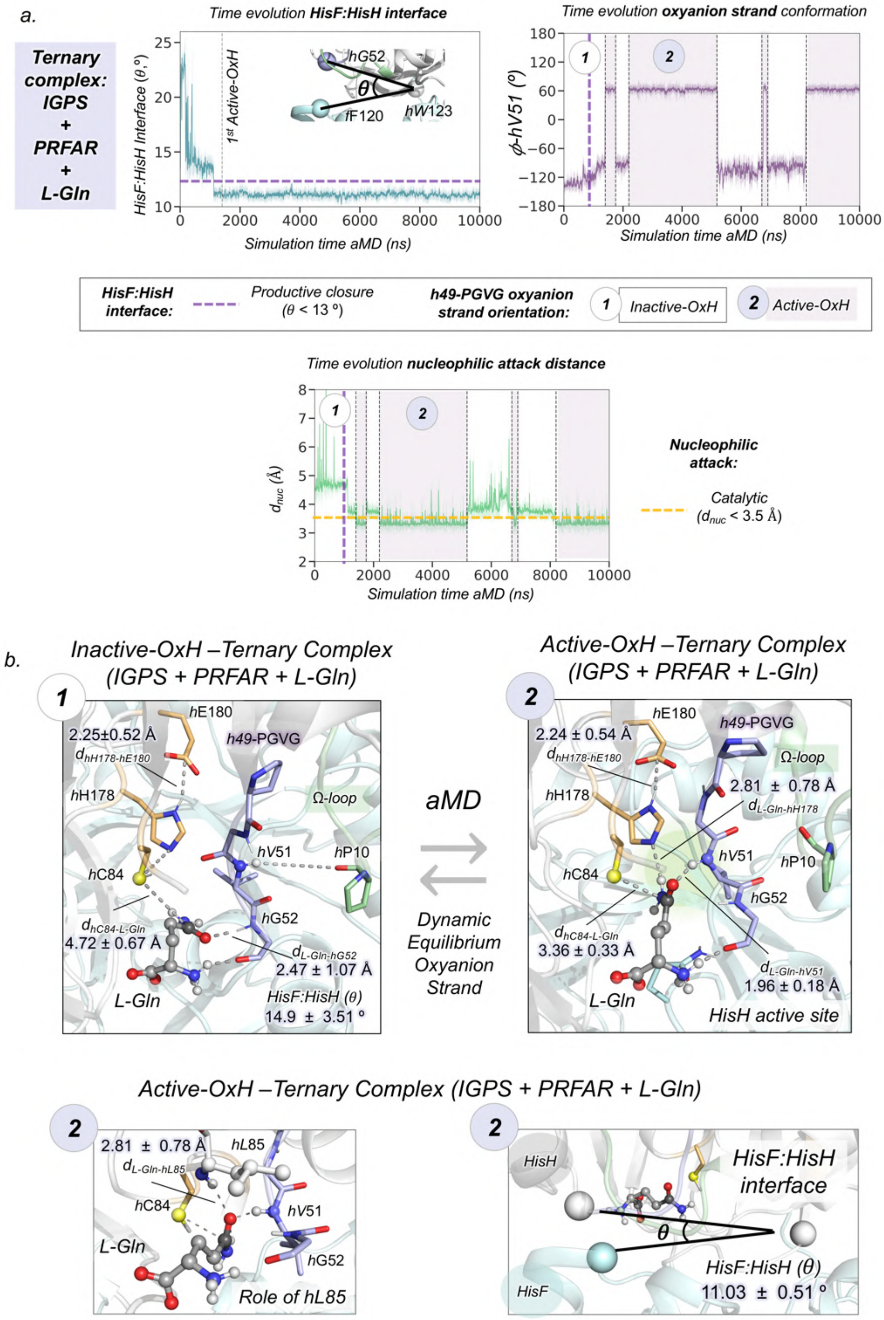


Figure 4.6. Allosteric activation of IGPS in the ternary complex. a) Plot of the HisF:HisH interface angle along the 10 µs-aMD simulations. Plot of the *h*V51 dihedral angle along the 10 µs-aMD simulations. Plot of the distance corresponding to the nucleophilic attack along the 10 µs-aMD simulations. The purple dashed line indicates when productive closure takes place, purple regions

indicate when Active-OxH state is populated, and white regions indicate when Inactive-OxH state is populated. **b)** Representative structures of the Inactive-OxH and Active-OxH states sampled in the aMD simulations of the IGPS ternary complex. The role of *h*L85 is shown for Active-OxH.

The pre-organized HisH active site with both the *L-Gln* bound and the *h*V51 oxyanion hole formed presents a similar structural rearrangement as the one characterized in the substrate-free form (Figures 4.3c and 4.6b). In the active ternary complex, an extensive network of non-covalent interactions is established between the substrate and the catalytic, HisF, HisH, and oxyanion strand residues. aMD simulations indicate the concerted formation of the *h*V51 oxyanion hole and reorientation of the *L-Gln* in the HisH active site. In the substrate-bound pose when the active-OxH is attained, the carbonyl oxygen establishes a hydrogen bond with the the H^N backbone of *h*V51 (1.96 ± 0.18 Å) instead of *h*G52 (now at 3.87 ± 0.55 Å). Moreover, the amido group of *L-Gln* interacts with catalytic *h*H178 (2.81 ± 0.78 Å). Considering all these rearrangements, the electrophilic carbon of *L-Gln* approaches the nucleophilic thiol group of *h*C84, with catalytically competent distances of 3.36 ± 0.33 Å. The rotation of the oxyanion strand exposes the H^N backbone of hL85, which establishes an additional hydrogen bond with the carbonyl of *L-Gln* (2.47 ± 0.39 Å) that creates the complete oxyanion hole that will stabilize the transient negative charges formed in the tetrahedral intermediate. Overall, the transition from inactive-OxH to active-OxH in the presence of the substrate significantly increases and strengthens the non-covalent interactions between *L-Gln* and active site residues (Figure A25), facilitating the nucleophilic attack, proton transfer, and stabilization of the tetrahedral intermediate required for efficiently hydrolysing glutamine. Remarkably, aMD simulations unraveled, without using *a priori* information, an active-OxH conformation of *wt*IGPS that show significant similarities with the *a posteriori* reported *h*C84A IGPS *X-ray* structure corresponding to the postulated allosterically active state of IGPS (Figure A29).[161] Thus, μs-aMD simulations unveiled the catalytically competent pose corresponding to the allosterically active ternary complex of *wt*IGPS.

By analyzing the 10 μs aMD simulation, we observed that the *h*V51 oxyanion hole forms up to 4 times within the simulation time (Figure 4.6a). This indicates that IGPS transitions between the active-OxH and inactive-OxH of the oxyanion strand, when the productive HisF:HisH closure has taken place, which points out a lower energy barrier for the oxyanion strand interconversion than in the PRFAR-free and substrate-free forms. More importantly, when both

L-Gln and PRFAR are bound the active-OxH conformation is clearly stabilized in comparison to the infrequent transient formations observed when only PRFAR is bound (in the absence of *L-Gln* substrate). Therefore, aMD simulations show that a population shift toward the active state takes place only when the ternary complex is assembled (Figure A30). Additionally, we explored the formation of the *h*V51 oxyanion hole in one out of five replicas of PRFAR-free aMD simulations (Figure A23). Interestingly, once the active-OxH is attained it remains in this state for the rest of the simulation time (up to 7 μs of aMD), thus suggesting a higher interconversion barrier when PRFAR is not present. The results obtained with aMD were validated using unconstrained Gaussian accelerated Molecular Dynamics (GaMD) simulations. Interestingly, GaMD simulations provided the same mechanism as in aMD simulations for the allosteric activation of IGPS ternary complex (Figure A31).

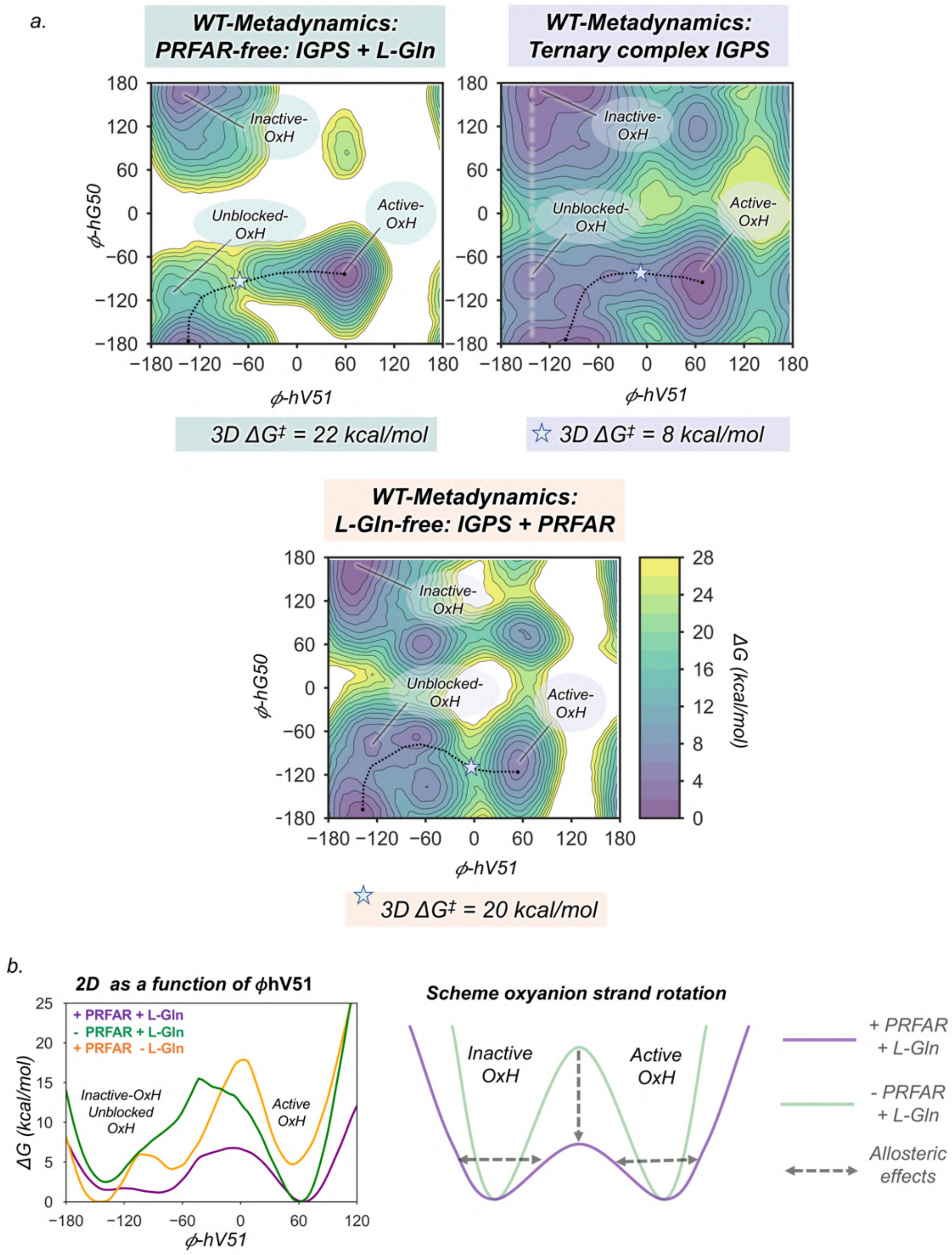


Figure 4.7. Free energy landscape reconstruction of allosteric activation. a) Free energy landscape (FEL) of the *h*49-PGVG in the PRFAR-free (only *L-Gln* bound), ternary complex (PRFAR and *L-Gln* bound), and *L-Gln*-free (only PRFAR bound) obtained from well-tempered metadynamics simulations. The star symbol represents the energy barrier between Inactive-OxH and Active-OxH states of the FEL. **b)** 2D free energy landscape of the Inactive-OxH/Unblocked-OxH to Active-OxH section) of *ϕ*-*h*V51 for *L-Gln*-free PRFAR-IGPS (+PRFAR -*L-Gln*, in orange), *L-Gln* bound PRFAR-free (-PRFAR + *L-Gln*, in green), and ternary complex (IGPS+PRFAR+*L-Gln*, in purple.). The 2D free

energy profile along *ϕ*-*h*V51 is calculated from the Boltzmann populations of all *ϕ*-*h*G50 at a certain value of *ϕ*-*h*V51 (as indicated with the light purple line). Since the inactive-OxH and unblocked-OxH states are found in the same range of values of *ϕ*-*h*V51, the 2D plot cannot differentiate them. The presence of PRFAR and *L-Gln* decreases the interconversion barrier and broadens the energy minima. Representative scheme of the oxyanion strand interconversion barrier in the PRFAR-free (green) and ternary complex (purple) systems.

We additionally carried out well-tempered metadynamics (WT-MetaD) simulations to unravel the underlying free-energy surface of the oxyanion strand reorientation in the presence of *L-Gln* in both PRFAR-free and PRFAR-bound states. WT-MetaD were performed using the two *ϕ*-hV51 and *ϕ*-hG50 dihedral angles as collective variables (Figure 4.7 and Appendix). To reconstruct the free-energy surface we used the multiple-walkers approach considering ten arbitrarily selected conformations (walkers) taken from the aMD simulations that present global and local features of the both inactive-OxH and active-OxH states in the IGPS ternary complex (Figures A32–A34). From these WT-MetaD simulations, we were able to determine the free-energy landscape (FEL) of the *h*49-PGVG oxyanion strand conformational dynamics (Figure 4.7a,b). Remarkable differences are observed between the FEL of the PRFAR-free and ternary complex (PRFAR+*L-Gln*) states. In the ternary complex, the interconversion barrier between the inactive-OxH and active-OxH states is of approximately 8 kcal/mol, while in the case of PRFAR-free, the barrier increases up to 22 kcal/mol. The FELs clearly show that the formation of the *h*V51 oxyanion hole is significantly slower when PRFAR is not present. Another important aspect that can be extracted from the FEL is that the relative stability between the inactive-OxH and active-OxH is preserved. This result suggests that both orientations of the oxyanion strand are similarly populated in the IGPS ternary complex. This facile interconversion can be important along the IGPS catalytic cycle. To better understand the coupled effect of the PRFAR and *L-Gln,* we calculated the FEL of the oxyanion strand interconversion only in the presence of PRFAR (see Figure 4.7a,b, orange, and Figure A33). As qualitatively observed in aMD simulations, PRFAR enhances the dynamism of the oxyanion strand with respect to the PRFAR-free, however, the absence of the substrate causes the destabilization of the active-OxH state (by 4 kcal/mol). Using this information, we can suggest that *L-Gln* induces a population shift toward active IGPS conformations (Figure A33). In addition, the energy barrier of the inactive-to-active transition is significantly decreased in the presence of both PRFAR and *L-Gln*. Taking all these results into consideration the coupled effect of both PRFAR and *L-Gln* is essential to promote the fully active conformation in IGPS under turnover conditions. Moreover, these results point out strong communication between both binding sites in the IGPS ternary complex (see the next section).

4.3.5. The Formation of the Active State Activates Correlated Motions in the Ternary Complex: Unveiling the Allosteric Activation Mechanism of IGPS

The question now is how the coupled effect of PRFAR (HisF) and *L-Gln* (HisH) binding activates correlated motions that regulate HisF:HisH interface and oxyanion strand conformational dynamics. To explore the dynamic allosteric communication pathways established between HisF and HisH binding sites, we used the Shortest Path Map (SPM) tool.[59,166] To capture the time evolution of dynamic networks of residues displaying correlated motions along the relevant steps of the allosteric activation process, we decided to divide aMD trajectories in concatenated time spans of 600 ns (that is, from 0 to 600 ns, from 300 to 900 ns, from 600 to 1200 ns, and so on). We chose time-spans of 600 ns because they differentiate the most relevant steps of the allosteric activation process. For each time-span, we performed the SPM analysis that we name as time-evolution SPM (te-SPM). We applied the te-SPM analysis to a 5 μs aMD simulation that captured all relevant states of the IGPS allosteric activation including: substrate binding, HisF:HisH productive closure, and subsequent *h*V51 oxyanion hole formation. This analysis unveils the fine-tuning of correlated motions and dynamic networks along the allosteric activation of IGPS (Figures 4.8, A35, and A36).

When the substrate *L-Gln* is still not bound (step 1 in Figure 4.8, with PRFAR already bound), correlated motions are generally found in the HisH subunit and HisF:HisH interface. Connections between key residues for allosteric activation are observed: *f*I93, *f*D98, *f*K99, *h*P10 and *h*N15. Subsequently, when the *L-Gln* binds the HisH active site and the HisF:HisH(θ) attains a partially closed state (step 2 in Figure 4.8), increased communication between the oxyanion strand (*h*G52 and *h*H53) and interface *f*α4 residues (*f*G121 and *f*S122) is observed. The first significant change regarding correlated motions takes place with productive HisF:HisH closure (step 3, in Figure 4.8). The closure of the IGPS interface activates concerted motions between

HisF and HisH subunits and multiple pathways that connect interface residues appear. For example, pathways that connect *h*H53 to *f*L153, catalytic *h*E180-*h*K181-*f*P76-*f*I75, or the anchor *h*W123-*f*A3 naturally arise. However, it is when the *h*V51 oxyanion hole formations start occurring (step 4 in Figure 4.8) that multiple allosteric pathways connecting HisF and HisH active sites appear, indicating the existence of functional correlated motions. When IGPS is allosterically activated, the dynamic coupling of HisH residues including *h*V51, catalytic residues *h*C84 and *h*H178, with HisF PRFAR binding site residues: *f*V12, *f*L50, *f*I102, *f*L222 is observed. Indeed, the PRFAR site and the glutaminase HisH site are connected through multiple pathways such as the network of hydrophobic residues that connect HisF PRFAR binding site with HisH oxyanion strand: *f*L50, *f*F49, *f*V79, *f*V100, *f*V125, *f*Q123, *f*G121, *h*G52 and *h*V51 (Figure 4.8b). Upon the first activation, the oxyanion strand attains a dynamic equilibrium between the active and inactive conformations of the oxyanion strand (step 5, Figure 4.8) that is correlated with the appearance of new networks of residues connecting the PRFAR active site with the HisF:HisH interface (*f*D130, *f*L169, *f*I199, *f*A220, *f*R5, *f*D45, *f*D98, and *h*N12). This change of dynamical networks suggests that the communication between effector and active site takes place in both directions. From this point, multiple dynamic pathways communicate both active sites that include catalytic residues of both HisF and HisH subunits: *f*D11 and *f*D130 and *h*C84 and *h*H178 (see Figure A36).

Several of the residues identified as important in the te-SPM analysis are involved in millisecond motions in the ternary complex as reported by NMR experiments. More insights on the allosteric communication mechanism of IGPS are obtained by analyzing the changes in community networks and monitoring the molecular interactions in the HisF and HisH subunits (see Appendix and Figures A37–A39 for a complete analysis of both local changes in HisF and HisH subunits in the IGPS ternary complex).

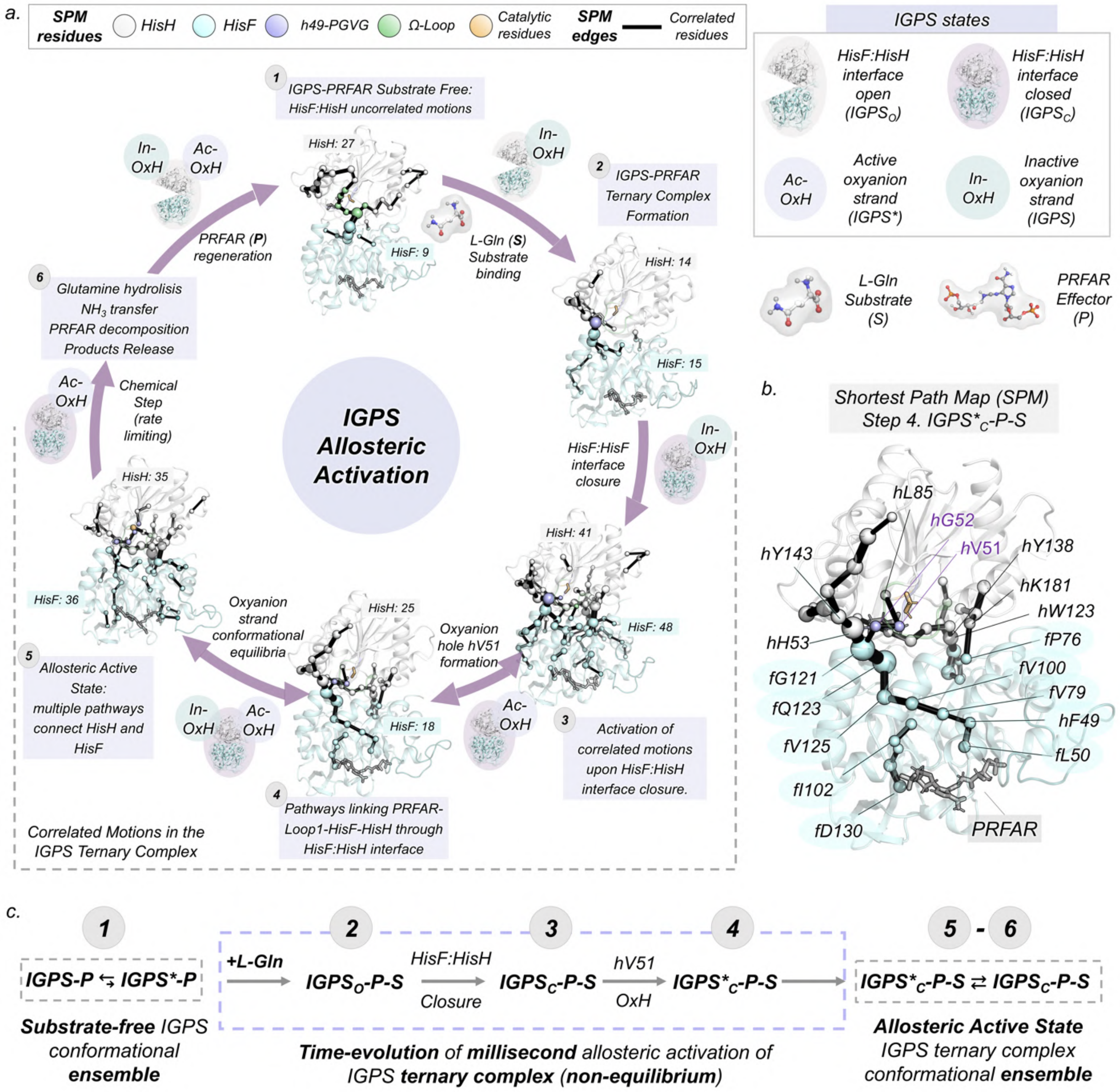

a.
SPM residues
HisH
HisF
h49-PGVG
Ω-Loop
Catalytic residues
SPM edges
Correlated residues
IGPS states
HisF:HisH interface open (IGPS$_O$)
HisF:HisH interface closed (IGPS$_C$)
Ac-OxH
Active oxyanion strand (IGPS*)
In-OxH
Inactive oxyanion strand (IGPS)
L-Gln Substrate (S)
PRFAR Effector (P)
1 IGPS-PRFAR Substrate Free: HisF:HisH uncorrelated motions
HisH: 27
HisF: 9
2 IGPS-PRFAR Ternary Complex Formation
L-Gln (S) Substrate binding
HisH: 14
HisF: 15
HisF:HisF interface closure
HisH: 41
HisF: 48
3 Activation of correlated motions upon HisF:HisH interface closure.
Oxyanion hole hV51 formation
HisH: 25
HisF: 18
4 Pathways linking PRFAR-Loop1-HisF-HisH through HisF:HisH interface
Oxyanion strand conformational equilibria
HisH: 35
HisF: 36
5 Allosteric Active State: multiple pathways connect HisH and HisF
Chemical Step (rate limiting)
6 Glutamine hydrolisis NH$_3$ transfer PRFAR decomposition Products Release
PRFAR (P) regeneration
IGPS Allosteric Activation
Correlated Motions in the IGPS Ternary Complex
In-OxH
Ac-OxH
b.
Shortest Path Map (SPM) Step 4. IGPS*$_C$-P-S
hL85
hG52
hV51
hY138
hY143
hK181
hW123
hH53
fP76
fG121
fV100
fV79
fQ123
hF49
fV125
fL50
fI102
fD130
PRFAR
c.
1
IGPS-P ⇆ IGPS*-P
Substrate-free IGPS conformational ensemble
2
3
4
+L-Gln
IGPS$_O$-P-S
HisF:HisH Closure
IGPS$_C$-P-S
hV51 OxH
IGPS*$_C$-P-S
Time-evolution of millisecond allosteric activation of IGPS ternary complex (non-equilibrium)
5 - 6
IGPS*$_C$-P-S ⇄ IGPS$_C$-P-S
Allosteric Active State IGPS ternary complex conformational ensemble

Figure 4.8. General scheme of the IGPS allosteric activation. a) Time-evolution Shortest-Path Map (SPM) analysis along the key states of the activation pathway. The sizes of the spheres and black edges are indicative of the importance of the position for the IGPS conformational dynamics. HisF (cyan), HisH (white), oxyanion strand (purple), Ω-loop (green), and catalytic (orange) residues are depicted in different colors. PRFAR, *L-Gln*, and catalytic *h*C84 are represented in sticks. The total number of residues included in the SPM of each subunit is highlighted in boxes. **b)** SPM map of the IGPS active ternary complex corresponding to step 4. **c)** General scheme of the IGPS allosteric activation. Gray and light purple boxes denote equilibrium and non-equilibrium.

4.4. Discussion

Understanding the mechanisms of allosterically regulated enzymes at the molecular level requires identifying and characterizing functionally relevant states in the ternary complex but also describing the time evolution of the dynamic conformational ensemble that connects these states.[124,130] In IGPS, the binding of an allosteric effector activates millisecond motions that are finely tuned by the binding of the substrate to promote the allosteric activation, which results in efficient glutamine hydrolysis in the HisH active site.[138,145,167] However, the fast glutamine turnover in the presence of the unstable PRFAR effector hampered the experimental detection of the allosterically active state in *wt*IGPS. From the computational perspective, it is challenging to capture the millisecond allosteric activation of IGPS. In this Chapter, we devised a computational strategy specifically designed to capture the time-evolution of millisecond time scale events that has been used to describe, step by step, the molecular mechanism of the allosteric activation of IGPS, connecting the inactive substrate-free form with the allosterically active ternary complex. Our simulations unveil a complex coupling between allosteric effector (PRFAR) and substrate glutamine (*L-Gln*) binding and the intrinsic HisF:HisH interface conformational dynamics that finely regulate the allosteric activation process of IGPS. Starting from the inactive conformation of IGPS and without using *a priori* information of the active state, our simulations spontaneously revealed a closed IGPS state that presents the HisH active site with the *h*V51 oxyanion hole formed and the substrate glutamine properly positioned in a catalytically productive pose. The insights provided by these simulations tie up the loose ends related to IGPS allosteric regulation mechanism.

We used microsecond conventional molecular dynamics (cMD) and accelerated molecular dynamics (aMD) simulations to understand how PRFAR binding impacts the substrate-free conformational ensemble in *wt*IGPS. A hidden conformation that presents the the *h*V51 oxyanion hole (active-OxH) spontaneously formed in the µs-cMD simulations, indicating that this state of the *h*49-PGVG oxyanion strand can exist in solution in the presence of PRFAR. The active-OxH state is characterized by the H^N backbone of *h*V51 oriented toward the catalytic *h*C84 pre-organizing the HisH active site for stabilizing the tetrahedral intermediate (see Figure 4.1c) formed in the glutaminase reaction. These results are in line with the hypothesis that the *h*V51 oxyanion hole can form in IGPS induced by PRFAR binding.[145,147,148] Our simulations provide structural explanations to NMR experiments that showed the broadening beyond the detection of *h*G50 and *h*G52 NH signals in the presence of PRFAR, suggesting the activation of μs-ms motions in the oxyanion strand residues.[145,147] The transient and infrequent population of the active-OxH state in μs-cMD simulations together with the intrinsic instability of PRFAR can help to understand why it has not been possible to obtain the *X-ray* structure of wild-type IGPS with PRFAR-bound and the *h*V51 oxyanion hole formed.[57,142,148,150] Recently, Yao and Hamelberg used μs-cMD simulations to characterize a transient *h*V51 oxyanion hole formation in *apo* IGPS, which points out that the active-OxH conformation is not stable in the absence of both the allosteric effector and substrate.[159]

With aMD simulations it was possible to access the microsecond–millisecond motions characteristic of substrate-free IGPS. These simulations captured multiple infrequent *h*V51 oxyanion hole formations and a rich HisF:HisH interface conformational dynamics, highlighting the existence of metastable closed state of the HisF:HisH interface. The transiently populated closed conformation involves the alignment of *h*α1 and *f*α3 helices and presents significant similarity to the substrate-bound *X-ray* *h*C84A IGPS.[161] Therefore, these aMD simulations show that a closed state of the IGPS can be sampled in solution, even when the substrate L-*Gln* or the effector PRFAR are not present. These observations are in line with the idea that IGPS should adopt a catalytically productive closed HisF:HisH interface to retain the substrate in the HisH active site during hydrolysis and keep the ammonia produced for internal transfer through the HisF tunnel. Consistent with these results, Kneuttinger *et al.,* introduced light-switchable non-natural amino acids that stabilize the partial closure of IGPS observing an enhancement of catalytic activity.[164]

In the absence of *L-Gln* substrate, we observed that the *h*V51 oxyanion hole formation and the closed HisF:HisH interface states are transiently populated in the PRFAR-IGPS conformational ensemble. MD simulations show that both the formation of the oxyanion hole and the closure of the interface are uncoupled from each other which is consistent with the NMR uncorrelated microsecond–millisecond motions that are activated upon PRFAR binding.[145] To unravel the effect of *L-Gln* substrate on the IGPS conformational ensemble toward the formation of the allosterically active ternary complex, we explored the sequence of events of *L-Gln* binding into the HisH active site and subsequent allosteric activation. Unconstrained enhanced sampling aMD simulations were used to simulate the spontaneous binding of *L-Gln* to the HisH active site in both PRFAR-free and PRFAR-bound IGPS. Binding events were captured in both cases with similar binding pathways: the substrate is initially recognized by the side chain of *f*Q123, then *L-Gln* evolves toward the HisH active site when the *h*49-PGVG oxyanion strand adopts the inactive-OxH conformation and the HisF:HisH region is found in the open conformation, and finally *L-Gln* is stabilized by several HisH active site residues including the formation of a hydrogen bond with the H^N amide of *h*G52. These observations are consistent with the similar K_M values for PRFAR-free and PRFAR-bound which is a signature of V-type allostery predominating in IGPS.[148] Importantly, aMD simulations show that adopting the inactive-OxH state is a prerequisite for glutamine binding in the HisH active site. In the active-OxH state, the *L-Gln* binding is blocked by the side chain of *h*L85, which prevents the access to catalytic *h*C84. Therefore, aMD simulations explain at the molecular level recent NMR studies of the *h*C84S IGPS mutant that detected substrate binding only when in inactive IGPS.[161] Our results indicate that even when PRFAR is bound, IGPS should be able to access both the inactive-OxH state of the oxyanion strand and the open HisF:HisH interface to recognize and bind the substrate in the HisH active site.

Upon *L-Gln* binding in the inactive HisH active site, while PRFAR is in the HisF binding site, a sequence of conformational rearrangements are triggered that reveal the allosterically active state of IGPS. We performed extensive μs-aMD simulations starting from the *L-Gln* bound conformation to unravel the time-evolution of the complete allosteric activation. These simulations show that the *h*49-PGVG oxyanion strand and the HisF:HisH interface conformational ensembles are significantly perturbed due to the coupled effect of PRFAR and *L-Gln*. After glutamine binding in the inactive IGPS conformation, the productive closure of the HisF:HisH interface is observed. This indicates that substrate binding in the presence of PRFAR is able to displace the IGPS conformational ensemble toward a closed HisF:HisH interface. The interdomain closure gates multiple formations of the *h*V51 oxyanion hole. The frequent and long-lived formations observed in aMD simulations indicate that the active-OxH and inactive-OxH should present similar relative stabilities and low oxyanion-strand interconversion barriers. In the ternary complex, well-tempered metadynamics (WT-MetaD) simulations reported an energy barrier of 8 kcal/mol for the oxyanion strand rotation with similar relative populations for the inactive-OxH and active-OxH populations. In addition, the oxyanion strand conformational ensemble is broadened in the ternary complex upon allosteric activation. All these observations are consistent with NMR experiments which reported a broadening of the IGPS ensemble without significantly changing the average conformations.[167] Overall, these results are in line with both mutagenesis and NMR experiments, indicating that glutamine hydrolysis and associated chemical steps are rate limiting in the presence of PRFAR.[167]

Once the active-OxH state is attained, *L-Gln* reorients in the pre-organized HisH active site to reach a nucleophilic attack catalytic distance of approximately 3.4 Å between the amide carbon of *L-Gln* and the thiol group of *h*C84. This distance is around 4.5 Å when the HisH active site is in the inactive-OxH. Therefore, proton abstraction and subsequent stabilization of the tetrahedral intermediate can only occur in the active-OxH state of the oxyanion strand. Based on our results, efficient glutamine hydrolysis requires both the pre-organization of the HisH active site and the closure of the HisF:HisH interface. Taking into account our simulations and experimental observations, we propose this conformation as the allosterically active state of IGPS.

aMD simulations indicate that both the active-OxH and inactive-OxH states are essential for the different steps of IGPS function. Therefore, the rapid interconversion between the inactive and active states of the oxyanion strand can be useful to drive the different steps of the glutamine hydrolysis along the catalytic cycle. The active state of wild-type IGPS could not be detected with NMR. We suggest that the reason for that is the rapid dynamic equilibrium between the two forms of the oxyanion strand. When PRFAR is not present (PRFAR-free), the relative populations are maintained but the interconversion barrier of the oxyanion strand increases up to 22 kcal/mol, which is 3-fold higher than in the ternary complex. This high-energy barrier points out that the conformational changes required to attain the active state are barely accessible by PRFAR-free IGPS at room temperature. This goes in line with the 4500-fold enhancement of glutaminase catalytic activity when the effector PRFAR is bound.[151] Overall, our results suggest that the coupled binding of PRFAR and *L-Gln* impacts the associated inactive-to-active transition barrier while keeping the relative stabilities of each state.

Finally, the allosteric communication pathways obtained from dynamical networks along the IGPS activation show communication between both active sites through multiple pathways. The time-evolution shortest path map (te-SPM) indicated the activation of concerted motions once the HisF:HisH interface productively closes. Subsequently, the concerted motions expand through the whole IGPS including the HisF subunit, the interface, and the HisH active site, triggering the formation of the *h*V51 oxyanion hole. Interestingly, the te-SPM analysis captured a wide number of residues that, according to NMR experiments, take part in concerted millisecond motions in the IGPS ternary complex. The work described in this Chapter represents a step forward with respect to previous studies where dynamical networks were performed on short nanosecond MD simulations that did not capture the complete allosteric activation process. The coupled closure of the HisF:HisH interface with the formation of the *h*V51 oxyanion hole and the activation of correlated motions captured in our simulations are consistent with millisecond motions characterized with NMR in the ternary complex.[145,168] Maschietto and coworkers recently reported different dynamic networks in bacteria and yeast forms of IGPS pointing out that allosteric communication pathways changed along IGPS evolution.[158] The presence of multiple allosteric communication pathways between binding sites and the activation of millisecond fluctuations are typical properties of dynamic allostery.[15] Gathering the information from the different analysis including: a) the existence of multiple communication pathways; and b) the observed broadening of the conformational ensemble of IGPS elicited by the binding of the coupled binding of substrate and effector, we propose that IGPS allosteric activation follows the violin model typical of dynamics-based allostery as observed in protein kinases.[169,170]

Chapter 5. Molecular basis of efficient and enantioselective biocatalytic C-N bond formation: from understanding to design

The following Chapter describes the combined experimental and computational work to design efficient and enantioselective biocatalytic platforms toward C-N bond formation. The results described in this Chapter were reported in *Nat. Chem.* **2021**, 13, 1166-1172 (including journal cover) and *ChemRxiv* **2022** (Calvó-Tusell et al. 10.26434/chemrxiv-2022-f02xh). Part of the material show is from: Liu, Z., Calvó-Tusell, C., Zhou, A.Z. et al. Dual-function enzyme catalysis for enantioselective carbon–nitrogen bond formation. *Nat. Chem.* 13, 1166–1172 (2021) published 2021, Springer Nature.

All the experiments described in this Chapter have been carried out by Zhen Liu, Andrew Zhou, and Kai Chen under the coordination of Frances Arnold (see 5.1.1 State of the art section and Results 5.3). We (Carla Calvó-Tusell with the coordination of Marc Garcia-Borràs) performed the complete computational modelling (see 5.3 and 5.4 Results section). All of the authors participated in the project design and discussion.

5.1. State of the art

Amine functional groups are commonly found in bioactive molecules.[171,172] Developing efficient and selective methods for incorporating chiral amines into molecular scaffolds through the formation of C-N bonds is still a challenge for chemistry and biochemistry. A wide number of strategies have been reported to form C–N bonds. Among them, the insertion of carbenes into N–H bonds have been used to efficiently construct nitrogen-containing molecules taking advantage of the high reactivity of carbene species.[173–177] Recently, highly selective and efficient heam-dependent enzymes have been repurposed using directed evolution to catalyze non-natural carbene and nitrene transfer reactions. In particular, cytochromes P450 have been evolved to insert carbene moieties into N–H bonds with significantly high catalytic efficiency (total turnover numbers (TTN)).[178–181] However, in comparison to other carbene-transfer reactions such as C-H insertion or alkene cyclopropanation, the insertion of carbenes into N-H bonds still lacks full stereocontrol.[73,182]

To achieve asymmetric carbene N–H insertion, several strategies based on small molecule catalysis and biocatalysis have been proposed. Transition-metal catalysts are commonly combined with a chiral proton-transfer catalyst (PTC) for the stepwise stereoselective N-H insertion.[183] In this process, first, the transition metal catalyst reacts with the carbene precursor to generate a metal–carbene species. Then, the amine substrate performs a *N*-nucleophilic addition to the carbene forming an ylide intermediate. Finally, the chiral PTC, such as a chiral phosphoric acid or amino-thiourea, is responsible for carrying the enantioselective protonation on the ylide intermediate to get the final product.[183,184] To achieve high asymmetric protonation, other proton sources must be avoided otherwise the selectivity is not controlled. Recently, Shaik *et al.*, used computational modelling to suggest a similar mechanism for N–H insertion reactions catalyzed by haem-dependent enzymes.[185] Since both enzyme protic amino acids and water molecules are potential proton sources, the challenge to achieve high enantioselectivity for these biocatalytic transformations relies on precisely controlling the proton transfer step once to the highly reactive ylide intermediate is formed. Recently, Fasan and coworkers reported a series of engineered myoglobins that catalyze the asymmetric N–H insertion with moderate enantiocontrol.[179,186]

The overall goal of this project is to design a dual-function biocatalytic platform for the enantioselective carbon-nitrogen bond formation. Similar to the dual-catalyst strategy used in small-molecule catalysis, the idea is to identify enzymes that can perform two distinct functions: 1) produce the carbene species for the efficient nucleophilic addition of the amine substrate; and 2) control the selective protonation of the ylide intermediate within the enzyme active site. In this way, a highly enantioselective biocatalytic N–H insertion reaction would be possible. First, experimental screening of previously evolved carbene and nitrene transfer enzymes will be used to identify variants that present high enantioselective N–H insertion toward one of the enantiomers (see State of the art 5.1.1). Then, computational modelling will be used to rationalize the molecular basis of the whole process providing a molecular explanation to the efficient and selective N–H insertion reaction (see Results 5.3.1). Finally, the mechanistic information obtained from computational modelling will be used to design enzyme variants that reverse the enantioselectivity toward the opposed enantiomer (see Results 5.3.2.).

5.1.1. Initial Screening and identification of high enantioselective N–H insertion variants.

The first objective was to discover enzymes that could perform the asymmetric N–H insertion with high efficiency and enantioselectivity. The chosen reaction for the enzyme screening was between lactone diazo and *N*-methyl aniline. This transformation is attractive because a biologically relevant α-amino lactone is generated.[187] Moreover, this is a challenging reaction because lactone-based carbenes present difficult stereocontrol in small molecular catalysis due to undesired β-hydride elimination processes.[188] Previously, the Arnold Lab harnessed the power of directed evolution to engineer haem proteins that accelerate the lactone-carbene-transfer process preventing undesired side reactions.[189,190] Therefore, the enzyme screening was focused on identifying enzyme variants that provide efficient transfer of the lactone-carbene species to the *N*-methyl aniline substrates and that, at the same time, control the enantioselective protonation of the formed ylide intermediate to generate the enantio-enriched product.

A total of 40 haem protein variants previously evolved for different carbene and nitrene transformations were screened toward the abovementioned transformation. This pool of enzyme variants includes engineered variants for lactone-carbene-transfer reactions including P411-**L7** and P411-**G8S** among others.[190,191] Most of the screened variants showed low activity (beyond 5% conversion). Interestingly, P411-**L7** was able to catalyze the N–H insertion with 81% yield and 94% *ee.* toward the *(R)*-enantiomer. P411-**L7** corresponds to P411 (the axial haem-ligating residue is a mutated serine instead of a cysteine) with the truncated FAD domain that contains 29 mutations with respect to wild-type $\mathrm{P450_{BM3}}$. This enzyme was previously engineered in Arnold Lab through several rounds of directed evolution to perform lactone-carbene C–H insertion starting from the P411-**C10** variant.[189,192] Attaching the reductase domain to P411-**L7** (**L7_FL**) enhances the catalytic performance up to 92% yield and 95% *ee*., which can be attributed to the increased stability of the full enzyme.[193]

By analyzing the P411-**L7** enzyme lineage (from **L1** to **L7**, one single mutation incorporated in each round) for the lactone-carbene N–H insertion, it was shown that **L6** and **L7** are excellent catalysts in terms of yield and enantioselectivity for this transformation. Thus, the significant jump in achieving high activity and selectivity for N–H insertion takes place when going from **L5** to **L6** which corresponds to the inclusion of A264S mutation. In particular, the impact on enantioselectivity is significant, from –21% in **L5** to 92% *ee*, in addition to significant yields increase. in **L6**. This points out that the residue at position 264 is key to trigger the carbene-transfer reaction offering excellent control on the stereoselective protonation step. To gain insight into the importance of this position, variants with different mutations of hotspot 264 were experimentally tested. When replaced with alanine and glycine, the activity is still high, but the selectivity is significantly lower, which suggests that the hydroxyl group of serine can play a key role in controlling enantioselectivity. However, if other polar residues are introduced such as aspartate, threonine, and cysteine both yield and stereocontrol are significantly affected. Therefore, it is not clear the role played by S264.

The molecular basis of this enzymatic transformation and the role of S264 in the asymmetric proton transfer step is still unknown. In addition, **L5** represents a promising branching point (-21 % *ee*. and 43% yield) for the enantioselective control toward one enantiomer or the other. Understanding the behavior of both **L6** and **L5** variants can provide relevant information for designing highly enantioselective variants.

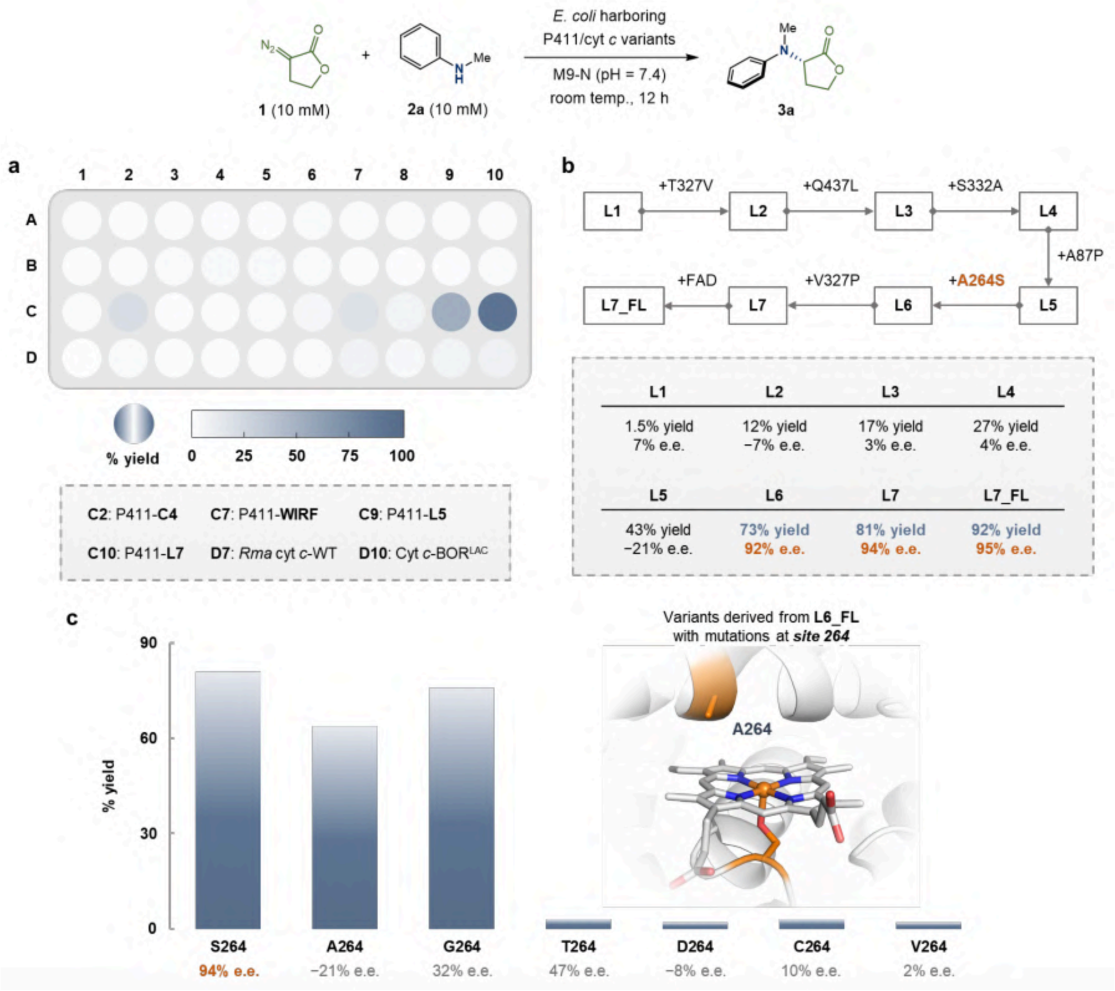


Figure 5.1. Screening for enzymatic N–H insertion with a haem protein collection and identification of A264S as the key mutation for achieving high activity and selectivity. a) Initial

screening performed with 40 haem protein variants. The screening led to the discovery of variant **L7** (in well C10), which originated from a previous lactone carbene C–H insertion project.[189] **b**) The lactone carbene C–H insertion lineage (**L1** to **L7**) was rescreened and variants with the A264S mutation (**L6** and **L7**) were found to be excellent catalysts for N–H insertion. **c**) Mutagenesis studies were performed to show the importance of residue S264 for N–H insertion. Replacing S264 with other amino acids led to low selectivities and diminished yields. In the figure, A264 in the active site of P411 variant **E10** (PDB ID:5UCW) is highlighted. **L7_FL** and **L6_FL** are **L7** and **L6** restored to their respective full-length P411 enzymes.

In this Chapter we will use a computational protocol tailored to unravel the molecular mechanism of the enantioselective N–H insertion in P411 variants. First, we will understand the role of A264 mutation and then, our goal is to harness this mechanistic information for design.

5.2. Computational Details and Protocols

In the following section, we will describe the computational strategy used to study the selected P411-**C10** lineage of variants (*i.e.,* P411-**L1** to P411-**L7**). This protocol combines molecular dynamics simulations, docking calculations, and quantum mechanics calculations and consists of the steps shown in Figure 5.2. To understand the different steps of the molecular mechanism of the N–H insertion, MD simulations are performed in the *apo*, lactone-carbene, substrate and ylide intermediate bound states. Mechanistic studies using density functional theory have been performed using truncated active site models based on these simulations.

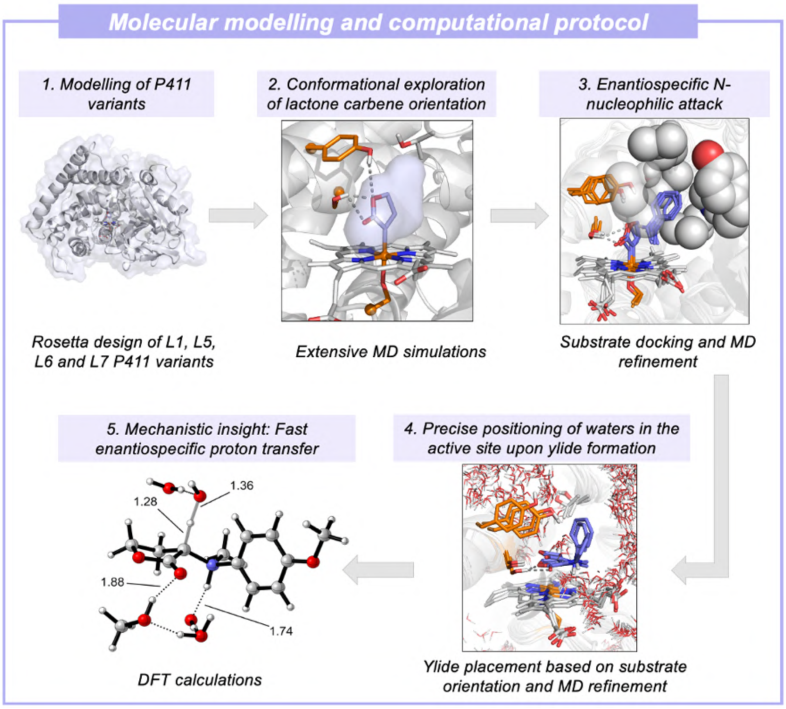

Molecular modelling and computational protocol
1. Modelling of P411 variants
2. Conformational exploration of lactone carbene orientation
3. Enantiospecific N-nucleophilic attack
Rosetta design of L1, L5, L6 and L7 P411 variants
Extensive MD simulations
Substrate docking and MD refinement
5. Mechanistic insight: Fast enantiospecific proton transfer
4. Precise positioning of waters in the active site upon ylide formation
1.28
1.36
1.88
1.74
DFT calculations
Ylide placement based on substrate orientation and MD refinement

Figure 5.2. Computational protocol. The computational strategy to study P411 variants for N–H insertions. The protocol is based on: **1)** constructing the P411 variants under study; **2)** performing MD simulations in the different states (*apo*, lactone-carbene, and substrate bound) to assess the conformational features that confer the high activity and specificity; **3)** docking calculations and MD refinement for substrate bound state; **4)** substrate bound state simulations were used to place the ylide intermediate and perform MD simulations; **5)** from the MD knowledge gathered in previous steps quantum mechanics calculations were performed to understand the different steps of the molecular mechanism of the N–H insertion. A more detailed description of the computational protocol can be found in Appendix B, Computational Protocols.

5.2.1. Preparation and Modeling of P411 variants: L5, L6, L7.

In the first place, we prepared the P411-**C10** (P411-**L1**) variant starting from the available crystal structure of the related P411-**E10** variant (PDB ID: 5UCW). To properly describe the effects of the 13 mutations with respect to P411-**E10**, Rosetta Design was used to generate a structure containing N70E, A74G, V78L, M118S, F162L, M177L, L263Y, H266V, A330Y, I401L, T436L, L437Q, S438T mutations. Then, the structure of P411-**L1** retrieved from Rosetta was refined with extensive MD simulations (5 replicas of 1,000 ns, gathering a total of 5 μs of simulation time). The MD simulations with the haem cofactor were run using the bonded model to describe the iron-porphyrin interaction. The MD trajectories of P411-**L1** were clustered based on the root-mean-square deviation (RMSD) of the protein backbone, selecting the most populated cluster as the representative structure of P411-**L1** *apo* state. This representative and refined MD structure of P411-**L1** was used as a template for preparing the lineage of P411 variants **L5**, **L6** and **L7**. Since the number of mutations between the **L1** and selected variants is low, the PyMOL mutagenesis tool was used. P411-**L5** presents four additional mutations with respect to **L1** (T327V, Q437L, S332A, A87P), **L6** variant includes one additional mutation respect **L5** (A264S), and **L7** presents one additional mutation respect to **L6** (V327P). The structures of **L5**, **L6**, and **L7** variants were subsequently refined with MD simulations. In this case, 3 replicas of 500 ns of MD simulations were run, accumulating 1.5 µs for each variant. Again, MD trajectories were clusterized based on the protein backbone RMSD to select the most populated cluster as the representative structure of *apo* state P411-**L5**, **L6**, and **L7** variants.

5.2.2. Conformational Exploration of Lactone Carbene Orientation

To generate the structure of the lactone carbene covalently bound to the haem iron of P411 variants, we manually place the lactone ring bound to the Fe, it in the active site of the representative structures of *apo* state **L5**, **L6** and **L7** variants. The lactone carbene is a small molecule where the carbene directly binds the iron of the porphyrin. From QM models it is known that the plane of the lactone carbene and the plane of the porphyrin rings are perpendicular. Therefore, we manually placed the lactone carbene in the same orientation as in the QM models. Then, MCPB.py[106] was used to obtain the bonded parameters of the lactone-carbene for MD simulations. Once the system was assembled, we ran 5 replicas of 500 ns of MD simulations using different orientations of the lactone-carbene as starting point, gathering 2.5 µs for each variant in the lactone-carbene state. We analyzed the conformations that the lactone carbene is able to explore in the active site and the interactions that it establishes with surrounding residues to capture the preferred orientation of the lactone-carbene in each of the selected variants. These MD trajectories were clusterized to identify the most relevant conformations of P411 variants based on protein backbone RMSD.

5.2.3. Analysis of the enantiospecific N-nucleophilic attack

To simulate the substrate bound pose and understand the molecular basis of the nucleophilic attack, we docked the amine substrate (N-methyl aniline for Section 5.3 and 4-Methoxy aniline for Section 5.4) in the active site of the lactone carbene-bound P411-**L5**, **L6**, and **L7** variants. Docking of the amine substrate was performed using Autodock Vina[194] for two different representative conformations of each system (the two most populated clusters of each variant obtained as described in section 5.2.2). Then, restrained-MD simulations were used to refine docking predictions of the substrate bound pose. In these MD simulations the distance between the nitrogen of the amine and the central carbon of the carbene was restrained up to 3.6 Å (this distance cannot be higher than this value in the MD simulations, see Computational Methods on Appendix B). For each system 3 replicas of 250 ns were performed using two different starting substrate-bound P411 structures, accumulating a total 2.5 µs for each variant. Again, substrate-bound MD simulations were clustered using protein backbone RMSD to identify the most relevant conformations visited along MD simulations. Importantly, these simulations offer relevant information of catalytically competent binding poses explored by the amine substrate, providing molecular details of the near attack conformation to perform the nucleophilic attack to the carbene.

5.2.4. Precise positioning of water molecules in the active site upon ylide formation.

Next, we prepared the ylide-bound P411 complex to explore the selective protonation step. Based on DFT calculations, the formation of the ylide breaks the covalent bond between the iron and the lactone ring (see Sections 5.2.5 and 5.3). Using the most populated cluster obtained from the lactone and substrate-bound MD simulations (see 5.2.3), the ylide intermediate was manually docked in the active site of P411-**L5**, **L6**, and **L7** variants. We used the previously characterized amine substrate binding pose as a template, superimposing the ylide with the amine and the lactone-carbene structures. Using a similar protocol as described before for the substrate, restrained-MD simulations were performed starting from the ylide-bound structures. In this case, the distance between the ylide central carbon atom and the haem iron was restrained up to 4.1 Å. By introducing this geometric constraint, we aimed to characterize the dissociated ylide dissociated complex in the enzyme active site to mimic the ylide complex characterized from model DFT calculations (see Figure 5.4 and Section 5.2.5). Three replicas of 100 ns of restrained-MD simulations were performed, accumulating 300 ns of simulation time for each variant. The distribution of potential proton sources (water molecules and amino acid residues) around the ylide intermediate was analyzed.

5.2.5. Mechanistic insight: fast enantiospecific proton transfer.

Finally, restrained-MD simulations of P411-**L6** and **L7** variants with ylide intermediate bound (from step described in section 5.2.4) were visually inspected. From the representative snapshots we prepared different truncated active site models that included the ylide intermediate, S264 residue, and key active site water molecules. These truncated models were used to study the possible ylide proton transfer pathways with DFT calculations.

Additionally, the complete energy profile for the lactone carbene N–H insertion has been characterized with DFT calculations. All DFT calculations carried out in this Chapter are performed using the protocol specified in Chapter 2. Computational Methods, section 2.6.

5.2.6. Evaluation of mechanistically-guided protein engineered P411 variants.

A similar computational protocol involving steps described in sections 5.2.1-5.2.4 was used to explain the molecular basis of newly designed variants toward the *(S)*-enantiomer. The starting point to model the new engineered variant P411-**L5_B3** was the P411-**L5** variant.

5.3. Results and Discussion

5.3.1. Origins of conformational control and enantioselectivity in P411-L6 and L7 variants

We applied the computational strategy described in 5.2 to unravel the molecular basis of this enzymatic transformation and understand the role of S264 in promoting the efficient and selective asymmetric carbene N–H insertion. First, we characterize the dynamic behavior of the P411-**L6** variant, which is the one that presents the A264S mutation. Extensive MD simulations for the **L6** variant with the lactone-carbene bound (in the absence of the substrate) pointed out that the lactone moiety mainly explores a single orientation in the active site (see Figure 5.3 and B1). To monitor the orientation of the lactone, we defined a dihedral angle, ∠(N–Fe–C1–C2), that describes the relative orientation explored by the carbene relative to the haem moiety. Despite starting with a different orientation, all replicas converged to the same orientation of the lactone with a dihedral angle around 90°. In this orientation, the lactone is stabilized by the S264 side chain through persistent hydrogen-bonding interactions between the hydroxyl group of S264 and the lactone ester group (see Figure 5.3 and B1). This orientation is further stabilized through a transient hydrogen bond between Y263 and lactone.

Interestingly, if the lactone is fixed to this orientation, only one face of the electrophilic carbene can be attacked by the nucleophilic amine substrate, which selectively yields the stereospecific reactive ylide intermediate. MD simulations of the variant **L7**, which also presents the A264S mutation, shows a similar behavior. In contrast to **L6** and **L7**, P411-**L5** variants contain an alanine at position 264. MD simulations of the **L5** variant show that, in the absence of S264, the lactone explores multiple conformations within the active site (see Figure B1). In addition, DFT calculations indicate that the hydrogen-bonding interactions involving the carbonyl group of the carbene enhance its electrophilic character making it more reactive (see Appendix Figure B10 and B11). Therefore, S264 participates in controlling the lactone orientation but also contributes to increasing the reactivity of the carbene species.

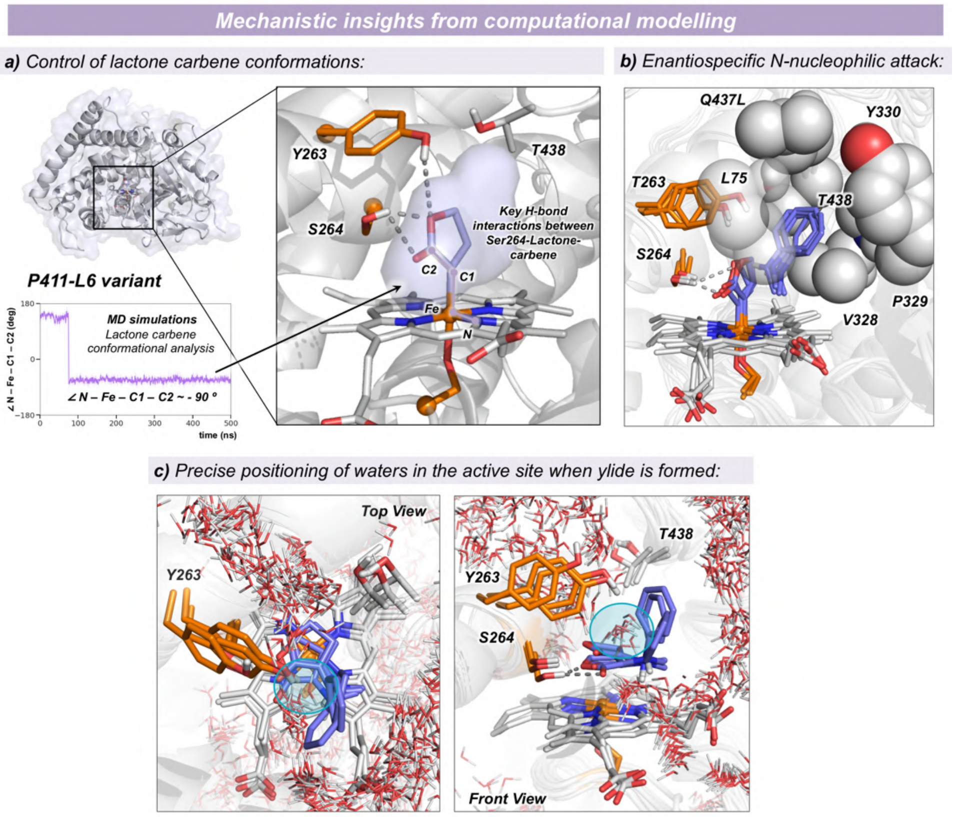

Mechanistic insights from computational modelling
a) Control of lactone carbene conformations:
b) Enantiospecific N-nucleophilic attack:
P411-L6 variant
MD simulations
Lactone carbene
conformational analysis
∠N – Fe – C1 – C2 ~ - 90 º
∠N – Fe – C1 – C2 (deg)
180
0
−180
0
100
200
300
400
500
time (ns)
Y263
T438
S264
Key H-bond interactions between Ser264-Lactone-carbene
C2
C1
Fe
N
Q437L
Y330
T263
L75
T438
S264
P329
V328
c) Precise positioning of waters in the active site when ylide is formed:
Top View
Y263
T438
Y263
S264
Front View

Figure 5.3. Origins of enantioselectivity in carbene transfer into N–H bonds catalyzed by P411-L6 from computational modelling. a) Representative snapshot from Molecular Dynamics (MD) simulations describing the conformations explored by the lactone carbene when formed in P411-**L6**. The ∠(N – Fe – C1 – C2) dihedral angle measured along the MD trajectory describes the relative orientation explored by the carbene (see Appendix B for additional replicas). Simulations show that the lactone preferentially explores a single conformation, which is stabilized by H-bond interactions established between the carbene ester group and S264 and Y263 (see Appendix B for additional details). **b)** Overlay of four representative snapshots obtained from restrained-MD simulations exploring near attack conformations for the N-nucleophilic attack of **2a** to the lactone carbene in **L6**. Hydrophobic interactions occurring between the aromatic ring of the aniline derivative and active

site residues (L75, V328, Q437L, P329) stabilize this binding mode, while H-bond interactions between the carbene ester group and S264 are maintained. These two factors are responsible for an enantiospecific nucleophilic addition. **c)** Overlay of 3 representative snapshots from restrained-MD simulations exploring **L6** active site arrangement when ylide **2a** is formed. Water molecules are precisely positioned on the top-face of the lactone ring, driven by Y263 and T438, and nearby the protonated ylide amine group. These water molecules are able to stereospecifically protonate the ylide intermediate from the pro-*S* face. Displayed water molecules are drawn from 25 random structures across the 100 ns MD trajectory.

We then modelled how the amine substrate approaches to the carbene and the subsequent formation of the ylide intermediate in the active site of P411-**L5**, **L6**, and **L7** variants. Restrained-MD simulations show that the amine substrate binding in a near-attack conformation for the *N*-nucleophilic addition to the lactone-carbene is stabilized by hydrophobic interactions between the substrate aromatic ring and active-site hydrophobic residues (L75, V328, L437, P329) while the hydrogen-bond between the lactone and S264 is maintained (see Figure 5.3 and Appendix Figure B2 and B3). In **L5**, the substrate adopts unproductive *N*-nucleophilic addition poses and shows improper packing in the active site. The simulations with the ylide intermediate also describe that the hydrogen bond between the lactone ester group and S264 is persistent in L6 and L7 variant. Next, we analyzed the presence of water molecules in the active site of **L6**. Based on the MD simulations, only a few water molecules are present in the active-site pocket, which are precisely funneled through two water channels, one formed by Y263 and T438 from the top face of the ylide lactone ring and the other one guided by the anionic haem carboxylates near the ylide amine group (see Figure 5.3 and Appendix B4 and B5). In P411-**L6** and P411-**L7**, water molecules are precisely positioned on the top-face of the lactone ring through *water channel 1* driven by Y263 and T438, and nearby the protonated ylide amine group through *water channel 2*. These water molecules are able to stereoselectively protonate the ylide intermediate from the *pro-S* face. In the P411-**L5** variant, water molecules cannot effectively access the top-face of the lactone ring, since *water channel 1* is blocked (see Figure B4).

5.3.2. Mechanistic insights into the N–H insertion reaction

We further investigated the intrinsic reaction mechanism of P411-catalyzed lactone-carbene N–H insertion to gain more insights to guide subsequent computational modeling and protein engineering. Here, we selected the 4-Methoxy aniline **2a** and Fe-lactone carbone as the model substrates (see Figure 5.1). Using a similar strategy as described before by Shaik and coworkers, we performed DFT calculations using **2a** as the amine substrate reacting with the lactone-carbene species. The *theozyme* (truncated enzyme) computational model includes a methanol molecule H-bonded to the ester group of LAC in order to mimic the role of S264 as H-bond donor in the enzyme active site as in **L6_FL** and **L7_FL** enzyme variants. Calculations showed that a plausible mechanism involves a first *N*-nucleophilic attack by the amine to the electrophilic iron lactone-carbene center (see Figure 5.4), forming an ylide intermediate covalently bound to the iron. The dissociation of the ylide from the iron was found to be energetically slightly uphill and barrierless. These results are in line with previous computational studies by the Shaik group on P450 (cysteine ligated)-catalyzed carbene insertion into N–H bonds considering the acyclic ethyl diazoacetate (EDA) as the carbene precursor.[185]

However, different from the acyclic carbene system, which is proposed to involve the formation of an enol via the direct intramolecular proton transfer from the protonated nitrogen to the oxygen of the carbonyl group during dissociation from Fe, our calculations indicated that the lactone ylide intermediate could directly dissociate from the iron center. With the lactone system, the 5-membered transition state required for the enol formation from the ylide intermediate becomes disfavored as compared to the aliphatic carbene system due to the geometric strain induced by the lactone ring (see Figure 5.4). Additionally, the H-bond interaction established by the carbonyl group and the external H-bond donor (*e.g.,* methanol in our model or S264 in the enzyme active site) is found to disfavor even more the formation of this enol intermediate (see Figure 5.4). Calculations also showed that the direct proton transfer from the protonated amine to the vicinal carbon via a three membered ring transition state is energetically highly disfavored.

Based on these observations, we proposed that the ylide rearrangement and the protonation of the carbon center should be facilitated by a water molecule. MD simulations carried out with the P411-**L6_FL** variant having the ylide intermediate bound in a conformation that mimics the just dissociated complex characterized from DFT calculations, revealed that only a few water molecules are present in the active site. These water molecules approach the ylide intermediate bearing interaction with S264 from the top face of the lactone ring, opposite face to the heme cofactor (see Figure 5.4). Considering this arrangement of water molecules around the ylide intermediate in the active site, truncated models were built and DFT calculations were carried out. Model calculations indicated that these water molecules could effectively protonate the carbon center in a fast proton transfer step, right after the ylide dissociates from the iron and before this reactive intermediate leaves the enzyme active site (see Figure 5.4b).

Therefore, the fast protonation step is proposed to be stereoselective, taking place from a specific face of the ylide intermediate and mediated by a precisely place water molecule. This step will ultimately depend on the first *N*-nucleophilic attack, which will determine which enantiotopic face of the ylide is oriented opposite to the heme cofactor and exposed to these active site water molecules.

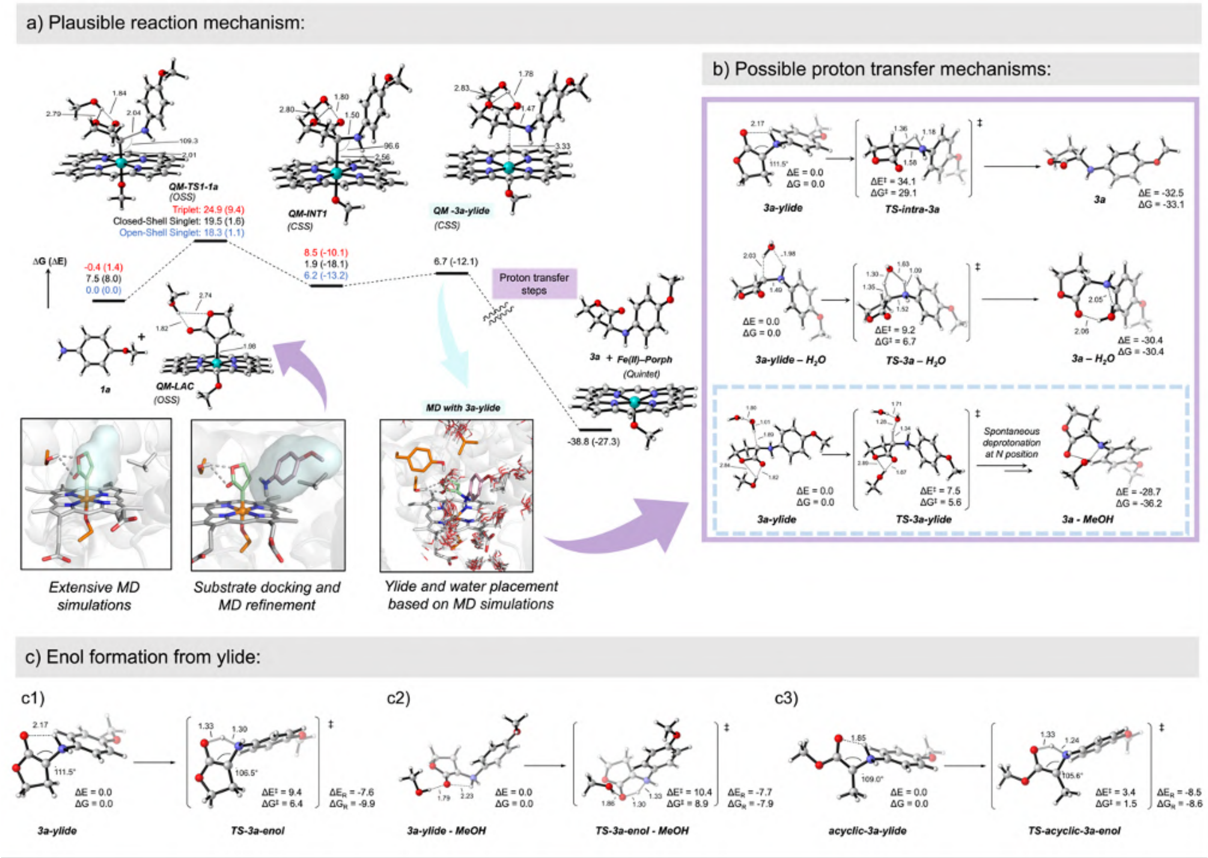


Figure 5.4. DFT *theozyme* calculations. Plausible reaction mechanism based on DFT calculations using a truncated computational model. a) DFT energy profile for lactone-carbene N–H insertion involving model substrate **1a.** A truncated model that includes a methanol molecule

to mimic P411-**L6** active site S264 residue based on substrate-bound MD simulations is used. Results obtained for energetically accessible electronic states are reported, and the lowest in energy optimized geometries for each stationary point are shown. **b**) Optimized model transition states (TSs) for stereoselective **3a** formation from **3a**-ylide. Computational models were built based on the conformations explored by the **3a**-ylide when formed in P411-**L6** active site and the arrangement of water molecules around the ylide as observed from MD simulations. **c**) Computational modeling of intramolecular ylide-enol tautomerization for: **c1**) **3a**-ylide as the model substrate; **c2**) **3a**-ylide as the model substrate and considering the H-bond interactions between the lactone carbonyl and a methanol molecule that mimics active site S264 residue in P411-**L6**; **c3**) acyclic-**3a**-ylide as the model substrate.

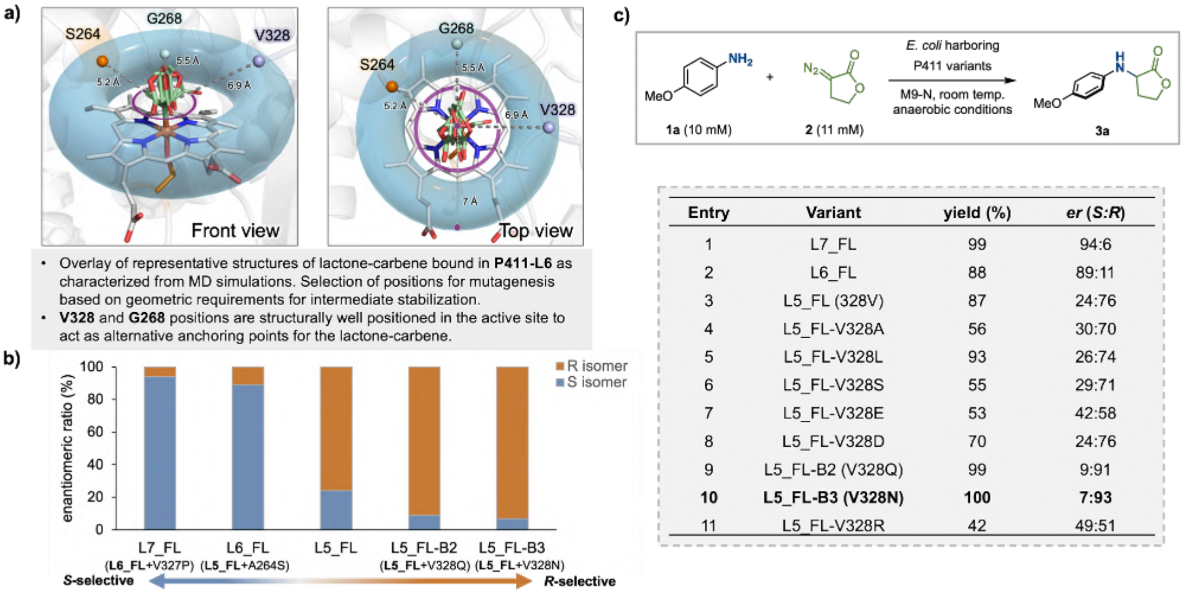
Collectively, the enantioselective formation of the ylide and the precise placement of water molecules in the active site for proton transfer enable the enzyme to control the selectivity of this N–H insertion reaction.

5.3.3. Computationally guided design of N–H insertion variants toward the *(R)*-enantiomer

With all these mechanistic and structural insights obtained from computational modeling, we sought to rationally engineer new enzyme variants to access opposite selectivities. To do so, we proposed to invert the orientation that the lactone-carbene (LAC) can explore in the enzyme active site, to force the *N*-nucleophilic attack to take place from the opposite enantiotopic face of the lactone-carbene ring, which will eventually lead to the opposite product enantiomer following a rapid stereospecific water-assisted protonation step.

To alter the LAC orientation in the enzyme's active site, we hypothesized that one could (1) replace the serine at position 264 in *(S)*-selective P411-**L6/L7_FL** variant with a non-polar residue to disrupt the original H-bond interaction; and (2) introduce a H-bond donor residue at the opposite side of the catalytic pocket that could serve as a new anchoring point for the LAC and invert its orientation in the enzyme active site. By analyzing the computational models generated for the LAC intermediate bound in the poorly selective **L5** and the selective **L6** variants (see 5.3.1), we identified two positions that could accommodate alternative anchoring points to invert the LAC orientation: positions 268 and 328 (see Figure 5.5). These positions were selected based on geometric requirements, their spatial disposition on the equatorial plane and their appropriate distance with respect to the LAC intermediate, in order to ensure an effective interaction between the new H-bond donor and the LAC ester group. We first applied a distance threshold of 7.0 Å to select amino acid side chains that are found to directly interact with the LAC intermediate in **L5** and **L6** variants during MD simulations. The selection was limited to positions on the equatorial plane of the LAC ring. Then, we analyzed specific MD snapshots in which the LAC intermediate explores the targeted conformation that we were aiming to stabilize. From there, we identified positions 268 and 328 as suitable hosts for polar side chains with the appropriate directionality to stabilize this particular conformation of the LAC intermediate (see Figure 5.5a). To minimize disruption of the active site environment evolved for efficient and selective *N*-nucleophilic addition and protonation steps, **L5_FL** bearing a non-polar alanine residue at position 264 was used as the starting point for the engineering campaign.

Using this information, Zhen Liu at the Arnold Lab carried out the following experiments. First, it was shown that **L7_FL** and **L6_FL** could catalyze the reaction between 4-methoxy aniline **1a** and lactone diazo 2, giving product **3a** in 94:6 and 89:11 *er*, respectively, favoring the formation of the *(S)*-enantiomer (see Figure 5.5). Performing site-saturation mutagenesis (SSM) at the 328 and 268 positions using **L5_FL** as the parent and screening the corresponding libraries led to two *(R)*-selective variants. In both variants, **L5_FL-B2** and **L5_FL-B3**, V328 was mutated to a protic residue, Q and N, respectively, which flipped the enantioselectivities to 9:91 and 7:93 *er*, respectively. Site-directed mutagenesis at 328 demonstrated that shorter polar residues (serine, Figure 5.5) or charged amino acids (glutamic acid, aspartic acid or arginine, Figure 5.5 entries 7, 8 and 11) led to significantly lower enantioselectivities and yields. No selectivity-enhancing mutations were found at position 268 (Figure 5.5), suggesting that polar residues at this position cannot establish an effective H-bond with the LAC intermediate while allowing a selective nucleophilic attack by the amine substrate.



- Overlay of representative structures of lactone-carbene bound in **P411-L6** as characterized from MD simulations. Selection of positions for mutagenesis based on geometric requirements for intermediate stabilization.
- **V328** and **G268** positions are structurally well positioned in the active site to act as alternative anchoring points for the lactone-carbene.

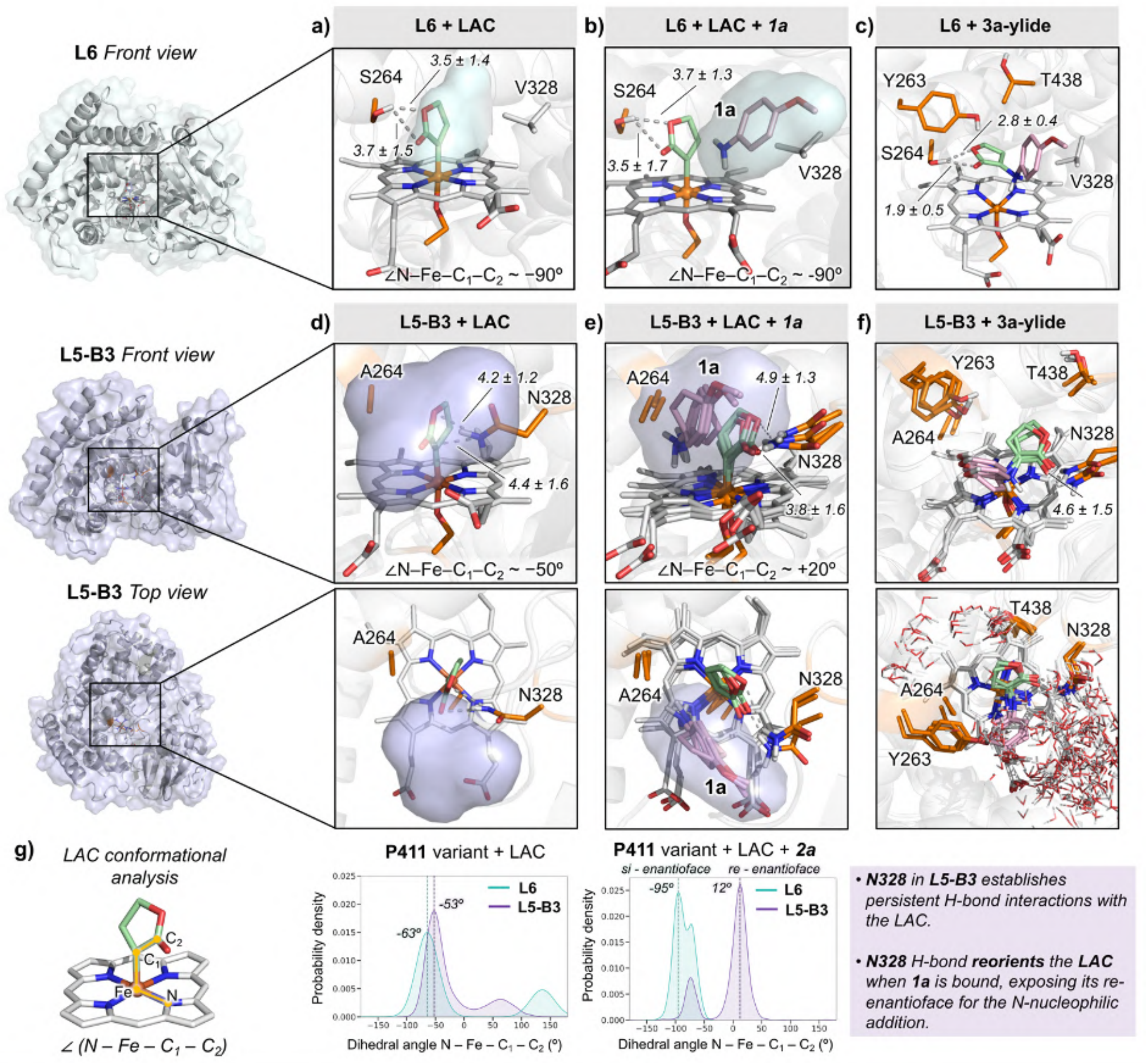
Entry	Variant	yield (%)	*er* (*S:R*)
1	L7_FL	99	94:6
2	L6_FL	88	89:11
3	L5_FL (328V)	87	24:76
4	L5_FL-V328A	56	30:70
5	L5_FL-V328L	93	26:74
6	L5_FL-V328S	55	29:71
7	L5_FL-V328E	53	42:58
8	L5_FL-V328D	70	24:76
9	L5_FL-B2 (V328Q)	99	9:91
10	**L5_FL-B3 (V328N)**	**100**	**7:93**
11	L5_FL-V328R	42	49:51

Figure 5.5. Computational and experimental strategies to revert enantioselectivity. a) Structural analysis of P411-L6 active site with lactone-carbene (LAC) bound as characterized from MD simulations, and identification of positions for alternative anchoring of LAC intermediate. **b)** Identification of *(R)*-selective variant through site-saturation mutagenesis (SSM) and screening at 268 and 328 sites. **c)** Comparison of residues at 328 position by site-directed mutagenesis.

With the best *(R)*-selective variant, **L5_FL-B3**, we performed MD simulations to unravel the role of mutation V328N in driving the *(R)*-selective carbene N–H insertion. First, the **L5-B3** variant was modeled considering the heme domain with the LAC bound. As expected, the newly introduced V328N residue establishes persistent H-bond interactions with the ester group of the LAC (see Figure 5.6), and it is placed at the opposite side in the active site as compared to S264 in the **L6** variant (Figure 5.3 and 5.6). The relative orientation of the LAC with respect to the heme (described by the ∠(N–Fe–C1–C2) dihedral angle) is similar in **L5-B3** and **L6** (Figure 5.6). Nevertheless, the interaction between the lactone and the newly introduced N328 side chain only makes accessible the *si* face of the LAC for the *N*-nucleophilic attack (Figure 5.6), which is opposite to the **L6** variant.

Next, the amine substrate **1a** bound in the **L5-B3** active site in the presence of the LAC was modeled, using contrained MD simulations to mimic near-attack conformations for the *N*-nucleophilic attack (see Appendix B. Computational Protocols for more details) as previously done for **L5**, **L6**, and **L7** variants. These simulations showed that the substrate bound in a catalytically relevant mode for the *N*-nucleophilic addition induces a slight reorientation of the LAC (rotation along the Fe–C bond from *ca.* -50° to *ca.* +15°, Figure 5.6) that keeps the H-bond between the lactone ester group and the amide of N328. This H-bond interaction was previously shown to be also important for enhancing the electrophilicity of the LAC.[195,196] Consequently, this binding mode of **1a** and LAC is expected to be more reactive than alternative ones lacking this H-bond interaction, thus biasing the reaction to happen from this characterized near-attack conformation. The amine substrate occupies the available space near the enantiotopic *si* face of the lactone and A264 (Figure 5.6, B21 and B23). This binding mode of the substrate is further stabilized by hydrophobic interactions with residues L75, P87, L437, and T438. All these factors synergistically favor the selective *N*-nucleophilic attack to the *si* face of the LAC ring. This is possible because the N328 side chain possesses the appropriate polarity and length to interact with the LAC via an H-bond in the absence but also in the presence of the substrate.

We then modeled the ylide intermediate in the **L5-B3** active site formed from the characterized near attack conformation, using restrained MD simulations to study how the ylide is accommodated in the active site when it dissociates from the iron. Simulations indicate that the ylide intermediate, once formed, can maintain the hydrogen bond with the N328 side chain. This helps stabilize the ylide in the active site within a major binding mode where the lactone and amine aromatic rings occupy similar positions as in the substrate-bound complex (see Figure 5.6), without significant fluctuations or conformational changes (see Figure 5.6). In line with the previous observations for the **L6** system, the general hydrophobicity of **L5-B3** active site is retained and only a few water molecules can access the active site from a predefined water channel (see Figure 5.6), from the top side of the lactone ring. Similar to the **L6** variant, it is proposed that these water molecules can rapidly protonate the ylide at the C position from the top face (*si* face) of the lactone ring, thus forming the *(R)*-enantiomer of the product selectively.



- **N328** in **L5-B3** establishes persistent H-bond interactions with the LAC.
- **N328** H-bond **reorients** the **LAC** when **1a** is bound, exposing its re-enantioface for the N-nucleophilic addition.

Figure 5.6. Computational modeling of L6 and L5-B3 variants based on MD simulations. a) Representative snapshot obtained from **L6** variant MD simulations describing the major conformation explored by the LAC bound. The ∠(N–Fe–C1–C2) dihedral angle describes the relative orientation explored by the carbene (see Appendix Figure B23 for additional details and replicas). The blue surface describes the available space in the active site cavity near the LAC intermediate. **b)** Representative snapshot from **L6** variant restrained-MD simulations describing the major near-attack conformation explored by the amine **1a** for the *N*-nucleophilic attack to the LAC intermediate. **c)** Representative snapshot obtained from restrained-MD simulations with **3a**-ylide formed in **L6** active site. **d)** Representative snapshot obtained from **L5_B3** variant MD simulations describing the major conformation explored by the LAC bound. The purple surface describes the available space in the active site cavity near the LAC intermediate. **e)** Overlay of representative snapshots from **L5-**

B3 variant restrained-MD simulations describing the major near-attack conformation explored by the amine **1a** for the *N*-nucleophilic attack to the LAC intermediate. **f)** Overlay of 3 representative snapshots obtained from restrained-MD simulations with **3a**-ylide formed in **L5-B3** active site. Water molecules shown are taken from 25 random structures selected across the 100 ns MD trajectory. **g)** Probability density plots describing the conformations explored by the LAC when bound in **L5-B3** and **L6** active sites, in the absence or presence of **1a** substrate, estimated from accumulated MD trajectories. Key distances and angles are given in Å and degrees (º).

Finally, the spontaneous binding pathway of amine **1a** from the bulk solvent to the active sites of **L5-B3** and **L6** variants with the LAC bound was characterized using extensive MD simulations (See Figure 5.7 and Appendix Figure B28 and B29). These simulations were performed without imposing any restriction or applying any potential bias. Starting with 4 molecules of the amine **1a** randomly placed in the bulk solvent, a total of 10 independent MD replicas for each variant were propagated for 250 ns, observing one spontaneous substrate binding event for **L5-B3** and two for **L6**, in which the substrate accesses into the protein scaffold. These successful binding trajectories were then propagated up to 1000 ns. We observed that, in both **L5-B3** and **L6** cases, the amine substrate accesses the active site from the top side of the P411 protein. This corresponds to the substrate entrance channel previously characterized for the parent P450$_{BM3}$ and related P450 enzymes, which is located between the F/G helices and FG loop region and the B' helix (see Figure 5.7 and B29).[197] Therefore, in the observed binding events, the substrate binding pathway is not altered by evolution or the presence of the reactive carbene species. However, more simulations will be required to completely determine the association rated and binding pathways. Further structural analysis of the binding pathway in **L5-B3** revealed that some of the mutations introduced in the previous engineering effort, A330Y and Q437L, participate in substrate recognition during binding (see Figure 5.7). Specifically, Q437L is found to act as a gate that, once the substrate accesses the binding pocket in a pre-catalytic binding mode, keeps the hydrophobicity of the active site and limits the access of water molecules into it (see Figure 5.7 and B30). This would be important for the enantioselective protonation of the ylide intermediate. Notably, the catalytically relevant binding poses characterized from these spontaneous substrate binding simulations (see Figure 5.7) are equivalent to the ones previously observed from the restrained MD simulations (see Figure 5.6). These results further validate the utility of the computational approach used to study the catalytically relevant substrate binding modes based on sequential (restrained-)MD simulations (a first set of MD simulations with LAC bound, which are followed by substrate docking calculations and refined by restrained-MD simulations, see Appendix B. Computational Protocols).

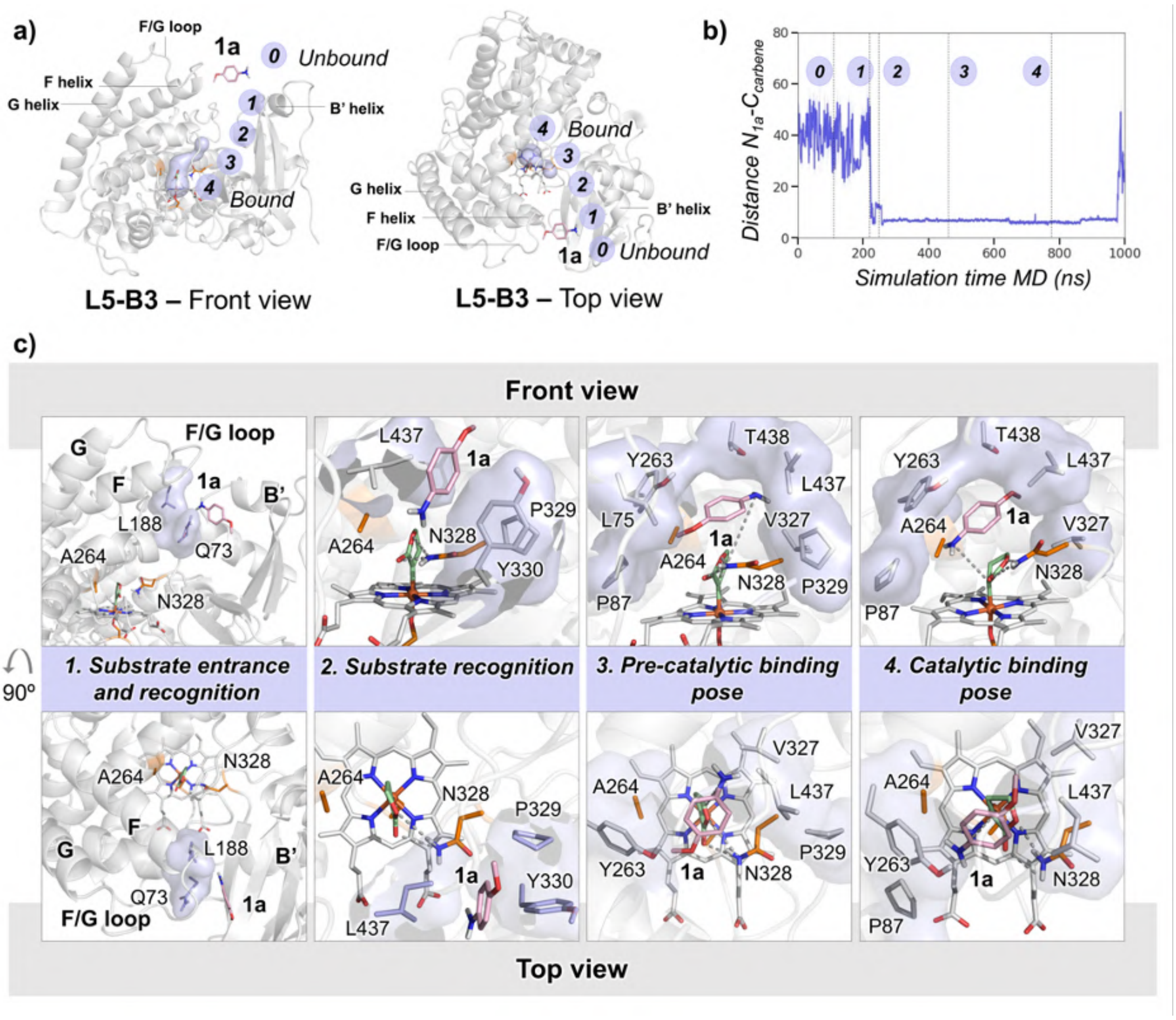


Figure 5.7. Spontaneous binding process of amine 1a in L5-B3 with LAC bound, as characterized from unbiased MD simulations. a) Schematic representation of the characterized substrate binding pathway. **b)** Binding process as described by the distance between the **1a** amine nitrogen atom (N_{1a}) and the LAC central carbon atom ($C_{carbene}$) along the MD simulation time. The N_{1a}–$C_{carbene}$ distance goes from large values (>50 Å) when the substrate is in the bulk solvent to small values (*ca.* 5 Å) when the substrate explores catalytically relevant binding modes for the *N*-

nucleophilic attack. **c)** Selected snapshots from the spontaneous binding pathway MD trajectory shown in 5b. See Appendix B and Figure B30 for further details.

Overall, with computational methods we explored the molecular basis of the biocatalytic enantiospecific N-H carbene insertion in a series of engineered P411. MD simulations and QM calculations of laboratory evolved variants indicate that the relative orientation of the LAC in the active site determines which enantiotopic face of the lactone-carbene is accessible for a selective N-nucleophilic attack by the amine substrate, prior to a final enantiospecific protonation step. The active site also precisely positions water molecules for rapid and stereoselective proton rearrangement before product release. Using this mechanistic information, we have developed an enantiodivergent enzymatic platform for carbene N–H insertion chemistry. A highly efficient, (R)-selective P411 variant, L5_FL-B3, was identified in a single round of protein engineering through a computation-assisted mechanism-guided approach. This work demonstrates that it is possible to geometrically control reactive carbene intermediates formed in enzyme active sites to modulate the selectivity of carbene transfer reactions.

Chapter 6. Molecular basis for the Selection of Formate Dehydrogenases with High Efficiency and Specificity toward NADP$^+$

The following Chapter describes the combined experimental and computational work to select formate dehydrogenases with high efficiency and specificity toward NADP$^+$ that has been published in *ACS Catal.* **2020**, 10, 14, 7512–7525. Reprinted with permission from *ACS Catal.* **2020**, 10, 14, 7512–7525 (https://pubs.acs.org/doi/full/10.1021/acscatal.0c01487). Copyright 2020 American Chemical Society. Further permissions to the material shown should be directed to the ACS.

All the experiments have been carried out by Liliana Calzadiaz-Ramirez, Gabriele M. M. Stoffel, and Steffen N. Lindner under the coordination of Tobias J. Erb, Arren Bar-Even, and Carlos G. Acevedo-Rocha (see 6.1.1 State of the art section). We (Carla Calvó-Tusell, Sílvia Osuna, and Marc Garcia-Borràs) performed the complete computational modelling and analysis (see 6.3 Results section).

6.1. State of the art

The efficient regeneration of cofactors is key for enzymatic processes occurring inside the cells or in cell-free systems.[198] There is a huge interest in developing and optimizing biocatalytic platforms that offer *in situ* cofactor regeneration. Relevant cofactors for industrial purposes include ATP, NADH, and NADPH.[199,200] Reducing agents such as formate (HCOOH) have been used *in vivo* and *in vitro* for the regeneration of NADH.[199,201] The reasons behind using formate include: (i) the existence of formate dehydrogenases (FDHs) in several organisms that efficiently use formate to obtain NADH from NAD$^+$, (ii) the essentially irreversible oxidation of formate, which drastically favors the regeneration of NADH, (iii) formate is a particularly small molecule (only five atoms) which facilitates crossing the membranes of the different cell components, and (iv) the oxidation of formate generates CO_2, which is easily removed from the system.[202]

A significant number of biocatalytic processes depend on NADPH instead of NADH.[202,203] In recent years there has been a growing interest in discovering FDHs that are able to naturally bind NADP$^+$ or evolving FDHs that originally accept NAD$^+$ toward recognizing the phosphorylated cofactor.[204–211] Despite significant advances in these directions, the catalytic efficiency of reported natural and evolved FDHs toward regenerating NADP$^+$ were still moderately low, $k_{cat}/K_M \leq 30$ s^{-1} mM^{-1}, and also the reported specificities of these enzymes toward NADP$^+$ with respect to NAD$^+$ were still far from ideal values, (k_{cat}/K_M)NADP/(k_{cat}/K_M)NAD

≤ 40. The low efficiency and specificity make the regeneration of NADP$^+$ in cell-free systems and *in vivo* more complicated. To compensate for the low catalytic efficiency, it is required to add high amounts of FDH to keep the NADPH production at a sufficient rate for cell-free biocatalytic processes. Inside cells, NADP$^+$ is found in significantly lower concentration than NAD$^+$ (approximately 100-fold).[212] This implies that using NADP-dependent FDHs with low specificity can hamper the efficient regeneration of NADPH cofactor. Considering all these aspects, the *in vivo* regeneration of NADP$^+$ would only take place in an efficient manner only when the specificity ratio (k_{cat}/K_M)NADP/(k_{cat}/K_M)NAD exceeds 100.[213] Another related problem for cells is that formate becomes toxic at high concentrations. Therefore, keeping a good affinity toward formate is also essential when selecting and evolving the most appropriate NADP-dependent FDHs. However, the affinities toward formate reported in previous studies are still relatively low (apparent K_M in the range of 50 and 200 mM).

The goal of this study is to design new efficient NADPH selective FDH variants. Computational modelling is used to generate a library for laboratory evolution of new FDH variants with high catalytic efficiency and specificity for regenerating NADPH. Experiments and computations have been combined to achieve this aim.

6.1.1. Experimental in Vivo selection and characterization of Formate Dehydrogenases with High Efficiency and Specificity toward NADP$^+$

The selected system to start with is wild-type *Pseudomonas* spp. 101 formate dehydrogenase (*Pse*FDH). This enzyme is a homodimer where each monomer consists of 400 residues and presents an active site for NAD$^+$ cofactor and formate binding. In the available *X-ray* structures of *Pse*FDH (either in the *apo* state or with NAD$^+$ bound), the C-terminal loop composed by residues S382-A393 that is found close to the active site is not completely solved, indicating a certain flexibility in this region (see Figure 6.1).[214] To have a complete *Pse*FDH structure, a computational model was built based on existing PDB structures (see section 6.2.1) and subsequently refined with molecular dynamics (MD) simulations (see section 6.3.1).

This model was used to perform a structural and dynamical analysis of the cofactor binding site of wild-type *Pse*FDH in terms of interactions between the active site residues and cofactor/substrate. This information was used to identify key residues that are expected to be important for both enzyme activity and cofactor specificity. Hydride transfer is mediated by catalytic R284 and H332. I122 and N146 participate in formate binding. Residues that directly interact with NAD^+ include: D221, H258, and E260 that interact with the adenosine ribose, S147, R201, I202, and S380 with the phosphodiester, and T282, D308, S334, and G335 establish hydrogen bonds with the nicotinamide group of the cofactor (see Figure 6.1).[215,216] Additionally, the web server CSR-SALAD (Cofactor Specificity Reversal – Structural Analysis and Library Design)[217] was used to identify furthers residues that can participate either in switching specificity from NAD^+ to $NADP^+$, which include residues D221, R222, and H223, and then, residues that can help to recover activity toward $NADP^+$, which include H258, E260, T261, H379, and S380. From previous experimental and structural studies, it is known that D221 plays a key role in switching cofactor specificity.[208] Indeed, the D221Q substitution has been reported to improve binding of $NADP^+$ by suppressing the repulsive interaction between the carboxylate group of D221 and the phosphate group of $NADP^+$ (see 6.3.1).[209,218] Based on comparing the sequence of *Pse*FDH with *Burkholderia* spp. FDH, other relevant mutations have been identified. First, A198G has been shown to improve cofactor binding when the specificity has already been improved by D221Q.[208] However, the A198G alone is not active for $NADP^+$. Second, C255I and C225V also improved the binding of the phosphorylated cofactor. However, these mutations alone are not sufficient to revert specificity, resulting in poor kinetics parameters for these variants.

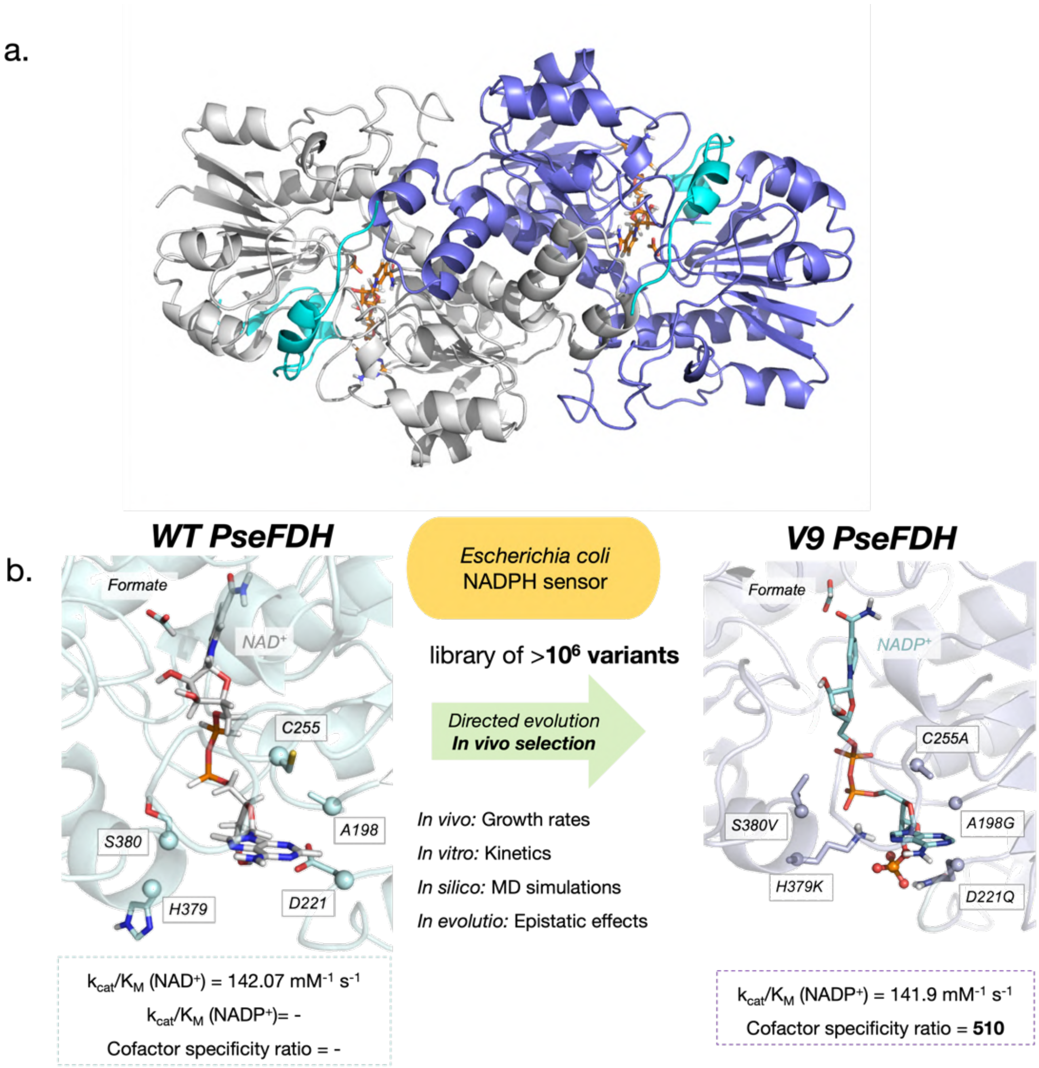

a.
b.
WT PseFDH
Formate
NAD+
C255
A198
S380
H379
D221
k_{cat}/K_M (NAD+) = 142.07 mM^{-1} s^{-1}
k_{cat}/K_M (NADP+)= -
Cofactor specificity ratio = -
Escherichia coli
NADPH sensor
library of >10^6 variants
Directed evolution
In vivo selection
In vivo: Growth rates
In vitro: Kinetics
In silico: MD simulations
In evolutio: Epistatic effects
V9 PseFDH
Formate
NADP+
C255A
S380V
A198G
H379K
D221Q
k_{cat}/K_M (NADP+) = 141.9 mM^{-1} s^{-1}
Cofactor specificity ratio = 510

Figure 6.1. Overview of *Pse*FDH enzyme and strategy followed to obtain Formate Dehydrogenases with high efficiency and specificity towards NADP$^+$. a) *Pse*FDH is a homodimeric complex formed by two non-covalently bound subunits (one subunit shown in grey and the other in purple). The *X-ray* structure used as starting point for our MD simulations is based on PDBs 2GO1 for the *apo* state and 2GUG for the *holo* state. In these *X-ray* structures, the cofactor NAD$^+$ and the important loop found in a region near cofactor binding (residues 375-400 depicted in cyan), were unsolved. This loop was reconstructed based on FDH *X-ray* structure PDB: 2NAD. **b)** Strategy followed to obtain a *Pse*FDH variant with high efficiency and specificity towards NADP$^+$. Starting with the WT enzyme, *in vivo, in vitro, in silico* and *in evolutio* techniques were applied to obtain the highly specific V9 *Pse*FDH variant.

Using the information from the different analysis provided a total of 10 positions for potential site-saturation mutagenesis including A198, D221, R222, H223, C255, H258, E260, T261, H379, and S380. Based on MD simulations (see Results 6.3.1), the R222 position was discarded. To further reduce the library size, the A198G variant was selected as the starting point for subsequent screenings, giving a total of 8 positions to screen. Moreover, the positions related to activity-recovering were divided into two groups (A: H379 and S380 and B: H258, E260, and T261), prioritizing group A because of its proximity to relevant D221 and H223 positions. Therefore, considering A198G was already present, five positions were finally screened: D221, H223, C255, H379 and S380, which gives a 3.3 x 10^7 library. Based on previous information, D221, H223 and H379 were randomized to 20 amino acids, C255 to 10 aliphatic amino acids, and S380 to 6 small side-chain residues, which results in a 3.5 × 10^6 library.

It is extremely challenging to screen a library of this size. To solve this problem, the power of natural selection can be harnessed to select the variants that are able to support NADPH regeneration. In this direction, the FDH enzymes composing the library were introduced in an engineered *E.coli* strain that deleted all enzymes producing NADPH (auxotrophic for NADPH) keeping only 6-phosphogluconate dehydrogenase for maintaining growth.[219] If gluconate is not added, then this strain can be used to assess which enzymes of the library provide *in vivo* regeneration of $NADP^+$ for keeping growth using formate as the only NADPH source. After five days, the 21 colonies that presented largest growth in a medium containing formate were kept for further analysis. By sequencing the 21 plasmids, seven different sequences were identified. Interestingly, all sequences contained D221Q and S380V indicating that these two residues play a key role in cofactor specificity.

The variant with the fastest growth is *Pse*FDH V9 which presents 5 mutations with respect to the wild-type enzyme: A198G, D221Q, C255A, H379K, and S380V. This FDH variant supports NADPH production with high specificity and catalytic efficiency. In terms of catalytic efficiency, the k_{cat}/K_M of V9 reached 142 s^{-1} mM^{-1}, which is 5-fold higher than previously reported variants. This is in part thanks to the low K_M value reported for $NADP^+$ (26 μM). On the other hand, the specificity toward $NADP^+$, (k_{cat}/K_M)NADP/(k_{cat}/K_M)NAD, in the V9 variant increased up to 510, which is 14-fold higher than previous variants. This variant also presents low apparent K_M for formate. To explore nonadditive epistatic effects in V9, several deconvoluted variants were generated and kinetic parameters evaluated. This analysis revealed that H379K mutation is critical to keep a high k_{cat} for NADP$^+$ while S380V is important for NADP$^+$ affinity. The combination of S38V with either C255A or H379K is key for obtaining a high specificity. Remarkably, none of the deconvoluted variants showed better kinetic parameters than *Pse*FDH V9, which points out that mutations act synergistically. Overall, this evolution strategy demonstrated the power of *in vivo* selection to select high specific and efficient enzyme variants with multiple mutations which would be highly complex to identify by means of conventional screening methods.

Despite the significant advances in terms of improving specificity and efficiency of FDH for the regeneration of NADPH, the molecular basis of the selection process remains unknown. To understand the preference of wild-type *Pse*FDH for NAD$^+$ and the specificity for NADP$^+$ of V9 variant at the molecular level, we performed extensive MD simulations of the WT, A198G, and V9 variants in apo, NAD$^+$-bound, and NADP$^+$-bound states.

6.2. Computational details and Protocols

6.2.1. Protein Preparation for MD simulations

*Pse*FDH is a dimeric enzyme complex composed of two identical subunits. The computational models were generated from the *apo* *Pse*FDH *X-ray* structure (PDB ID: 2GO1) and the *holo* *Pse*FDH *X-ray* structure (PDB ID 2GUG with formate in the active site). These *X-ray*s present some segments unsolved including residues 257-262 (PDB 2GUG), which are near the cofactor, and residues 375-400 (PDB 2GO1 and 2GUG) corresponding to the loop C-terminus. The missing regions were reconstructed employing PDB 2NAD as a reference to generate a complete *Pse*FDH model using Phyre2[220] and Robetta[221] web servers. NAD$^+$ and NADP$^+$ cofactors were positioned in the active site of *Pse*FDH (with formate bound) by aligning PDB 2GUG with PDB 2NAD (NAD$^+$ cofactor and azide bound, analog of the transition state). These structures were used to analyze the residues that directly interact with the cofactor (see 6.1.1).

Based on this protocol, the following models were generated for subsequent MD simulations: wild-type enzyme (WT-*apo*, WT-NAD$^+$, and WT-NADP$^+$), A198G variant (A198G-*apo*, A198G-NAD$^+$, and A198G-NADP$^+$) and V9 variant (V9-*apo*, V9-NAD$^+$, and V9-NADP$^+$). The protonation states were assigned using the H++ server, at a pH of 7.4.[222] The mutations corresponding to

A198G and V9 variants were introduced to the template wild-type *Pse*FDH structure using the mutagenesis tool of PyMOL.[223] From these coordinates, we started the MD simulations.

6.2.2. Protocol for MD Simulations

MD simulations were performed to elucidate the molecular basis of cofactor specificity in PseFDH. MD simulations were performed as described in Appendix B. Computational Protocols. The parameters for carrying out the MD simulations in the presence of the formate were generated as described in Appendix B. Computational Protocols, while the parameters for the cofactors NAD^+ and $NADP^+$ were retrieved from the Manchester parameter database (http://research.bmh.manchester.ac.uk/bryce/amber/). A total of three replicas of 500 ns MD simulations were run for each system (WT-*apo*, WT-NAD^+, WT-$NADP^+$, A198G-*apo*, A198G-NAD^+, A198G-$NADP^+$, V9-*apo*, V9-NAD^+, and V9-$NADP^+$) gathering a total of 13.5 μs of simulation time.

6.2.3. Quantum Mechanics (QM) Calculation Details

A truncated DFT computational model of the transition state (TS) structure corresponding to the cofactor reduction reaction was used to determine the optimal reaction geometric parameters. The truncated model includes the formate and only the nicotinamide ring of the cofactor. Geometry optimizations and frequency calculations were performed using the (U)B3LYP functional with the 6-31+G* basis set. Only one negative force constant corresponding to the hydrogen transfer was identified in the frequency calculation.

6.3. Results and Discussion

The results section is divided into three different parts: 1) the study of the cofactor specificity in wild-type *Pse*FDH for library generation; 2) the effect of A198G mutation; and 3) the impact of mutations in V9 variant to revert cofactor specificity toward $NADP^+$.

6.3.1. Computational Characterization of Wild-type Variant

To unravel the molecular basis of NAD^+ cofactor specificity in wild-type *Pse*FDH, we carried out MD simulations comparing the active site conformational dynamics with either NAD^+ or $NADP^+$ bound (see Figure 6.2) These simulations pointed out the stability of the natural *Pse*FDH-NAD^+ complex and the instability of the $NADP^+$ in the active site of wild-type *Pse*FDH. To understand cofactor preference, we monitored key interactions between the binding pocket and the cofactor along MD simulations. In particular, we observed that the active site residues are rearranged due to the presence of the 2′-phosphate group of $NADP^+$. Specifically, the repulsion between the negatively charged 2′-phosphate group and the carboxylate group of D221 (distance 4.7 ± 1.0 Å, see Figure 6.2) causes a change in the orientation of the adenosine ring of the cofactor. In contrast to what is observed for $NADP^+$, D221 establishes a strong hydrogen bond interaction with the 2′-hydroxyl group of NAD^+ (2.5 ± 1.2 Å, see Figure 6.2 and Appendix C1-C6) These simulations indicate that mutagenesis of negatively charged residue D221 to a neutral or positively charged amino acid is required to accommodate the negatively charged 2′-phosphate group of $NADP^+$ in the *Pse*FDH active site.

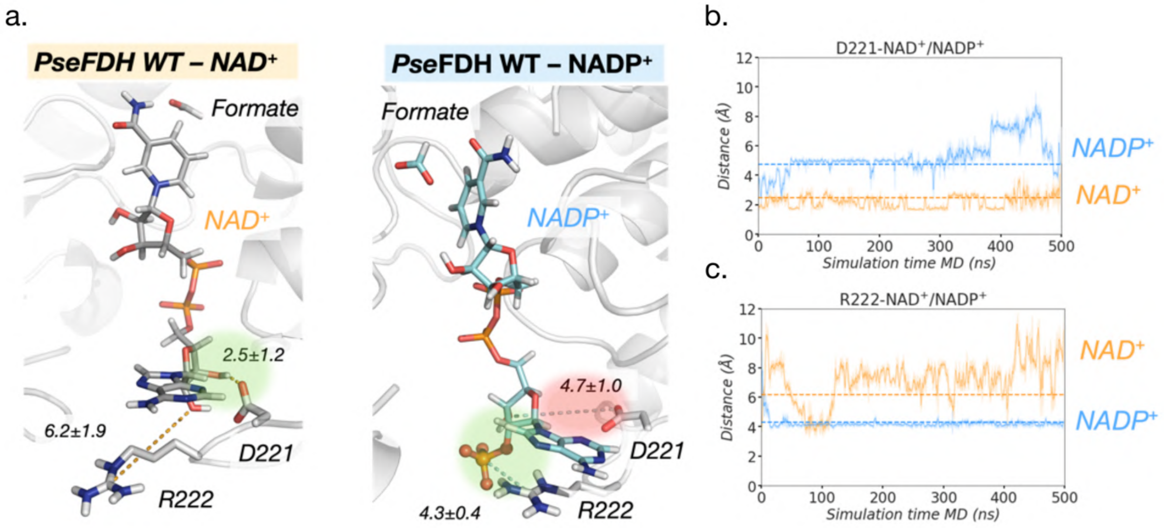


Figure 6.2. *Pse*FDH WT NAD$^+$/NADP$^+$ conformational dynamics. a) Representative structures of a *Pse*FDH wild-type (WT) active site in the presence of NAD$^+$ (gray, left) or NADP$^+$ (cyan, right) and formate extracted from MD simulations (most populated clusters). The presence of the 2′-phosphate group of NADP$^+$ causes a rearrangement of binding pocket residues. In WT-NAD$^+$, the hydrogen bond interaction between D221 and the hydrogen of the 2′-OH group of NAD$^+$ is highlighted in green. In WT-NADP$^+$, the repulsive interaction between D221 and the 2′-phosphate group of NADP$^+$ is shown in red and the salt-bridge interaction between R222 and the 2′-phosphate group of NADP$^+$ in green. Relevant average distances (in Å) obtained from MD simulations are also depicted. **b)** Plot of the distance between the carbon of the carboxylate group of D221 and either the 2′-OH group of NAD$^+$ (orange) or the 2′-phosphate group of NADP$^+$ (blue) along representative 500 ns replicas of MD simulations. Average distances (dashed orange line for WT-NAD$^+$ and dashed blue line for WT-NADP$^+$) of 2.5 ± 1.2 and 4.7 ± 1.0 Å are also shown, respectively. **c)** Plot of the distance between the carbon of the guanidinium group of R222 and either the oxygen of the 2′-OH group of NAD$^+$ (orange) or the 2′-phosphate group of NADP$^+$ (blue) along representative 500 ns replicas of MD simulations. Average distances (dashed orange line for WT-NAD$^+$ and dashed blue line for WT-NADP+) of 6.2 ± 1.9 and 4.3 ± 0.4 Å are also included, respectively. All distances are represented in Å.

Adjacent to D221, there is residue R222 with a positively charged guanidinium group. In the *apo* MD simulations, the side chain of R222 presents significant flexibility alternating between two conformations (see Appendix Figure C1). Based on the analysis of *Pse*FDH-NAD$^+$ crystal structure (PDB 2NAD), it was suggested that R222 is responsible to keep the optimal conformation of the active site, but it is not directly involved in cofactor NAD^+ binding. Wild-type *Pse*FDH-NAD^+ complex MD simulations are consistent with this idea, since the interaction between R222 and NAD^+ is not explored (see Figure 6.2 and C1). Despite the instability of $NADP^+$ in the *Pse*FDH active site, the simulations of the *Pse*FDH-$NADP^+$ complex indicated that electrostatic and cation-π interactions are established between R222 and the adenine ring of $NADP^+$ (see Figure 6.2 and C1). These results suggest that the mutagenesis of R222 might be detrimental when trying to revert the specificity of the cofactor from NAD^+ to $NADP^+$. For this reason, this position was excluded from the enzyme library (see section 6.1.1).

To get insights into the importance of *Pse*FDH active site pre-organization for the cofactor regeneration reaction, we performed QM calculations to characterize the transition state for the cofactor reduction (transfer of hydride from formate to NAD^+/$NADP^+$). This calculation provides information about the optimal geometric parameters for the reaction to occur and can be compared with the geometric features explored in MD simulations. Our results show that the optimal distance for hydride transfer is 1.38 Å and the proper angle between the substrate and the cofactor is 113° at the TS geometry. We used these geometric criteria as a reference to study the Near Attack Conformations (NAC) explored by the substrate-cofactor pair in our MD simulations. NAC will give us an idea of the catalytically competent conformations sampled along MD simulations. The NAC is monitored in all simulations that contain both the cofactor and formate in the active site. Since we started from the optimal arrangement, in both WT *Pse*FDH-NAD^+ and WT *Pse*FDH-$NADP^+$, catalytically competent distances are sampled. Interestingly, for the NAD^+ cofactor, a narrow distribution of distances and angles close to the reference catalytically competent conformation are observed. This indicates that for WT *Pse*FDH-NAD^+, the substrate and cofactor are properly positioned for catalysis during most of the simulation time. In the case of $NADP^+$, a wider range of angles and distances are sampled due to the loss of conformational stability in the binding pocket observed in MD simulations indicating the lower propensity to achieve the proper orientation for catalysis, thus suggesting lower catalytic efficiency.

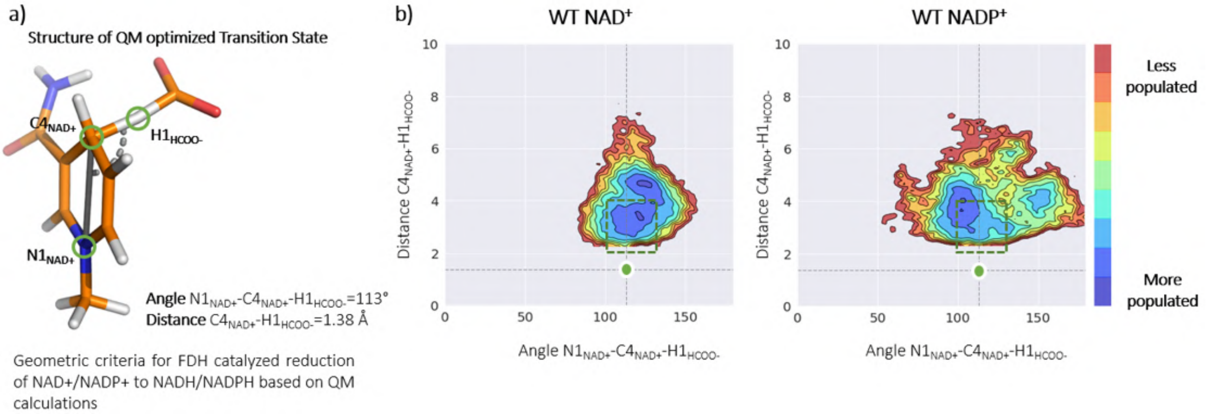


Figure 6.3. QM studies and conformational population analysis based on the QM-derived geometric criteria. a) Structure of the QM optimized Transition State for *Pse*FDH catalyzed reduction of NAD^+/$NADP^+$ with the optimal angle and distance for hydride transfer reaction. A truncated computational model of the cofactor was used in the TS calculations (see computational details). **b)** Conformational population analysis based on the geometric criteria (hydride transfer distance versus angle) for *Pse*FDH hydride transfer in the case of WT-NAD^+ (left) and in the case of WT-$NADP^+$ (right). The plots have been constructed using the angle $N1_{NAD+/NADP+}$-$C4_{NAD+/NADP+}$-$H1_{HCOO-}$ and the distance $C4_{NAD+/NADP+}$-$H1_{HCOO-}$ sampled along 3 replicas of 500 ns MD simulations for WT-NAD^+ and V9-$NADP^+$. The catalytic distance (1.38 Å, represented by a horizontal dashed black line; value obtained from QM calculation) and the proper angle (ca. 113°, represented by a vertical dashed black line; value obtained from QM calculation) required for hydride transfer is represented by a green dot. The range of distances and angles considered as catalytically relevant in our MD simulations are those found within the green box (distances that range from 2 to 4 Å and angles from 100° to 130°).

6.3.2. Computational Characterization of A198G variant

The A198G mutation is included in all variants generated in the enzyme library. The sequence motif GxGxxG is a specific feature of NAD^+ dependent dehydrogenases. However, formate dehydrogenases from bacteria (*Pse*FDH) and fungi (*Mor*FDH) contain an Alanine instead of a Glycine in the first position of the motif (position 198 in *Pse*FDH). Alekseeva and coworkers hypothesized that this mutation could decrease the conformational tension of the cofactor binding pocket region. Indeed, the substitution A198G reduces the K_M value from 0.053 to 0.035

mM for NAD$^+$, keeping K_M for formate almost unchanged. This is indicative of a better NAD$^+$ binding. When combined with other mutations it is also known to improve NADP$^+$ binding.

To explore the molecular basis of this enhancement of affinity, we carried out MD simulations for A198G-*apo*, A198G-NAD$^+$, A198G-NADP$^+$. These simulations provide the molecular and atomistic details to validate the hypothesis which says that the reduction of the conformational tension stabilizes the binding pose of the cofactor NAD$^+$. We observed that the rearrangement of active site interactions induced by A198G mutation allows the proper orientation of the Rossmann fold secondary structure elements for enhanced NAD$^+$ binding. Comparing both WT and A198G systems, we observe that the distance between D221 and 2'- and 3'- hydroxyl groups of the adenosine ribose remain formed in both WT and A198G systems. However, the impact of this mutation is represented by the rearrangement of the network of backbone interactions in the β-strand of the Rossmann fold structure located next to the cofactor (see Figure 6.4). The higher flexibility of glycine allows the proper orientation of the secondary structure for NAD$^+$ binding. Interestingly, the network of backbone interactions contains D308 residue, which is directly involved in stabilizing the amide group of the nicotinamide ring. Based on these MD simulations, we suggest that the interaction between D308 and NAD$^+$ is important to further stabilize the nicotinamide ring in the proper conformation required for the reaction to occur.

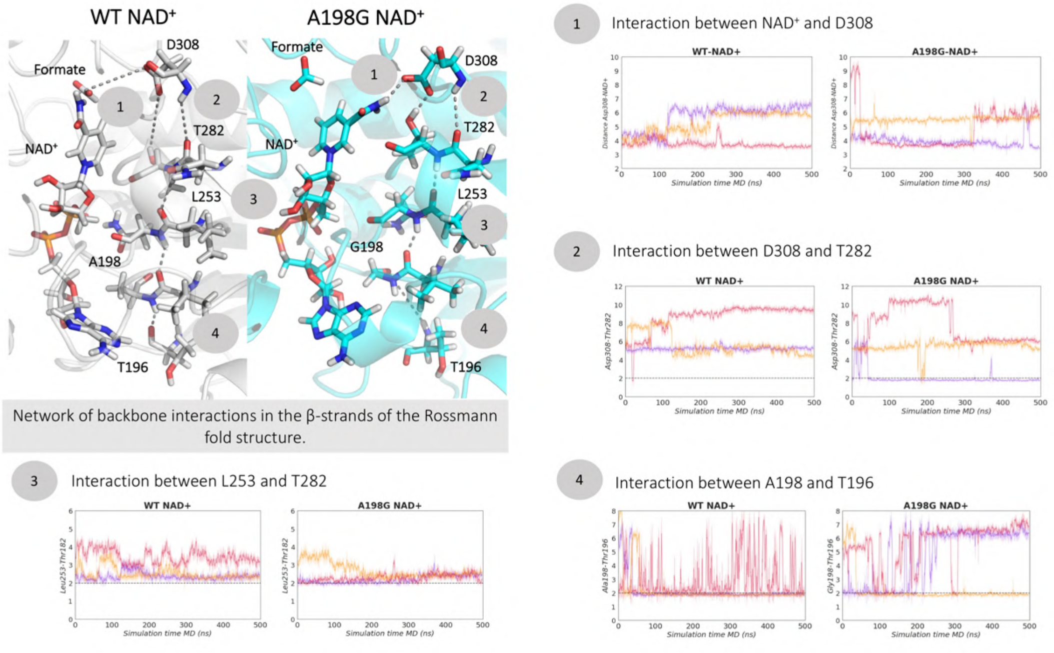


Figure 6.4. Molecular basis of the role of A198G. MD simulations show that the reduction of the conformational tension stabilizes the binding pose of the cofactor NAD$^+$. Introducing A198G mutation allows for high mobility and the proper reorganization of the secondary structure for NAD$^+$ binding. This network of backbone interactions contains D308, which is involved in stabilizing the amide of the nicotinamide ring.

Since the A198G substitution generates more space for cofactor binding, it was suggested that this additional space would help to more effectively bind the 2'-phosphate group of the NADP$^+$ cofactor. The MD simulations of A198G with NADP$^+$ bound still show a strong repulsion between the carboxylate D221 and the 2'-phosphate group, which causes the displacement of NADP$^+$. The change in orientations prevents the 2'-phosphate group from binding near G198. These results are in line with kinetic parameters that show that the A198G variant is not active toward NADP$^+$, indicating that additional mutations are required. Further analysis of the role of A198G mutations combined with the mutations introduced in the V9 variant will be key to elucidate the molecular basis of NADP$^+$ specificity.

6.3.3. Computational Characterization of V9 variant

To gain molecular insights for the high affinity and catalytic activity displayed by the *Pse*FDH V9 toward NADP$^+$, we performed MD simulations to explore the interactions of V9 active site residues with formate and NADP$^+$. The results obtained for WT *Pse*FDH were used as a reference for rationalizing the impact of the introduced mutation. In V9, D221 is substituted by a glutamine, which eliminates the repulsive interaction between the phosphate group of NADP$^+$ observed in WT *Pse*FDH. The amide group of the introduced D221Q can establish hydrogen bonds with both the 2′-phosphate and the 3′-OH group of NADP$^+$, which are frequently observed along the MD simulations (5.2 ± 1.7 Å, see Figure 6.5 and C1-C6). By eliminating the repulsive interaction allows the proper binding of the adenine ring of NADP$^+$ in a similar orientation as observed for the natural NAD$^+$ cofactor in WT *Pse*FDH. As mentioned before for WT, keeping R222 in V9 allows stabilizing the NADP$^+$ cofactor in the binding pocket through a salt bridge interaction with the 2′-phosphate group (see Figure 6.5 and C1).

In *Pse*FDH V9, H379 is mutated to a lysine, thus introducing a long side chain with a positive charge. According to the MD simulations, H379K establishes frequent but transient interactions with the negatively charged 2′-phosphate group of $NADP^+$ (4.8 ± 2.0 Å; see Figure 6.5 and C4). On the other hand, the distance between H379 and the 2′-phosphate group increases up to 9.2 ± 2.5 Å in the WT MD simulations. Additionally, the side chain of H379K interacts through a salt-bridge with the linker 4′-phosphate group of the $NADP^+$ (6.9 ± 2.7, see Figure 6.5). Again, this interaction is not observed in WT *Pse*FDH (11.0 ± 1.2 Å). Therefore, H379K plays an essential role in stabilizing the phosphate groups of the $NADP^+$ cofactor in the V9 variant.

The substitution of a cysteine by an alanine in position C255A (located in a loop surrounding the cofactor), provides an explanation for the improved binding of $NADP^+$ adenine ring in *Pse*FDH V9 (see Figure C5). Since the cysteine is mutated to a smallar alanine, extra room is generated in the binding pocket, thus providing space for the adenine ring to establish a cation–π interaction with the guanidinium group of R222. In addition, the alanine side chain establishes hydrophobic CH–π interactions with the adenine ring further stabilizing this moiety of $NADP^+$ in the active site (4.7 ± 0.7 Å; see Figure 6.5 and C5). Finally, the S380V mutation increases the hydrophobic character of this region of the binding pocket. The side chain of V380 directly interacts with the side chain of P256 that is found in the loop surrounding the cofactor next to C255A (see Figure 6.5, C5 and C6). When this interaction is established, $NADP^+$ is wrapped in the binding pocket, which was not observed in the presence of $NADP^+$ WT *Pse*FDH.

Overall, the mutations incorporated in V9 *Pse*FDH have reshaped the active site in terms of electrostatics to stabilize $NADP^+$ cofactor in an orientation similar to WT *Pse*FDH with the natural cofactor NAD^+. Additionally, MD simulations indicated that the nicotinamide ring of the $NADP^+$ in V9 shows slightly higher flexibility than NAD^+ in WT *Pse*FDH, exploring different orientations in the active site. To evaluate the impact of these motions in the catalytic efficiency when the cofactor specificity is switched, we analyzed the near attack conformations (NACs) explored by formate with respect to the cofactor nicotinamide ring for productive hydride transfer along the MD trajectories, as similarly done with the WT enzyme (see 6.3.1). From these analysis, we observed that formate bound to V9 *Pse*FDH with $NADP^+$ has the ability to explore catalytically competent poses in the MD simulations ($C4_{NADP+}$–$H1_{HCOO-}$ distance below 4 Å; and $N1_{NADP+}$–$C4_{NADP+}$–$H1_{HCOO-}$ attack angles of *ca.* 100–130°, see Appendix Figure C7). V9 *Pse*FDH with cofactor $NADP^+$ explores a wider range of near attack angles for hydride transfer than in

WT *Pse*FDH-NAD$^+$, which can be attributed to the higher flexibility of the nicotinamide ring. This could explain the slightly lower k_{cat} value of *Pse*FDH V9 with NADP$^+$ with respect to that for *Pse*FDH WT with NAD$^+$ and the significant increase in apparent K_M values by formate in *Pse*FDH V9 in comparison to *Pse*FDH WT.

Therefore, combining molecular dynamics simulations with experimental kinetic data, we can rationalize the contribution of each mutation to the superior kinetics and specificity of this variant.

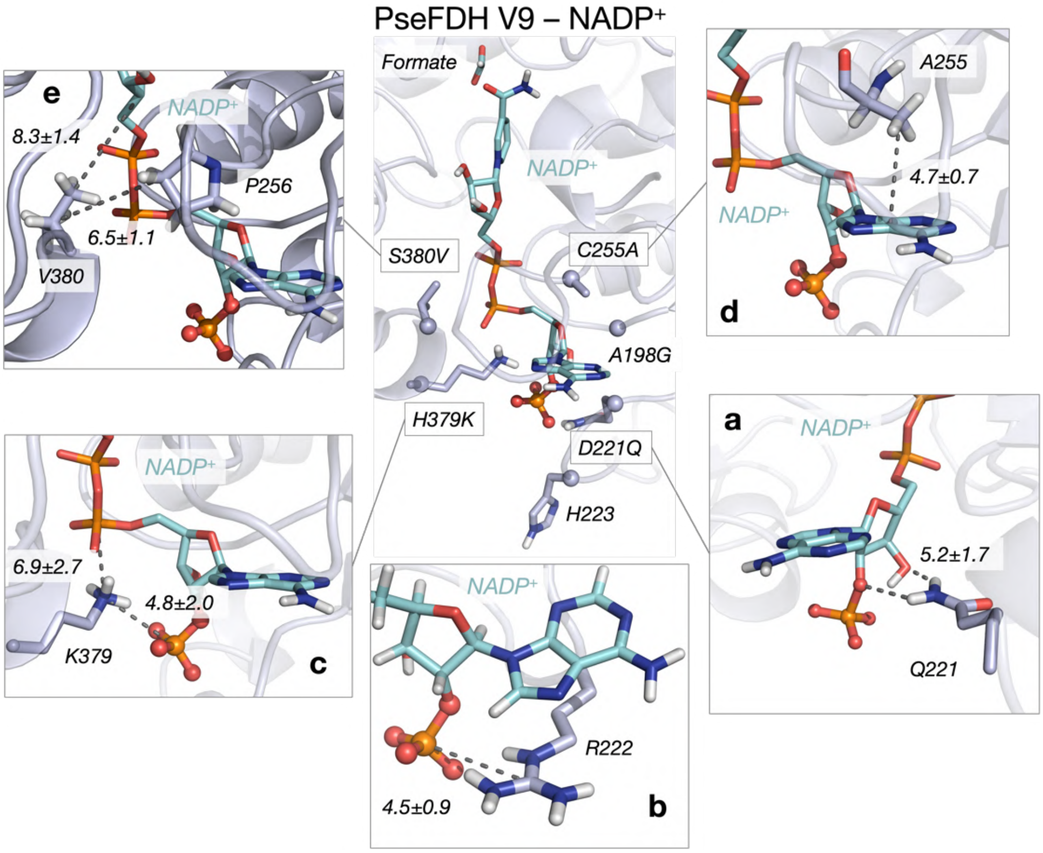


Figure 6.5. Conformational dynamics of *Pse*FDH V9. A representative structure of the reshaped active site with NADP$^+$ (cyan) and formate extracted from MD simulations (most populated cluster) is shown in the center with introduced mutations (Cα atoms depicted as spheres). **a)** Representative structure of hydrogen bonds between D221Q and 2′-phosphate and the 3′-OH group of NADP$^+$. The average distance between the nitrogen of the amide group of D221Q and 3′-OH group of NADP$^+$ (5.2 ± 1.7 Å) is shown. To calculate this distance, nitrogen is selected because the hydrogens of the amide group can interconvert through rotation of the Q221 side chain. For clarity, a gray dash line representative of hydrogen bonds is depicted. **b)** Representative structure of the salt-bridge

interaction between the guanidinium group of R222 and the 2′-phosphate group of NADP$^+$ (mean distance of 4.5 ± 0.9 Å) and the cation–π interaction between the guanidinium group of R222 and the adenine group of NADP$^+$. **c)** Representative structure of the salt-bridge interaction between the amino group of H379K and the 2′-phosphate group of NADP$^+$ (4.8 ± 2.0 Å) and the salt-bridge interaction between the amino group of H379K and the linker 4′-phosphate group of NADP$^+$ (6.9 ± 2.7 Å). **d)** Representative structure of the CH–π interaction between the adenine ring of NADP$^+$ and the β-carbon of the side chain of C255A. The average distance between the center of mass (COM) of the NADP$^+$ adenine ring and the side chain of C255A (4.7 ± 0.7 Å) is depicted. **e)** Representative structure of the interactions between the side chain of S380V and the side chain of P256 (with an average distance of 6.5 ± 1.1 Å) and the interactions between the side chain of S380V and the nicotinamide ribose group of NADP$^+$ (8.3 ± 1.4 Å). All representative structures are extracted from the most populated clusters of three replicas of 500 ns of MD simulations for V9-NADP$^+$. All distances are represented in Å.

Chapter 7. Conclusions

In this thesis, we have explored the molecular basis of biochemical and biocatalytic processes using computational methods. In particular, we have designed computational strategies that combine multiscale methods such as molecular dynamics simulations, enhanced sampling techniques, correlation-based tools, and quantum mechanics to provide relevant information of enzyme function and to give insights for design. In general, the studies of this thesis put forward the importance of understanding enzyme function at molecular level to harness this information for rational enzyme design.

The main conclusions of the projects described in **Chapters 4-6** are described as follows:

In **Chapter 4**, we characterized the millisecond allosteric activation of imidazole glycerol phosphate synthase (IGPS). To this end, we have designed a computational strategy tailored to reconstruct millisecond time scale events, that combines molecular dynamics simulations, enhanced sampling techniques and dynamical networks. We captured the essential molecular details of the time evolution of the millisecond allosteric activation of IGPS in the ternary complex. Based on these extensive conformational sampling simulations, we suggested a general scheme for describing the IGPS allosteric activation pathway taking place prior to the chemical step. First, the *h*V51 oxyanion hole formation and closure of the HisF:HisH interface pre-exist in solution in the substrate-free form, although both are high-energy states in the IGPS-PRFAR conformational ensemble. Second, substrate recognition occurs in the IGPS open HisF:HisH interface state, while the oxyanion strand attains an inactive conformation. Third, the interdomain region productively closes to retain the glutamine substrate in the HisH active site. Finally, formation of the *h*V51 oxyanion hole couples with the repositioning of the substrate in a catalytically productive pose to finally form the allosterically active state. The formation of this allosterically active state is controlled by fine-tuned correlated motions connecting the PRFAR effector and HisH binding sites that are activated throughout the whole process.

The proposed model of the allosteric activation pathway of IGPS based on the millisecond time scale computational strategy developed provides multiple unprecedented molecular insights not previously identified by means of *X-ray* crystallography, solution NMR experiments, and short time scale MD simulations. Most importantly, it also answers many of the open questions existing in IGPS allosteric regulation and function. This computational strategy can be used to decipher the molecular basis of allosteric mechanisms in related enzymes, which is key for developing new therapeutics and engineering novel enzymatic functions in IGPS and related systems.

In **Chapter 5**, we have studied the molecular basis of biocatalytic enantiospecific N-H carbene insertion and the development of an enzymatic platform for enantiodivergent carbene N–H insertion. The engineered P411 enzyme **L7_FL** acted as a dual-function biocatalyst that promoted the transfer of the lactone-carbene to amines and exerted excellent stereocontrol in the subsequent protonation step. Computational studies based on MD simulations and DFT calculations elucidated the detailed mechanism of this fascinating process, explaining the critical role of the serine residue at position 264 for achieving high activity and selectivity. The engineered active site controls the conformation of the lactone-carbene, yielding to an enantioselective *N*-nucleophilic attack for the ylide formation; it also precisely positions water molecules for rapid and stereoselective proton rearrangement before product release.

Using mechanistic information, we have developed an enantiodivergent enzymatic platform for carbene N–H insertion chemistry. A highly efficient, *(R)*-selective P411 variant, **L5_FL-B3**, was identified in a single round of protein engineering through a computation-assisted mechanism-guided approach. This variant complements the previously engineered *(S)*-selective mutants. Computational modeling was used to investigate the key LAC intermediates formed in the active site. These models served as starting points to search and characterize key positions for controlling the orientation of the LAC intermediate via H-bond interactions. The relative orientation of the LAC in the active site determines which enantiotopic face of the lactone-carbene is accessible for a selective *N*-nucleophilic attack by the amine substrate, prior to a final enantiospecific protonation step. MD simulations were employed to elucidate the origin of enantioselectivity and high activity of **L5_FL-B3**, and to characterize the amine binding process. This is the first time that substrate binding pathways in carbene transferases have been fully characterized. We also showed that **L5_FL-B3** could accept a broad scope of substrates with excellent yields (up to >99% yield, 12,300 TTN) and good enantiocontrol (up to 7:93 *er*).

This work demonstrates that it is possible to geometrically control reactive carbene intermediates formed in enzyme active sites to modulate the selectivity of carbene transfer reactions. Beyond our example, there have been many more biocatalytic transformations, natural or non-natural, recruiting similar hydrogen bonds in enzymes' active sites to drive stereoselectivity, but very few have demonstrated proper protein engineering to introduce a different hydrogen bond-anchoring point to reaction intermediates could alter the stereo- or site-selectivity. Therefore, we hope our study will inspire more mechanism-driven protein engineering efforts, aiming to control key biocatalytic intermediates formed in enzyme active sites to enhance activity and control selectivity.

In **Chapter 6**, we have described the molecular basis for the selection of formate dehydrogenases with high efficiency and specificity toward NADP$^+$. Using MD simulations, first we were able to guide efficient library construction by providing essential structural insights for the wild-type enzyme and its preference for NAD$^+$ over NADP$^+$ binding. Then, we were able to uncover the molecular basis of the increased activity and selectivity of the new engineered variants. We further determined the existence of strong non-additive epistatic effects, which are difficult to predict via rational design or iterative SSM but are essential to overcome activity and selectivity tradeoffs. Only a few studies used MD simulations to understand the interactions occurring between NAD(P)$^+$ and FDHs variants. In our work, MD simulations highlighted the important role of the mutations in *Pse*FDH V9 that act together to reshape and modulate the polarity of the binding pocket of the enzyme, allowing the formation of new polar interactions with NADP$^+$. In particular, H379K and R222 were found to be instrumental for stabilizing the additional negatively charged 2′-phosphate group of NADP$^+$, whereas D221Q reduced the electrostatic repulsion generated by the original aspartate residue of the WT enzyme. C255A and S380V decreased the polarity of the active site while simultaneously reshaped the binding pocket.

References

(1) Copeland, R. A. *Enzymes: A Practical Introduction to Structure, Mechanism, and Data Analysis*, 2nd ed.; Wiley-VCH, Ed.; New York, 200AD.

(2) Robinson, P. K. Enzymes: Principles and Biotechnological Applications. *Essays Biochem.* **2015**, *59*, 1–41. https://doi.org/10.1042/BSE0590001.

(3) Alberts, B.; Johnson, A.; Lewis, J.; Raff, M.; Roberts, K.; Water, P. *Molecular Biology of the Cell*, 5th ed.; Garland Science, 2008.

(4) Bottaro, S.; Lindorff-Larsen, K. Biophysical Experiments and Biomolecular Simulations: A Perfect Match? *Science (80-.).* **2018**, *361* (6400), 355–360. https://doi.org/10.1126/science.aat4010.

(5) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2.

(6) Kerns, J. S.; Agafonov, R. V.; Cho, Y.-J.; Pontiggia, F.; Otten, R.; Pachov, D. V.; Kutter, S.; Phung, L. A.; Murphy, P. N.; Thai, V.; Alber, T.; Hagan, M. F.; Kern, D. The Energy Landscape of Adenylate Kinase during Catalysis. *Nat. Struct. Mol. Biol.* **2015**, *22* (2), 124–131. https://doi.org/10.1038/nsmb.2941.

(7) Peracchi, A. The Limits of Enzyme Specificity and the Evolution of Metabolism. *Trends Biochem. Sci.* **2018**, *43* (12), 984–996. https://doi.org/10.1016/j.tibs.2018.09.015.

(8) Hedstrom, L. Enzyme Specificity and Selectivity. *eLS* **2010**, No. February. https://doi.org/10.1002/9780470015902.a0000716.pub2.

(9) Casalino, L.; Nierzwicki, L.; Jinek, M.; Palermo, G. Catalytic Mechanism of Non-Target DNA Cleavage in CRISPR- Cas9 Revealed by Ab Initio Molecular Dynamics. *ACS Catal.* **2020**, *10* (22), 13596–13605. https://doi.org/10.1021/acscatal.0c03566.

(10) Crean, R. M.; Biler, M.; Van Der Kamp, M. W.; Hengge, A. C.; Kamerlin, S. C. L. Loop

Dynamics and Enzyme Catalysis in Protein Tyrosine Phosphatases. *J. Am. Chem. Soc.* **2021**, *143* (10), 3830–3845. https://doi.org/10.1021/jacs.0c11806.

(11) Tokuriki, N.; Tawfik, D. S. Protein Dynamism and Evolvability. *Science* **2009**, *324* (April), 203–207.

(12) Henzler-Wildman, K.; Kern, D. Dynamic Personalities of Proteins. *Nature*. 2007. https://doi.org/10.1038/nature06522.

(13) Nussinov, R. Introduction to Protein Ensembles and Allostery. *Chem. Rev.* **2016**, *116* (11), 6263–6266. https://doi.org/10.1021/acs.chemrev.6b00283.

(14) Ma, B.; Nussinov, R. Conformational Footprints. *Nat. Chem. Biol.* **2016**, *12* (11), 890–891. https://doi.org/10.1038/nchembio.2212.

(15) Guo, J.; Zhou, H. X. Protein Allostery and Conformational Dynamics. *Chemical Reviews*. 2016. https://doi.org/10.1021/acs.chemrev.5b00590.

(16) Lisi, G. P.; Loria, J. P. Solution NMR Spectroscopy for the Study of Enzyme Allostery. *Chem. Rev.* **2016**, *116* (11), 6323–6369. https://doi.org/10.1021/acs.chemrev.5b00541.

(17) Orozco, M. A Theoretical View of Protein Dynamics. *Chem. Soc. Rev.* **2014**, *43* (14), 5051–5066. https://doi.org/10.1039/c3cs60474h.

(18) Maria-Solano, M. A.; Serrano-Hervás, E.; Romero-Rivera, A.; Iglesias-Fernández, J.; Osuna, S. Role of Conformational Dynamics in the Evolution of Novel Enzyme Function. *Chem. Commun.* **2018**, *54* (50), 6622–6634. https://doi.org/10.1039/c8cc02426j.

(19) Papaleo, E.; Saladino, G.; Lambrughi, M.; Lindorff-Larsen, K.; Gervasio, F. L.; Nussinov, R. The Role of Protein Loops and Linkers in Conformational Dynamics and Allostery. *Chem. Rev.* **2016**, *116* (11), 6391–6423. https://doi.org/10.1021/acs.chemrev.5b00623.

(20) Grossfield, A.; Patrone, P. N.; Roe, D. R.; Schultz, A. J.; Siderius, D. W.; Zuckerman, D. M. Best Practices for Quantification of Uncertainty and Sampling Quality in Molecular Simulations. *Living J. Comput. Mol. Sci.* **2018**, *1* (1). https://doi.org/doi:10.33011/livecoms.1.1.5067.

(21) Osuna, S.; Jiménez-Osés, G.; Noey, E. L.; Houk, K. N. Molecular Dynamics Explorations of Active Site Structure in Designed and Evolved Enzymes. *Acc. Chem. Res.* **2015**, *48* (4), 1080–1089. https://doi.org/10.1021/ar500452q.

(22) Zuckerman, D. M. Equilibrium Sampling in Biomolecular Simulations. *Annu. Rev.*

Biophys. **2011**, *40* (1), 41–62. https://doi.org/10.1146/annurev-biophys-042910-155255.

(23) Dror, R. O.; Dirks, R. M.; Grossman, J. P.; Xu, H.; Shaw, D. E. Biomolecular Simulation: A Computational Microscope for Molecular Biology. *Annu. Rev. Biophys.* **2012**, *41* (1), 429–452. https://doi.org/10.1146/annurev-biophys-042910-155245.

(24) Duan, Y.; Kollman, P. A. Pathways to a Protein Folding Intermediate Observed in a 1-Microsecond Simulation in Aqueous Solution. *Science (80-.).* **1998**, *282* (5389), 740–744. https://doi.org/10.1126/science.282.5389.740.

(25) Markwick, P. R. L.; McCammon, J. A. Studying Functional Dynamics in Bio-Molecules Using Accelerated Molecular Dynamics. *Phys. Chem. Chem. Phys.* **2011**, *13* (45), 20053–20065. https://doi.org/10.1039/c1cp22100k.

(26) Husic, B. E.; Pande, V. S. Markov State Models: From an Art to a Science. *J. Am. Chem. Soc.* **2018**, *140* (7), 2386–2396. https://doi.org/10.1021/jacs.7b12191.

(27) Ingram, J. R.; Knockenhauer, K. E.; Markus, B. M.; Mandelbaum, J.; Ramek, A.; Shan, Y.; Shaw, D. E.; Schwartz, T. U.; Ploegh, H. L.; Lourido, S. Allosteric Activation of Apicomplexan Calcium-Dependent Protein Kinases. *Proc. Natl. Acad. Sci.* **2015**. https://doi.org/10.1073/pnas.1505914112.

(28) Zuckerman, D. M.; Chong, L. T. Weighted Ensemble Simulation: Review of Methodology, Applications, and Software. *Annu. Rev. Biophys.* **2017**, *46* (February), 43–57. https://doi.org/10.1146/annurev-biophys-070816-033834.

(29) Sohraby, F.; Nunes-Alves, A. Advances in Computational Methods for Ligand Binding Kinetics. *Trends Biochem. Sci.* **2022**, *xx* (xx), 1–13. https://doi.org/10.1016/j.tibs.2022.11.003.

(30) Sztain, T.; Ahn, S. H.; Bogetti, A. T.; Casalino, L.; Goldsmith, J. A.; Seitz, E.; McCool, R. S.; Kearns, F. L.; Acosta-Reyes, F.; Maji, S.; Mashayekhi, G.; McCammon, J. A.; Ourmazd, A.; Frank, J.; McLellan, J. S.; Chong, L. T.; Amaro, R. E. A Glycan Gate Controls Opening of the SARS-CoV-2 Spike Protein. *Nat. Chem.* **2021**, *13* (10), 963–968. https://doi.org/10.1038/s41557-021-00758-3.

(31) Barducci, A.; Bonomi, M.; Parrinello, M. Metadynamics. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *1* (5), 826–843. https://doi.org/10.1002/wcms.31.

(32) Valleau, J. P.; J.M., T. Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling. *J. Comput. Phys.* **1977**, *23* (2), 187–199.

(33) Laio, A.; Parrinello, M. Escaping Free-Energy Minima. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99* (20), 12562–12566. https://doi.org/10.1073/pnas.202427399.

(34) Doshi, U.; Hamelberg, D. Towards Fast, Rigorous and Efficient Conformational Sampling of Biomolecules: Advances in Accelerated Molecular Dynamics. *Biochim. Biophys. Acta - Gen. Subj.* **2015**, *1850* (5), 878–888. https://doi.org/10.1016/j.bbagen.2014.08.003.

(35) Miao, Y.; McCammon, J. A. Gaussian Accelerated Molecular Dynamics: Theory, Implementation, and Applications. *Annu. Rep. Comput. Chem.* **2017**, *13*, 213–278. https://doi.org/10.1016/bs.arcc.2017.06.005.

(36) Sultan, M. M.; Pande, V. S. Automated Design of Collective Variables Using Supervised Machine Learning. *J. Chem. Phys.* **2018**, *149* (9). https://doi.org/10.1063/1.5029972.

(37) Changeux, J. P.; Christopoulos, A. Allosteric Modulation as a Unifying Mechanism for Receptor Function and Regulation. *Cell*. 2016. https://doi.org/10.1016/j.cell.2016.08.015.

(38) Vanwart, A. T.; Eargle, J.; Luthey-Schulten, Z.; Amaro, R. E. Exploring Residue Component Contributions to Dynamical Network Models of Allostery. *J. Chem. Theory Comput.* **2012**. https://doi.org/10.1021/ct300377a.

(39) Motlagh, H. N.; Wrabl, J. O.; Li, J.; Hilser, V. J. The Ensemble Nature of Allostery. *Nature* **2014**, *508* (7496), 331–339. https://doi.org/10.1038/nature13001.

(40) Gunasekaran, K.; Ma, B.; Nussinov, R. Is Allostery an Intrinsic Property of All Dynamic Proteins? *Proteins Struct. Funct. Genet.* **2004**, *57* (3), 433–443. https://doi.org/10.1002/prot.20232.

(41) Greener, J. G.; Sternberg, M. J. Structure-Based Prediction of Protein Allostery. *Curr. Opin. Struct. Biol.* **2018**, *50*, 1–8. https://doi.org/10.1016/j.sbi.2017.10.002.

(42) Nussinov, R.; Tsai, C. J. Allostery in Disease and in Drug Discovery. *Cell*. 2013. https://doi.org/10.1016/j.cell.2013.03.034.

(43) Liu, J.; Nussinov, R. Allostery: An Overview of Its History, Concepts, Methods, and Applications. *PLoS Computational Biology*. 2016. https://doi.org/10.1371/journal.pcbi.1004966.

(44) Cui, Q.; Karplus, M. Allostery and Cooperativity Revisited. *Protein Sci.* **2008**. https://doi.org/10.1110/ps.03259908.

(45) Jacob, F.; Monod, J. Genetic Regulatory Mechanisms in the Synthesis of Proteins. *J. Mol. Biol.* **1961**, *3* (3), 318–356. https://doi.org/10.1016/S0022-2836(61)80072-7.

(46) Perutz, M. F.; Rossmann, M. G.; Cullis, A. F.; Muirhead, H.; Will, G. Structure of Haemoglobin. *Nature* **1960**, *185*, 416–422.

(47) Monod, J.; Wyman, J.; Changeux, J. P. On the Nature of Allosteric Transitions: A Plausible Model. *J. Mol. Biol.* **1965**, *12* (1), 88–118. https://doi.org/10.1016/S0022-2836(65)80285-6.

(48) Koshland, D. E.; NCmethy, G.; Filmer, D. Models in Proteins Containing Subunits. *Biochem. Biophys. Res. Commun* **1965**, *12* (1), 880.

(49) Changeux, J. P. Allostery and the Monod-Wyman-Changeux Model after 50 Years. *Annu. Rev. Biophys.* **2012**, *41* (1), 103–133. https://doi.org/10.1146/annurev-biophys-050511-102222.

(50) KOSHLAND, D. E. Enzyme Flexibility and Enzyme Action. *J. Cell. Comp. Physiol.* **1959**, *54*, 245–258. https://doi.org/10.1002/jcp.1030540420.

(51) Kovermann, M.; Grundström, C.; Elisabeth Sauer-Eriksson, A.; Sauer, U. H.; Wolf-Watz, M. Structural Basis for Ligand Binding to an Enzyme by a Conformational Selection Pathway. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114* (24), 6298–6303. https://doi.org/10.1073/pnas.1700919114.

(52) Cooper, A.; Dryden, D. T. F. Allostery without Conformational Change - A Plausible Model. *Eur. Biophys. J.* **1984**, *11* (2), 103–109. https://doi.org/10.1007/BF00276625.

(53) Malmstrom, R. D.; Kornev, A. P.; Taylor, S. S.; Amaro, R. E. Allostery through the Computational Microscope: CAMP Activation of a Canonical Signalling Domain. *Nat. Commun.* **2015**, *6* (May). https://doi.org/10.1038/ncomms8588.

(54) Wagner, J. R.; Lee, C. T.; Durrant, J. D.; Malmstrom, R. D.; Feher, V. A.; Amaro, R. E. Emerging Computational Methods for the Rational Discovery of Allosteric Drugs. *Chem. Rev.* **2016**, *116* (11), 6370–6390. https://doi.org/10.1021/acs.chemrev.5b00631.

(55) Lisi, G. P.; Loria, J. P. Allostery in Enzyme Catalysis. *Curr. Opin. Struct. Biol.* **2017**, *47*, 123–130. https://doi.org/10.1016/j.sbi.2017.08.002.

(56) Fenton, A. W. Allostery: An Illustrated Definition for the "Second Secret of Life." https://doi.org/10.1016/j.tibs.2008.05.009.

(57) Lipchock, J. M.; Loria, J. P. Nanometer Propagation of Millisecond Motions in V-Type Allostery. *Structure* **2010**, *18* (12), 1596–1607. https://doi.org/10.1016/j.str.2010.09.020.

(58) Formoso, E.; Limongelli, V.; Parrinello, M. Energetics and Structural Characterization of the Large-Scale Functional Motion of Adenylate Kinase. *Sci. Rep.* **2015**, *5*, 8425. https://doi.org/10.1038/srep08425.

(59) Romero-Rivera, A.; Garcia-Borràs, M.; Osuna, S. Role of Conformational Dynamics in the Evolution of Retro-Aldolase Activity. *ACS Catal.* **2017**, *7* (12), 8524–8532. https://doi.org/10.1021/acscatal.7b02954.

(60) Van Wart, A. T.; Durrant, J.; Votapka, L.; Amaro, R. E. Weighted Implementation of Suboptimal Paths (WISP): An Optimized Algorithm and Tool for Dynamical Network Analysis. *J. Chem. Theory Comput.* **2014**. https://doi.org/10.1021/ct4008603.

(61) La Sala, G.; Decherchi, S.; De Vivo, M.; Rocchia, W. Allosteric Communication Networks in Proteins Revealed through Pocket Crosstalk Analysis. *ACS Cent. Sci.* **2017**, *3* (9), 949–960. https://doi.org/10.1021/acscentsci.7b00211.

(62) Maria-Solano, M. A.; Iglesias-Fernández, J.; Osuna, S. Deciphering the Allosterically Driven Conformational Ensemble in Tryptophan Synthase Evolution. *J. Am. Chem. Soc.* **2019**, *141* (33). https://doi.org/10.1021/jacs.9b03646.

(63) Sethi, A.; Eargle, J.; Black, A. A.; Luthey-Schulten, Z. Dynamical Networks in TRNA: Protein Complexes. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106* (16), 6620–6625. https://doi.org/10.1073/pnas.0810961106.

(64) Vu, P. J.; Yao, X. Q.; Momin, M.; Hamelberg, D. Unraveling Allosteric Mechanisms of Enzymatic Catalysis with an Evolutionary Analysis of Residue-Residue Contact Dynamical Changes. *ACS Catal.* **2018**, *8* (3), 2375–2384. https://doi.org/10.1021/acscatal.7b04263.

(65) Mak, W. S.; Siegel, J. B. Computational Enzyme Design: Transitioning from Catalytic Proteins to Enzymes. *Current Opinion in Structural Biology*. 2014. https://doi.org/10.1016/j.sbi.2014.05.010.

(66) Boyle, J. Lehninger Principles of Biochemistry (4th Ed.): Nelson, D., and Cox, M. *Biochem. Mol. Biol. Educ.* **2005**, *33* (1). https://doi.org/10.1002/bmb.2005.494033010419.

(67) Truppo, M. D. Biocatalysis in the Pharmaceutical Industry: The Need for Speed. *ACS*

Med. Chem. Lett. **2017**, *8* (5). https://doi.org/10.1021/acsmedchemlett.7b00114.

(68) Chica, R. A.; Doucet, N.; Pelletier, J. N. Semi-Rational Approaches to Engineering Enzyme Activity: Combining the Benefits of Directed Evolution and Rational Design. *Current Opinion in Biotechnology*. 2005. https://doi.org/10.1016/j.copbio.2005.06.004.

(69) Reetz, M. T.; Wang, L. W.; Bocola, M. Directed Evolution of Enantioselective Enzymes: Iterative Cycles of CASTing for Probing Protein-Sequence Space. *Angew. Chemie - Int. Ed.* **2006**, *45* (8). https://doi.org/10.1002/anie.200502746.

(70) Reetz, M. T. Laboratory Evolution of Stereoselective Enzymes: A Prolific Source of Catalysts for Asymmetric Reactions. *Angewandte Chemie - International Edition*. 2011. https://doi.org/10.1002/anie.201000826.

(71) Ebert, M. C.; Pelletier, J. N. Computational Tools for Enzyme Improvement: Why Everyone Can – and Should – Use Them. *Current Opinion in Chemical Biology*. 2017. https://doi.org/10.1016/j.cbpa.2017.01.021.

(72) Reetz, M. T.; Garcia-Borràs, M. The Unexplored Importance of Fleeting Chiral Intermediates in Enzyme-Catalyzed Reactions. *Journal of the American Chemical Society*. 2021. https://doi.org/10.1021/jacs.1c04551.

(73) Liu, Z.; Arnold, F. H. New-to-Nature Chemistry from Old Protein Machinery: Carbene and Nitrene Transferases. *Current Opinion in Biotechnology*. NIH Public Access June 1, 2021, pp 43–51. https://doi.org/10.1016/j.copbio.2020.12.005.

(74) Moody, P. C. E.; Raven, E. L. The Nature and Reactivity of Ferryl Heme in Compounds i and II. *Acc. Chem. Res.* **2018**, *51* (2). https://doi.org/10.1021/acs.accounts.7b00463.

(75) Yang, Y.; Arnold, F. H. Navigating the Unnatural Reaction Space: Directed Evolution of Heme Proteins for Selective Carbene and Nitrene Transfer. *Acc. Chem. Res.* **2021**, *54* (5), 1209–1225. https://doi.org/10.1021/acs.accounts.0c00591.

(76) Coelho, P. S.; Brustad, E. M.; Kannan, A.; Arnold, F. H. Olefin Cyclopropanation via Carbene Transfer Catalyzed by Engineered Cytochrome P450 Enzymes. *Science (80-.).* **2013**, *339* (6117). https://doi.org/10.1126/science.1231434.

(77) Singh, R.; Bordeaux, M.; Fasan, R. P450-Catalyzed Intramolecular Sp3 C-H Amination with Arylsulfonyl Azide Substrates. *ACS Catal.* **2014**, *4* (2). https://doi.org/10.1021/cs400893n.

(78) McIntosh, J. A.; Coelho, P. S.; Farwell, C. C.; Wang, Z. J.; Lewis, J. C.; Brown, T. R.; Arnold, F. H. Enantioselective Intramolecular C-H Amination Catalyzed by Engineered Cytochrome P450 Enzymes in Vitro and in Vivo. *Angew. Chemie - Int. Ed.* **2013**, *52* (35). https://doi.org/10.1002/anie.201304401.

(79) Zanghellini, A.; Jiang, L.; Wollacott, A. M.; Cheng, G.; Meiler, J.; Althoff, E. A.; Röthlisberger, D.; Baker, D. New Algorithms and an in Silico Benchmark for Computational Enzyme Design. *Protein Sci.* **2006**, *15* (12). https://doi.org/10.1110/ps.062353106.

(80) Bolon, D. N.; Mayo, S. L. Enzyme-like Proteins by Computational Design. *Proc. Natl. Acad. Sci. U. S. A.* **2001**, *98* (25). https://doi.org/10.1073/pnas.251555398.

(81) Kries, H.; Blomberg, R.; Hilvert, D. De Novo Enzymes by Computational Design. *Current Opinion in Chemical Biology*. 2013. https://doi.org/10.1016/j.cbpa.2013.02.012.

(82) Yeh, A. H. W.; Norn, C.; Kipnis, Y.; Tischer, D.; Pellock, S. J.; Evans, D.; Ma, P.; Lee, G. R.; Zhang, J. Z.; Anishchenko, I.; Coventry, B.; Cao, L.; Dauparas, J.; Halabiya, S.; DeWitt, M.; Carter, L.; Houk, K. N.; Baker, D. De Novo Design of Luciferases Using Deep Learning. *Nature* **2023**, *614* (7949), 774–780. https://doi.org/10.1038/s41586-023-05696-3.

(83) Tantillo, D. J.; Chen, J.; Houk, K. N. Theozymes and Compuzymes: Theoretical Models for Biological Catalysis. *Curr. Opin. Chem. Biol.* **1998**, *2* (6). https://doi.org/10.1016/S1367-5931(98)80112-9.

(84) Kiss, G.; Çelebi-Ölçüm, N.; Moretti, R.; Baker, D.; Houk, K. N. Computational Enzyme Design. *Angew. Chemie Int. Ed.* **2013**, *52* (22), 5700–5725. https://doi.org/10.1002/anie.201204077.

(85) Hilvert, D. Design of Protein Catalysts. *Annu. Rev. Biochem.* **2013**, *82* (1), 447–470. https://doi.org/10.1146/annurev-biochem-072611-101825.

(86) Privett, H. K.; Kiss, G.; Lee, T. M.; Blomberg, R.; Chica, R. A.; Thomas, L. M.; Hilvert, D.; Houk, K. N.; Mayo, S. L. Iterative Approach to Computational Enzyme Design. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109* (10). https://doi.org/10.1073/pnas.1118082108.

(87) Röthlisberger, D.; Khersonsky, O.; Wollacott, A. M.; Jiang, L.; DeChancie, J.; Betker, J.; Gallaher, J. L.; Althoff, E. A.; Zanghellini, A.; Dym, O.; Albeck, S.; Houk, K. N.; Tawfik, D. S.; Baker, D. Kemp Elimination Catalysts by Computational Enzyme Design. *Nature*

2008, *453* (7192). https://doi.org/10.1038/nature06879.

(88) Jiang, L.; Althoff, E. A.; Clemente, F. R.; Doyle, L.; Röthlisberger, D.; Zanghellini, A.; Gallaher, J. L.; Betker, J. L.; Tanaka, F.; Barbas, C. F.; Hilvert, D.; Houk, K. N.; Stoddard, B. L.; Baker, D. De Novo Computational Design of Retro-Aldol Enzymes. *Science (80-.).* **2008**, *319* (5868). https://doi.org/10.1126/science.1152692.

(89) Siegel, J. B.; Zanghellini, A.; Lovick, H. M.; Kiss, G.; Lambert, A. R.; St.Clair, J. L.; Gallaher, J. L.; Hilvert, D.; Gelb, M. H.; Stoddard, B. L.; Houk, K. N.; Michael, F. E.; Baker, D. Computational Design of an Enzyme Catalyst for a Stereoselective Bimolecular Diels-Alder Reaction. *Science (80-.).* **2010**, *329* (5989). https://doi.org/10.1126/science.1190239.

(90) Richter, F.; Blomberg, R.; Khare, S. D.; Kiss, G.; Kuzin, A. P.; Smith, A. J. T.; Gallaher, J.; Pianowski, Z.; Helgeson, R. C.; Grjasnow, A.; Xiao, R.; Seetharaman, J.; Su, M.; Vorobiev, S.; Lew, S.; Forouhar, F.; Kornhaber, G. J.; Hunt, J. F.; Montelione, G. T.; Tong, L.; Houk, K. N.; Hilvert, D.; Baker, D. Computational Design of Catalytic Dyads and Oxyanion Holes for Ester Hydrolysis. *J. Am. Chem. Soc.* **2012**, *134* (39). https://doi.org/10.1021/ja3037367.

(91) Davidi, D.; Longo, L. M.; Jabłońska, J.; Milo, R.; Tawfik, D. S. A Bird's-Eye View of Enzyme Evolution: Chemical, Physicochemical, and Physiological Considerations. *Chemical Reviews*. 2018. https://doi.org/10.1021/acs.chemrev.8b00039.

(92) Giger, L.; Caner, S.; Obexer, R.; Kast, P.; Baker, D.; Ban, N.; Hilvert, D. Evolution of a Designed Retro-Aldolase Leads to Complete Active Site Remodeling. *Nat. Chem. Biol.* **2013**, *9* (8), 494–498. https://doi.org/10.1038/nchembio.1276.

(93) Obexer, R.; Godina, A.; Garrabou, X.; Mittl, P. R. E.; Baker, D.; Griffiths, A. D.; Hilvert, D. Emergence of a Catalytic Tetrad during Evolution of a Highly Active Artificial Aldolase. *Nat. Chem.* **2017**, *9* (1). https://doi.org/10.1038/nchem.2596.

(94) Zeymer, C.; Zschoche, R.; Hilvert, D. Optimization of Enzyme Mechanism along the Evolutionary Trajectory of a Computationally Designed (Retro-)Aldolase. *J. Am. Chem. Soc.* **2017**, *139* (36). https://doi.org/10.1021/jacs.7b05796.

(95) De Raffele, D.; Martí, S.; Moliner, V. QM/MM Theoretical Studies of a de Novo Retro-Aldolase Design. *ACS Catal.* **2019**, *9* (3). https://doi.org/10.1021/acscatal.8b04457.

(96) Liu, Z.; Calvó-Tusell, C.; Zhou, A. Z.; Chen, K.; Garcia-Borràs, M.; Arnold, F. H. Dual-

Function Enzyme Catalysis for Enantioselective Carbon–Nitrogen Bond Formation. *Nat. Chem.* **2021**, *13* (12), 1166–1172. https://doi.org/10.1038/s41557-021-00794-z.

(97) Braun, E.; Gilmer, J.; Mayes, H. B.; Mobley, D. L.; Monroe, J. I.; Prasad, S.; Zuckerman, D. M. Best Practices for Foundations in Molecular Simulations [Article v1.0]. *Living J. Comput. Mol. Sci.* **2019**, *1* (1). https://doi.org/10.33011/livecoms.1.1.5957.

(98) Leach, A. R. *Molecular Modelling: Principles and Applications*, 2nd ed.; Pearson Education, Ed.; 2001.

(99) Brooks, B. R.; Brooks, C. L.; Mackerell, A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M. CHARMM: The Biomolecular Simulation Program. *J. Comput. Chem.* **2009**, *30* (10). https://doi.org/10.1002/jcc.21287.

(100) Case, David A., D. S. Cerutti, Thomas Cheatham, T. D. *Amber 2017, University of California, San Francisco*; 2017.

(101) Christen, M.; Hünenberger, P. H.; Bakowies, D.; Baron, R.; Bürgi, R.; Geerke, D. P.; Heinz, T. N.; Kastenholz, M. A.; Kräutler, V.; Oostenbrink, C.; Peter, C.; Trzesniak, D.; Van Gunsteren, W. F. The GROMOS Software for Biomolecular Simulation: GROMOS05. *Journal of Computational Chemistry*. 2005. https://doi.org/10.1002/jcc.20303.

(102) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. Ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from Ff99SB. *J. Chem. Theory Comput.* **2015**, *11* (8), 3696–3713. https://doi.org/10.1021/acs.jctc.5b00255.

(103) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79* (2), 926–935. https://doi.org/10.1063/1.445869.

(104) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, *25* (9), 1157–1174. https://doi.org/10.1002/jcc.20035.

(105) Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. A Well-Behaved Electrostatic

Potential Based Method Using Charge Restraints for Deriving Atomic Charges: The RESP Model. *J. Phys. Chem.* **1993**, *97* (40), 10269–10280. https://doi.org/10.1021/j100142a004.

(106) Li, P.; Merz, K. M. MCPB.Py: A Python Based Metal Center Parameter Builder. *J. Chem. Inf. Model.* **2016**, *56* (4), 599–604. https://doi.org/10.1021/acs.jcim.5b00674.

(107) Li, P.; Song, L. F.; Merz, K. M. Parameterization of Highly Charged Metal Ions Using the 12-6-4 LJ-Type Nonbonded Model in Explicit Water. *J. Phys. Chem. B* **2015**, *119* (3). https://doi.org/10.1021/jp505875v.

(108) Åqvist, J.; Warshel, A. Free Energy Relationships in Metalloenzyme-Catalyzed Reactions. Calculations of the Effects of Metal Ion Substitutions in Staphylococcal Nuclease. *J. Am. Chem. Soc.* **1990**, *112* (8). https://doi.org/10.1021/ja00164a003.

(109) Duarte, F.; Bauer, P.; Barrozo, A.; Amrein, B. A.; Purg, M.; Åqvist, J.; Kamerlin, S. C. L. Force Field Independent Metal Parameters Using a Nonbonded Dummy Model. *J. Phys. Chem. B* **2014**, *118* (16). https://doi.org/10.1021/jp501737x.

(110) Ringnér, M. What Is Principal Component Analysis? *Nature Biotechnology*. 2008. https://doi.org/10.1038/nbt0308-303.

(111) Naritomi, Y.; Fuchigami, S. Slow Dynamics of a Protein Backbone in Molecular Dynamics Simulation Revealed by Time-Structure Based Independent Component Analysis. *J. Chem. Phys.* **2013**, *139* (21). https://doi.org/10.1063/1.4834695.

(112) Pérez-Hernández, G.; Paul, F.; Giorgino, T.; De Fabritiis, G.; Noé, F. Identification of Slow Molecular Order Parameters for Markov Model Construction. *J. Chem. Phys.* **2013**, *139* (1). https://doi.org/10.1063/1.4811489.

(113) Sittel, F.; Stock, G. Perspective: Identification of Collective Variables and Metastable States of Protein Dynamics. *J. Chem. Phys.* **2018**, *149* (15). https://doi.org/10.1063/1.5049637.

(114) Zuckerman, D. M. *Statistical Physics of Biomolecules: An Introduction*, 1st ed.; CRC Press, Ed.; 2010.

(115) Calvó-Tusell, C.; Maria-Solano, M. A.; Osuna, S.; Feixas, F. Time Evolution of the Millisecond Allosteric Activation of Imidazole Glycerol Phosphate Synthase. *J. Am. Chem. Soc.* **2022**, *144* (16), 7146–7159. https://doi.org/10.1021/jacs.1c12629.

(116) Valsson, O.; Tiwary, P.; Parrinello, M. Enhancing Important Fluctuations: Rare Events and Metadynamics from a Conceptual Viewpoint. *Annu. Rev. Phys. Chem.* **2016**, *67*. https://doi.org/10.1146/annurev-physchem-040215-112229.

(117) Miao, Y.; Feixas, F.; Eun, C.; McCammon, J. A. Accelerated Molecular Dynamics Simulations of Protein Folding. *J. Comput. Chem.* **2015**, *36* (20). https://doi.org/10.1002/jcc.23964.

(118) Wang, J.; Arantes, P. R.; Bhattarai, A.; Hsu, R. V; Pawnikar, S.; Huang, Y. ming M.; Palermo, G.; Miao, Y. Gaussian Accelerated Molecular Dynamics: Principles and Applications. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2021**, *11* (5), e1521. https://doi.org/10.1002/WCMS.1521.

(119) Cohen, A. J.; Mori-Sánchez, P.; Yang, W. Challenges for Density Functional Theory. *Chemical Reviews*. 2012. https://doi.org/10.1021/cr200107z.

(120) Calzadiaz-Ramirez, L.; Calvó-Tusell, C.; Stoffel, G. M. M.; Lindner, S. N.; Osuna, S.; Erb, T. J.; Garcia-Borràs, M.; Bar-Even, A.; Acevedo-Rocha, C. G. In Vivo Selection for Formate Dehydrogenases with High Efficiency and Specificity toward NADP+. *ACS Catal.* **2020**, *10* (14). https://doi.org/10.1021/acscatal.0c01487.

(121) Himo, F.; de Visser, S. P. Status Report on the Quantum Chemical Cluster Approach for Modeling Enzyme Reactions. *Commun. Chem.* **2022**, *5* (1), 20–23. https://doi.org/10.1038/s42004-022-00642-2.

(122) Sousa, S. F.; Ribeiro, A. J. M.; Neves, R. P. P.; Brás, N. F.; Cerqueira, N. M. F. S. A.; Fernandes, P. A.; Ramos, M. J. Application of Quantum Mechanics/Molecular Mechanics Methods in the Study of Enzymatic Reaction Mechanisms. *Wiley Interdisciplinary Reviews: Computational Molecular Science*. 2017. https://doi.org/10.1002/wcms.1281.

(123) Wodak, S. J.; Paci, E.; Dokholyan, N. V.; Berezovsky, I. N.; Horovitz, A.; Li, J.; Hilser, V. J.; Bahar, I.; Karanicolas, J.; Stock, G.; Hamm, P.; Stote, R. H.; Eberhardt, J.; Chebaro, Y.; Dejaegere, A.; Cecchini, M.; Changeux, J.-P.; Bolhuis, P. G.; Vreede, J.; Faccioli, P.; Orioli, S.; Ravasio, R.; Yan, L.; Brito, C.; Wyart, M.; Gkeka, P.; Rivalta, I.; Palermo, G.; McCammon, J. A.; Panecka-Hofman, J.; Wade, R. C.; Di Pizio, A.; Niv, M. Y.; Nussinov, R.; Tsai, C.-J.; Jang, H.; Padhorny, D.; Kozakov, D.; McLeish, T. Allostery in Its Many Disguises: From Theory to Applications. *Structure* **2019**, *27* (4), 566–578. https://doi.org/10.1016/j.str.2019.01.003.

(124) Tsai, C. J.; Nussinov, R. A Unified View of "How Allostery Works." *PLoS Comput. Biol.* **2014**, *10* (2), e1003394. https://doi.org/10.1371/journal.pcbi.1003394.

(125) Bozovic, O.; Zanobini, C.; Gulzar, A.; Jankovic, B.; Buhrke, D.; Post, M.; Wolf, S.; Stock, G.; Hamm, P. Real-Time Observation of Ligand-Induced Allosteric Transitions in a PDZ Domain. *Proc. Natl. Acad. Sci. U. S. A.* **2020**, *117* (42), 26031–26039. https://doi.org/10.1073/pnas.2012999117.

(126) Nussinov, R.; Tsai, C. J. Allostery without a Conformational Change? Revisiting the Paradigm. *Curr. Opin. Struct. Biol.* **2015**, *30*, 17–24. https://doi.org/10.1016/j.sbi.2014.11.005.

(127) Jiménez-Osés, G.; Osuna, S.; Gao, X.; Sawaya, M. R.; Gilson, L.; Collier, S. J.; Huisman, G. W.; Yeates, T. O.; Tang, Y.; Houk, K. N. The Role of Distant Mutations and Allosteric Regulation on LovD Active Site Dynamics. *Nat. Chem. Biol.* **2014**, *10* (6), 431–436. https://doi.org/10.1038/nchembio.1503.

(128) Fenton, A. W. Allostery: An Illustrated Definition for the 'Second Secret of Life.' *Trends Biochem. Sci.* **2008**, *33* (9), 420–425. https://doi.org/10.1016/j.tibs.2008.05.009.

(129) Dokholyan, N. V. Controlling Allosteric Networks in Proteins. *Chem. Rev.* **2016**, *116* (11), 6463–6487. https://doi.org/10.1021/ACS.CHEMREV.5B00544.

(130) Mehrabi, P.; Schulz, E. C.; Dsouza, R.; Müller-Werkmeister, H. M.; Tellkamp, F.; Dwayne Miller, R. J.; Pai, E. F. Time-Resolved Crystallography Reveals Allosteric Communication Aligned with Molecular Breathing. *Science (80-.).* **2019**, *365* (6458), 1167–1170. https://doi.org/10.1126/science.aaw9904.

(131) Fraser, J. S.; Clarkson, M. W.; Degnan, S. C.; Erion, R.; Kern, D.; Alber, T. Hidden Alternative Structures of Proline Isomerase Essential for Catalysis. *Nature* **2009**, *462* (7273), 669–673. https://doi.org/10.1038/nature08615.

(132) Aviram, H. Y.; Pirchi, M.; Mazal, H.; Barak, Y.; Riven, I.; Haran, G. Direct Observation of Ultrafast Large-Scale Dynamics of an Enzyme under Turnover Conditions. *Proc. Natl. Acad. Sci. U. S. A.* **2018**, *115* (13), 3243–3248. https://doi.org/10.1073/pnas.1720448115.

(133) Dasgupta, M.; Budday, D.; de Oliveira, S. H. P.; Madzelan, P.; Marchany-Rivera, D.; Seravalli, J.; Hayes, B.; Sierra, R. G.; Boutet, S.; Hunter, M. S.; Alonso-Mori, R.; Batyuk, A.; Wierman, J.; Lyubimov, A.; Brewster, A. S.; Sauter, N. K.; Applegate, G. A.; Tiwari, V.

K.; Berkowitz, D. B.; Thompson, M. C.; Cohen, A. E.; Fraser, J. S.; Wall, M. E.; van den Bedem, H.; Wilson, M. A. Mix-and-Inject XFEL Crystallography Reveals Gated Conformational Dynamics during Enzyme Catalysis. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116* (51), 25634–25640. https://doi.org/10.1073/pnas.1901864116.

(134) East, K. W.; Newton, J. C.; Morzan, U. N.; Narkhede, Y. B.; Acharya, A.; Skeens, E.; Jogl, G.; Batista, V. S.; Palermo, G.; Lisi, G. P. Allosteric Motions of the CRISPR–Cas9 HNH Nuclease Probed by NMR and Molecular Dynamics. *J. Am. Chem. Soc.* **2019**, *142* (3), 1348–1358. https://doi.org/10.1021/JACS.9B10521.

(135) Beismann-Driemeyer, S.; Sterner, R. Imidazole Glycerol Phosphate Synthase from Thermotoga Maritima. Quaternary Structure, Steady-State Kinetics, and Reaction Mechanism of the Bienzyme Complex. *J. Biol. Chem.* **2001**. https://doi.org/10.1074/jbc.M102012200.

(136) Chaudhuri, B. N.; Lange, S. C.; Myers, R. S.; Davisson, V. J.; Smith, J. L. Toward Understanding the Mechanism of the Complex Cyclization Reaction Catalyzed by Imidazole Glycerolphosphate Synthase: Crystal Structures of a Ternary Complex and the Free Enzyme. *Biochemistry* **2003**, *42* (23), 7003–7012. https://doi.org/10.1021/bi034320h.

(137) Lisi, G. P. P.; Manley, G. A. A.; Hendrickson, H.; Rivalta, I.; Batista, V. S. S.; Loria, J. P. Dissecting Dynamic Allosteric Pathways Using Chemically Related Small-Molecule Activators. *Structure* **2016**, *24* (7), 1155–1166. https://doi.org/10.1016/j.str.2016.04.010.

(138) Lisi, G. P.; East, K. W.; Batista, V. S.; Loria, J. P. Altering the Allosteric Pathway in IGPS Suppresses Millisecond Motions and Catalytic Activity. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114* (17), E3414--E3423. https://doi.org/10.1073/pnas.1700448114.

(139) Rivalta, I.; Lisi, G. P.; Snoeberger, N. S.; Manley, G.; Loria, J. P.; Batista, V. S. Allosteric Communication Disrupted by a Small Molecule Binding to the Imidazole Glycerol Phosphate Synthase Protein-Protein Interface. *Biochemistry* **2016**. https://doi.org/10.1021/acs.biochem.6b00859.

(140) Lisi, G. P.; Currier, A. A.; Loria, J. P. Glutamine Hydrolysis by Imidazole Glycerol Phosphate Synthase Displays Temperature Dependent Allosteric Activation. *Front. Mol. Biosci.* **2018**, *5* (FEB), 4. https://doi.org/10.3389/fmolb.2018.00004.

(141) Reisinger, B.; Sperl, J.; Holinski, A.; Schmid, V.; Rajendran, C.; Carstensen, L.; Schlee,

S.; Blanquart, S.; Merkl, R.; Sterner, R. Evidence for the Existence of Elaborate Enzyme Complexes in the Paleoarchean Era. *J. Am. Chem. Soc.* **2014**, *136* (1), 122–129. https://doi.org/10.1021/JA4115677/SUPPL_FILE/JA4115677_SI_001.PDF.

(142) Chaudhuri, B. N.; Lange, S. C.; Myers, R. S.; Chittur, S. V.; Davisson, V. J.; Smith, J. L. Crystal Structure of Imidazole Glycerol Phosphate Synthase: A Tunnel through a (β/α)8barrel Joins Two Active Sites. *Structure* **2001**. https://doi.org/10.1016/S0969-2126(01)00661-X.

(143) Amaro, R. E.; Myers, R. S.; Davisson, V. J.; Luthey-Schulten, Z. A. Structural Elements in IGP Synthase Exclude Water to Optimize Ammonia Transfer. *Biophys. J.* **2005**, *89* (1), 475–487. https://doi.org/10.1529/biophysj.104.058651.

(144) Klem, T. J.; Chen, Y.; Davisson, V. J. Subunit Interactions and Glutamine Utilization by Escherichia Coli Imidazole Glycerol Phosphate Synthase. *J. Bacteriol.* **2001**. https://doi.org/10.1128/JB.182.3.989-996.2001.

(145) Lipchock, J. M.; Loria, J. P. Nanometer Propagation of Millisecond Motions in V-Type Allostery. *Structure* **2010**. https://doi.org/10.1016/j.str.2010.09.020.

(146) List, F.; Vega, M. C.; Razeto, A.; Häger, M. C.; Sterner, R.; Wilmanns, M. Catalysis Uncoupling in a Glutamine Amidotransferase Bienzyme by Unblocking the Glutaminase Active Site. *Chem. Biol.* **2012**, *19* (12), 1589–1599. https://doi.org/10.1016/j.chembiol.2012.10.012.

(147) Rivalta, I.; Sultan, M. M.; Lee, N.-S.; Manley, G. A.; Loria, J. P.; Batista, V. S. Allosteric Pathways in Imidazole Glycerol Phosphate Synthase. *Proc. Natl. Acad. Sci.* **2012**. https://doi.org/10.1073/pnas.1120536109.

(148) Chaudhuri, B. N.; Lange, S. C.; Myers, R. S.; Davisson, V. J.; Smith, J. L. Toward Understanding the Mechanism of the Complex Cyclization Reaction Catalyzed by Imidazole Glycerolphosphate Synthase: Crystal Structures of a Ternary Complex and the Free Enzyme. *Biochemistry* **2003**. https://doi.org/10.1021/bi034320h.

(149) Thoden, J. B.; Miran, S. G.; Phillips, J. C.; Howard, A. J.; Raushel, F. M.; Holden, H. M. Carbamoyl Phosphate Synthetase: Caught in the Act of Glutamine Hydrolysis. *Biochemistry* **1998**, *37* (25), 8825–8831. https://doi.org/10.1021/bi9807761.

(150) List, F.; Vega, M. C.; Razeto, A.; Häger, M. C.; Sterner, R.; Wilmanns, M. Catalysis Uncoupling in a Glutamine Amidotransferase Bienzyme by Unblocking the Glutaminase

Active Site. *Chem. Biol.* **2012**. https://doi.org/10.1016/j.chembiol.2012.10.012.

(151) Wurm, J. P.; Sung, S.; Kneuttinger, A. C.; Hupfeld, E.; Sterner, R.; Wilmanns, M.; Sprangers, R. Molecular Basis for the Allosteric Activation Mechanism of the Heterodimeric Imidazole Glycerol Phosphate Synthase Complex. *Nat. Commun.* **2021**, *12* (1). https://doi.org/10.1038/s41467-021-22968-6.

(152) Rivalta, I.; Sultan, M. M.; Lee, N. S.; Manley, G. A.; Loria, J. P.; Batista, V. S. Allosteric Pathways in Imidazole Glycerol Phosphate Synthase. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109* (22). https://doi.org/10.1073/pnas.1120536109.

(153) Ribeiro, A. A. S. T.; Ortiz, V. Determination of Signaling Pathways in Proteins through Network Theory: Importance of the Topology. *J. Chem. Theory Comput.* **2014**, *10* (4), 1762–1769. https://doi.org/10.1021/ct400977r.

(154) Negre, C. F. A.; Morzan, U. N.; Hendrickson, H. P.; Pal, R.; Lisi, G. P.; Patrick Loria, J.; Rivalta, I.; Ho, J.; Batista, V. S. Eigenvector Centrality for Characterization of Protein Allosteric Pathways. *Proc. Natl. Acad. Sci. U. S. A.* **2018**, *115* (52), E12201--E12208. https://doi.org/10.1073/pnas.1810452115.

(155) Botello-Smith, W. M.; Luo, Y. Robust Determination of Protein Allosteric Signaling Pathways. *J. Chem. Theory Comput.* **2019**, *15* (4), 2116–2126. https://doi.org/10.1021/acs.jctc.8b01197.

(156) Gheeraert, A.; Pacini, L.; Batista, V. S.; Vuillon, L.; Lesieur, C.; Rivalta, I. Exploring Allosteric Pathways of a V-Type Enzyme with Dynamical Perturbation Networks. *J. Phys. Chem. B* **2019**, acs.jpcb.9b01294. https://doi.org/10.1021/acs.jpcb.9b01294.

(157) Lake, P. T.; Davidson, R. B.; Klem, H.; Hocky, G. M.; McCullagh, M. Residue-Level Allostery Propagates through the Effective Coarse-Grained Hessian. *J. Chem. Theory Comput.* **2020**, *16* (5), 3385–3395. https://doi.org/10.1021/acs.jctc.9b01149.

(158) Maschietto, F.; Gheeraert, A.; Piazzi, A.; Batista, V. S.; Rivalta, I. Distinct Allosteric Pathways in Imidazole Glycerol Phosphate Synthase from Yeast and Bacteria. *Biophys. J.* **2022**, *121* (1), 119–130. https://doi.org/10.1016/J.BPJ.2021.11.2888.

(159) Yao, X. Q.; Hamelberg, D. Residue-Residue Contact Changes during Functional Processes Define Allosteric Communication Pathways. *J. Chem. Theory Comput.* **2022**, *18* (2), 1173–1187. https://doi.org/10.1021/acs.jctc.1c00669.

(160) Mouilleron, S.; Golinelli-Pimpaneau, B. Conformational Changes in Ammonia-

Channeling Glutamine Amidotransferases. *Curr. Opin. Struct. Biol.* **2007**, *17* (6), 653–664. https://doi.org/10.1016/j.sbi.2007.09.003.

(161) Wurm, J. P.; Sung, S.; Kneuttinger, A. C.; Hupfeld, E.; Sterner, R.; Wilmanns, M.; Sprangers, R. Molecular Basis for the Allosteric Activation Mechanism of the Heterodimeric Imidazole Glycerol Phosphate Synthase Complex. *Nat. Commun.* **2021**, *12* (1), 2748. https://doi.org/10.1038/s41467-021-22968-6.

(162) Hamelberg, D.; Mongan, J.; McCammon, J. A. Accelerated Molecular Dynamics: A Promising and Efficient Simulation Method for Biomolecules. *J. Chem. Phys.* **2004**, *120* (24), 11919–11929. https://doi.org/10.1063/1.1755656.

(163) Hamelberg, D.; De Oliveira, C. A. F.; McCammon, J. A. Sampling of Slow Diffusive Conformational Transitions with Accelerated Molecular Dynamics. *J. Chem. Phys.* **2007**, *127* (15), 155102. https://doi.org/10.1063/1.2789432.

(164) Kneuttinger, A. C.; Rajendran, C.; Simeth, N. A.; Bruckmann, A.; König, B.; Sterner, R. Significance of the Protein Interface Configuration for Allostery in Imidazole Glycerol Phosphate Synthase. *Biochemistry* **2020**, *59* (29), 2729–2742. https://doi.org/10.1021/acs.biochem.0c00332.

(165) Myers, R. S.; Amaro, R. E.; Luthey-Schulten, Z. A.; Davisson, V. J. Reaction Coupling through Interdomain Contacts in Imidazole Glycerol Phosphate Synthase. *Biochemistry* **2005**. https://doi.org/10.1021/bi050706b.

(166) Osuna, S. The Challenge of Predicting Distal Active Site Mutations in Computational Enzyme Design. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2021**, *11* (3), e1502. https://doi.org/10.1002/WCMS.1502.

(167) Lisi, G. P. P.; Manley, G. A. A.; Hendrickson, H.; Rivalta, I.; Batista, V. S. S.; Loria, J. P. Dissecting Dynamic Allosteric Pathways Using Chemically Related Small-Molecule Activators. *Structure* **2016**. https://doi.org/10.1016/j.str.2016.04.010.

(168) Lisi, G. P.; East, K. W.; Batista, V. S.; Loria, J. P. Altering the Allosteric Pathway in IGPS Suppresses Millisecond Motions and Catalytic Activity. *Proc. Natl. Acad. Sci.* **2017**. https://doi.org/10.1073/pnas.1700448114.

(169) Ahuja, L. G.; Kornev, A. P.; Mcclendon, C. L.; Veglia, G.; Taylor, S. S. Mutation of a Kinase Allosteric Node Uncouples Dynamics Linked to Phosphotransfer. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114* (6), E931--E940. https://doi.org/10.1073/pnas.1620667114.

(170) Ahuja, L. G.; Taylor, S. S.; Kornev, A. P. Tuning the "Violin" of Protein Kinases: The Role of Dynamics-based Allostery. *IUBMB Life* **2019**, *71* (6), 685–696. https://doi.org/10.1002/iub.2057.

(171) Hili, R.; Yudin, A. K. Making Carbon-Nitrogen Bonds in Biological and Chemical Synthesis. *Nature Chemical Biology*. 2006. https://doi.org/10.1038/nchembio0606-284.

(172) Froidevaux, V.; Negrell, C.; Caillol, S.; Pascault, J. P.; Boutevin, B. Biobased Amines: From Synthesis to Polymers; Present and Future. *Chemical Reviews*. 2016. https://doi.org/10.1021/acs.chemrev.6b00486.

(173) Bariwal, J.; Van Der Eycken, E. C-N Bond Forming Cross-Coupling Reactions: An Overview. *Chemical Society Reviews*. 2013. https://doi.org/10.1039/c3cs60228a.

(174) Kim, J. E.; Choi, S.; Balamurugan, M.; Jang, J. H.; Nam, K. T. Electrochemical C–N Bond Formation for Sustainable Amine Synthesis. *Trends in Chemistry*. 2020. https://doi.org/10.1016/j.trechm.2020.09.003.

(175) Kohls, H.; Steffen-Munsberg, F.; Höhne, M. Recent Achievements in Developing the Biocatalytic Toolbox for Chiral Amine Synthesis. *Current Opinion in Chemical Biology*. 2014. https://doi.org/10.1016/j.cbpa.2014.02.021.

(176) Hauer, B. Embracing Nature's Catalysts: A Viewpoint on the Future of Biocatalysis. *ACS Catal.* **2020**, *10* (15), 8418–8427. https://doi.org/10.1021/acscatal.0c01708.

(177) Wu, S.; Snajdrova, R.; Moore, J. C.; Baldenius, K.; Bornscheuer, U. T. Biocatalysis: Enzymatic Synthesis for Industrial Applications. *Angewandte Chemie - International Edition*. 2021. https://doi.org/10.1002/anie.202006648.

(178) Wang, Z. J.; Peck, N. E.; Renata, H.; Arnold, F. H. Cytochrome P450-Catalyzed Insertion of Carbenoids into N-H Bonds. *Chem. Sci.* **2014**, *5* (2). https://doi.org/10.1039/c3sc52535j.

(179) Sreenilayam, G.; Fasan, R. Myoglobin-Catalyzed Intermolecular Carbene N-H Insertion with Arylamine Substrates. *Chem. Commun.* **2015**, *51* (8). https://doi.org/10.1039/c4cc08753d.

(180) Steck, V.; Carminati, D. M.; Johnson, N. R.; Fasan, R. Enantioselective Synthesis of Chiral Amines via Biocatalytic Carbene N-H Insertion. *ACS Catal.* **2020**, *10* (19), 10967–10977. https://doi.org/10.1021/acscatal.0c02794.

(181) Steck, V.; Sreenilayam, G.; Fasan, R. Selective Functionalization of Aliphatic Amines via Myoglobin-Catalyzed Carbene N-H Insertion. *Synlett* **2020**, *31* (3). https://doi.org/10.1055/s-0039-1690007.

(182) Chen, K.; Arnold, F. H. Engineering New Catalytic Activities in Enzymes. *Nature Catalysis*. Nature Research March 1, 2020, pp 203–213. https://doi.org/10.1038/s41929-019-0385-5.

(183) Li, M. L.; Yu, J. H.; Li, Y. H.; Zhu, S. F.; Zhou, Q. L. Highly Enantioselective Carbene Insertion into N–H Bonds of Aliphatic Amines. *Science (80-.).* **2019**, *366* (6468), 990–994. https://doi.org/10.1126/science.aaw9939.

(184) Xu, B.; Zhu, S. F.; Xie, X. L.; Shen, J. J.; Zhou, Q. L. Asymmetric Ni-H Insertion Reaction Cooperatively Catalyzed by Rhodium and Chiral Spiro Phosphoric Acids. *Angew. Chemie - Int. Ed.* **2011**, *50* (48), 11483–11486. https://doi.org/10.1002/anie.201105485.

(185) Sharon, D. A.; Mallick, D.; Wang, B.; Shaik, S. Computation Sheds Insight into Iron Porphyrin Carbenes' Electronic Structure, Formation, and N-H Insertion Reactivity. *J. Am. Chem. Soc.* **2016**, *138* (30), 9597–9610. https://doi.org/10.1021/jacs.6b04636.

(186) Nam, D.; Tinoco, A.; Shen, Z.; Adukure, R. D.; Sreenilayam, G.; Khare, S. D.; Fasan, R. Enantioselective Synthesis of α-Trifluoromethyl Amines via Biocatalytic N-H Bond Insertion with Acceptor-Acceptor Carbene Donors. *J. Am. Chem. Soc.* **2022**, *144* (6), 2590–2602. https://doi.org/10.1021/jacs.1c10750.

(187) Pavlović, D.; Mutak, S.; Andreotti, D.; Biondi, S.; Cardullo, F.; Paio, A.; Piga, E.; Donati, D.; Lociuro, S. Synthesis and Structure-Activity Relationships of α-Amino-γ-Lactone Ketolides: A Novel Class of Macrolide Antibiotics. *ACS Med. Chem. Lett.* **2014**, *5* (10). https://doi.org/10.1021/ml500279k.

(188) Deangelis, A.; Dmitrenko, O.; Fox, J. M. Rh-Catalyzed Intermolecular Reactions of Cyclic α-Diazocarbonyl Compounds with Selectivity over Tertiary C-H Bond Migration. *J. Am. Chem. Soc.* **2012**, *134* (26). https://doi.org/10.1021/ja3046712.

(189) Zhou, A. Z.; Chen, K.; Arnold, F. H. Enzymatic Lactone-Carbene C-H Insertion to Build Contiguous Chiral Centers. *ACS Catal.* **2020**, *10* (10), 5393–5398. https://doi.org/10.1021/acscatal.0c01349.

(190) Chen, K.; Zhang, S. Q.; Brandenberg, O. F.; Hong, X.; Arnold, F. H. Alternate Heme Ligation Steers Activity and Selectivity in Engineered Cytochrome P450-Catalyzed

Carbene-Transfer Reactions. *J. Am. Chem. Soc.* **2018**, *140* (48). https://doi.org/10.1021/jacs.8b09613.

(191) Zhou, A. Z.; Chen, K.; Arnold, F. H. Enzymatic Lactone-Carbene C-H Insertion to Build Contiguous Chiral Centers. *ACS Catal.* **2020**, *10* (10), 5393–5398. https://doi.org/10.1021/acscatal.0c01349.

(192) Coelho, P. S.; Wang, Z. J.; Ener, M. E.; Baril, S. A.; Kannan, A.; Arnold, F. H.; Brustad, E. M. A Serine-Substituted P450 Catalyzes Highly Efficient Carbene Transfer to Olefins in Vivo. *Nat. Chem. Biol.* **2013**, *9* (8). https://doi.org/10.1038/nchembio.1278.

(193) Brandenberg, O. F.; Chen, K.; Arnold, F. H. Directed Evolution of a Cytochrome P450 Carbene Transferase for Selective Functionalization of Cyclic Compounds. *J. Am. Chem. Soc.* **2019**, *141* (22). https://doi.org/10.1021/jacs.9b02931.

(194) Trott, O.; Olson, A. J. AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *J. Comput. Chem.* **2009**, *31* (2), NA-NA. https://doi.org/10.1002/jcc.21334.

(195) Garcia-Borràs, M.; Kan, S. B. J.; Lewis, R. D.; Tang, A.; Jimenez-Osés, G.; Arnold, F. H.; Houk, K. N. Origin and Control of Chemoselectivity in Cytochrome c Catalyzed Carbene Transfer into Si-H and N-H Bonds. *J. Am. Chem. Soc.* **2021**, *143* (18), 7114–7123. https://doi.org/10.1021/jacs.1c02146.

(196) Khade, R. L.; Zhang, Y. Catalytic and Biocatalytic Iron Porphyrin Carbene Formation: Effects of Binding Mode, Carbene Substituent, Porphyrin Substituent, and Protein Axial Ligand. *J. Am. Chem. Soc.* **2015**, *137* (24), 7560–7563. https://doi.org/10.1021/jacs.5b03437.

(197) Acevedo-Rocha, C. G.; Li, A.; D'Amore, L.; Hoebenreich, S.; Sanchis, J.; Lubrano, P.; Ferla, M. P.; Garcia-Borràs, M.; Osuna, S.; Reetz, M. T. Pervasive Cooperative Mutational Effects on Multiple Catalytic Enzyme Traits Emerge via Long-Range Conformational Dynamics. *Nat. Commun.* **2021**, *12* (1), 1–13. https://doi.org/10.1038/s41467-021-21833-w.

(198) Claassens, N. J.; Burgener, S.; Vögeli, B.; Erb, T. J.; Bar-Even, A. A Critical Comparison of Cellular and Cell-Free Bioproduction Systems. *Current Opinion in Biotechnology*. 2019. https://doi.org/10.1016/j.copbio.2019.05.003.

(199) Hummel, W.; Gröger, H. Strategies for Regeneration of Nicotinamide Coenzymes

Emphasizing Self-Sufficient Closed-Loop Recycling Systems. *J. Biotechnol.* **2014**, *191*. https://doi.org/10.1016/j.jbiotec.2014.07.449.

(200) Andexer, J. N.; Richter, M. Emerging Enzymes for ATP Regeneration in Biocatalytic Processes. *ChemBioChem*. 2015. https://doi.org/10.1002/cbic.201402550.

(201) Babel, W. The Auxiliary Substrate Concept: From Simple Considerations to Heuristically Valuable Knowledge. *Engineering in Life Sciences*. 2009. https://doi.org/10.1002/elsc.200900027.

(202) Tishkov, V. I.; Popov, V. O. Protein Engineering of Formate Dehydrogenase. *Biomolecular Engineering*. 2006. https://doi.org/10.1016/j.bioeng.2006.02.003.

(203) Aslan, S.; Noor, E.; Bar-Even, A. Holistic Bioengineering: Rewiring Central Metabolism for Enhanced Bioproduction. *Biochemical Journal*. 2017. https://doi.org/10.1042/BCJ20170377.

(204) Gul-Karaguler, N.; Sessions, R. B.; Clarke, A. R.; Holbrook, J. J. A Single Mutation in the NAD-Specific Formate Dehydrogenase from Candida Methylica Allows the Enzyme to Use NADP. *Biotechnol. Lett.* **2001**, *23* (4). https://doi.org/10.1023/A:1005610414179.

(205) Serov, A. E.; Popova, A. S.; Fedorchuk, V. V.; Tishkov, V. I. Engineering of Coenzyme Specificity of Formate Dehydrogenase from Saccharomyces Cerevisiae. *Biochem. J.* **2002**, *367* (3). https://doi.org/10.1042/BJ20020379.

(206) Andreadeli, A.; Platis, D.; Tishkov, V.; Popov, V.; Labrou, N. E. Structure-Guided Alteration of Coenzyme Specificity of Formate Dehydrogenase by Saturation Mutagenesis to Enable Efficient Utilization of NADP+. *FEBS J.* **2008**, *275* (15). https://doi.org/10.1111/j.1742-4658.2008.06533.x.

(207) Hatrongjit, R.; Packdibamrung, K. A Novel NADP+-Dependent Formate Dehydrogenase from Burkholderia Stabilis 15516: Screening, Purification and Characterization. *Enzyme Microb. Technol.* **2010**, *46* (7). https://doi.org/10.1016/j.enzmictec.2010.03.002.

(208) Hoelsch, K.; Sührer, I.; Heusel, M.; Weuster-Botz, D. Engineering of Formate Dehydrogenase: Synergistic Effect of Mutations Affecting Cofactor Specificity and Chemical Stability. *Appl. Microbiol. Biotechnol.* **2013**, *97* (6). https://doi.org/10.1007/s00253-012-4142-9.

(209) Ihara, M.; Kawano, Y.; Urano, M.; Okabe, A. Light Driven CO2 Fixation by Using Cyanobacterial Photosystem I and NADPH-Dependent Formate Dehydrogenase. *PLoS*

One **2013**, *8* (8). https://doi.org/10.1371/journal.pone.0071581.

(210) Fogal, S.; Beneventi, E.; Cendron, L.; Bergantino, E. Structural Basis for Double Cofactor Specificity in a New Formate Dehydrogenase from the Acidobacterium Granulicella Mallensis MP5ACTX8. *Appl. Microbiol. Biotechnol.* **2015**, *99* (22). https://doi.org/10.1007/s00253-015-6695-x.

(211) Alpdağtaş, S.; Yücel, S.; Kapkaç, H. A.; Liu, S.; Binay, B. Discovery of an Acidic, Thermostable and Highly NADP+ Dependent Formate Dehydrogenase from Lactobacillus Buchneri NRRL B-30929. *Biotechnol. Lett.* **2018**, *40* (7). https://doi.org/10.1007/s10529-018-2568-6.

(212) Bennett, B. D.; Kimball, E. H.; Gao, M.; Osterhout, R.; Van Dien, S. J.; Rabinowitz, J. D. Absolute Metabolite Concentrations and Implied Enzyme Active Site Occupancy in Escherichia Coli. *Nat. Chem. Biol.* **2009**, *5* (8). https://doi.org/10.1038/nchembio.186.

(213) Warnecke, T.; Gill, R. T. Organic Acid Toxicity, Tolerance, and Production in Escherichia Coli Biorefining Applications. *Microbial Cell Factories*. 2005. https://doi.org/10.1186/1475-2859-4-25.

(214) Lamzin, V. S.; Dauter, Z.; Popov, V. O.; Harutyunyan, E. H.; Wilson, K. S. High Resolution Structures of Holo and Apo Formate Dehydrogenase. *J. Mol. Biol.* **1994**, *236* (3). https://doi.org/10.1006/jmbi.1994.1188.

(215) Tishkov, V. I.; Matorin, A. D.; Rojkova, A. M.; Fedorchuk, V. V.; Savitsky, P. A.; Dementieva, L. A.; Lamzin, V. S.; Mezentzev, A. V.; Popov, V. O. Site-Directed Mutagenesis of the Formate Dehydrogenase Active Centre: Role of the His332-Gln313 Pair in Enzyme Catalysis. *FEBS Lett.* **1996**, *390* (1). https://doi.org/10.1016/0014-5793(96)00641-2.

(216) Galkin, A. G.; Kutsenko, A. S.; Bajulina, N. P.; Esipova, N. G.; Lamzin, V. S.; Mesentsev, A. V.; Shelukho, D. V.; Tikhonova, T. V.; Tishkov, V. I.; Ustinnikova, T. B.; Popov, V. O. Site-Directed Mutagenesis of the Essential Arginine of the Formate Dehydrogenase Active Centre. *Biochim. Biophys. Acta - Protein Struct. Mol. Enzymol.* **2002**, *1594* (1). https://doi.org/10.1016/S0167-4838(01)00297-7.

(217) Cahn, J. K. B.; Werlang, C. A.; Baumschlager, A.; Brinkmann-Chen, S.; Mayo, S. L.; Arnold, F. H. A General Tool for Engineering the NAD/NADP Cofactor Preference of Oxidoreductases. *ACS Synth. Biol.* **2017**, *6* (2).

https://doi.org/10.1021/acssynbio.6b00188.

(218) Alekseeva, A. A.; Fedorchuk, V. V.; Zarubina, S. A.; Sadykhov, E. G.; Matorin, A. D.; Savin, S. S.; Tishkov, V. I. The Role of Ala198 in the Stability and Coenzyme Specificity of Bacterial Formate Dehydrogenases. *Acta Naturae* **2015**, *7* (1). https://doi.org/10.32607/20758251-2015-7-1-60-69.

(219) Lindner, S. N.; Ramirez, L. C.; Krüsemann, J. L.; Yishai, O.; Belkhelfa, S.; He, H.; Bouzon, M.; Döring, V.; Bar-Even, A. NADPH-Auxotrophic E. Coli: A Sensor Strain for Testing in Vivo Regeneration of NADPH. *ACS Synth. Biol.* **2018**, *7* (12). https://doi.org/10.1021/acssynbio.8b00313.

(220) Kelley, L. A.; Mezulis, S.; Yates, C. M.; Wass, M. N.; Sternberg, M. J. E. The Phyre2 Web Portal for Protein Modeling, Prediction and Analysis. *Nat. Protoc.* **2015**, *10* (6). https://doi.org/10.1038/nprot.2015.053.

(221) Kim, D. E.; Chivian, D.; Baker, D. Protein Structure Prediction and Analysis Using the Robetta Server. *Nucleic Acids Res.* **2004**, *32* (WEB SERVER ISS.). https://doi.org/10.1093/nar/gkh468.

(222) Gordon, J. C.; Myers, J. B.; Folta, T.; Shoja, V.; Heath, L. S.; Onufriev, A. H++: A Server for Estimating PKas and Adding Missing Hydrogens to Macromolecules. *Nucleic Acids Res.* **2005**, *33* (SUPPL. 2). https://doi.org/10.1093/nar/gki464.

(223) WL, D. The PyMOL Molecular Graphics System. *CCP4 Newsl. Protein Crystallogr.* **2002**, *40* (1), 82–92.

Appendix A

Supplementary information Chapter 4. Reprinted with permission from *J. Am. Chem. Soc.* **2022**, 144, 7146-159 (https://pubs.acs.org/doi/10.1021/jacs.1c12629). Copyright 2022 American Chemical Society. Further permissions to the material shown should be directed to the ACS.

Appendix A Extended Text

Computational protocols

Protein preparation. The computational structural models of IGPS were based on the crystal structure of the *apo* complex from *Thermotoga maritima* IGPS at 2.4 Å resolution (PDB:1GPW) reported by Douangamath and coworkers.[1] To generate the structural model of IGPS, chains A and B of PDB 1GPW were used. In chain B of 1GPW, the *h*49-PGVG oxyanion strand is found in an inactive conformation (Inactive-OxH). It is postulated that the C-terminal loop (*f*Loop1) of chain A is found in a closed (assumed active) conformation. The original bacterial crystal structure presents an active site mutation (*f*D11N) that was mutated back to its original residue using PyMOL. The crystal structure of the PRFAR-bound complex from *Saccharomyces cerevisiae* at 2.5 Å resolution (PDB:1OX5), crystallized with the effector PRFAR, was used to generate the PRFAR-bound state. The coordinates of the effector PRFAR were aligned to two phosphate groups from the chain A of the PDB 1GPW. These phosphate groups were suggested to belong to an unresolved PRFAR molecule since the effector was present in the solution but not in the crystal during the crystallization procedure. Following the same system preparation described in previous works, the crystallographic waters of 1GPW were kept for the molecular dynamics simulations.[2] According to previous works,[2–4] a δ-nitrogen (HID) protonation state was assigned to residues *f*H84, *f*H209, *f*H244 of HisF subunit and *h*H73, *h*H120, *h*H141, and *h*H178 in HisH subunit; ε-nitrogen (HIE) protonation state for residues *f*H228 and *h*H53 and both δ-nitrogen and ε-nitrogen (HIP) of residue *f*H151 were protonated. The catalytic residues *h*C84, *h*H178, and *h*E180 were treated as protonated thiol group (-SH), δ-nitrogen protonated (HID), and deprotonated carboxylate group (COO^-). The remaining residues were protonated at pH 7.4, using the H++ server.

Ligand parametrization: allosteric effector (PRFAR) and substrate (L-Gln). PRFAR initial structure was obtained from PDB 1OX5 (from IGPS *Saccharomyces cerevisiae*). Parameters for MD simulations for PRFAR and L-Gln were generated with antechamber module of AMBER16[5] using the generalized AMBER force field (GAFF),[6] with partial charges set to fit the electrostatic potential generated at HF/6-31G* level of theory by restrained electrostatic potential (RESP) model.[7] The atomic charges were calculated according to the Merz–Singh–Kollman[7] scheme using Gaussian 09.[8]

Conventional molecular dynamics (cMD) simulations: Substrate-Free IGPS. Molecular Dynamics simulations of all IGPS complexes (*apo* and PRFAR effector-bound) were performed in explicit water using AMBER16 package. AMBER-ff14SB force field[9] was used to describe the protein, GAFF for PRFAR and L-Gln and TIP3P for water molecules.[10] Each system was solvated in a pre-equilibrated cubic box with a 12 Å buffer of TIP3P water molecules and was neutralized by addition of explicit sodium and chloride counterions (Na^+ or Cl^-). Subsequently, a two-stage geometry optimization approach was performed. First, a short minimization of the water molecules positions, with positional restraints on solute by a harmonic potential with a force constant of 500 kcal mol^{-1} $Å^{-2}$ was done. The second stage was an unrestrained minimization of all the atoms in the simulation cell. Then, the systems were gently heated using six 50 ps steps, incrementing the temperature 50 K each step (0-300 K) under constant-volume, periodic-boundary conditions and the particle-mesh Ewald approach to introduce long-range electrostatic effects.[11] For these steps, an 11 Å cut-off was applied to Lennard-Jones and electrostatic interactions. Bonds involving hydrogen were constrained with the SHAKE algorithm. Harmonic restraints of 10 kcal mol^{-1} were applied to the solute, and the Langevin equilibration scheme is used to control and equalize the temperature. The time step was kept at 2 fs during the heating stages, allowing potential inhomogeneities to self-adjust. Each system was then equilibrated for 4 ns with a 2 fs timestep at a constant pressure of 1 atm to relax the density of the system. After the systems were equilibrated in the NPT ensemble, MD simulations were performed under the NVT ensemble and periodic-boundary conditions using our Galatea cluster at the University of Girona (composed by a total of 178 GTX1080 GPUs). PRFAR-bound state simulations were carried out applying soft distance restrains between the effector phosphate groups and the amide backbone of residues *f*T104/*f*A224 and between the carboxylate group of *f*D130 with the hydroxyl group of the ribose ring of PRFAR. These soft restraints retain PRFAR in the HisF binding site while allowing for certain flexibility.

The cMD simulations used for the analysis of substrate-free IGPS conformational dynamics consist of 10 replicas of 1.5 µs for the *apo* state and 9 replicas of 1.5 µs and 1 replica of 4 µs for the PRFAR-IGPS state. The analysis of the distances, angles, and dihedral angles is performed using the *cpptraj* MD analysis program.[12] The PyEMMA 2.5 software was used for constructing the conformational landscape and for clustering and principal component analysis (PCA).[13]

A. Analysis of Conventional Molecular Dynamics Simulations.

Oxyanion strand conformational dynamics. Additional unblocked-OxH states of the oxyanion strand presenting similar characteristics to one described in Figure 4.2 of Chapter 4 are sampled in cMD simulations. These conformations are represented by the rotation of other dihedral angles of the oxyanion strand (see Figure A2 and A3) such as ϕ hG52. The analysis of individual cMD trajectories show that the Inactive-OxH and Unblocked-OxH states can interconvert in the microsecond time scale (2 out 10 replicas for ϕ hG50, while ϕ hG52 transition occurs more frequently). The conformation of the unblocked state associated with the rotation of ϕ hG52 resembles the one identified by Kneuttinger and coworkers by means of MD simulations.[1]

Loop1 conformational dynamics and hydrophobic cluster. To assess the changes in global backbone flexibility, the Root-mean-square fluctuations (RMSF) of all Cα atoms was computed for *apo* and PRFAR-bound states. RMSF analysis show no significant global backbone rearrangements, both *apo* and PRFAR display similar patterns (see Figure A8). The main conformational differences are located in Loop1 of the HisF subunit (*f*R16-*f*D31), which presents enhanced flexibility in the presence of PRFAR. In the *X-ray* (PDB 1GPW), Loop1 is formed by two small β-sheets strands that are stabilized by a hydrogen bond network of conserved residues in this enzyme family. Interestingly significant differences arise after one microsecond of simulation time when Loop1 loses its secondary structure and moves away from the cyclase active site (see Figure A8). Loop1 adopts a conformational structure similar to some IGPS *X-ray* structures (*e.g.,* chain E of PDB 1GPW) where the loop is disordered and partially unsolved. This conformational transition was not described previously and occurs in three out of ten PRFAR bound replicas indicating that PRFAR binding enhances microsecond Loop1 motions compared to *apo*. In the PRFAR bound replica where the formation of the oxyanion hole is observed, the conformational change of Loop1 is correlated with the rotation of the oxyanion strand and preceded by the disruption of the hydrophobic cluster (*f*F23 and *f*I52 hydrophobic interaction, see Figure A9). Upon the conformational change, both *f*F23 and *f*R27 are pointing towards the solvent while *f*K19 establishes transient interactions with the glycerol phosphate group (gP), in contrast to the *apo* state simulations where *f*K19 is exposed to the bulk during the whole simulation time. The hydrophobic cluster remains unformed and flexible when IGPS presents the oxyanion hole formed. The inner flexibility of Loop1 is related to facile proteolysis by trypsin at *f*R27 position.[2] Upon hV51 oxyanion hole formation, Loop1 remains relatively stable for 1 μs of MD simulation time while subtle changes occur upon oxyanion hole deactivation. However, in other replicas the Loop1 conformational rearrangements are uncoupled to changes on the oxyanion strand dynamics indicating the existence of uncorrelated motions. Our hypothesis is that the ordered *X-ray* conformation of Loop1 (used as a starting point in most IGPS MD simulations) is not the most stable in solution. These observations corroborate previous NMR and computational studies that described that PRFAR binding alters Loop1 dynamics.

Salt Bridge network between fα2, fα3 and hα1 and Heterodimer Interface network. The binding of PRFAR gated a series of conformational rearrangements on the HisF and HisH subunits besides the formation of the oxyanion hole and motions in Loop1 (see Figure A9). In particular, the salt bridge network between *f*α2, *f*α3 and *h*α1 helices presents some alterations in the presence of PRFAR. In the early steps of the allosteric activation, the salt bridge interactions between *f*E67 and both *f*R95 and *h*R18 are strengthened compared to *apo* while the interaction between *f*E71 and *h*R18 is weakened (see Figure A9). *h*Arg18 can adopt two major conformations, one pointing towards the dimer interface and another pointing towards the solvent. The reshape of these interactions enhances the communication between the *f*α2, *f*α3 and *h*α1 structural motifs and are key to unlocked changes on the interdomain region that precede the oxyanion strand formation. These rearrangements impact orientation of *h*N13 and *h*N15 in *h*α1. In particular, *h*N15 backbone establishes a transient hydrogen bond with the *h*Arg18 side chain that helps orient the side chain of *h*N15 towards the interdomain region and HisH active site. These rearrangements precede the rotation of *h*P10, which leads to the breaking of *h*P10-*h*V51 hydrogen bond and the displacement of the Ω-loop (see Figure A4). The breaking of *h*P10-*h*V51 interaction is, thus, a prerequisite for oxyanion hole formation and for the activation of HisH for catalysis. However, in the *apo* state simulations the hydrogen bond between *h*N15 and *h*Arg18 is not established and hN15 leaves move away from the interdomain region while keeping the Ω-loop stable in the inactive form. Finally, at the same time, the formation of the oxyanion hole alters the interdomain salt bridge between the side chains of *f*D98 and *h*K181 that has been shown to be key for allosteric communication*.* Upon oxyanion hole formation the salt bridge is weakened compared to *apo* and PRFAR inactive states (see Figure A9). *h*K181 gains mobility establishing interactions with catalytic *h*E180 when the breathing motion is more compressed while *f*D98 alters between *h*Y138, *h*N15, *h*K181 residues.

Therefore, PFRAR alters the electrostatic environment of the interdomain region. However, the sequence of these events differs among replicas indicating the predominance of uncorrelated motions in these non-equilibrium MD simulations.

HisF:HisH interface. All these local rearrangements impact the global dynamics of IGPS. One of the unanswered questions is whether IGPS is able to attain a closed state of the HisF:HisH interface to retain ammonia. When analyzed globally, the HisF:HisF interface conformational dynamics is not showing significant differences in the *apo* and PRFAR bound states (see Figure A10). Angles between 15° and 35° are sampled and similar distributions are observed in both cases. However, displacement towards shorter angles is observed in the replica where the oxyanion hole is formed. A deeper analysis of the MD simulation where the oxyanion hole formation is observed reveals that the HisF:HisH conformational dynamics becomes slightly restrained when the active state is sampled (values below 20 Å are frequently sampled and stabilized). This is consistent with the idea of a population shift towards a closed state of the interdomain region when PRFAR is present. However, the values sampled (24.8 ± 3.2 º) in the cMD simulations are still far from the values observed in the *X-ray h*C84A IGPS structure (HisF:HisH interface angle of 9.7 º).

B. Analysis of Accelerated Molecular Dynamics Simulations.

To unravel the effect of PRFAR on the conformational dynamics of IGPS beyond microsecond time scale, we ran ten replicas of 1 μs accelerated molecular dynamics (aMD) simulations in the *apo* and PRFAR bound states. To explore global conformational motions in deeper detail, we performed PCA analysis on the aMD simulations (see Figure A13). Interestingly, PC1 describes the counter rotation of HisH and HisF subunits pointing out an enhanced communication between subunits. On the other hand, PC2 captures motions in *f*Loop1 and changes on the HisF:HisH interface. Overall, a number of metastable states are identified that present different degrees of HisF:HisH rotation and interdomain closure/occlusion. In more detail, PRFAR unlocks the rotation of HisF:HisH subunits compared to *apo*, flexibilizing the interdomain region. These results are in line with the enhanced millisecond motions upon PRFAR binding observed in NMR studies. See Figure A13 legend for a more detailed description.

C. Analysis of WT-Metadynamics Simulations

To estimate the energy barrier of the oxyanion strand reorientation, we performed well-tempered metadynamics (WT-Metadynamics) simulations using the multiple-walkers approach. Defining a reaction coordinate without knowing the end state can be difficult. However, the inactive and active states of IGPS identified in the aMD simulations can be used as a reference for accurate metadynamics simulations. Therefore, we use the aMD simulations to extract relevant conformations to seed the starting conformations (*i.e.,* walker replicas staring points) for metadynamics simulations (see Figure A32-34 for more details). The selected states encompass global and local features of inactive and active states respectively. In this case, we started five walkers in the Active-OxH and five walkers in the Inactive-OxH states to completely reconstruct the free energy landscape of the oxyanion strand conformational dynamics.

D. Ternary complex IGPS Conformational Dynamics: HisF and community networks

Additional complementary insights are gained by further tracing the changes in the dynamic network of interactions in the HisF and HisH subunits, including alterations in the open-to-closed exchange of the interdomain region (see Figure A37). The analysis is performed by monitoring the changes in particular interactions along one of the aMD simulations that captured the complete allosteric activation (see Results section of Chapter 4) and comparing the results obtained with the *X-ray* structure of the hC84A IGPS (PDB: 7AC8, chains E and F).

In the presence of PRFAR, fK19 side chain makes permanent contacts with the glycerol phosphate group (gP) of PRFAR (distance of ca. 3.87 ± 0.63 Å), as in the hC84A IGPS X-ray. If this residue is mutated the catalytic activity of the ternary complex significantly decreases.[3] This interaction remains formed most of the simulation time and helps orienting PRFAR in the cyclase active site. Among the residues that form the hydrophobic cluster, the interactions between fL50-fI52 and fV48-fL50 are persistent along the simulation (see Figure A37). When IGPS attains the HisF:HisH closed state, a series of dynamic interactions propagate through the two active sites of HisH and HisF and interdomain regions. In particular, we observed in the aMD simulation that the productive closure is preceded by the breaking of fE67-hR18 salt bridge in the presence of PRFAR. This interaction occurs in the interdomain HisF:HisH region and seems to play a key role in controlling the dynamics of interface residues, which exhibit a higher flexibility in the presence of PRFAR. This interaction is not formed in PDB 7AC8, thus our simulations show the important role of this interaction in controlling HisF:HisH conformational dynamics. On the other hand, the *f*E67-*f*R95 salt bridge remains quite formed and stable as well as the *f*E91-*f*R95 and *f*R59-*f*E91 displaying similar interactions as in the *X-ray*. Simultaneously, in the interdomain region, the new orientation of *h*R18 points towards the side chain of *h*N15 at the Ω loop in the productive closed state of HisF:HisH altering the interface dynamics (see Figure A37). *h*N15 is directly interacting with *h*R18 in the *X-ray* structure. As a consequence, the side chain of *h*N15 alternates between *h*K181 and *h*R18 residues, while *h*N12 side chain establishes a relatively stable hydrogen bond with amide backbone of *h*N15 (3.1 ± 0.71 Å). Eventually the *h*K181 and *f*D98 salt bridge is strengthened upon productive HisF:HisH closure enhancing the communication between HisF subunit and HisH active site. This pair of residues is key for allosteric communication and, in fact, the mutation of *f*D98 disrupts the millisecond motions in IGPS.[3] Finally, the *h*Pro10-*h*V51 hydrogen bond completely breaks before the productive closure of HisF:HisH interface.

<u>fLoop1 conformational dynamics in the ternary complex</u>. The principal difference between *X-ray h*C84A IGPS (PDB 7AC8 chains E/F) and the structures sampled in aMD simulations is observed in the conformation of *f*Loop1 (see Figure A29). In our simulations the *f*Loop1 is disordered and is not covering the PRFAR binding site as shown in the *X-ray* (see Figure A29). However, despite the different conformation, *f*K19 (*f*Loop1) still makes permanent contacts with the glycerol phosphate group of PRFAR as shown in the *X-ray* structure. This discrepancy may be explained by the different environment of IGPS in the *X-ray* structure and in solution). By analyzing the crystal packing of PDB 7AC8 chains E/F, it can be seen that *f*Loop1 is making persistent hydrogen bond interactions and salt bridges with other IGPS chains in the crystal that may be important to stabilize the closed conformation of *f*Loop1 in the crystal (see Figure A38). Considering that the additional IGPS chains are not present in solution in our MD simulations, it is possible that *f*Loop1 can explore alternative conformations. As shown by some of us for the LovD enzyme[4] and also by Gervasio and coworkers in p38alpha protein kinase,[5] the conformation of some loops could be an artifact of crystal contacts when compared with the behavior of the protein in solution. A similar case could happen in IGPS, because in all MD simulations we observed intrinsic flexibility in *f*Loop1.

<u>PRFAR dynamics in the ternary complex.</u> In our MD simulations, we used the natural effector PRFAR as an allosteric effector of wild-type IGPS. We observed that PRFAR presents significant flexibility when bound in the HisF active site (see Figure A37) which is in line with experimental observations that show that PRFAR is unstable.[6] Some of these conformations resemble the orientation of ProFAR in PDB 7AC8, however considering the more unstable nature of PRFAR a direct comparison using a single snapshot cannot be established. As described in SI Methods section A, to retain the effector in the HisF and ensure the full allosteric effect, we used soft distance restraints between PRFAR and HisF active site residues.

Moreover, MD simulations are carried out in the presence of sodium ions to neutralize the system. Considering that the global charge of PRFAR is -4, sodium ions are constantly interacting with PRFAR phosphate groups contributing to PRFAR flexibility and orientation in the active site, which also involves some changes on the orientation HisF active site residues, including *f*D11, *f*K19, and other relevant residues. Considering that PRFAR is retained in the HisF active site in all simulations, we captured the expected allosteric effect elicited by PRFAR in our simulations. Indeed, the impact of PRFAR on the HisH active site corresponding to the allosteric activation is the formation of the *h*V51 oxyanion hole (as shown by the overlay between the computational predictions and PDB 7AC8 chain E/F, see Figure A29).

Community network analysis. The analysis of the evolution of community networks along the allosteric activation show that when the open-to-closed transition of the HisF:HisH interface takes place the communication between HisF and HisH communities is significantly enhanced (Figure A39). Indeed, the residues of the HisF:HisH interface form a unique community together with oxyanion strand residues when the allosteric activation is completed. These results are in agreement with the ones obtained using the Shortest Path Map tool (Figure 4.6 in Chapter 4 and Figures A35-36).

Computational Strategy (Figure A1)

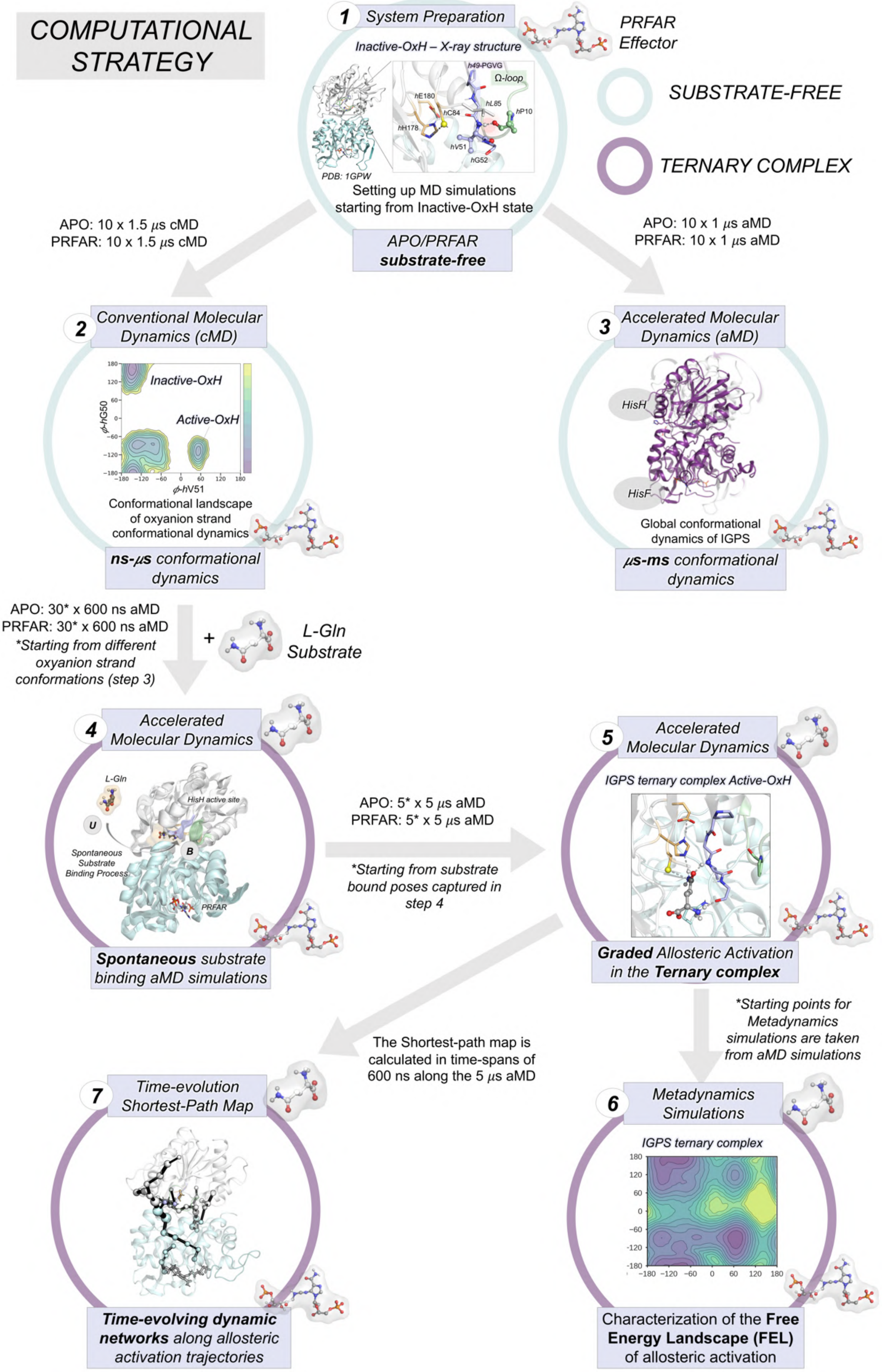
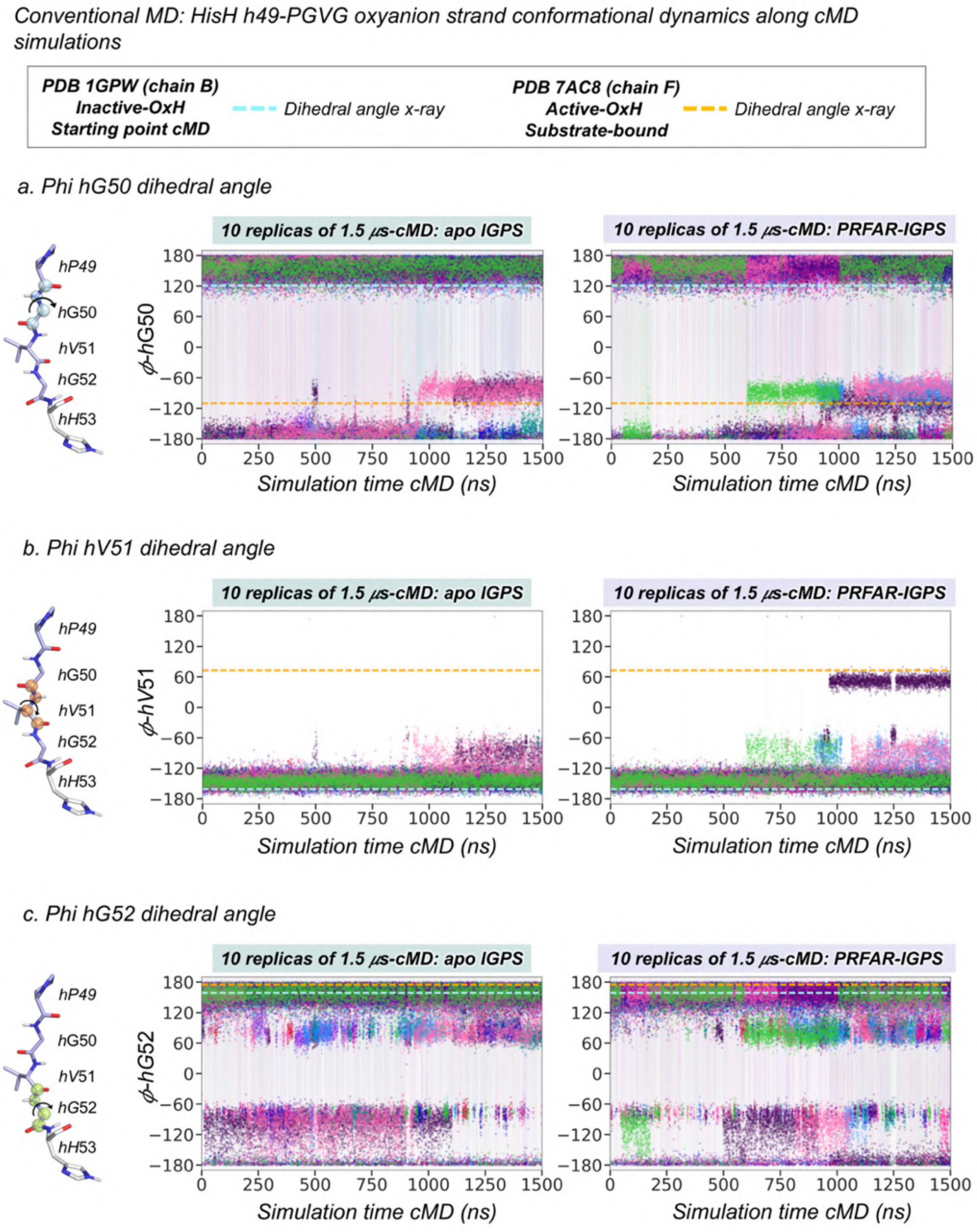

COMPUTATIONAL STRATEGY
1 System Preparation
Inactive-OxH – X-ray structure
hE180
hC84
hH178
h49-PGVG
Ω-loop
hL85
hP10
hV51
hG52
PDB: 1GPW
Setting up MD simulations starting from Inactive-OxH state
APO/PRFAR substrate-free
PRFAR Effector
SUBSTRATE-FREE
TERNARY COMPLEX
APO: 10 x 1.5 μs cMD
PRFAR: 10 x 1.5 μs cMD
APO: 10 x 1 μs aMD
PRFAR: 10 x 1 μs aMD
2 Conventional Molecular Dynamics (cMD)
Inactive-OxH
Active-OxH
ϕ-hG50
ϕ-hV51
180
120
60
0
−60
−120
−180
Conformational landscape of oxyanion strand conformational dynamics
ns-μs conformational dynamics
3 Accelerated Molecular Dynamics (aMD)
HisH
HisF
Global conformational dynamics of IGPS
μs-ms conformational dynamics
APO: 30* x 600 ns aMD
PRFAR: 30* x 600 ns aMD
*Starting from different oxyanion strand conformations (step 3)
+
L-Gln Substrate
4 Accelerated Molecular Dynamics
L-Gln
HisH active site
U
B
Spontaneous Substrate Binding Process
PRFAR
Spontaneous substrate binding aMD simulations
APO: 5* x 5 μs aMD
PRFAR: 5* x 5 μs aMD
*Starting from substrate bound poses captured in step 4
5 Accelerated Molecular Dynamics
IGPS ternary complex Active-OxH
Graded Allosteric Activation in the Ternary complex
*Starting points for Metadynamics simulations are taken from aMD simulations
The Shortest-path map is calculated in time-spans of 600 ns along the 5 μs aMD
7 Time-evolution Shortest-Path Map
Time-evolving dynamic networks along allosteric activation trajectories
6 Metadynamics Simulations
IGPS ternary complex
180
120
60
0
−60
−120
−180
−180 −120 −60 0 60 120 180
Characterization of the Free Energy Landscape (FEL) of allosteric activation

Figure A1. Summary of the computational strategy used to characterize the molecular basis of the millisecond allosteric activation of *wt*IGPS. All simulations were performed starting from *X-ray* structure of IGPS in the inactive state (PDB ID 1GPW (chains A and B)).

Conventional Molecular Dynamics Simulations IGPS: substrate-free (Figures A2-A10)



Conventional MD: HisH h49-PGVG oxyanion strand conformational dynamics along cMD simulations

a. Phi hG50 dihedral angle

b. Phi hV51 dihedral angle

c. Phi hG52 dihedral angle

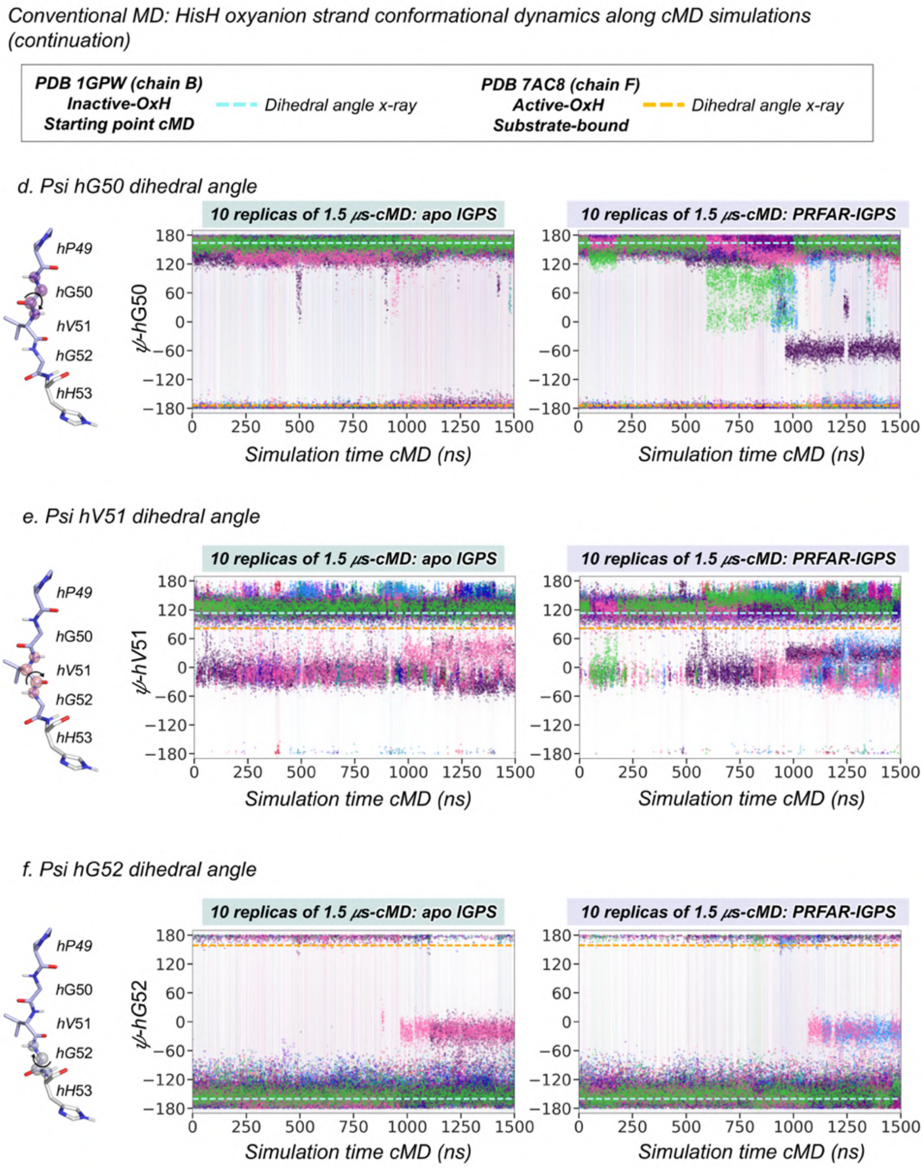
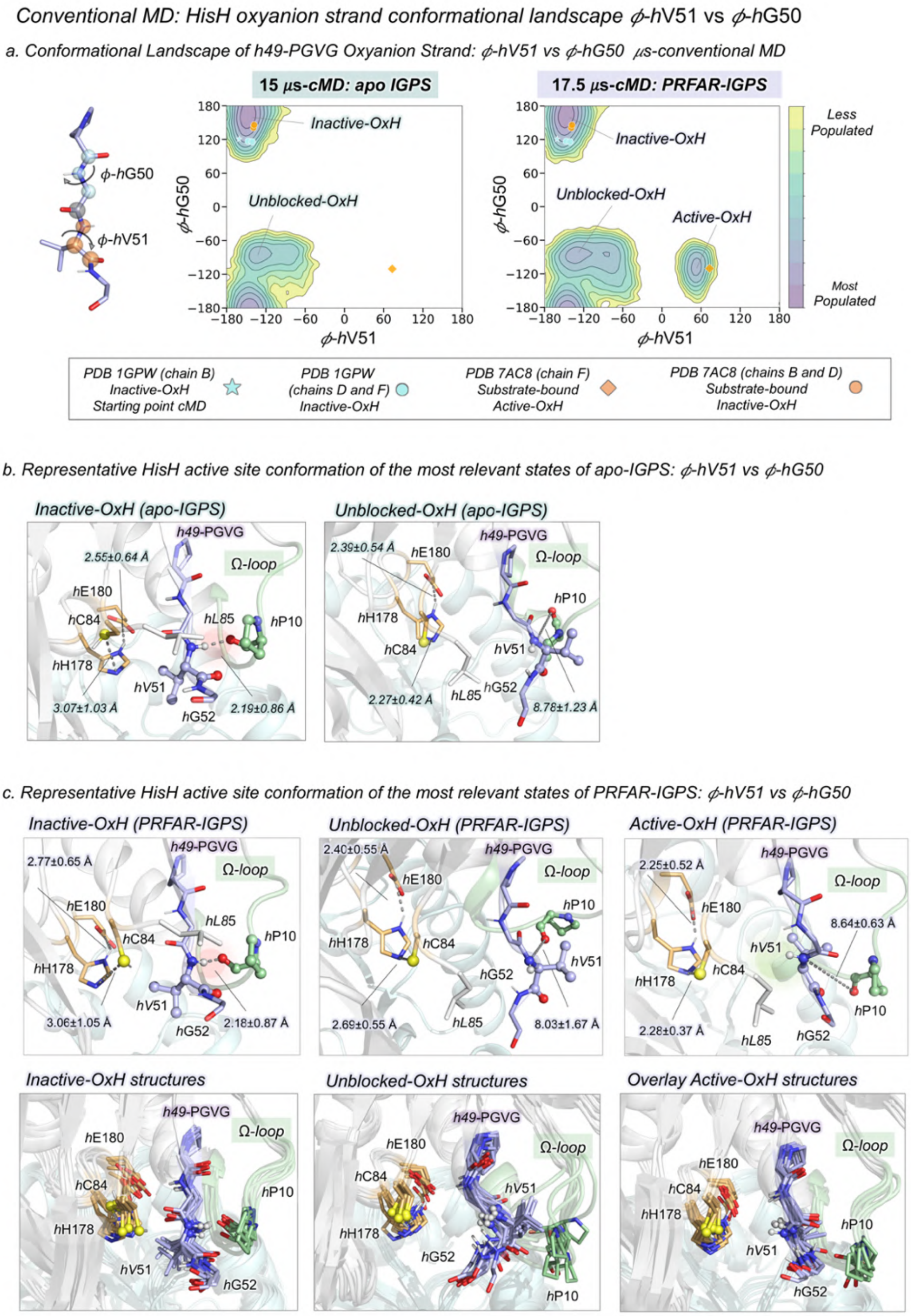


Figure A2. HisH *h*49-PVGV conformational dynamics along cMD simulations. Plot of the most relevant dihedral angles of the *h*49-PGVG oxyanion strand for ten replicas of 1.5 μs conventional molecular dynamics (cMD) simulations in the *apo*-IGPS and PRFAR-IGPS states. Each replica is depicted in a different color. Horizontal cyan dashed lines indicate the value of the dihedral angle corresponding to the *X-ray* structure (PDB 1GPW (chain B)) used as starting point for cMD simulations. Horizontal orange dashed lines indicate the value of the dihedral angle corresponding to the *X-ray* structure of substrate-bound *h*C84A IGPS (PDB 7AC8 (chain F)) that displays an active conformation of the h49-PVGV oxyanion strand. The oxyanion strand residues are shown in light purple and the atoms involved in each dihedral angle are represented as spheres of different color. The cMD trajectory of PRFAR-IGPS where the Active-OxH state is sampled is represented in deep purple in all plots. (a) ϕ dihedral angle of *h*G50; (b) ϕ dihedral angle of *h*V51; (c) ϕ dihedral angle of *h*V51; (d) ψ dihedral angle of *h*G50; (e) ψ dihedral angle of *h*V51; (f) ψ dihedral angle of *h*G52.

All dihedrals display a certain degree of flexibility along the 1.5 μs cMD simulations. Multiple orientations with respect to the *X-ray* dihedral angle are observed in most cases, being ϕ-*h*G52 and ψ-*h*V51 the ones displaying more transitions in the nanosecond timescale. The dihedral angles ϕ-*h*G50 and ϕ-*h*V51 were selected as the most relevant ones for monitoring the formation of the Active-OxH state and the allosteric activation of IGPS. See Figure A3 for a molecular representation of the most relevant states of the *h*49-PGVG oxyanion strand.



Conventional MD: HisH oxyanion strand conformational landscape ϕ-hV51 vs ϕ-hG50

a. Conformational Landscape of h49-PGVG Oxyanion Strand: ϕ-hV51 vs ϕ-hG50 μs-conventional MD

b. Representative HisH active site conformation of the most relevant states of apo-IGPS: ϕ-hV51 vs ϕ-hG50

c. Representative HisH active site conformation of the most relevant states of PRFAR-IGPS: ϕ-hV51 vs ϕ-hG50

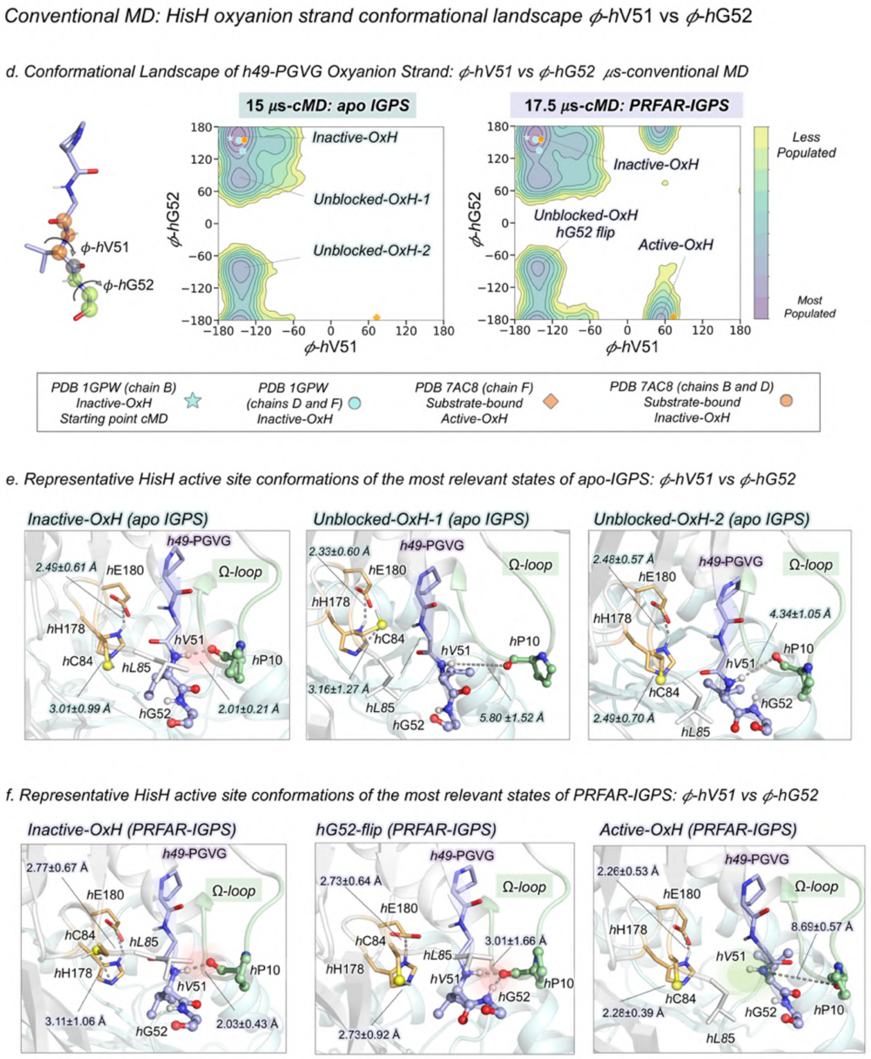


Figure A3. Conformational landscape of *h*49-PGVG oxyanion strand obtained from μs-conventional Molecular Dynamics (cMD) simulations. The conformational landscape of the *h*49-PGVG oxyanion strand of *apo*-IGPS is constructed from an accumulated time of 15 μs of cMD simulations (10 replicas of 1.5 μs) while the oxyanion strand conformational landscape of PRFAR-IGPS is built from 17.5 μs of cMD simulation time (9 replicas of 1.5 μs and one replica of 4 μs). The conformational landscape of each state is clusterized into 20 different clusters. The clusters corresponding to the most populated regions are selected for further analysis (*e.g.,* calculation of average distances). The *h*49-PGVG conformational landscape and the most representative conformations of each of the major conformational states have been obtained for both *ϕ*-*h*V51/*ϕ*-*h*G50 and *ϕ*-*h*V51/*ϕ*-*h*G52 pairs of dihedral angles. **a)** Conformational landscape of *apo* and PRFAR-IGPS constructed using the *ϕ* dihedral angles of *h*V51 and *h*G50. The values of the *ϕ* dihedral angles of *h*V51 and *h*G50 found in the *X-ray* structures corresponding to the three chains of PDB 1GPW are depicted in cyan and the three chains of PDB 7AC8 are represented in orange, respectively. The conformation used as starting point for cMD simulations is shown using the cyan star symbol. The conformation corresponding to the active oxyanion strand (Active-OxH) observed

in *h*C84A IGPS is depicted using the orange diamond symbol. **b)** Representative HisH active site structures of most populated states in *apo*-IGPS conformational landscape: Inactive-OxH and Unblocked-OxH. **c)** Representative HisH active site structures of most populated states in PRFAR-IGPS conformational landscape: Inactive-OxH, Unblocked-OxH, and Active-OxH. Overlay of eight representative structures corresponding to each conformational state of the oxyanion strand in PRFAR-IGPS. **d)** Conformational landscape of *apo* and PRFAR-IGPS constructed using the ϕ dihedral angles of *h*V51 and *h*G52. The values of the ϕ dihedral angles of *h*V51 and *h*G52 found in the *X-ray* structures corresponding to the three chains of PDB 1GPW are depicted in cyan and the three chains of PDB 7AC8 are represented in orange, respectively. The conformation used as starting point for cMD simulations is shown using the cyan star symbol. The conformation corresponding to the active oxyanion strand (Active-OxH) observed in *h*C84A IGPS is depicted using the orange diamond symbol. **e)** Representative HisH active site structures of most populated states in *apo*-IGPS conformational landscape: Inactive-OxH, Unblocked-OxH-1, Unblocked-OxH-2. **f)** Representative HisH active site structures of most populated states in PRFAR-IGPS conformational landscape: Inactive-OxH, *h*G52-flip-OxH, and Active-OxH. Relevant average distances (in Å) of each conformational state are depicted in green and purple for *apo* and PRFAR-bound states, respectively. The average distances are calculated considering all the structures included in each cluster. The HisH catalytic residues are highlighted in orange, Ω-loop residues in green, and the residues of the *h*49-PGVG oxyanion strand in purple. Other relevant HisF and HisH residues are shown in cyan and white, respectively. The atoms of *h*V51 and *h*P10 are shown as spheres. The red surface is used to show when the *h*P10-*h*V51 hydrogen bond is shown, blocking the formation of the *h*V51 oxyanion hole. The green surface is used to show when H^N of *h*V51 is pointing toward the active site (Active-OxH with *h*V51 formed).

The Active-OxH state of the *h*49-PGVG oxyanion strand is only sampled in PRFAR-IGPS. In the Active-OxH state, the average values of the ϕ-*h*V51 and ϕ-*h*G50 are 51.3 ± 8.2 ° and -107.5 ± 5.5 °. These values slightly deviate from the *X-ray* ϕ-*h*V51 and ϕ-*h*G50 dihedral angles of substrate-bound *h*C84A IGPS (PDB 7AC8 (chain F)) which are 73.2 ° and -110.5 °, respectively. This displacement is associated with the presence of the substrate (see Figure A24). The three catalytic HisH active site residues display more flexibility in the Inactive-OxH state than in the Unblocked-OxH and Active-OxH states of the oxyanion-strand (overlay Figure A3c). The distances between catalytic residues are monitored in Figure A4.

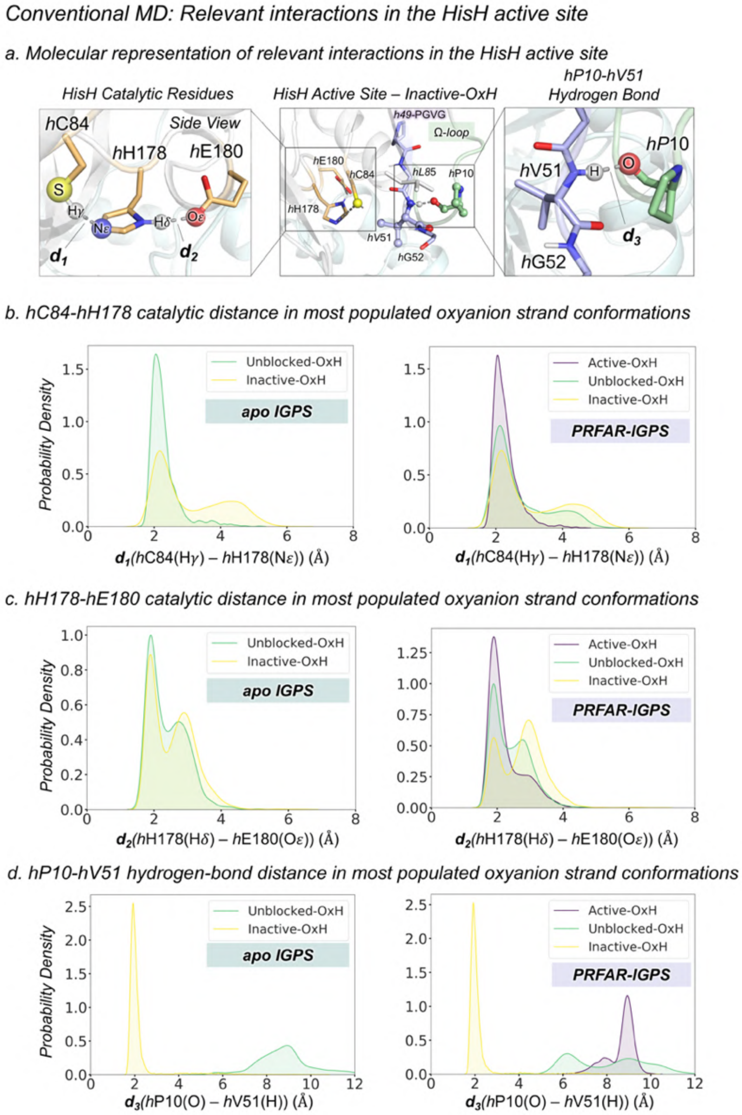
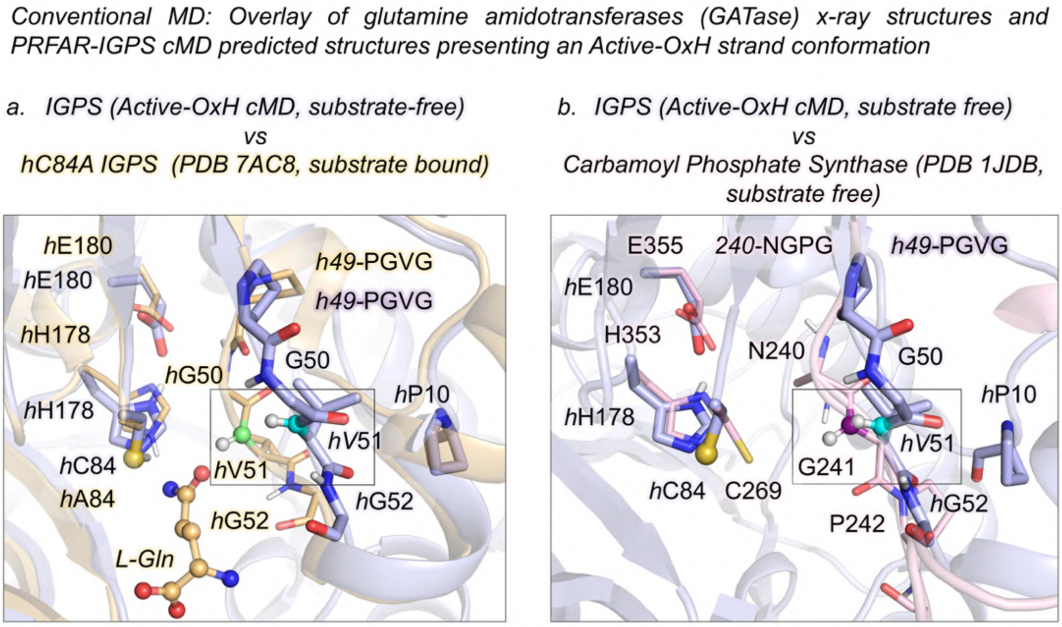


Figure A4. Relevant interactions in the HisH active site in cMD simulations. a) Structural representation of the relevant interactions in the HisH active site: interaction between catalytic residues *h*C84-*h*H178 (*d₁*) and *h*H178-*h*E180 (*d₂*), and the hydrogen between the Ω-loop residue *h*P10 and the oxyanion strand residue *h*V51 (*d₃*). **b)** Probability density distribution of the *h*C84-*h*H178 distance in the most populated states of the oxyanion strand conformational landscape (see A3) of *apo* and PRFAR-IGPS. The distances are calculated considering all the structures included in the representative cluster of each state. The distance is monitored between the thiol hydrogen (Hγ) of *h*C84 and the ε nitrogen (Nε) of *h*H178. **c)** Probability density distribution of the *h*H178-*h*E180 distance. The distance is monitored between the δ hydrogen (Hδ) of *h*H178 and the oxygen of the carboxylate group (Oε) of *h*E180. (d) Probability density distribution of the *h*P10-*h*V51 distance. The distance is monitored between the amide backbone hydrogen (H^N) of *h*V51 and the backbone oxygen (O) of *h*P10. The distances corresponding to the Inactive-OxH, Unblocked-OxH, and Active-OxH are shown in yellow, green, and purple respectively. All distances are in Å. The average distances of each state are shown in Figure A3.



Conventional MD: Overlay of glutamine amidotransferases (GATase) x-ray structures and PRFAR-IGPS cMD predicted structures presenting an Active-OxH strand conformation

a. IGPS (Active-OxH cMD, substrate-free) vs hC84A IGPS (PDB 7AC8, substrate bound)

b. IGPS (Active-OxH cMD, substrate free) vs Carbamoyl Phosphate Synthase (PDB 1JDB, substrate free)

Figure A5. Overlay of class I glutamine amidotransferase (GATase) *X-ray* structures and PRFAR-IGPS cMD structures. a) Overlay of a representative substrate-free Active-OxH PRFAR-IGPS (in purple) structure extracted from cMD simulations with the *X-ray* structure of the postulated active state of substrate-bound *h*C84A IGPS (PDB 7AC8 (chain F), in orange). The NH backbone of *h*V51 in the Active-OxH PRFAR-IGPS (cMD) is highlighted in cyan while the NH backbone of *h*V51 corresponding to *h*C84A IGPS is highlighted in green. The glutamine (*L-Gln*) substrate present in the *X-ray* structure is highlighted with spheres. In both cases, the NH backbone is pointing toward the HisH catalytic residues (*h*C84/*h*A84). The conformation of the *h*49-PGVG oxyanion strand in the HisH active site show some differences due to the presence of the substrate in PDB 7AC8 (chain F). **b)** Overlay of a representative substrate-free Active-OxH PRFAR-IGPS (in purple) structure extracted from cMD simulations with the *X-ray* structure of the active state substrate-free carbamoyl phosphate synthase (PDB 1JDB (chain B), in light pink). In carbamoyl phosphate synthase the oxyanion strand is formed by 240-NGPG residues being G241 the residue responsible of forming the oxyanion hole (equivalent to *h*V51 in IGPS). The NH backbone of *h*V51 in the Active-OxH PRFAR-IGPS is highlighted in cyan while the NH backbone of G241 corresponding to carbamoyl phosphate synthase is highlighted in magenta. In both cases, the NH backbone of *h*V51 is pointing toward the catalytic residues.

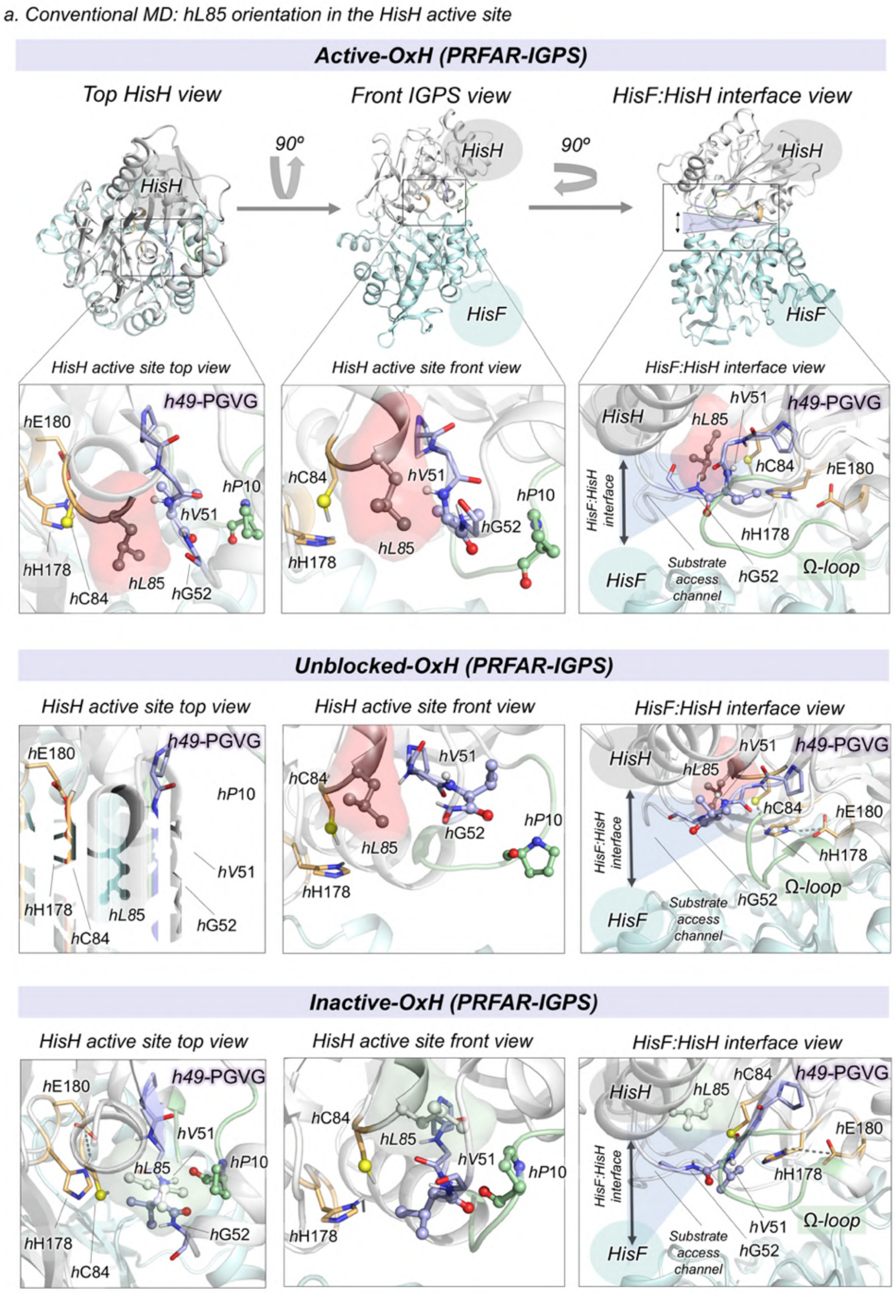


Figure A6. Orientation of *h*L85 in the HisH active site. a) Three different points of view of IGPS and HisH active site: top view (HisH is located above HisF), front view, and HisF:HisH interface view. Representative HisH active site conformations for the Active-OxH, Unblocked-OxH, and Inactive-OxH *h*49-PGVG oxyanion strand states of PRFAR-IGPS sampled in cMD simulations. *h*L85 is highlighted as a red surface when it is blocking the access to the HisH active site (Active-OxH and Unblocked-OxH states) and as a green surface when is not oriented toward the HisH active site (Inactive-OxH). In the Active-OxH and Unblocked-OxH states, the *h*L85 side chain is positioned between the catalytic and oxyanion strand residues blocking the substrate access. In the Inactive-OxH conformation the side chain of *h*V51 occupies the active site while *h*L85 is placed above the oxyanion strand residues (in the selected HisH active site views). The HisH catalytic residues are

highlighted in orange, Ω-loop residues in green, and the residues of the *h*49-PGVG oxyanion strand in purple. Other relevant HisF and HisH residues are shown in cyan and white, respectively. The atoms of *h*L85, *h*V51, and *h*P10 are shown as spheres. In the HisF:HisH interface view, the substrate access channel is shown in blue.

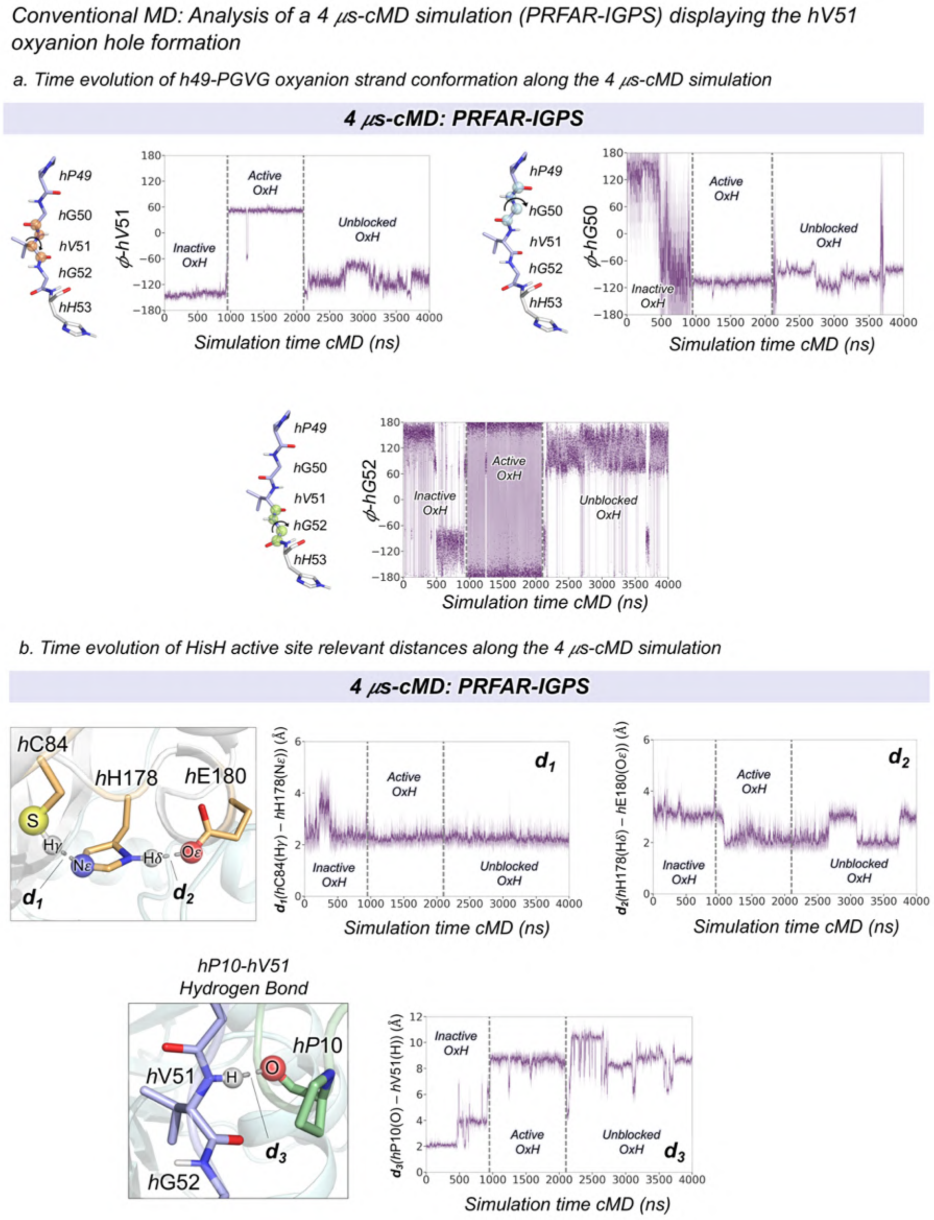


Figure A7. Analysis of a representative 4 µs-cMD simulation displaying the *hV*51 oxyanion hole formation. a) Plot of the most relevant dihedral angles of the *h*49-PGVG oxyanion strand along a 4 µs-cMD simulation: ϕ dihedral angle of *h*G50; ϕ dihedral angle of *h*V51; ϕ dihedral angle of *h*V51. Vertical gray dashed lines indicate the *h*V51 oxyanion hole formation. **b)** Plot of the most relevant HisH active site distances (see Figure A4) along the 4 µs-cMD simulation. Interaction between catalytic residues *h*C84-*h*H178 (d_1) and *h*H178-*h*E180 (d_2), and the hydrogen between the Ω-loop residue *h*P10 and the oxyanion strand residue *h*V51 (d_3) is also shown. The formation of the Active-OxH is preceded by the disruption of the *h*P10 and *h*V51 hydrogen bond. All distances are in Å.

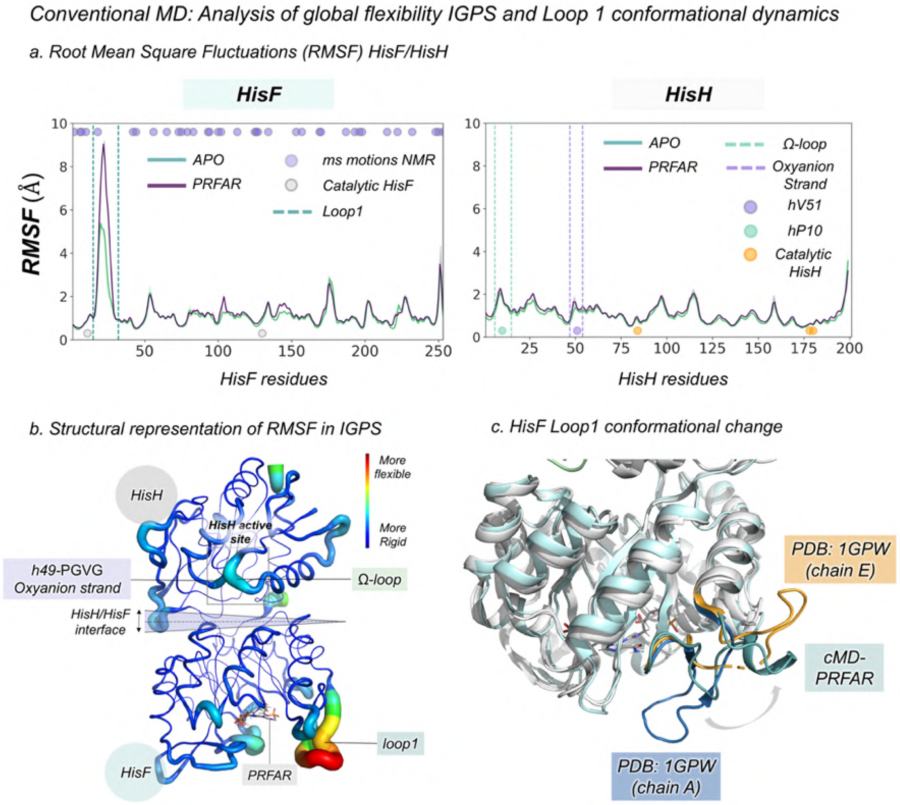
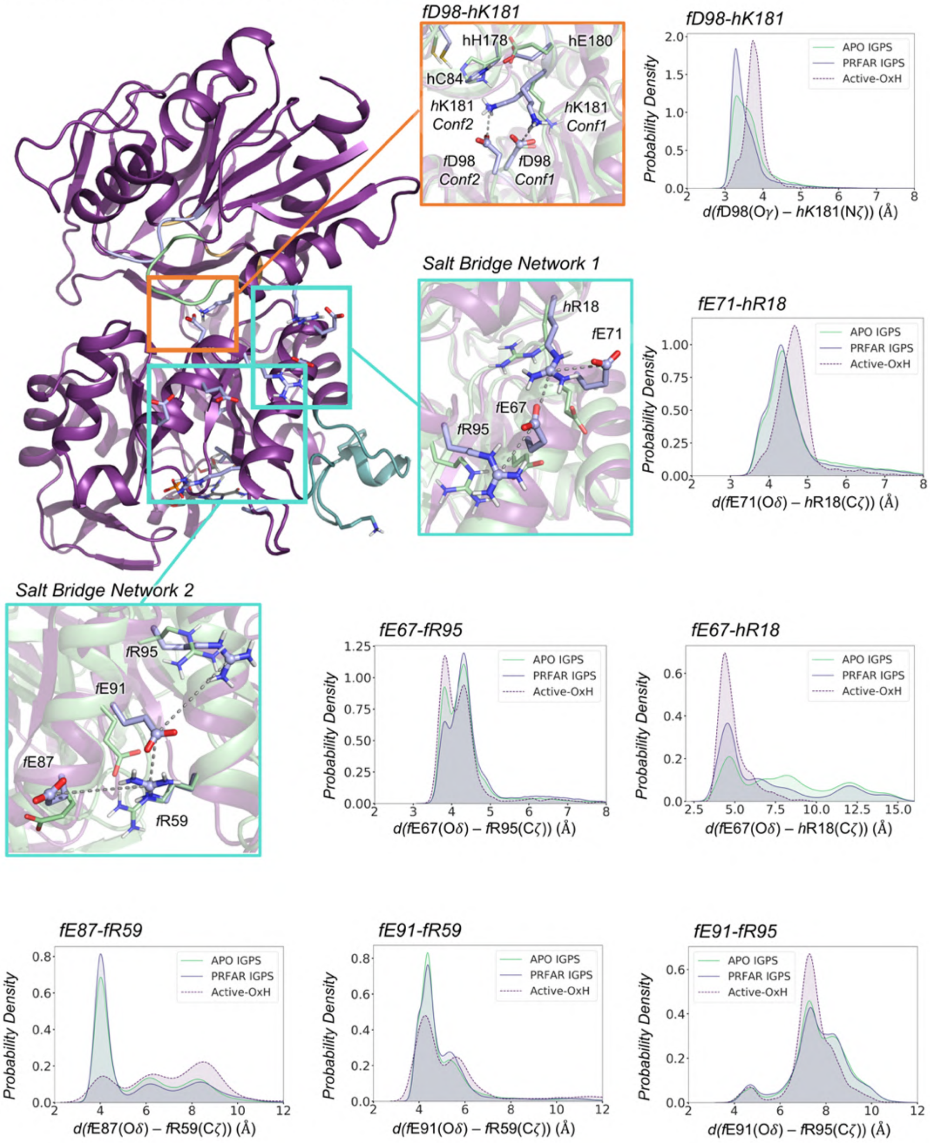


Figure A8. Analysis of global flexibility of IGPS and Loop1 conformational dynamics. a) Plot of the Root-Mean-Square-Fluctuation (RMSF, in Å) for *apo* (green) and PRFAR-IGPS (purple) obtained from ten replicas of 1.5 μs cMD simulations. The plot is divided into HisF (left) *f*1-*f*253 residues and HisH (right) *h*1-*h*201 residues. The most relevant catalytic residues are highlighted in gray and in orange for HisF and HisH, respectively. The residues displaying NMR millisecond motions in HisF in the presence of PRFAR reported by Lisi and Loria[7] are highlighted in purple in the top of the plot. In HisH, the positions of *h*P10 and *h*V51 are shown in green and purple respectively. Vertical green and purple dashed lines indicate the position of the Ω-loop and oxyanion strand, respectively. In terms of global flexibility, the most significant differences are in *f*Loop1. **b)** Structural representation of RMSF in the IGPS structure. The most flexible regions are represented in red (thicker loops), and the least flexible in blue (thinner loops). **c)** Representative conformation of Loop1 extracted from cMD simulations (teal). Overlay of IGPS structures with a closed Loop1 conformation (PDB 1GPW chain A, in blue) and open Loop1 conformation (PDB 1GPW chain E, in orange). Loop1 transitions from closed to open conformation in both *apo* and PRFAR-IGPS cMD simulations.

Conventional MD: HisF conformational dynamics

a. HisF:HisH interface and HisF salt bridge network interactions



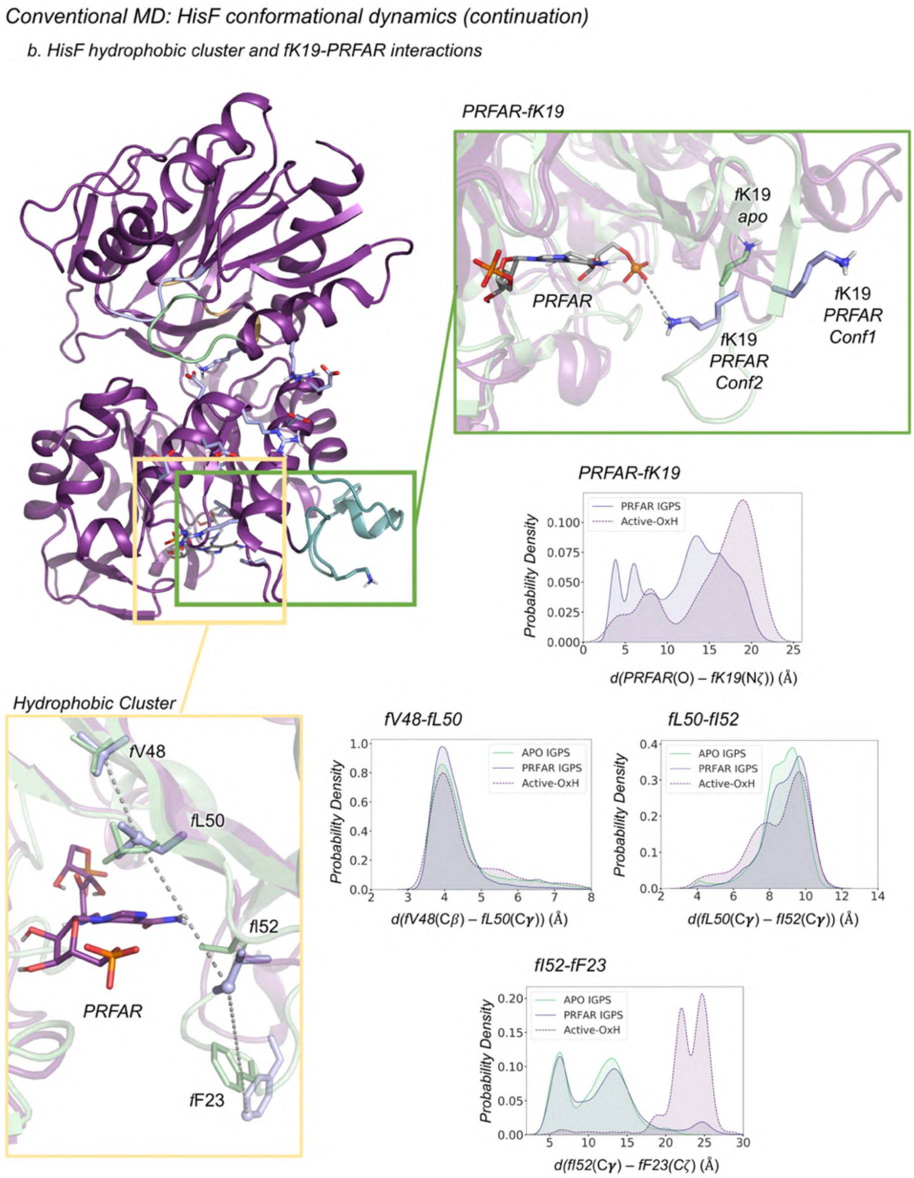


Figure A9. HisF conformational dynamics in cMD simulations. Analysis of the most relevant interactions in HisF subunit in the *apo* and PRFAR-IGPS states along the cMD simulations. The selected distances were described by Rivalta et al.[8] and shown to be relevant for analyzing the effects of PRFAR. The global conformation of PRFAR-IGPS is shown in deep purple. Loop1 is represented in teal, the Ω-loop in green, and the oxyanion strand in light purple. Overlay of the side chains of representative conformations in the *apo* and IGPS-PRFAR states are shown in green and light purple, respectively. **a)** Molecular representation of HisF salt bridge network (highlighted in cyan squares) and *f*D98-*h*K181 HisF:HisH interaction (highlighted in an orange square). Probability density distribution of the most relevant distances of the salt bridge network and *f*D98-*h*K181 distance. The distances of the salt bridge network are calculated between the carbon atom of the carboxylate group of the glutamate side chain and the carbon atom of the guanidinium group of the arginine residues. The distance of the *f*D98-*h*K181 interaction is calculated between the carbon atom of the carboxylate group of *f*D98 side chain and the nitrogen of the side chain of *h*K181. **b)**

Molecular representation of HisF hydrophobic cluster (highlighted in a yellow square) and PRFAR-*f*K19 HisF:HisH interaction (highlighted in a green square). Probability density distribution of the most relevant distances of the hydrophobic cluster and PRFAR-*f*K19 distance. The distances between the residues forming the hydrophobic cluster are monitored between the β-carbon of *f*V48, the γ-carbon of *f*L50, the γ-carbon of *f*I52, and the ζ-carbon of *f*F23. The distance of the PRFAR-*h*K19 interaction is calculated between the phosphorus atom of PRFAR and the nitrogen of the side chain of *h*K19. In the probability density plots the *apo*, PRFAR-IGPS, and Active-OxH PRFAR-IGPS distances are shown in green, purple, and dashed purple lines, respectively. All distances are in Å. See Appendix A Extended text for a complete description of the results.

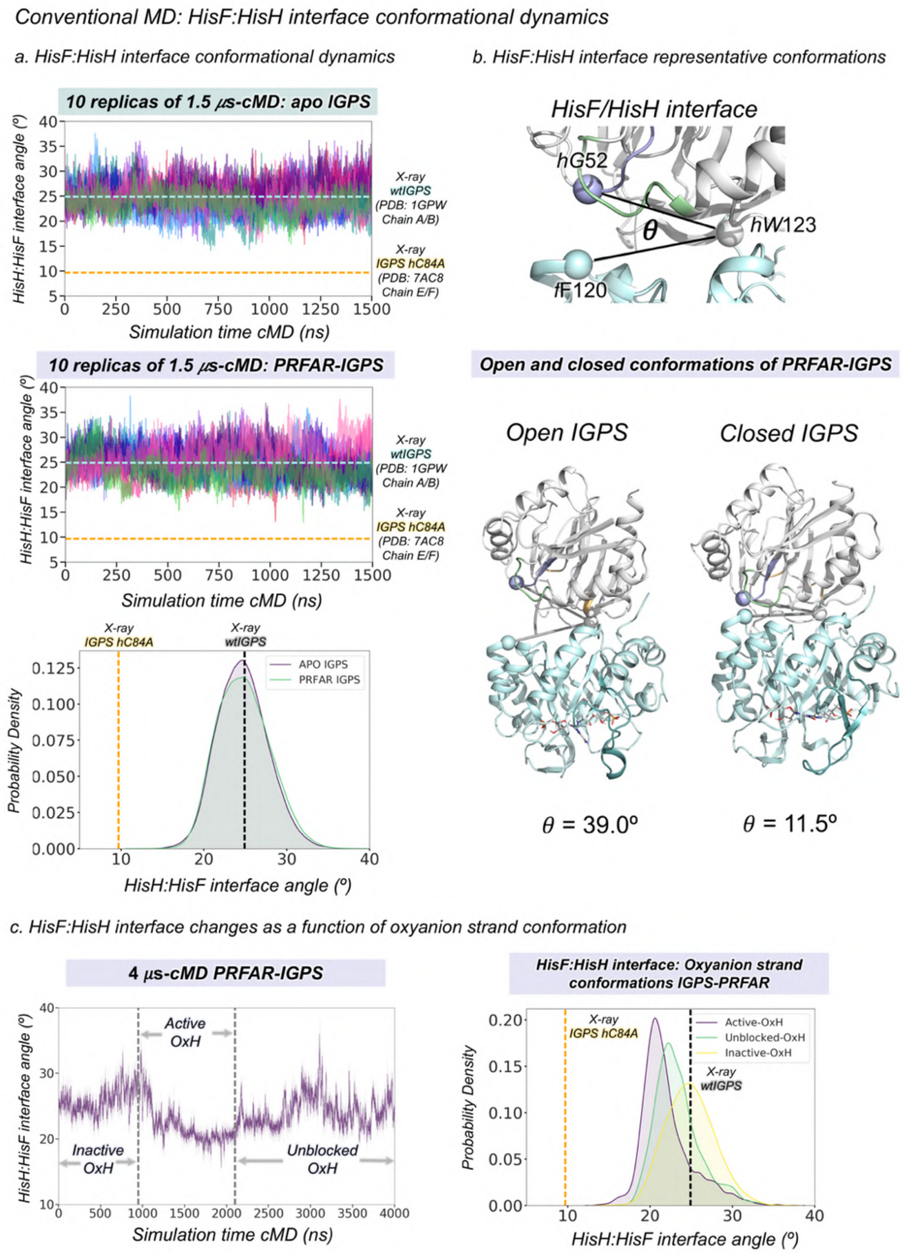


Figure A10. HisF:HisH interface conformational dynamics in cMD simulations. a) Plot of the HisF:HisH interface angle (θ, in degrees) for ten replicas of 1.5 μs cMD simulations in the *apo*-IGPS and PRFAR-IGPS states. Each replica is depicted in a different color. Horizontal cyan dashed lines indicate the HisF:HisH interface angle (θ = 24.9 º) found in the *X-ray* structure (PDB 1GPW chains A/B) used as starting point for cMD simulations. Horizontal orange dashed lines indicate the HisF:HisH interface angle (θ = 9.7 º) found in the *X-ray* structure of substrate-bound *h*C84A IGPS (PDB 7AC8 chains E/F) that displays an active conformation of the oxyanion strand. Probability density distribution for the HisF:HisH interface angle in the *apo* and PRFAR-IGPS states. Vertical gray dashed line corresponds to the 1GPW (chains A/B) *X-ray* HisF:HisH interface angle and the vertical dashed orange line to the 7AC8 (chains E/F) *X-ray* HisF:HisH interface angle. **b)**

Representative conformations of an open and closed conformations of the HisF:HisH interface sampled in PRFAR-IGPS. **c)** Plot of the HisF:HisH interace angle along a representative 4 μs-cMD simulation that displays the formation of the *h*V51 oxyanion hole. Vertical dashed gray lines indicate the range of the Active-OxH state of the oxyanion strand. The formation of the *h*V51 oxyanion hole is correlated with a partial closure of the HisF:HisH interface (see Appendiz A Extended text for a complete description) Probability density distribution of the HisF:HisH interface angle for the different states of the oxyanion strand obtained in cMD simulations. The population of the Active-OxH oxyanion strand decreases the HisF:HisH interface angle with respect to the Inactive-OxH and Unblocked-OxH states. However, the values of the HisF:HisH interface angle are still far from the productive closure observed in PDB 7AC8 (chains E/F). The angle of the HisF:HisH interface is calculated from the Cα of *f*F120, *h*W123 and *h*G52 as indicated by Rivalta and coworkers.[8]

Accelerated Molecular Dynamics Simulations IGPS: substrate-free (Figures A11-A14)

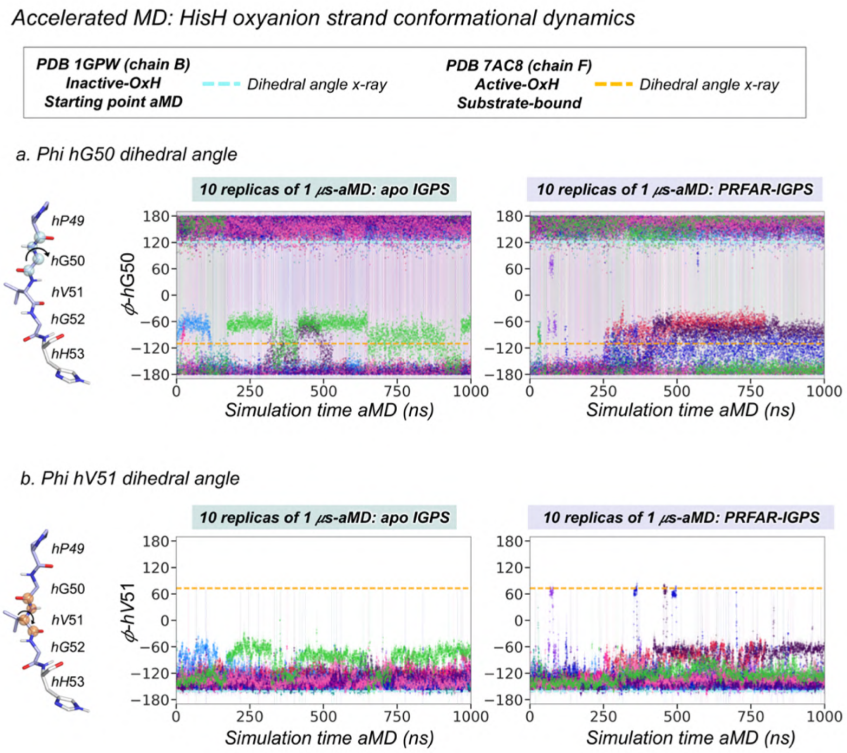


Figure A11. HisH *h*49-PVGV conformational dynamics along aMD simulations. Plot of the most relevant dihedral angles of the *h*49-PGVG oxyanion strand for ten replicas of 1 μs accelerated molecular dynamics (aMD) simulations in the *apo*-IGPS and PRFAR-IGPS states. Each replica is depicted in a different color. Horizontal cyan dashed lines indicate the value of the dihedral angle corresponding to the *X-ray* structure (PDB 1GPW (chain B)) used as starting point for aMD simulations. Horizontal orange dashed lines indicate the value of the dihedral angle corresponding to the *X-ray* structure of substrate-bound *h*C84A IGPS (PDB 7AC8 (chain F)) that displays an active conformation of the h49-PVGV oxyanion strand. The oxyanion strand residues are shown in light purple and the atoms involved in each dihedral angle are represented as spheres of different color. (a) ϕ dihedral angle of *h*G50; (b) ϕ dihedral angle of *h*V51. In aMD, multiple short-lived formations of the *h*V51 oxyanion hole are observed. See Figure A12 for a molecular representation of the most relevant states.

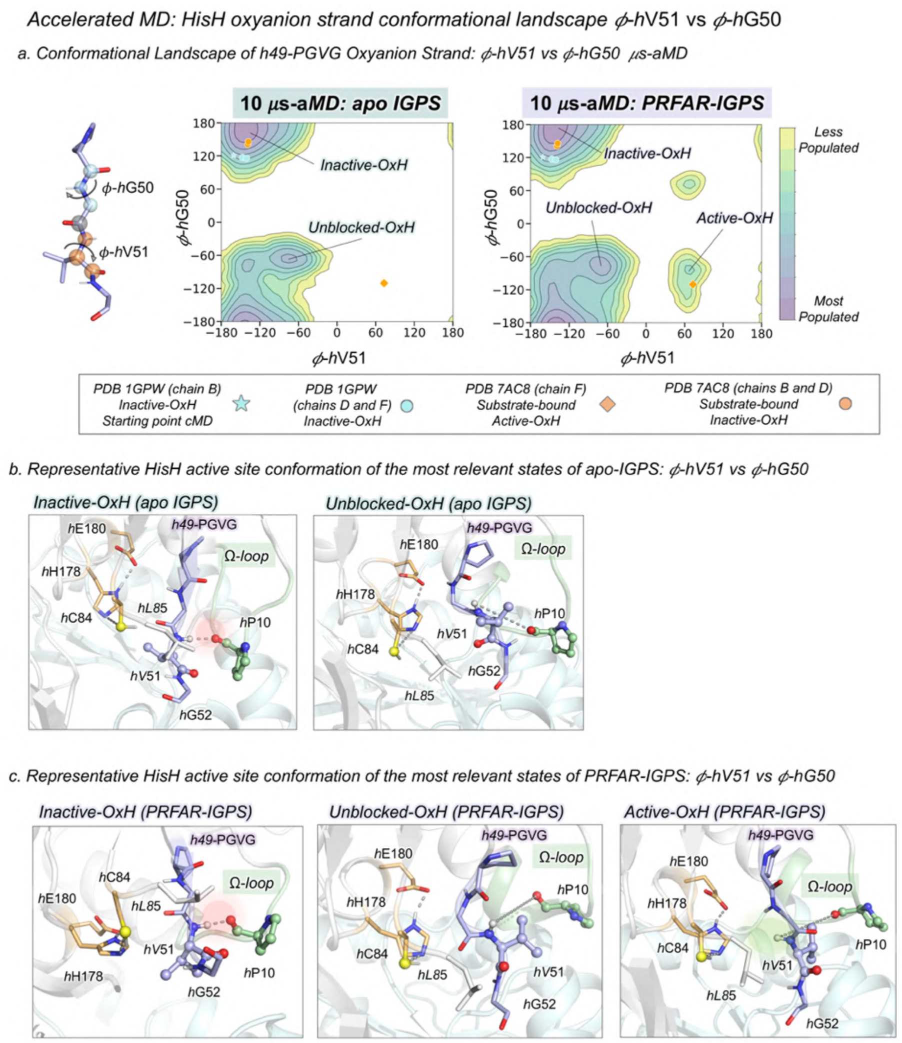
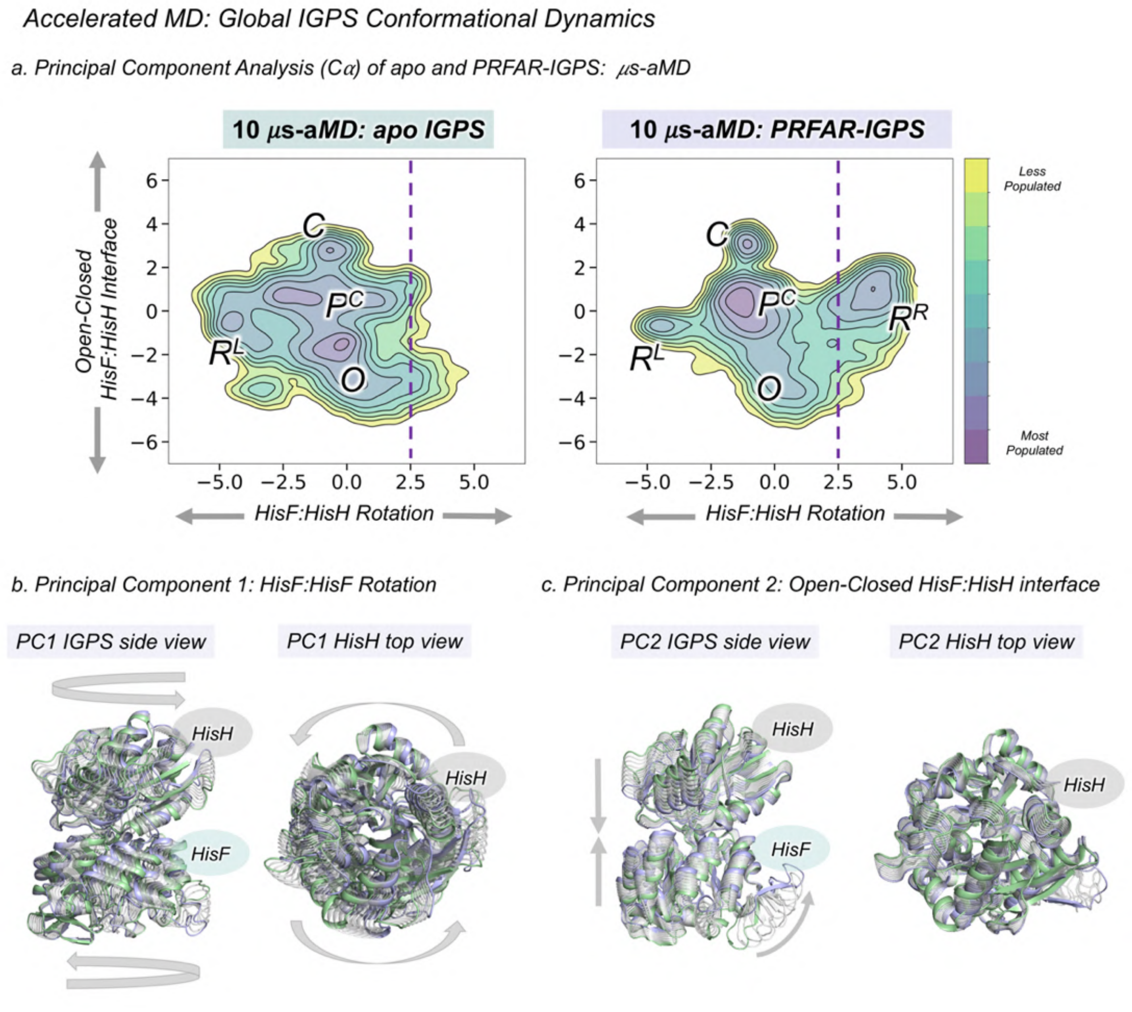


Figure A12. Conformational Landscape of *h*49-PGVG Oxyanion Strand obtained from accelerated Molecular Dynamics (aMD) simulations. The conformational landscape of the *h*49-PGVG oxyanion strand of *apo* and PRFAR-IGPS is constructed from an accumulated time of 10 μs of aMD simulations (10 replicas of 1 μs). The conformational landscape of each state is clusterized into 20 different clusters. **a)** Conformational landscape of *apo* and PRFAR-IGPS constructed using the ϕ dihedral angles of *h*V51 and *h*G50. The values of the ϕ dihedral angles of *h*V51 and *h*G50 found in the *X-ray* structures corresponding to the three chains of PDB 1GPW are depicted in cyan and the three chains of PDB 7AC8 are represented in orange, respectively. The conformation used as starting point for cMD simulations is shown using the cyan star symbol. The conformation corresponding to the active oxyanion strand (Active-OxH) observed in *h*C84A IGPS is depicted using the orange diamond symbol. **b)** Representative HisH active site structures of most populated states in *apo*-IGPS conformational landscape: Inactive-OxH and Unblocked-OxH. (c) Representative HisH active site structures of most populated states in PRFAR-IGPS conformational landscape: Inactive-OxH, Unblocked-OxH, and Active-OxH.


Accelerated MD: Global IGPS Conformational Dynamics
a. Principal Component Analysis (Cα) of apo and PRFAR-IGPS: μs-aMD
10 μs-aMD: apo IGPS
10 μs-aMD: PRFAR-IGPS
Open-Closed HisF:HisH Interface
HisF:HisH Rotation
C
P^C
R^L
O
R^R
6
4
2
0
−2
−4
−6
−5.0
−2.5
0.0
2.5
5.0
Less Populated
Most Populated
b. Principal Component 1: HisF:HisF Rotation
c. Principal Component 2: Open-Closed HisF:HisH interface
PC1 IGPS side view
PC1 HisH top view
PC2 IGPS side view
PC2 HisH top view
HisH
HisF

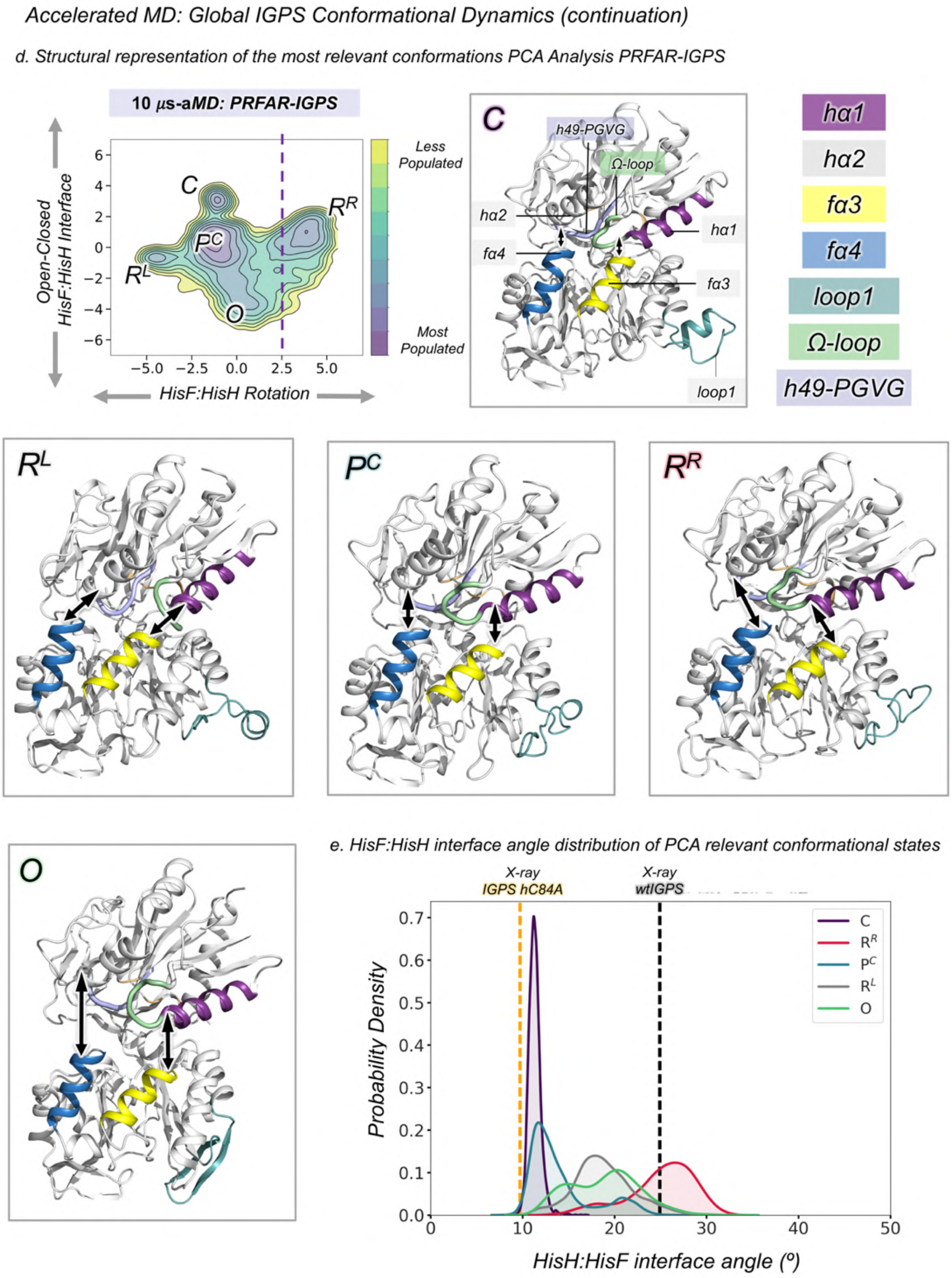
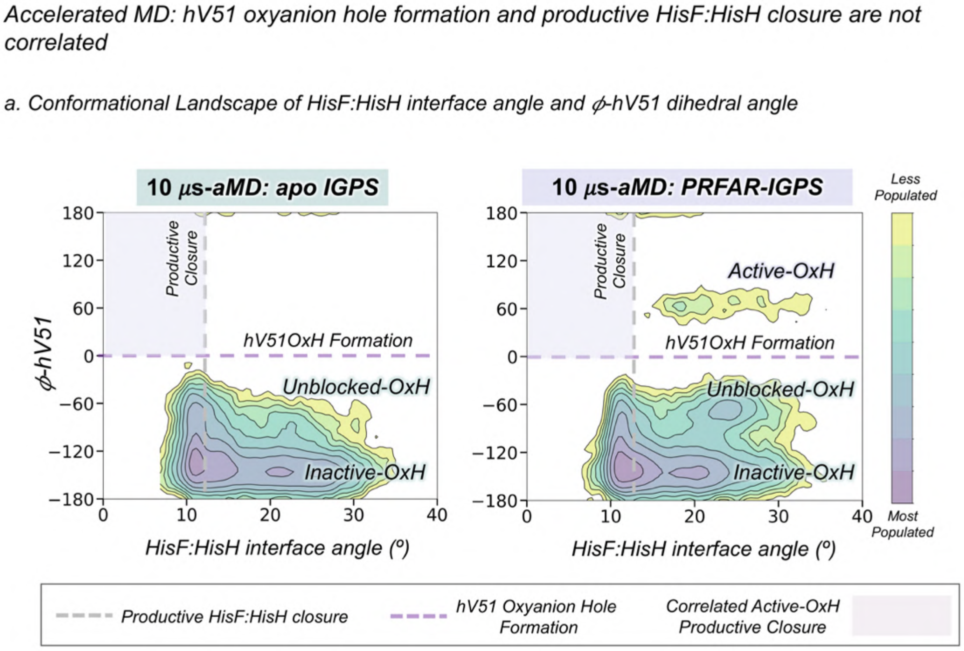


Figure A13. Global IGPS conformational dynamics in aMD simulations. Principal component analysis of aMD simulations. a) Principal Component (PC) analysis considering all alpha-carbons PRFAR-IGPS states reconstructed from ten replicas of 1 μs aMD simulations. PC1 indicates the counter-clock rotation of HisF and HisH subunits with respect to the HisF:HisH interface while PC2 represents the open-closed transition of the HisF:HisH interface. The *apo*-IGPS simulations are projected into the PC space of PRFAR-IGPS for a direct comparison. The vertical purple dashed line indicate the region of PC1 space not visited in *apo*-IGPS simulations. **b)** Structural representation of PC1 and PC2 motions from side and top views of IGPS (see Figure A6). The arrows indicate the direction of motions when going from negative (green cartoon) to positive (purple

cartoon) values of the PC space. **c)** Structural representation of PC2 motion. The arrows indicate the direction of motions when going from negative (green cartoon) to positive (purple cartoon) values of the PC space. **d)** Representative conformations of the most populated states of the PC landscape of PRFAR-IGPS. The *h*α1, *h*α2, *f*α3, and *f*α4 helices are shown in purple, gray, yellow, and blue, respectively. Loop1 is represented in teal, the Ω-loop in green, and the oxyanion strand in light purple. **e)** Probability density distribution of the HisF:HisH interface angle in the most relevant states visited in the aMD simulations. Vertical gray dashed line indicate the HisF:HisH interface angle found in the *X-ray* structure (PDB 1GPW chains A/B) used as starting point for aMD simulations. Vertical orange dashed lines indicate the HisF:HisH interface angle found in the *X-ray* structure of substrate bound *h*C84A IGPS (7AC8 chain E/F) that displays an active conformation of the oxyanion strand. The angle of the HisF:HisH interface is calculated from the Cα of *f*F120, *h*W123 and *h*G52. aMD simulations show that IGPS can attain the productive closure (C state) even when the substrate is not present.

Several orientations of the HisF:HisH subunits are found to be relatively stable along PC1 that highlight different closures of IGPS interdomain regions. In the most populated one (P^C, see Figure A13), the oxyanion strand loop interacts with the top of *f*α4 residues (*f*T119) while the Ω-loop and the bottom of *h*α1 establish interactions with the top of *f*α3. In this state, the HisF:HisH interface angle decreases to average values of 14.0 ± 3.5 °. The two additional states along PC1 correspond to different degrees of rotation of the HisF:HisH subunits. In R^L, the Ω-loop interacts with the top of *f*α4 residues and represents the displacement of HisH towards the left with respect to HisF. In R^R, the HisH subunit rotates towards the right with respect to HisF, with the oxyanion strand residues topping the *f*α3. This degree of rotation is not captured in *apo* IGPS. PC2 captures transitions in Loop1 and the closure of the HisF:HisH interface. We identified a state that displays productive closure (C, HisF:HisH interface angle of 11.3 ± 0.6 º), as observed in the *X-ray h*C84A IGPS structure, PDB 7AC8 (interface angle of 9.7 º). In this state, the amide backbone of *h*H53 establishes a hydrogen bond with the carbonyl backbone of *f*T119, the Ω-loop collapses over *f*α3 and the *h*α1 and *f*α3 are perfectly aligned. We have identified a potential productively closed state of IGPS that can be key for catalytic activity. In general, the conformational ensemble is displaced towards shorter angles of the HisF:HisH interface (see Figure A13e). However, the closure of the subunit is not stabilized through the simulation time and is not correlated with other motions. Similar closed states are sampled in the *apo* state simulations indicating that the efficient closure of IGPS is not limited to PRFAR simulations. PRFAR releases tension in the interdomain region facilitating the rotation of HisH and HisF subunits and the closure of the interdomain region.



Accelerated MD: hV51 oxyanion hole formation and productive HisF:HisH closure are not correlated

a. Conformational Landscape of HisF:HisH interface angle and ϕ-hV51 dihedral angle

Figure A14. Uncorrelated HisF:HisH interface and oxyanion-strand conformational dynamics in aMD simulations. Conformational landscape constructed using the HisF:HisH interface angle and ϕ dihedral angle of *h*V51 obtained from accelerated Molecular Dynamics (aMD) simulations of *apo* and PRFAR-IGPS states. Vertical gray dashed line indicates productive closure (HisF:HisH interface angle below 12º). Horizontal purple dashed line indicate *h*V51 oxyanion hole formation (ϕ-*h*V51 above 0º). The purple area in the plot indicates the region of the conformational landscape with a productively closed HisF:HisH interface and a *h*V51 oxyanion hole formed. As the results show, the two events are not correlated in substrate-free aMD simulations.

Accelerated Molecular Dynamics Simulations IGPS: spontaneous substrate binding (Figures A15-A22)

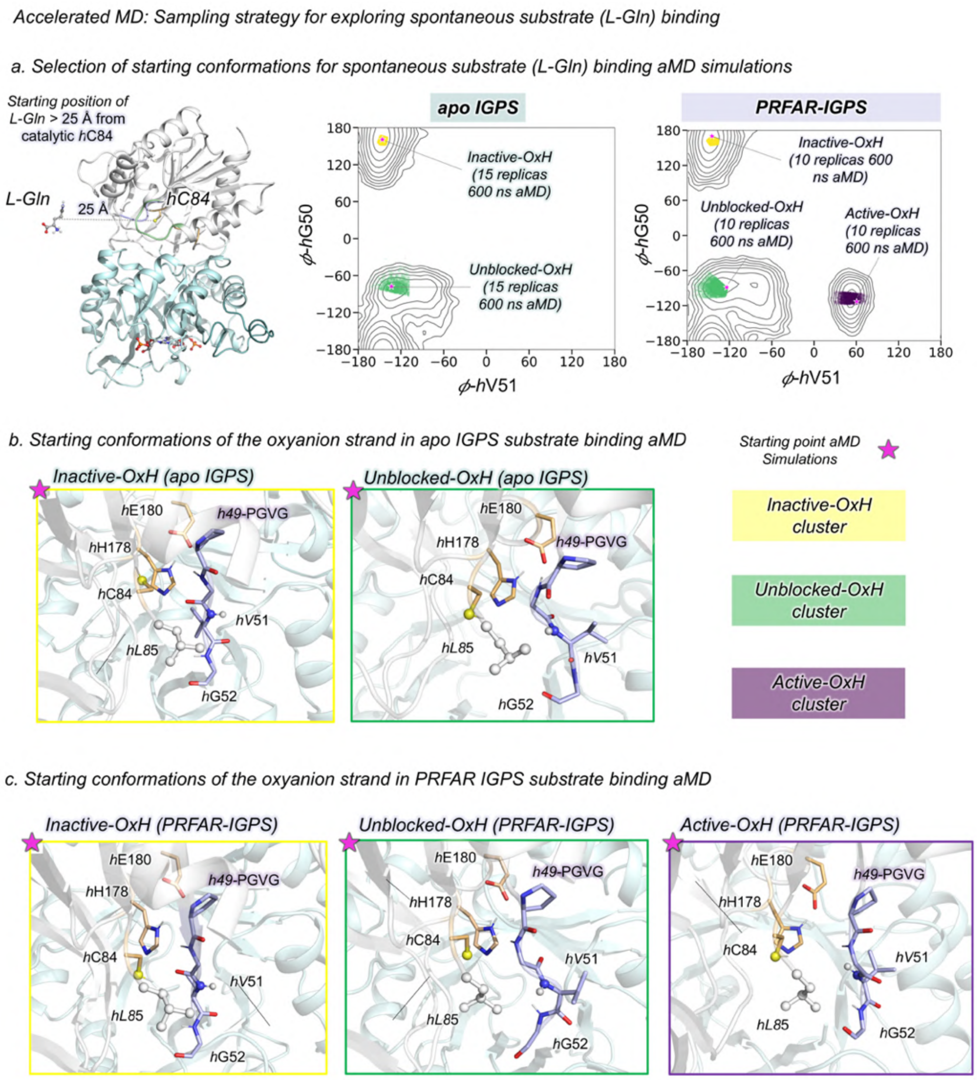

Accelerated MD: Sampling strategy for exploring spontaneous substrate (L-Gln) binding
a. Selection of starting conformations for spontaneous substrate (L-Gln) binding aMD simulations
Starting position of L-Gln > 25 Å from catalytic hC84
L-Gln
25 Å
hC84
apo IGPS
Inactive-OxH (15 replicas 600 ns aMD)
Unblocked-OxH (15 replicas 600 ns aMD)
ϕ-hG50
ϕ-hV51
180
120
60
0
−60
−120
−180
PRFAR-IGPS
Inactive-OxH (10 replicas 600 ns aMD)
Unblocked-OxH (10 replicas 600 ns aMD)
Active-OxH (10 replicas 600 ns aMD)
b. Starting conformations of the oxyanion strand in apo IGPS substrate binding aMD
Inactive-OxH (apo IGPS)
Unblocked-OxH (apo IGPS)
Starting point aMD Simulations
Inactive-OxH cluster
Unblocked-OxH cluster
Active-OxH cluster
hE180
h49-PGVG
hH178
hC84
hV51
hL85
hG52
c. Starting conformations of the oxyanion strand in PRFAR IGPS substrate binding aMD
Inactive-OxH (PRFAR-IGPS)
Unblocked-OxH (PRFAR-IGPS)
Active-OxH (PRFAR-IGPS)

Figure A15. Spontaneous *L-Gln* substrate binding sampling strategy. a) General scheme of spontaneous substrate binding process in the *apo* and PRFAR-IGPS states. The ligand (glutamine, *L-Gln*) is arbitrarily positioned 25 Å away from the catalytic *h*C84 residue at the HisH active site. Conformational landscape of *apo* and PRFAR-IGPS constructed using the ϕ dihedral angles of *h*V51 and *h*G50 with the relevant oxyanion strand conformations highlighted. The pink stars indicate the IGPS structures used as starting point for substrate binding aMD simulations. The cluster of conformations of Inactive-OxH, Unblocked-OxH, and Active-OxH are shown in yellow, green and purple, respectively. In *apo*-IGPS, 15 replicas of 600 ns are carried out starting from both the Inactive-OxH and Unblocked-OxH states. In PRFAR-IGPS, 10 replicas of 600 ns are performed starting from the Inactive-OxH, Unblocked-OxH, and Active-OxH states. **b)** Structures of IGPS with the corresponding HisH active site conformation selected as starting point for substrate binding aMD simulations in the *apo* state. (c) Structures of IGPS with the corresponding HisH active site

conformation selected as starting point for substrate binding aMD simulations in the PRFAR-IGPS state.

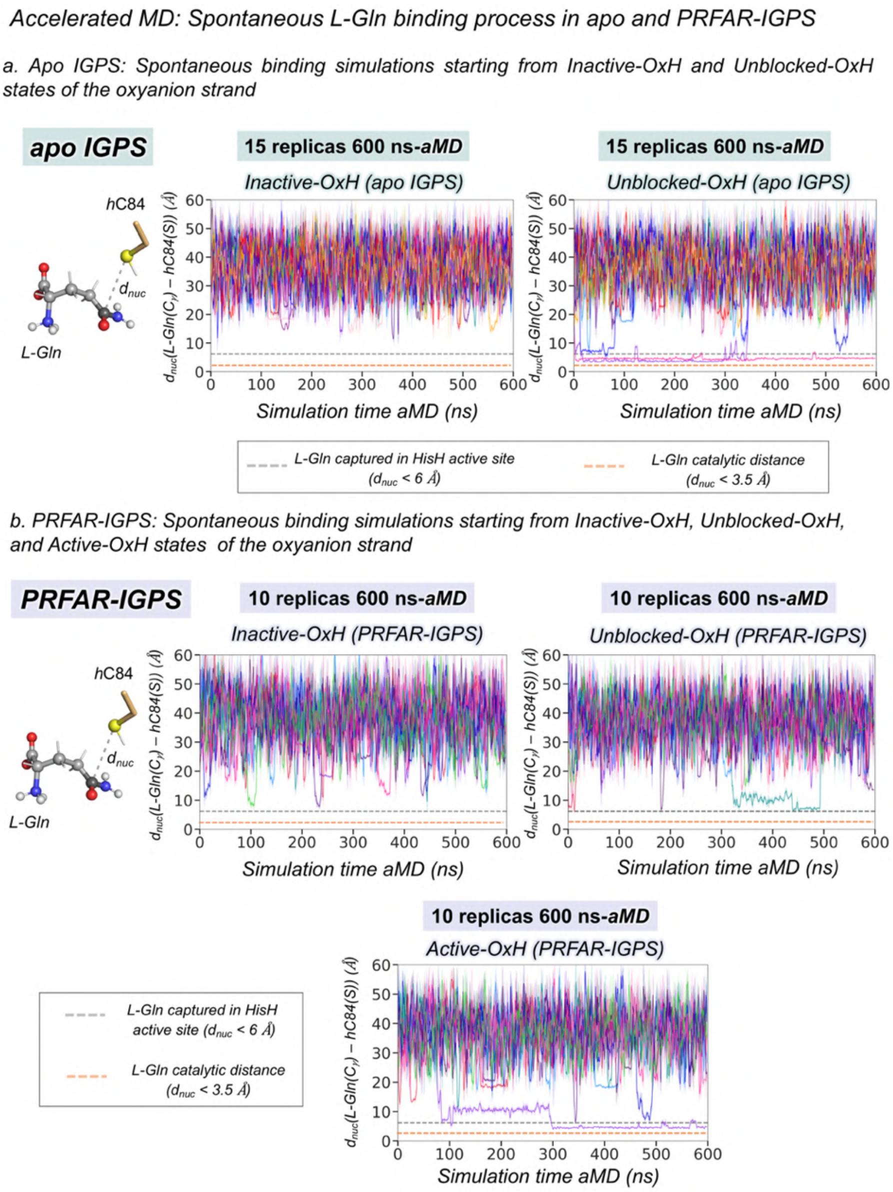


Figure A16. Spontaneous *L-Gln* binding process in *apo* and PRFAR-IGPS states. a) Plot of the distance (d_{nuc}) between the γ-carbon of *L-Gln* and the sulfur of *h*C84 for fifteen replicas of 600 ns aMD simulations starting from the Inactive-OxH and Unblocked-OxH states of the oxyanion strand in *apo*-IGPS. Each replica is depicted in different color. *L-Gln* binding occurs in 0/15 and 2/15 (magenta and purple) replicas that started with Inactive-OxH and Unblocked-OxH states, respectively. **b)** Plot of the distance (d_{nuc}) between the γ-carbon of *L-Gln* and the sulfur of *h*C84 for ten replicas of 600 ns aMD simulations simulations starting from the Inactive-OxH, Unblocked-OxH, and Active-OxH states in PRFAR-IGPS. *L-Gln* binding occurs in 0/10, 0/10, and 1/10 (purple) replicas that started with Inactive-OxH, Unblocked-OxH, and Active-OxH states, respectively. Despite the starting orientation, binding always occur when the oxyanion strand attains the Inactive-OxH state (see below). Horizontal gray dashed line indicates the distance when *L-Gln* is captured in the HisH active site (d_{nuc} below 6 Å). Horizontal orange dashed line indicate the distance when *L-Gln* is at a catalytic distance of *h*C84 (d_{nuc} below 3.5 Å). All distances are in Å.

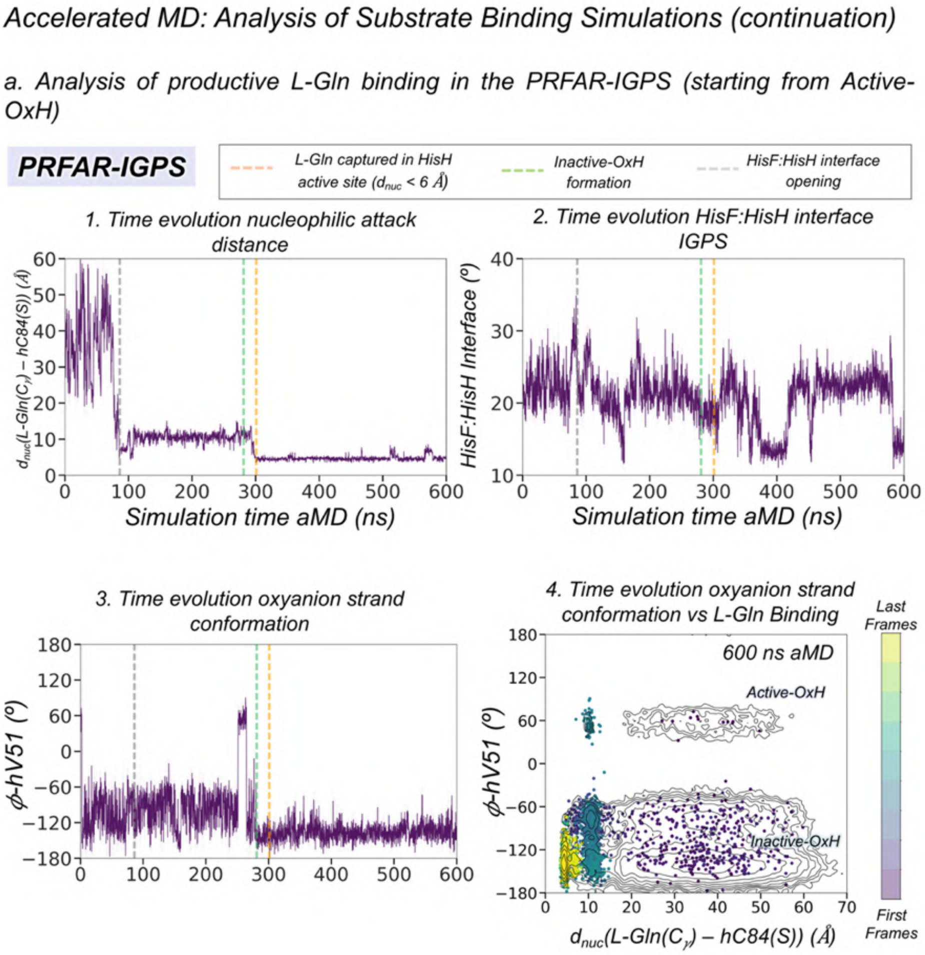


Figure A17. Analysis of substrate binding simulations in PRFAR-IGPS. Analysis of a representative aMD simulation where *L-Gln* binding in the HisH active site is observed in PRFAR-IGPS. **a)** Plot of the most significant distances for ligand binding aMD simulations in the PRFAR-IGPS. Vertical orange dashed line indicates when *L-Gln* is captured for the first time in the HisH active site. Vertical gray dashed line indicates when HisF:HisH interface expands to capture *L-Gln*. Vertical green dashed line indicates the deactivation of the oxyanion strand from Active-OxH to Inactive-OxH. **1.** Plot of the nucleophilic attack distance between the amide carbon of *L-Gln* and the sulfur of the side chain of *h*C84. **2.** Plot of the HisF:HisH interface angle along the simulation time. **3.** Plot of the ϕ dihedral angle of *h*V51. **4.** Projection of a representative aMD trajectory on the conformational landscape obtained from the nucleophilic attack distance between the thiol group of catalytic *h*C84 and the amide carbon of *L-Gln*, and the ϕ dihedral angle of *h*V51 (see Figure 4.4 of Chapter 4). The time evolution of the ligand binding pathway is represented in a color scale ranging from purple for the first frames to yellow for the last frames of the aMD trajectory.

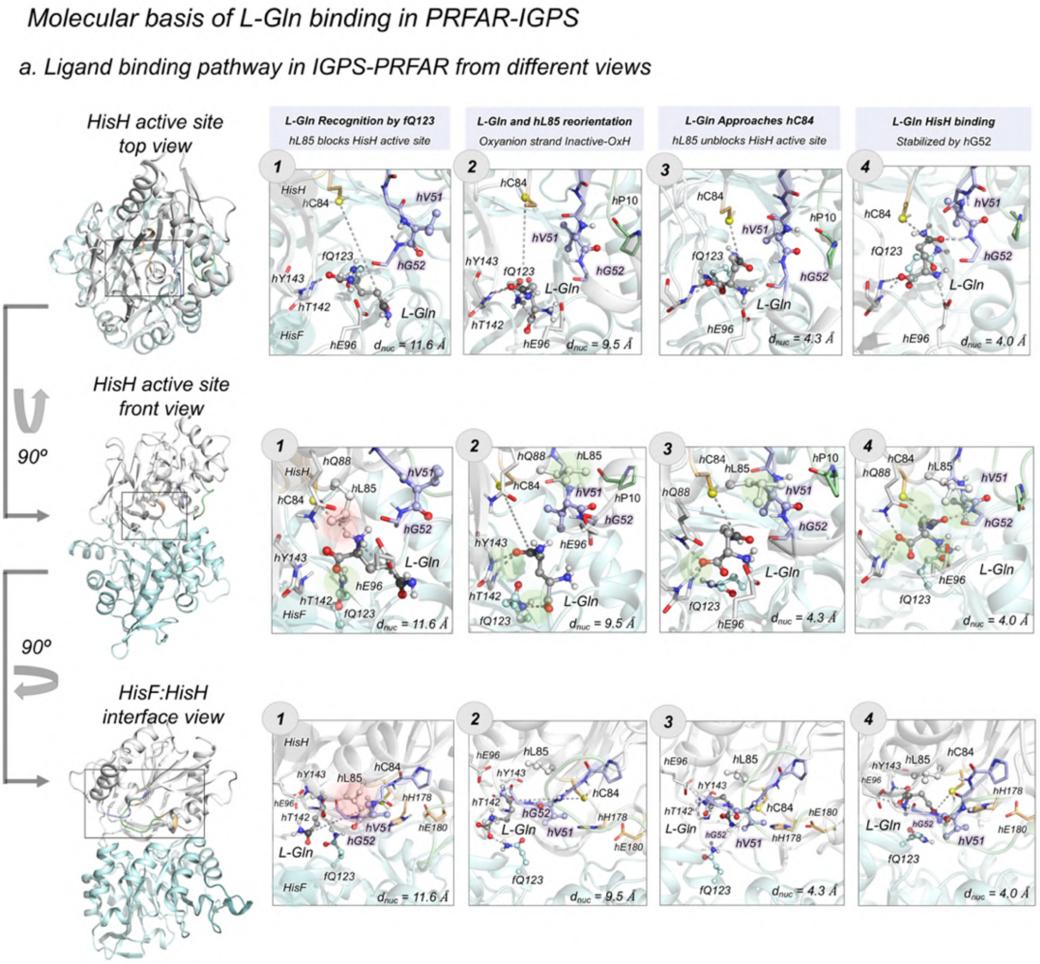

Molecular basis of L-Gln binding in PRFAR-IGPS
a. Ligand binding pathway in IGPS-PRFAR from different views
HisH active site top view
HisH active site front view
HisF:HisH interface view
90º
90º
L-Gln Recognition by fQ123
hL85 blocks HisH active site
L-Gln and hL85 reorientation
Oxyanion strand Inactive-OxH
L-Gln Approaches hC84
hL85 unblocks HisH active site
L-Gln HisH binding
Stabilized by hG52
1
2
3
4
d_{nuc} = 11.6 Å
d_{nuc} = 9.5 Å
d_{nuc} = 4.3 Å
d_{nuc} = 4.0 Å
HisH
HisF
hC84
hV51
hG52
fQ123
hY143
hT142
hE96
hP10
hQ88
hL85
hH178
hE180
L-Gln

Figure A18. Molecular basis of *L-Gln* binding in PRFAR-IGPS. a) Molecular representation of the four most relevant steps in the binding pathway of *L-Gln* into the HisH active site obtained from aMD trajectories. The selected snapshots are depicted from different views: top, front and HisF:HisH interface views (see Figure A6 for a complete description of the different points of view). The HisH catalytic residues are highlighted in orange, Ω-loop residues in green, and the residues of the *h*49-PGVG oxyanion strand in purple. Other relevant HisF and HisH residues are shown in cyan and white, respectively. The atoms of *h*L85, *f*Q123, *h*V51, and *h*P10 are shown as spheres. *h*L85 is not shown in the top view to facilitate the visual analysis of *L-Gln* conformation along the binding pathway. The green surfaces indicate the establishment of non-covalent interactions. The red surface indicates when *h*L85 blocks the access to the nucleophilic *h*C84. The nucleophilic attack distance (d_{nuc}) between the amide carbon of *L-Gln* and the sulfur of the side chain of *h*C84 is specified for each snapshot.

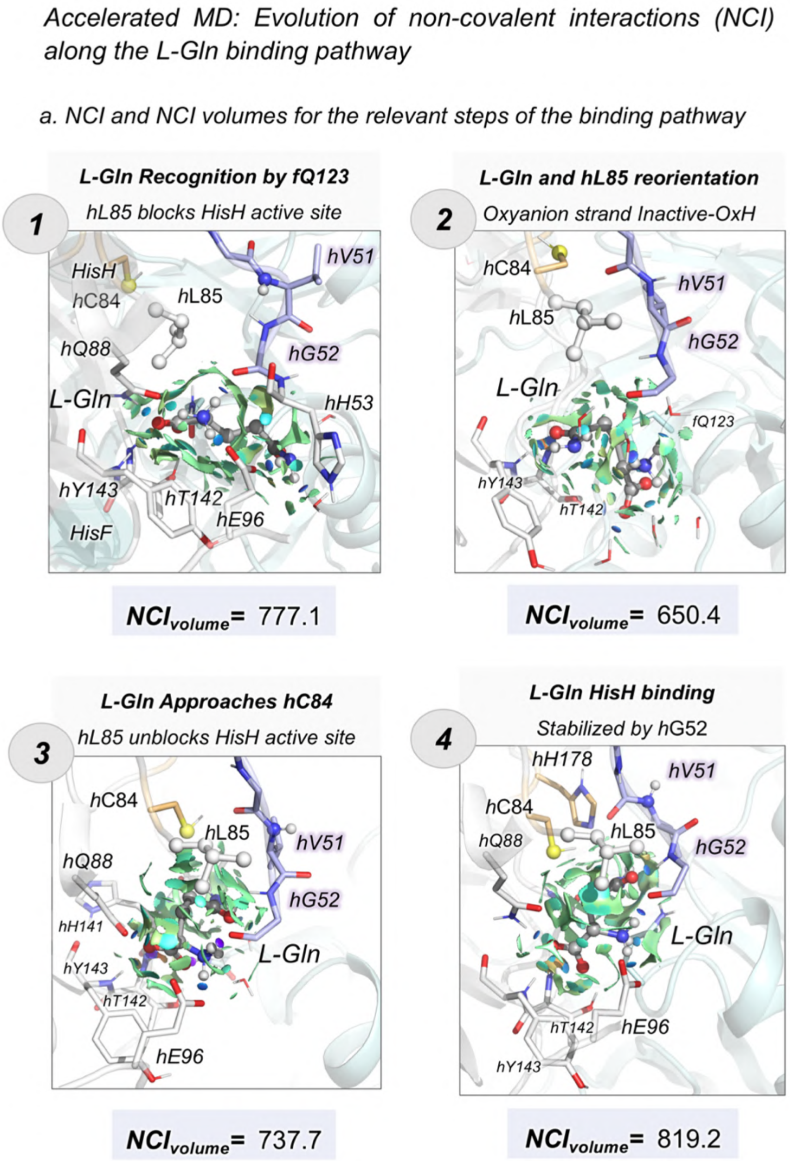


Figure A19. Evolution of non-covalent interactions (NCI) along the ligand binding pathway. Schematic representation of non-covalent interactions for the four most relevant steps of the *L-Gln* binding process into the HisH active site in PRFAR-IGPS calculated with the NCI plot.[9] Blue, green, and red surfaces indicate strong, weak, and repulsive non-covalent interactions, respectively. The integrated volumes of non-covalent interactions are provided for each step. Higher volumes indicate overall stronger NCI interactions.

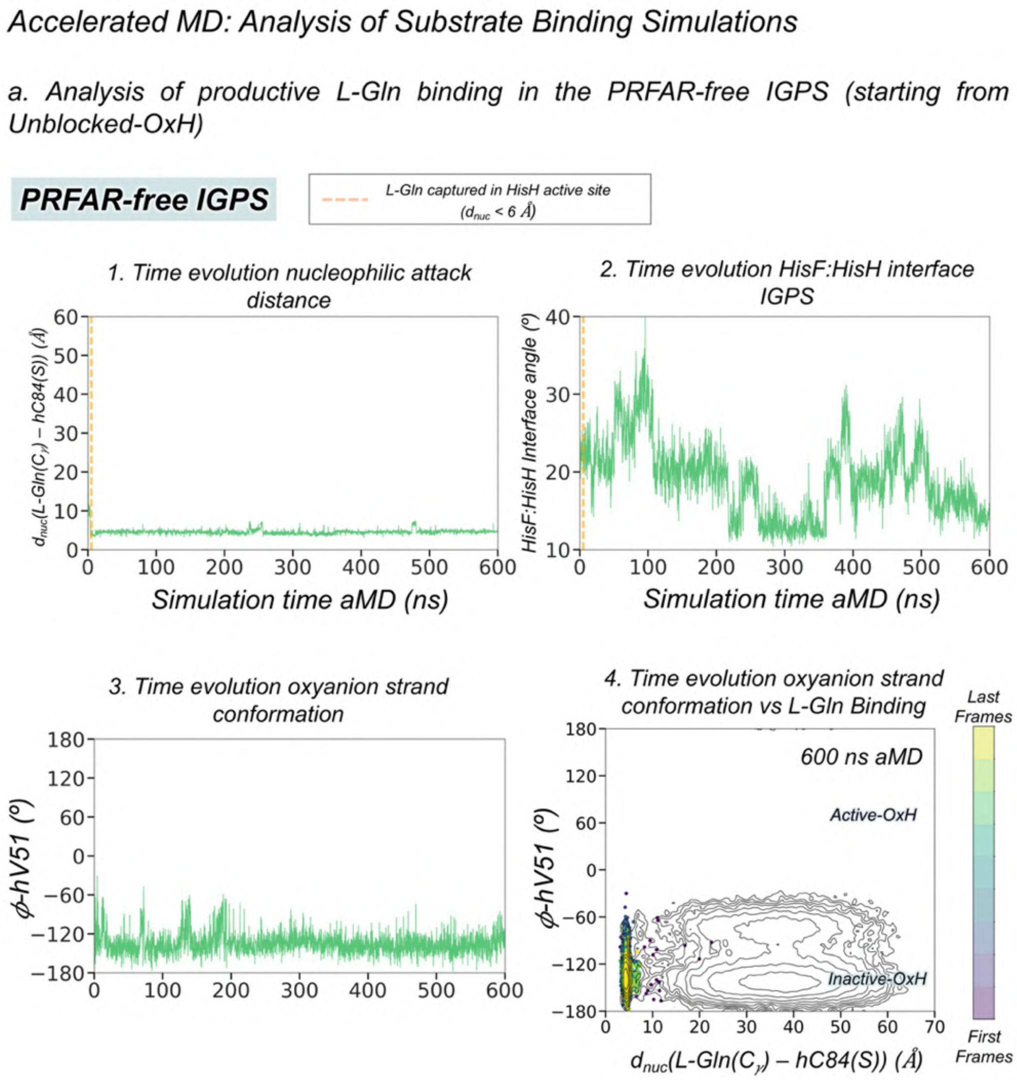


Figure A20. Analysis of substrate binding simulations in PRFAR-free IGPS. Analysis of a representative aMD simulation where *L-Gln* binding in the HisH active site is observed in PRFAR-free IGPS. **a)** Plot of the most significant distances for ligand binding aMD simulations in the PRFAR-free IGPS. Vertical orange dashed line indicates when *L-Gln* is captured for the first time in the HisH active site. **1.** Plot of the nucleophilic attack distance between the amide carbon of *L-Gln* and the sulfur of the side chain of *h*C84. **2.** Plot of the HisF:HisH interface angle along the simulation time. **3.** Plot of the ϕ dihedral angle of *h*V51. **4.** Projection of a representative aMD trajectory on the conformational landscape obtained from the nucleophilic attack distance between the thiol group of catalytic *h*C84 and the amide carbon of *L-Gln*, and the ϕ dihedral angle of *h*V51 (see Figure 4.4). The time evolution of the ligand binding pathway is represented in a color scale ranging from purple for the first frames to yellow for the last frames of the aMD trajectory.

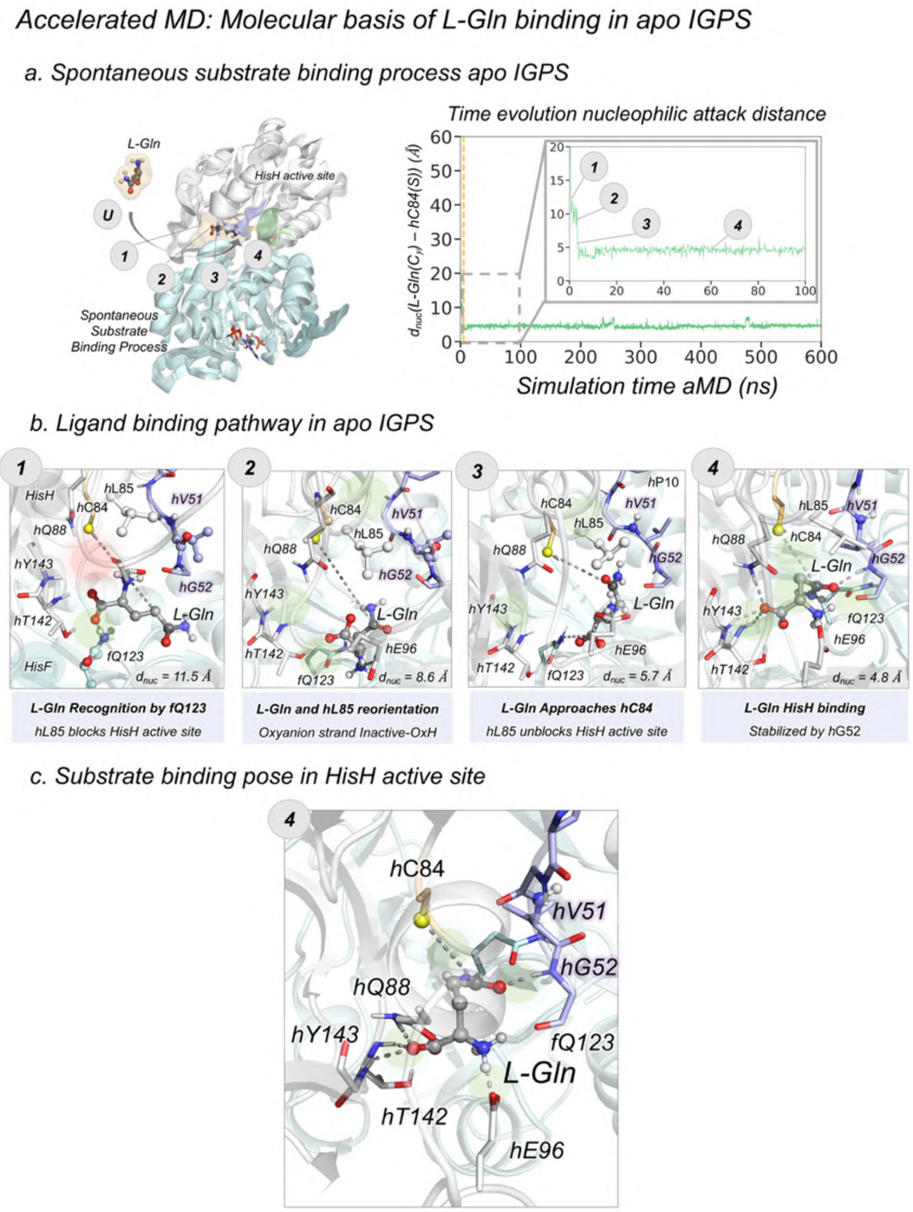

Accelerated MD: Molecular basis of L-Gln binding in apo IGPS
a. Spontaneous substrate binding process apo IGPS
L-Gln
HisH active site
U
1
2
3
4
Spontaneous Substrate Binding Process
Time evolution nucleophilic attack distance
d_{nuc}(L-Gln(C_γ) – hC84(S)) (Å)
60
50
40
30
20
10
0
0
100
200
300
400
500
600
Simulation time aMD (ns)
b. Ligand binding pathway in apo IGPS
1
HisH
hL85
hV51
hC84
hQ88
hY143
hG52
hT142
L-Gln
HisF
fQ123
d_{nuc} = 11.5 Å
L-Gln Recognition by fQ123
hL85 blocks HisH active site
2
hC84
hV51
hL85
hQ88
hG52
hY143
L-Gln
hE96
hT142
fQ123
d_{nuc} = 8.6 Å
L-Gln and hL85 reorientation
Oxyanion strand Inactive-OxH
3
hP10
hC84
hV51
hL85
hQ88
hG52
L-Gln
hY143
hE96
hT142
fQ123
d_{nuc} = 5.7 Å
L-Gln Approaches hC84
hL85 unblocks HisH active site
4
hL85
hV51
hQ88
hC84
L-Gln
hG52
hY143
fQ123
hE96
hT142
d_{nuc} = 4.8 Å
L-Gln HisH binding
Stabilized by hG52
c. Substrate binding pose in HisH active site
4
hC84
hV51
hG52
hQ88
hY143
fQ123
L-Gln
hT142
hE96

Figure A21. Molecular basis of *L-Gln* binding in PRFAR-free IGPS. a) General scheme of spontaneous substrate binding process in PRFAR-free IGPS. The numbers indicate the different steps of the substrate binding process. Plot of the distance corresponding to the nucleophilic attack along the 600 ns of aMD simulation for a representative replica. **b)** Structural representation of selected key conformational states of the *L-Gln* binding pathway in *apo* IGPS. The substrate is shown in gray, the oxyanion strand residues in purple, the catalytic residues in orange, the Ω-loop in green and other relevant HisH and HisF residues in white and cyan, respectively. **c)** Structural representation of the *L-Gln* binding pose in the HisH active site of PRFAR-free IGPS.

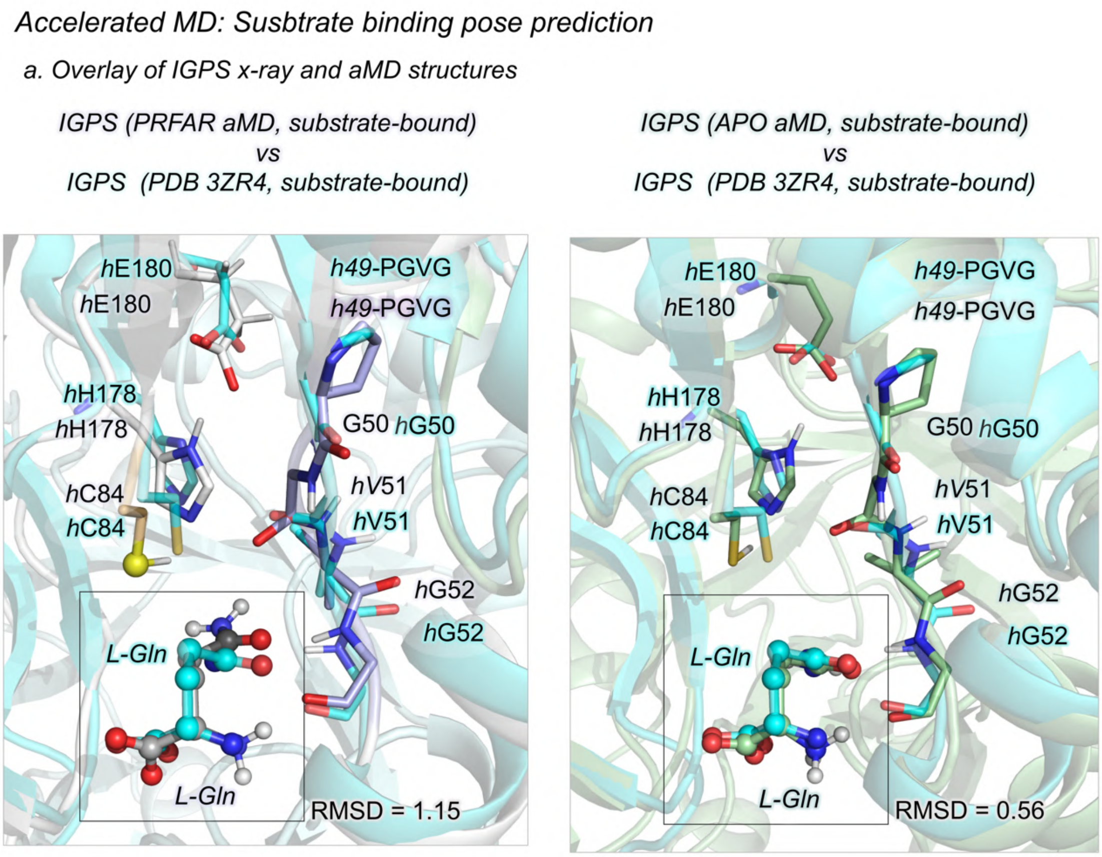

Accelerated MD: Susbtrate binding pose prediction
a. Overlay of IGPS x-ray and aMD structures
IGPS (PRFAR aMD, substrate-bound)
vs
IGPS (PDB 3ZR4, substrate-bound)
IGPS (APO aMD, substrate-bound)
vs
IGPS (PDB 3ZR4, substrate-bound)
hE180
hE180
h49-PGVG
h49-PGVG
hH178
hH178
G50 hG50
hC84
hC84
hV51
hV51
hG52
hG52
L-Gln
L-Gln
RMSD = 1.15
RMSD = 0.56

Figure A22. Substrate binding pose prediction and *X-ray* comparison. a) Overlay of a representative substrate-bound Inactive-OxH PRFAR-IGPS (in purple) structure extracted from the aMD simulations with the substrate-bound IGPS *X-ray* structure (PDB: 3ZR4, in cyan)**. b)** Overlay of a representative substrate-bound Inactive-OxH PRFAR-free IGPS (in green) structure extracted from the aMD simulations with the substrate-bound IGPS *X-ray* structure (PDB: 3ZR4, in cyan).

Accelerated Molecular Dynamics Simulations IGPS: ternary complex (Figures A23-A31)

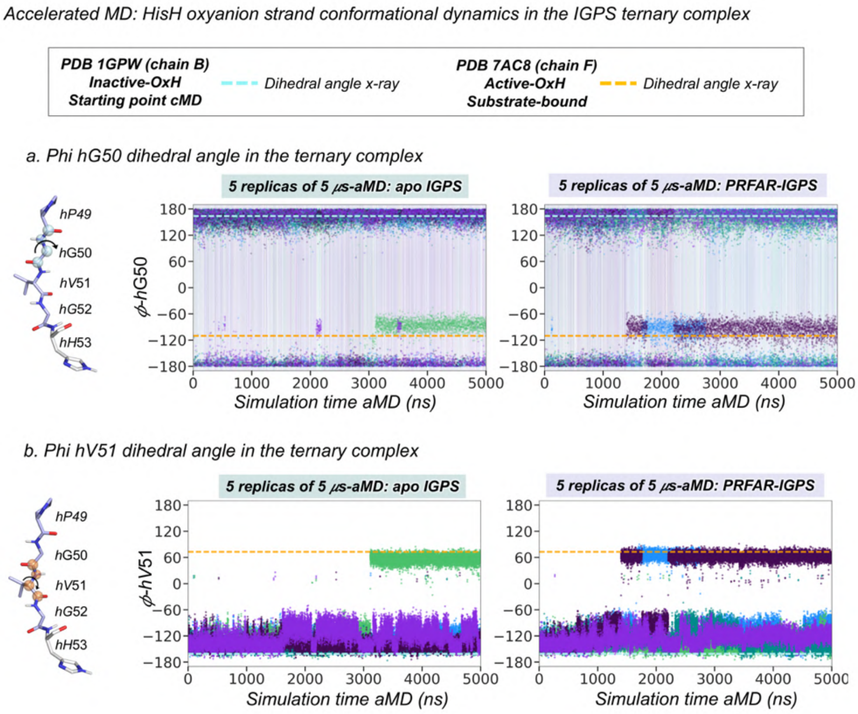


Figure A23. HisH *h*49-PVGV conformational dynamics in the IGPS ternary complex. Plot of the most relevant dihedral angles of the *h*49-PGVG oxyanion strand for five replicas of 5 μs accelerated MD simulations in the PRFAR-free IGPS and PRFAR-IGPS states. Each replica is depicted in a different color. Horizontal cyan dashed lines indicate the dihedral angle found in the *X-ray* structure (1GPW chain B) used as starting point for cMD simulations. Horizontal orange dashed lines indicate the dihedral angle found in the *X-ray* structure of *h*C84A IGPS (7AC8 chain F) that displays an active conformation of the oxyanion strand. (a) ϕ dihedral angle of *h*G50; (b) ϕ dihedral angle of *h*V51. See Figure A24 for a molecular representation of the most relevant states.

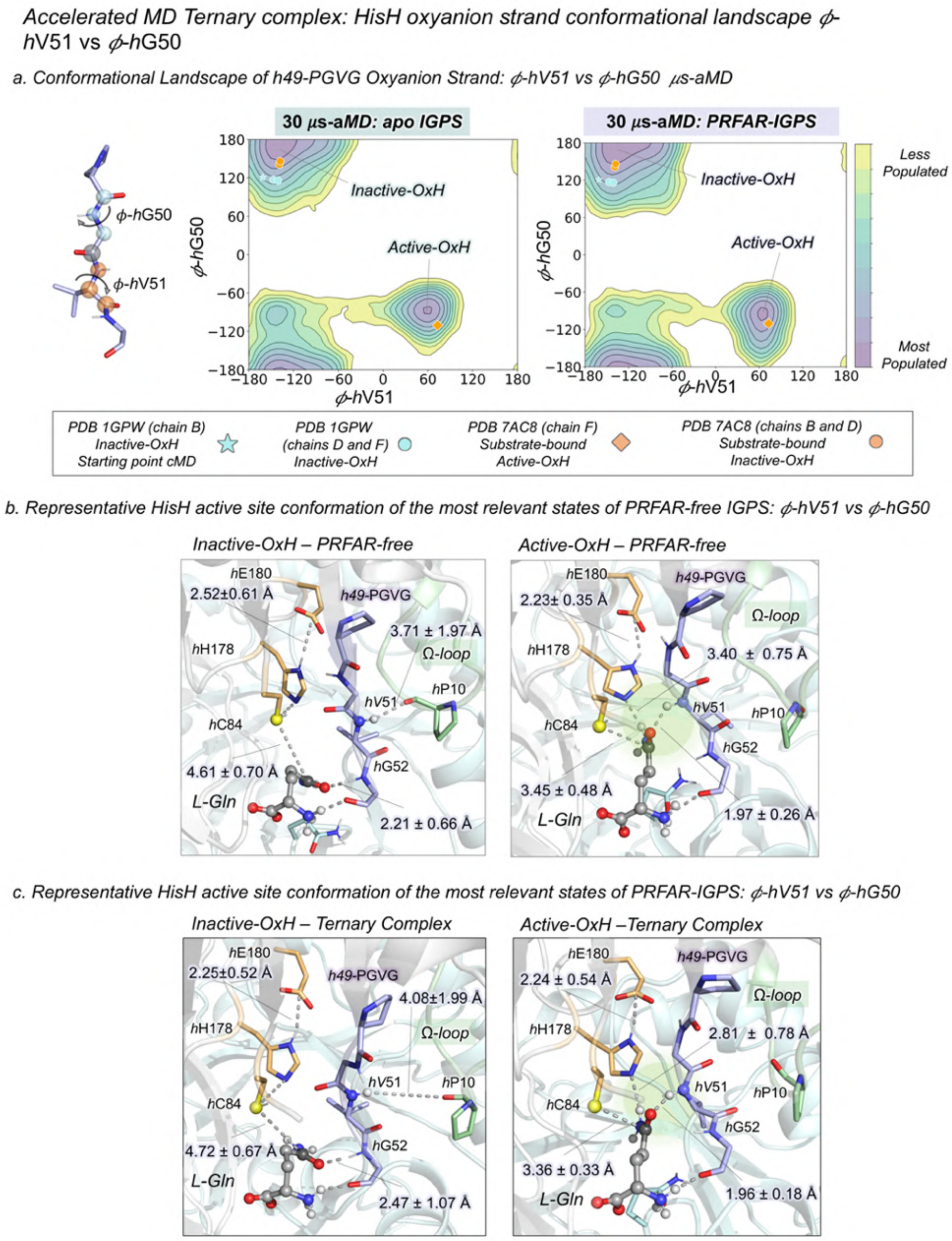
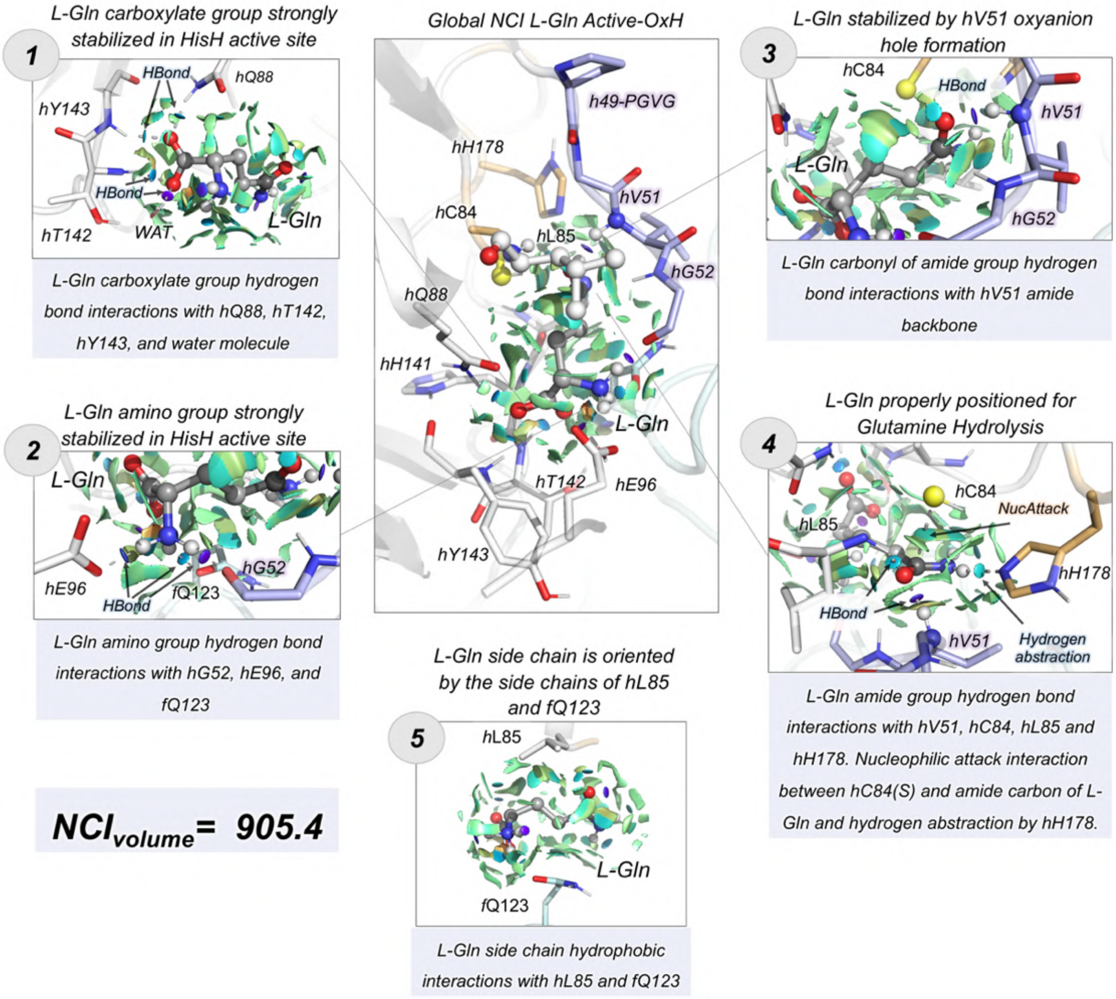


Figure A24. Conformational Landscape of *h*49-PGVG Oxyanion Strand in the ternary complex. The PRFAR-free IGPS and PRFAR-IGPS conformational landscapes are constructed from a total of 30 µs of aMD simulations in each case. **a)** Conformational landscape of *apo* and PRFAR-IGPS constructed using the ϕ dihedral angles of *h*V51 and *h*G50. The values of the ϕ dihedral angles of *h*V51 and *h*G50 corresponding to the three chains of PDB 1GPW are depicted in cyan and the three chains of PDB 7AC8 are represented in orange. The conformation used as starting point for cMD simulations is shown using the star symbol. The conformation corresponding to the active oxyanion strand observed in *h*C84A IGPS is depicted using the diamond symbol. **b)** Representative HisH active site structures of most populated states in PRFAR-free IGPS conformational landscape. **c)** Representative HisH active site structures of most populated states in PRFAR-IGPS conformational

landscape. The HisH catalytic residues are highlighted in orange, Ω-loop residues in green, and the residues of the *h*49-PGVG oxyanion strand in purple. Other relevant HisF and HisH residues are shown in cyan and white respectively. The atoms of *L-Gln* are shown as spheres.

Accelerated MD: Non-covalent interactions (NCI) in the HisH active site of the Ternary Complex

a. NCI and NCI volumes for the Active-OxH ternary complex



L-Gln carboxylate group strongly stabilized in HisH active site

L-Gln carboxylate group hydrogen bond interactions with hQ88, hT142, hY143, and water molecule

Global NCI L-Gln Active-OxH

L-Gln stabilized by hV51 oxyanion hole formation

L-Gln carbonyl of amide group hydrogen bond interactions with hV51 amide backbone

L-Gln amino group strongly stabilized in HisH active site

L-Gln amino group hydrogen bond interactions with hG52, hE96, and fQ123

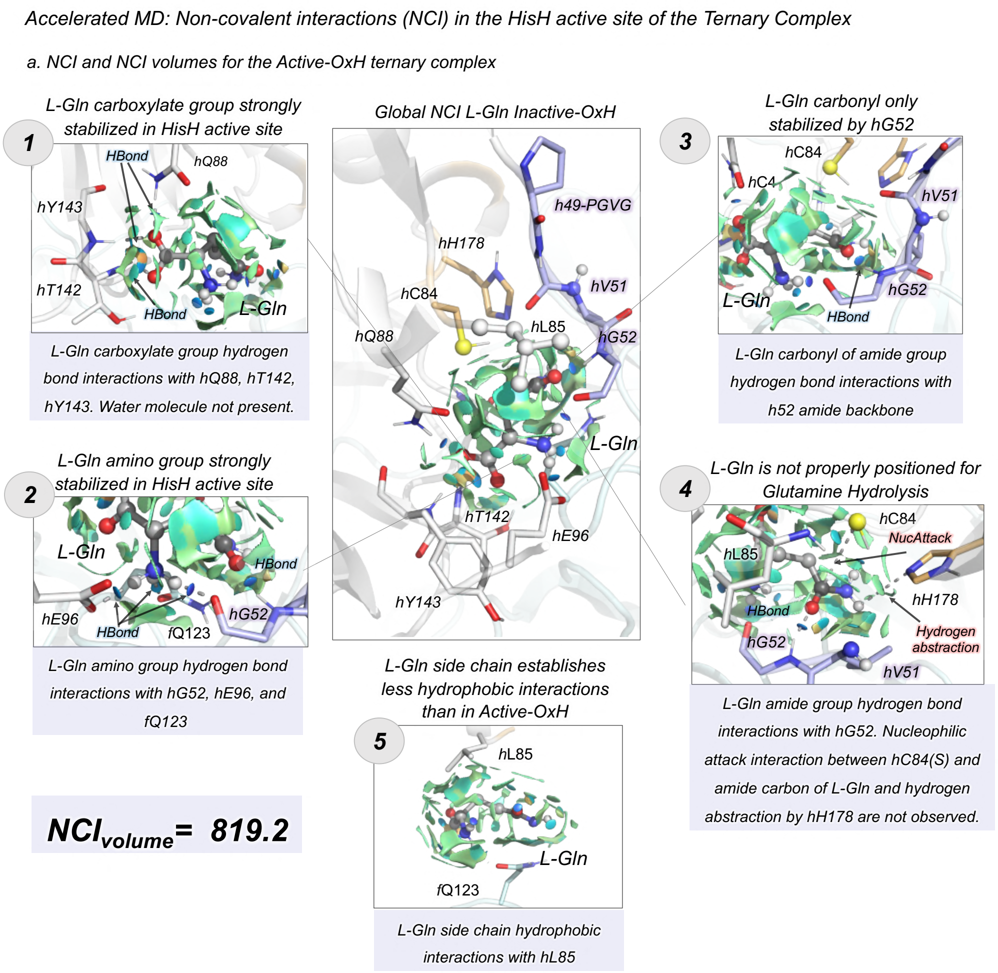
L-Gln properly positioned for Glutamine Hydrolysis

L-Gln amide group hydrogen bond interactions with hV51, hC84, hL85 and hH178. Nucleophilic attack interaction between hC84(S) and amide carbon of L-Gln and hydrogen abstraction by hH178.

L-Gln side chain is oriented by the side chains of hL85 and fQ123

L-Gln side chain hydrophobic interactions with hL85 and fQ123

NCI$_{volume}$= 905.4


Accelerated MD: Non-covalent interactions (NCI) in the HisH active site of the Ternary Complex

a. NCI and NCI volumes for the Active-OxH ternary complex

1 L-Gln carboxylate group strongly stabilized in HisH active site

L-Gln carboxylate group hydrogen bond interactions with hQ88, hT142, hY143. Water molecule not present.

2 L-Gln amino group strongly stabilized in HisH active site

L-Gln amino group hydrogen bond interactions with hG52, hE96, and fQ123

Global NCI L-Gln Inactive-OxH

3 L-Gln carbonyl only stabilized by hG52

L-Gln carbonyl of amide group hydrogen bond interactions with h52 amide backbone

4 L-Gln is not properly positioned for Glutamine Hydrolysis

L-Gln amide group hydrogen bond interactions with hG52. Nucleophilic attack interaction between hC84(S) and amide carbon of L-Gln and hydrogen abstraction by hH178 are not observed.

5 L-Gln side chain establishes less hydrophobic interactions than in Active-OxH

L-Gln side chain hydrophobic interactions with hL85

$NCI_{volume} = 819.2$

Figure A25. Non-covalent interactions (NCI) in the HisH active site of the IGPS Ternary Complex. Schematic representation of non-covalent interactions for the *L-Gln* bound to Active-OxH (a) and Inactive-OxH (b) states of PRFAR-IGPS calculated with the NCI plot.[9] Blue, green, and red surfaces indicate strong, weak, and repulsive non-covalent interactions, respectively. The integrated volumes of non-covalent interactions are provided for each step. Higher volumes indicate stronger NCI interactions.

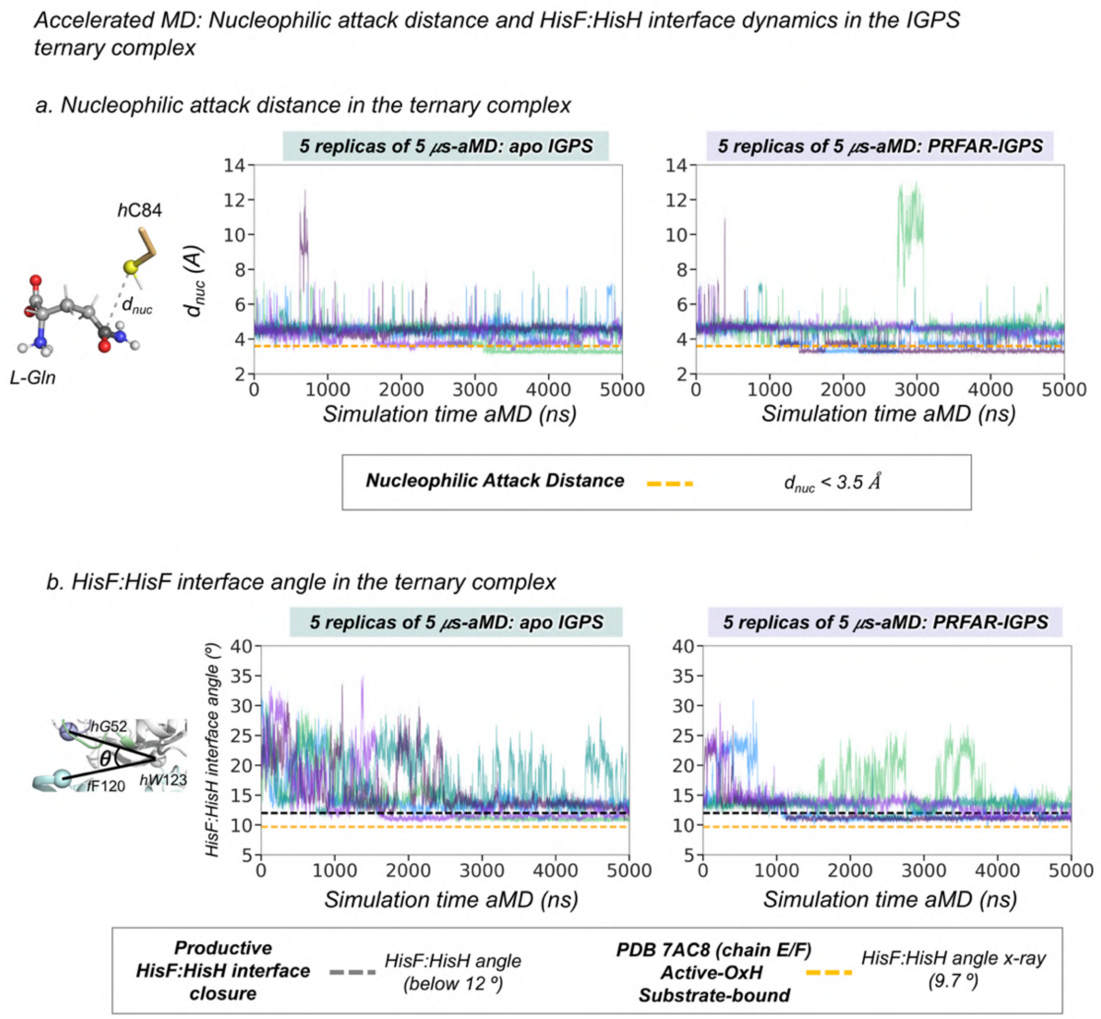


Figure A26. HisH *h*49-PVGV conformational dynamics in the IGPS ternary complex. Plot of the nucleophilic attack distance (a) and HisF:HisH interface angle (b) for five replicas of 5 μs aMD simulations in the PRFAR-free IGPS and PRFAR-IGPS states. Each replica is depicted in a different color. (a) Horizontal orange dashed line indicate that indicate nucleophilic attack distance at catalytic distance. (b) Horizontal black dashed lines indicate the HisF:HisH angle is below 12º indicative of productive interface closure. Horizontal orange dashed lines indicate the HisF:HisH interface angle found in the *X-ray* structure of *h*C84A IGPS (7AC8 chains E/F) that displays an active conformation of the oxyanion strand.

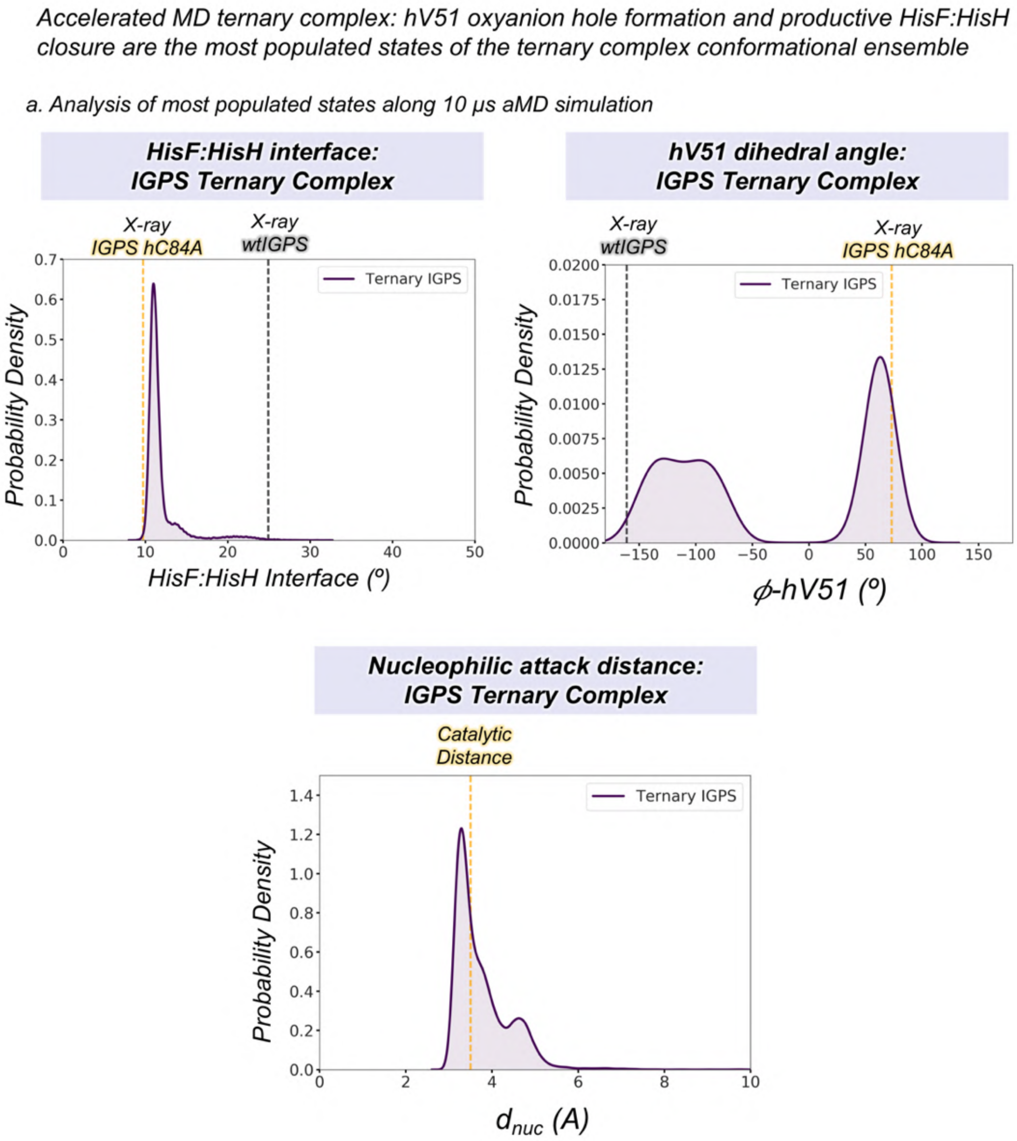


Figure A27. Conformational ensemble of IGPS Ternary Complex. Probability density distribution for the HisF:HisH interface angle, ϕ dihedral angle of *h*V51, and nucleophilic attack distance between the amide carbon of *L-Gln* and the sulfur of the side chain of *h*C84. Vertical gray dashed lines indicate the HisF:HisH interface angle and ϕ dihedral angle of *h*V51 found in the *X-ray* structure (PDB 1GPW chains A/B) used as starting point for cMD simulations. Vertical orange dashed lines indicate the HisF:HisH interface angle and ϕ-*h*V51 found in the *X-ray* structure of substrate-bound *h*C84A IGPS (PDB 7AC8 chains E/F) that displays an active conformation of the oxyanion strand. Vertical orange line indicates the catalytically productive distance (3.5 Å).

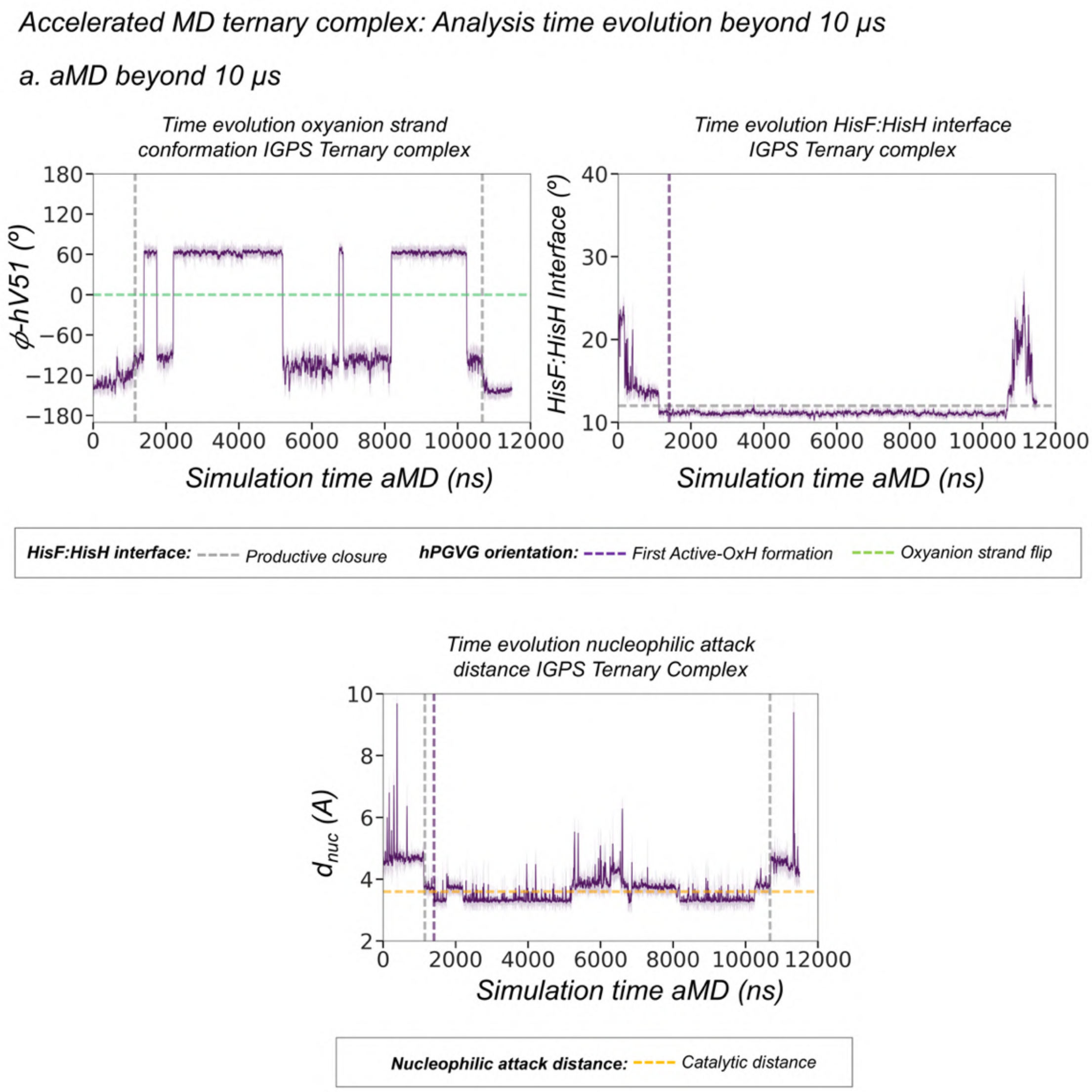


Figure A28. Ternary complex conformational dynamics beyond 10 μs. a) Plot of the HisF:HisH interface angle along 11.5 μs-aMD simulations. Plot of the *h*V51 dihedral angle along the 11.5 μs-aMD simulations. Plot of the distance corresponding to the nucleophilic attack along the 15 μs-aMD simulations. Gray dashed line indicates the range of HisF:HisH productive closure. Purple dashed line indicates the moment when the first *h*V51 oxyanion hole formation occurs. Green dashed line indicates the oxyanion hole transition, whereas orange dashed line indicate catalytic distance.

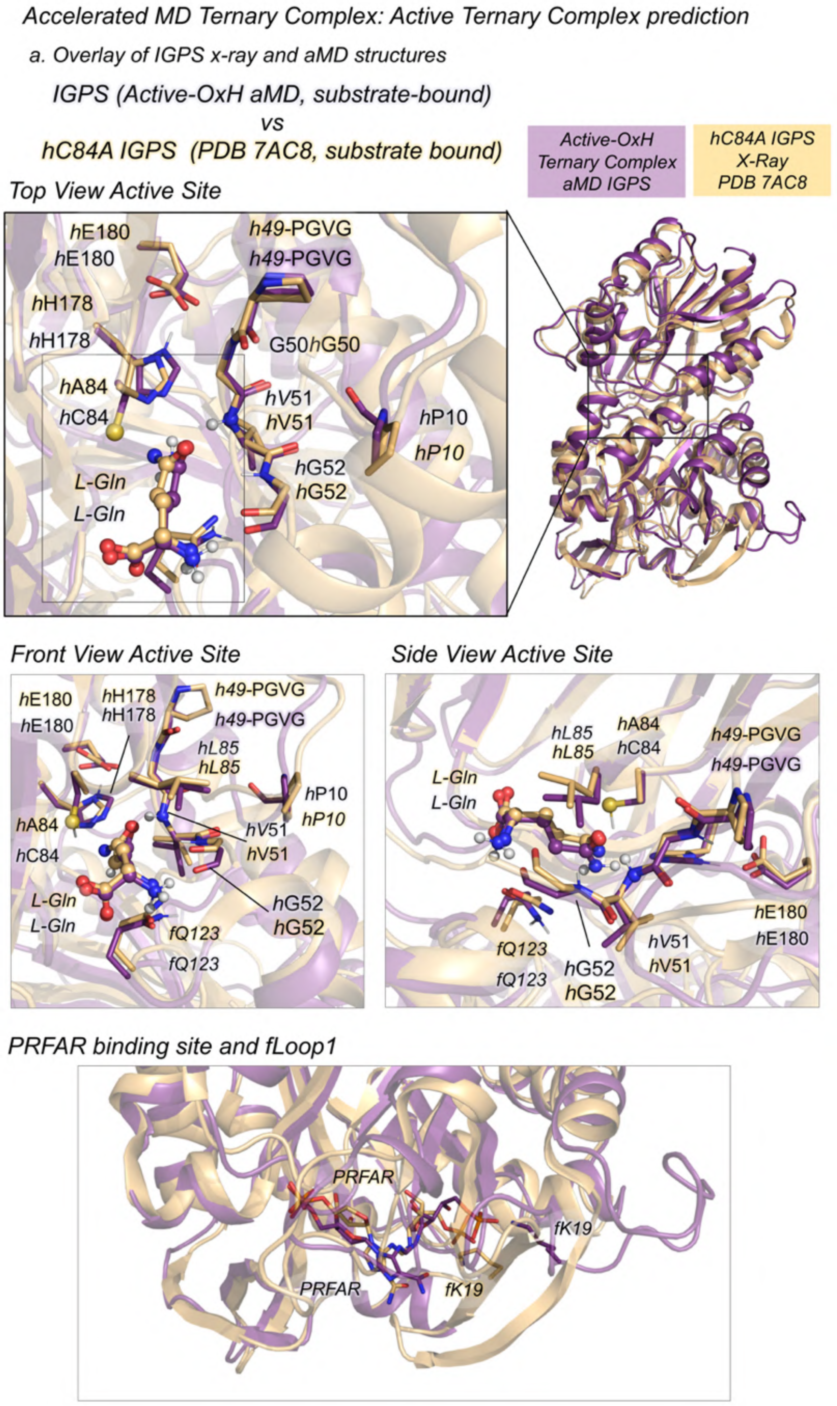


Figure A29. Active ternary complex predicted from aMD simulations. a) Overlay of a representative substrate-bound Active-OxH PRFAR-IGPS (in purple) structure extracted from the PRFAR-IGPS conformational landscape with the substrate-bound *h*C84A IGPS (PDB: 7AC8 (chain F), in orange) from different views.

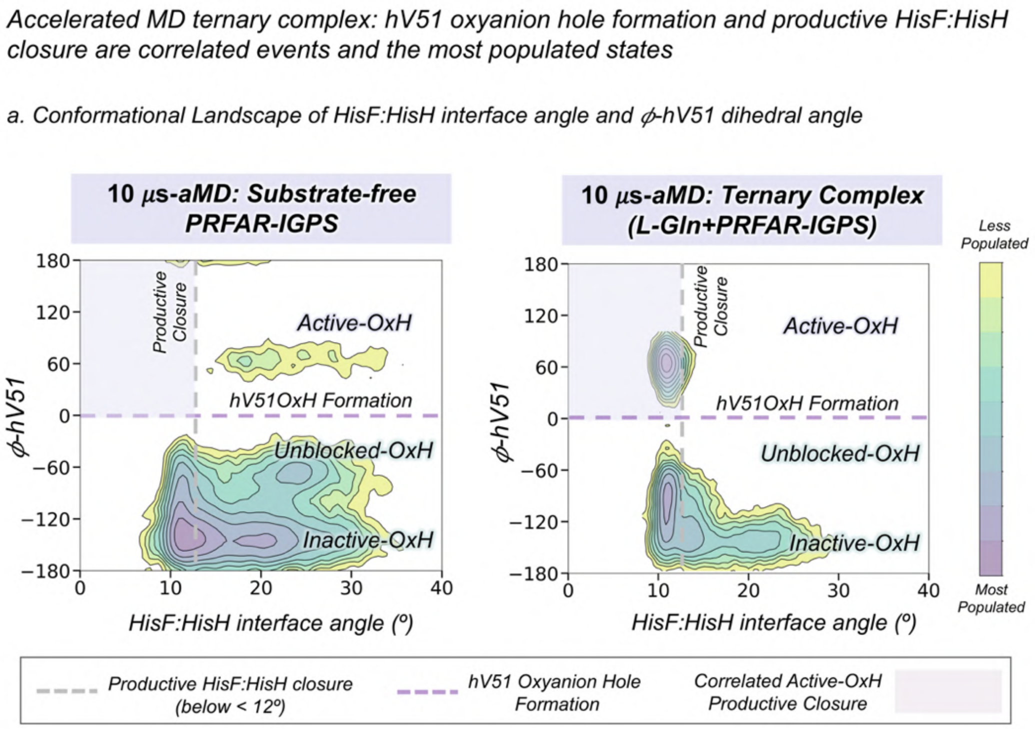


Figure A30. Correlated HisF:HisH interface and oxyanion-strand conformational dynamics in aMD simulations of the IGPS ternary complex. Conformational landscape constructed using the HisF:HisH interface angle and ϕ dihedral angle of *h*V51 obtained from accelerated Molecular Dynamics (aMD) simulations of substrate-free PRFAR-IGPS and ternary complex (*L-Gln*+PRFAR+IGPS). The purple area in the plot indicates the region of the conformational landscape with a productively closed HisF:HisH interface and a *h*V51 oxyanion hole formed. In the case of the IGPS ternary complex both events are coupled.

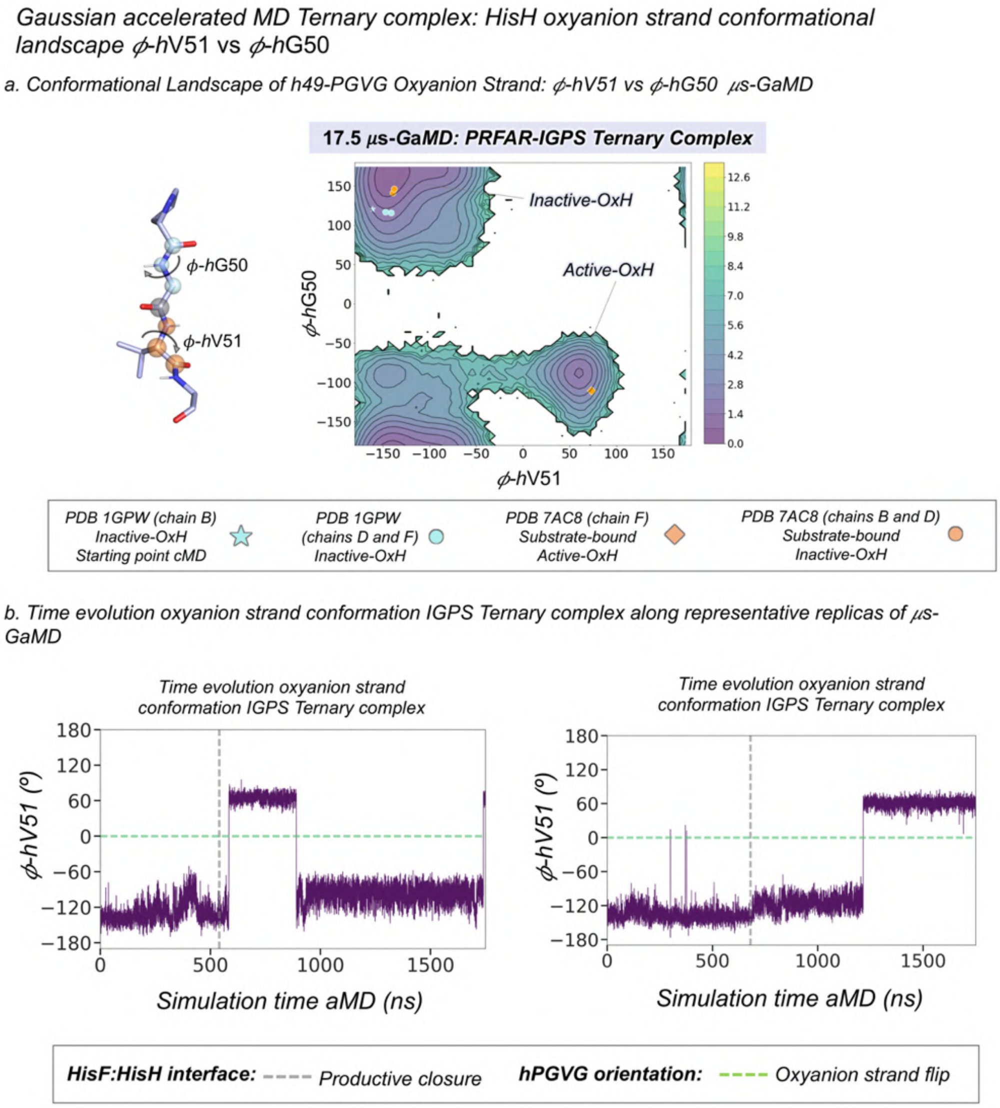


Figure A31. Gaussian Accelerated Molecular Dynamics in the IGPS ternary complex. a) The PRFAR-IGPS free energy landscape (FEL) is constructed from a total of 17.5 μs of GaMD simulations from ten independent replicas of 1.75 μs as described in the methods section. The FEL show similar relative stabilities for active-OxH and inactive-OxH. **b)** The formation of the allosteric active state (*h*V51 oxyanion hole formed and HisF:HisH interface closed) occurs in two out of ten replicas. The oxyanion hole formation takes place more than one time in a single GaMD simulation indicating dynamic equilibrium between the different states of the oxyanion strand. Plot of the *h*V51 dihedral angle along the 1.75 μs-GaMD simulations for two representative replicas. Gray dashed line indicates the range of HisF:HisH productive closure. Green dashed line indicates the oxyanion hole transition.

Metadynamics Simulations IGPS: ternary complex (Figure A32-A34)

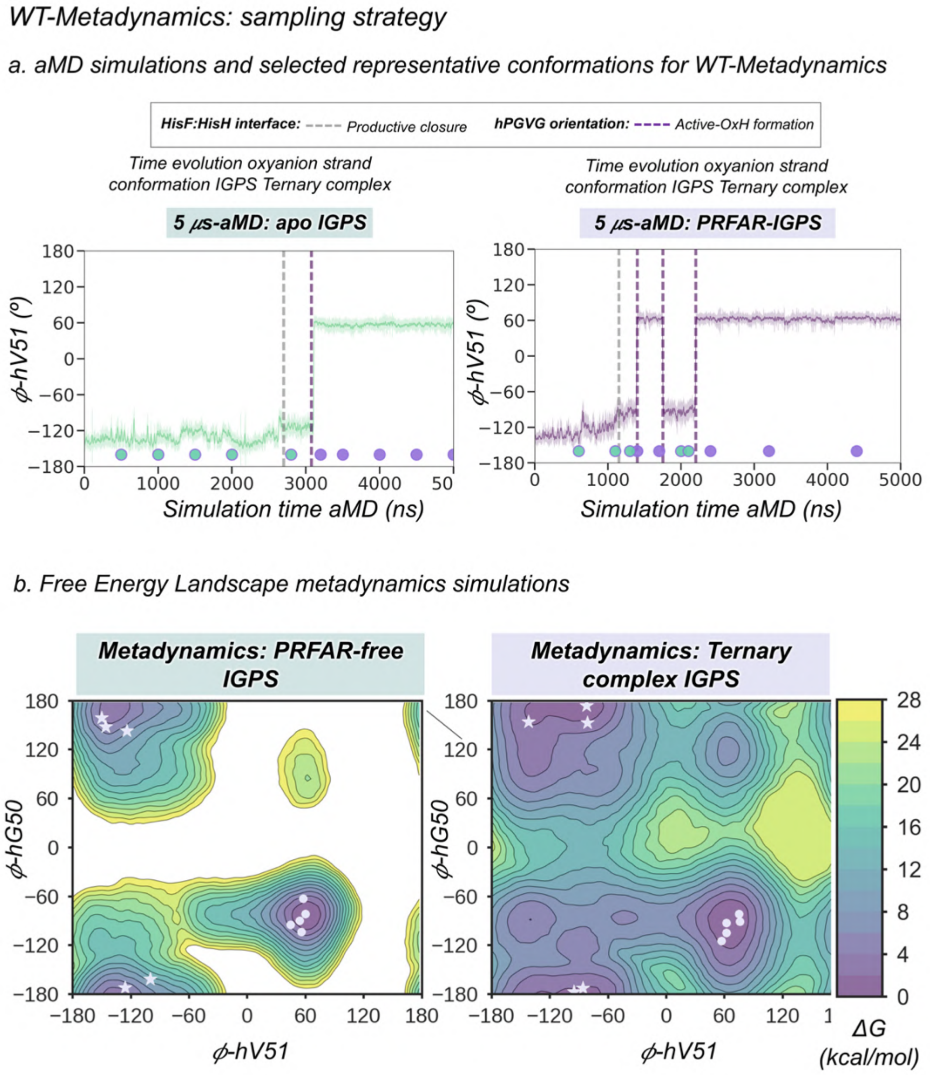


Figure A32. WT-Metadynamics sampling strategy. a) Plot of the *h*V51 dihedral angle along the 5 μs-aMD simulations in PRFAR-free (green) and PRFAR-IGPS (purple). The five representative aMD structures obtained from the Inactive-OxH and Active-OxH states used as starting points for the WT-metadynamics simulations are shown as green and purple circles, respectively. **b)** Free energy landscape of the *h*49-PGVG in the *apo* and PRFAR-IGPS states obtained from WT-tempered metadynamics simulations. Stars indicate the coordinates of the five starting points corresponding to Inactive-OxH walker replicas used for the WT-metadynamics while circles indicate the coordinates of the five starting points corresponding to the Active-OxH walker replicas. Note that the green and purple circles shown in (a) corresponds to the stars and circles shown in (b), respectively.

The FEL obtained from metadynamics simulations show remarkable differences in the PRFAR-free and PRFAR bound states. In the PRFAR bound state, the formation of the oxyanion hole presents a surmountable energy barrier of 8 kcal/mol while in the PRFAR-free state this value rises to 22 kcal/mol. These results are in line with experimental k$_{cat}$ values. Further, the relative stability of the inactive and active forms is maintained which indicates that both states are accessible in PRFAR and may be important for enzyme catalysis (binding and chemical step). These transitions take place in the IGPS closed state without changes in the HisF:HisH interface. Most importantly, while in the presence of PRFAR the oxyanion hole can easily arrange and disarrange in the closed state, the oxyanion hole cannot form in the PRFAR-free state hampering the catalytic activity.

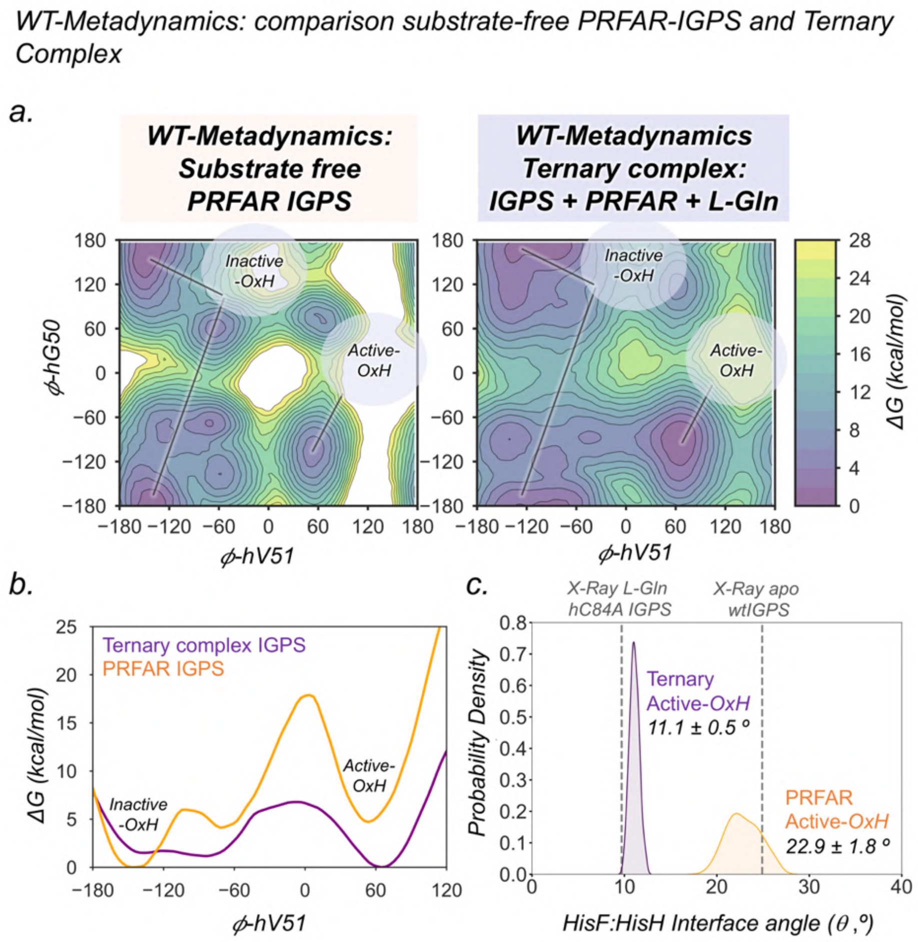


Figure A33. Communication between PRFAR and *L-Gln* triggers the allosteric activation of IGPS. a) Free energy landscape of the h49-PGVG oxyanion strand in PRFAR IGPS (only PRFAR bound) and ternary complex IGPS (PRFAR and *L-Gln* bound) obtained from well-tempered metadynamics (WT-MetaD) simulations. **b)** 2D free energy landscape representation focusing on the *ϕ*-*h*V51 dihedral angle. **c)** Probability density distribution of the HisF:HisH interface angle (in º) estimated in the Active-OxH conformations sampled in the WT-MetaD simulations. The average angle values are also shown. The angle (θ) of the HisF:HisH interface is calculated from the alpha-carbons of *f*F120, *h*W123 and *h*G52. The vertical dashed gray lines corresponds to the *h*C84A IGPS (PDB: 7AC8 (chains E/F)) and wtIGPS (PDB:1GPW (chains A/B)) *X-ray* HisF:HisH interface angles. Note that in the absence of *L-Gln*, the productive closure of the HisF:HisH interface is not sampled, the active-OxH state is destabilized *ca.* 4 kcal/mol and the energy barrier of the allosteric activation rises substantially.

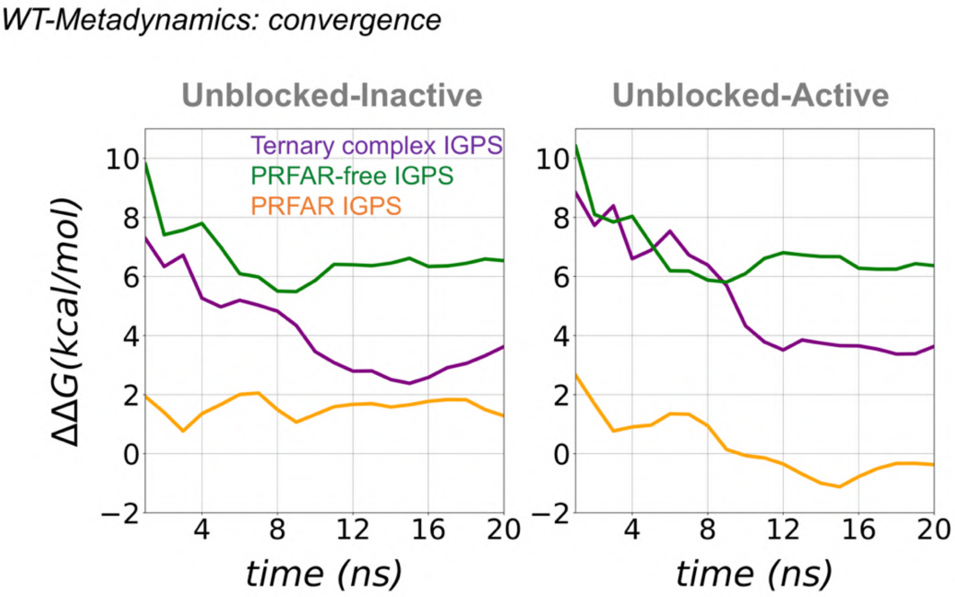


Figure A34. Estimate of the free energy differences between the selected regions of the free energy surface. The lines represent the mean ΔΔG value of the 10 walker replicas along the simulation time. The unblocked-inactive (on the left) and the unblocked-active (on the right) energy differences of ternary complex (PRFAR and *L-Gln* bound), PRFAR-free (only *L-Gln* bound) and PRFAR (only PRFAR bound) are depicted in purple, green and orange, respectively.

Dynamical-network analysis IGPS (Shortest-Path Map) and HisF conformational dynamics: ternary complex (Figures A35-A39)

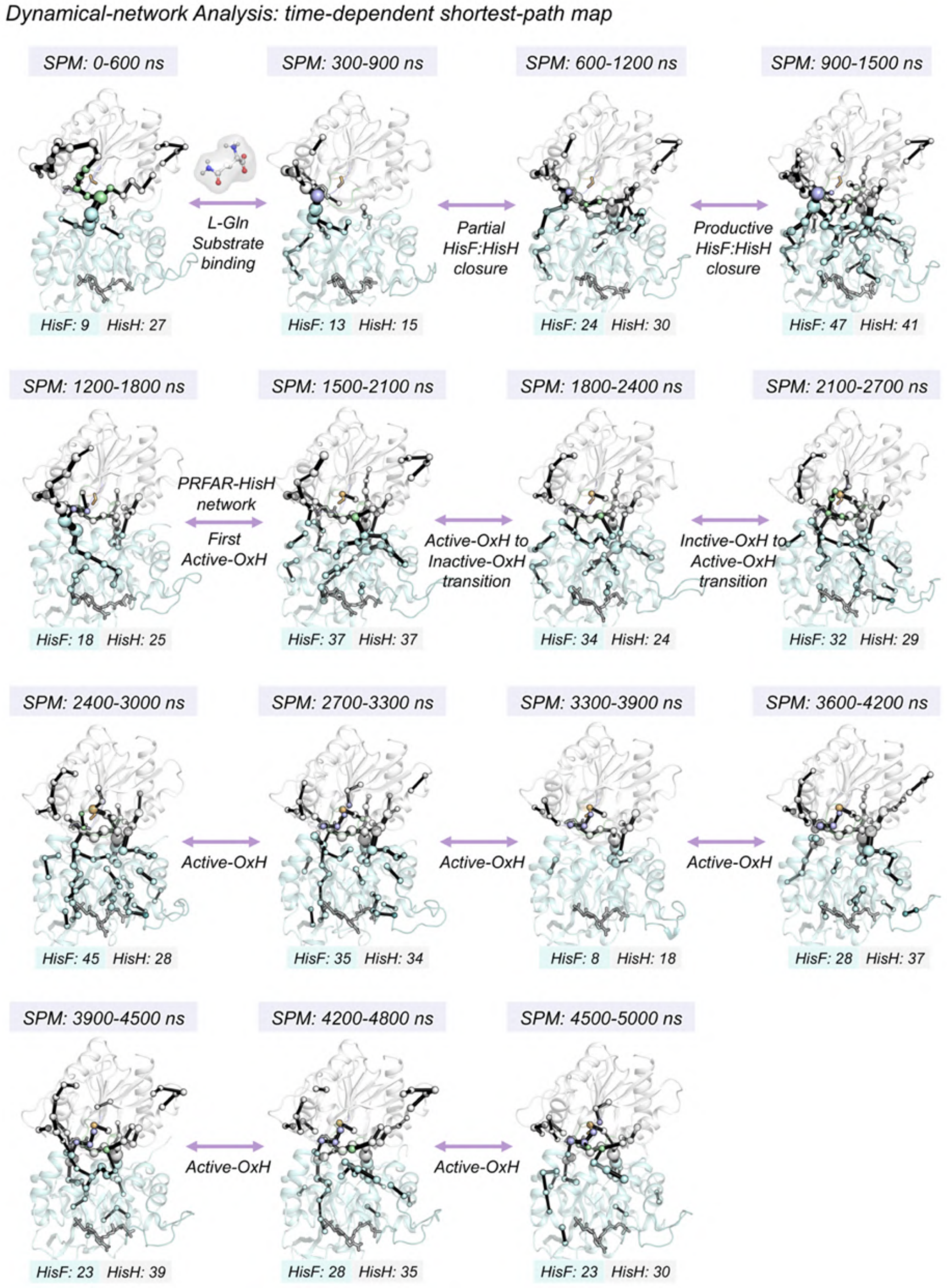
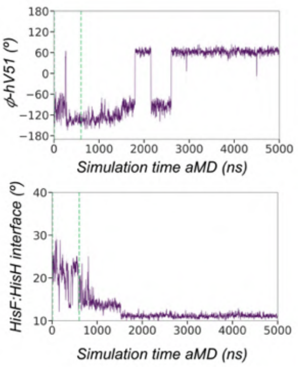


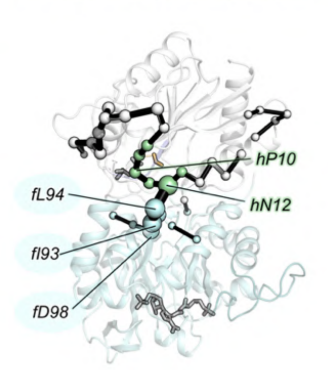
Figure A35. Dynamical Network Analysis: time-evolution Shortest-Path Map analysis. Identification of the amino acids that contribute to the propagation of the allosteric activation in IGPS. Residues belonging to HisF and HisH subunits are highlighted in cyan and white, respectively. oxyanion strand, Ω-loop, catalytic residues, and loop1 are colored in purple, green, orange, and teal, respectively.

Dynamical Network Analysis: Shortest Path Map

a. SPM 0-600 ns

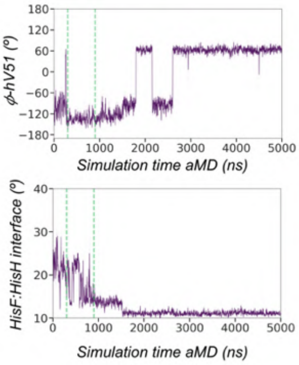


L-Gln binding 400 ns aMD

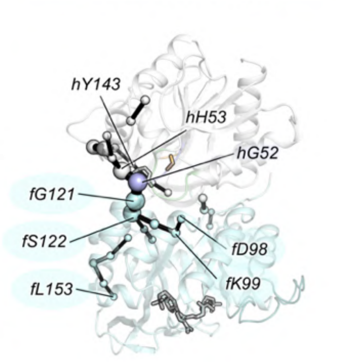


HisF (9/253)	**HisH (27/201)**		
fA3	fV8	hV30	hW123
fV66	fG9	hS31	
fI93	fP10	hE36	
fL94	fG11	hS37	
fA97	fN12	hG55	
fD98	fI13	hE56	
fK99	fH14	hR59	
fS122	fY17	hR62	
fA124	fR18	hE63	
	fK21	hN64	
	fS24	hE95	
	fF27	hE96	
	fG28	hA97	

b. SPM 300-900 ns

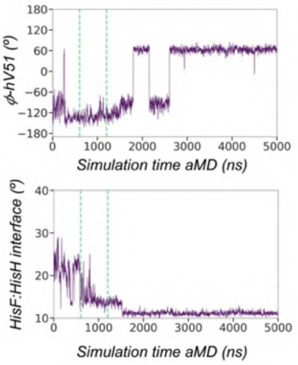


Partial Closure 900 ns

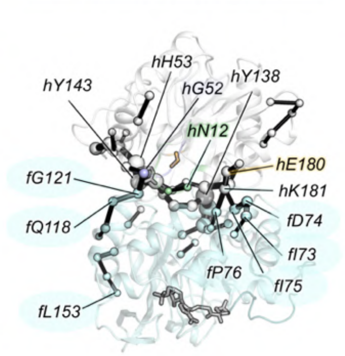


HisF (15/253)		**HisH (14/201)**	
fA3	fV246	hG52	fY143
fD98	fN247	hH53	
fK99		hG55	
fG121		hE56	
fS122		hR59	
fA124		hL61	
fL153		hL66	
fW156		hE95	
fE159		hA97	
fV160		fV111	
fA165		fW123	
fG166		fV140	
fE167		fV141	

c. SPM 600-1200 ns



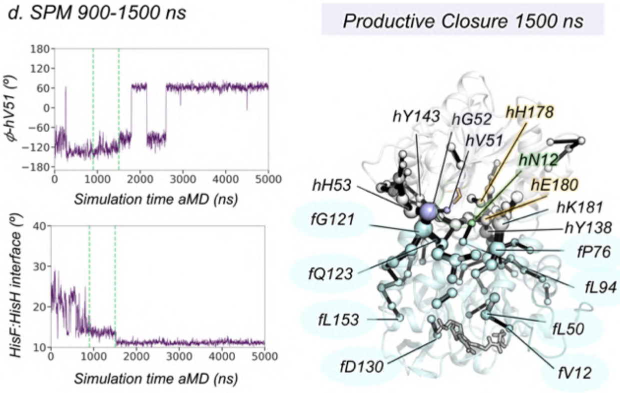
Partial Closure 900 ns



HisF (24/253)		**HisH (34/201)**		
fA3	fL153	fG11	fL61	fV140
fK4	fW156	fN12	fL66	fH141
fR5	fE159	fG19	fE95	fT142
fD45	fV160	fA23	fA97	fY143
fV69	fR163	fS24	fV110	fE180
fI73	fT195	fF27	fV111	fK181
fD74	fI199	fE28	fR117	fS182
fI75	fD219	fV30	fH120	fG186
fP76	fA220	fG52	fM121	
fF77	fV246	fH53	fG122	
fT114	fN247	fG55	fW123	
fQ118		fE56	fY138	
fG121		fR59	fF139	

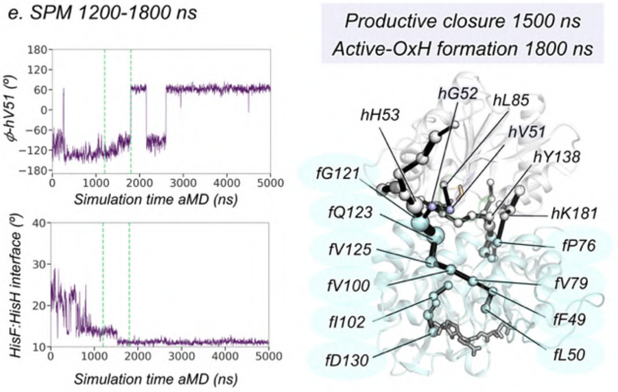
Dynamical Network Analysis: Shortest Path Map

d. SPM 900-1500 ns



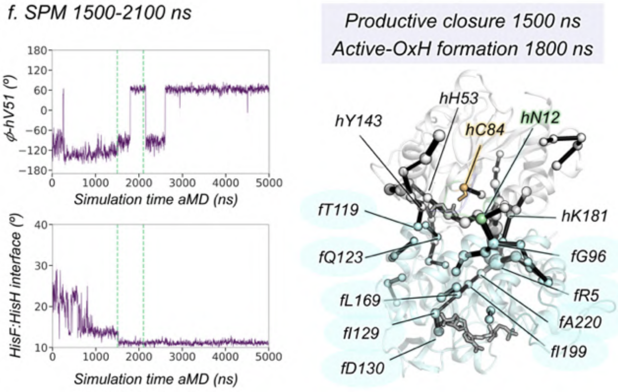
HisF (48/253)			HisH (41/201)		
fA3	fA97	fR163	hN12	hE95	hV140
fK4	fK99	fL193	hS24	hA97	hH141
fR5	fV100	fT194	hF27	hV111	hY143
fV12	fS101	fT195	hE28	hK112	hF177
fY39	fI102	fL196	hV30	hL113	hH178
fI44	fT114	fS201	hE36	hS115	hE180
fD45	fA117	fA218	hS37	hR117	hK181
fE46	fG121	fD219	hV51	hW123	hS182
fF49	fQ123	fA220	hG52	hN124	hG186
fL50	fA124	fA221	hH53	hE125	
fI73	fI129	fL222	hG55	hV126	
fI75	fD130	fD233	hE56	hI127	
fP76	fL153	fE236	hM58	hF128	
fF77	fW156	fL241	hR59	hY137	
fT78	fE159	fV246	hR62	hY138	
fL94	fV160	fN247	hE63	hF139	

e. SPM 1200-1800 ns



HisF (18/253)		HisH (25/201)	
fA3	fA128	hP38	hW123
fK4	fI129	hV51	hN124
fR5	fD130	hG52	hE125
fF49	fV246	hH53	hV126
fL50	fN247	hG55	hY137
fP76		hE56	hY138
fV79		hR59	hF139
fV100		hL61	hH141
fS101		hL66	hY143
fI102		hL85	hK181
fG121		hE95	hS182
fQ123		hA97	hG186
fV125		hV111	

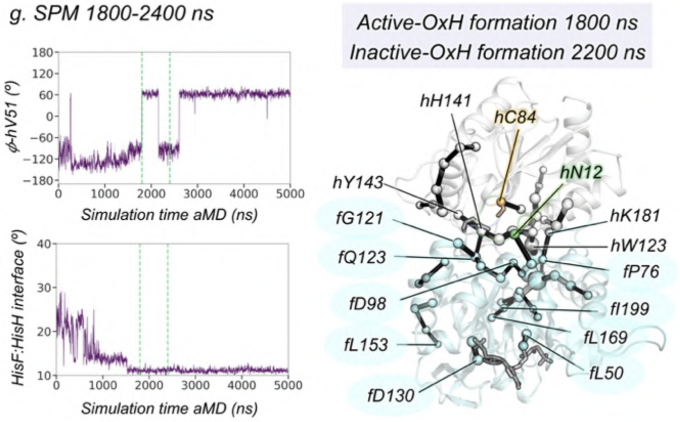
f. SPM 1500-2100 ns



HisF (37/253)			HisH (37/201)		
fA3	fG96	fW156	hN12	hC84	hF128
fK4	fA97	fE159	hM14	hE95	hY137
fR5	fK99	fV160	hS24	hA97	hY138
fV33	fV100	fR163	hF27	hV111	hF139
fG36	fS101	fE167	hE28	hK112	hV140
fY39	fI102	fL169	hV30	hL113	hH141
fI44	fT114	fL196	hS31	hS115	hY143
fD45	fA117	fI199	hH53	hR117	hF177
fL50	fT119	fA220	hG55	hW123	hK181
fD51	fG121	fV246	hE56	hN124	hS182
fV69	fQ123	fN247	hR59	hE125	hG186
fA70	fI129		hL61	hV126	
fP76	fD130		hL66	hI127	

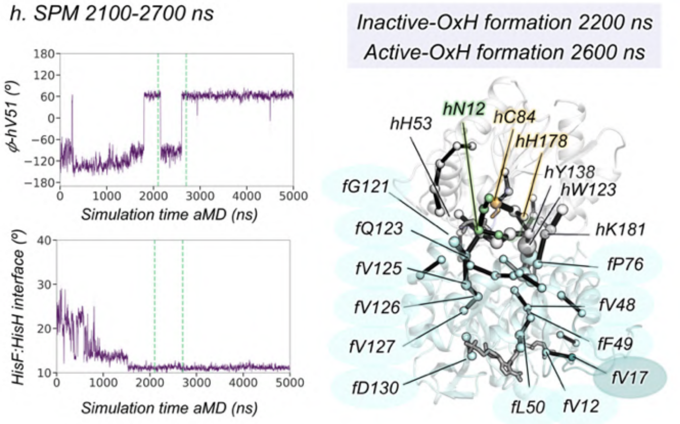
Dynamical Network Analysis: Shortest Path Map

g. SPM 1800-2400 ns



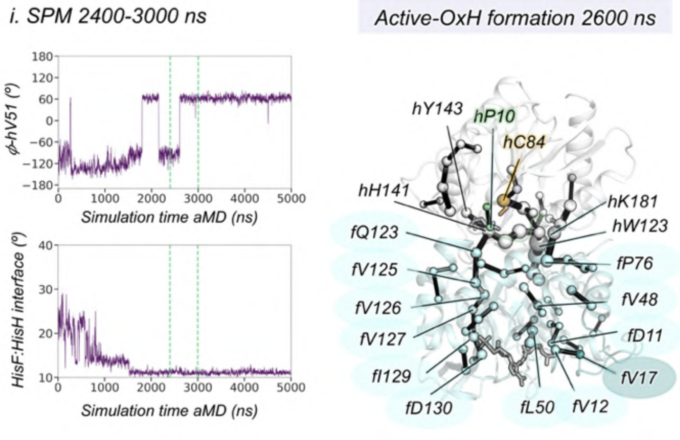
HisF (34/253)			HisH (24/201)	
fA3	fA97	fV160	hN12	hI127
fK4	fD98	fR163	hG55	hF128
fR5	fK99	fL169	hE56	hY137
fY39	fT114	fL199	hR59	hY138
fI44	fA117	fA218	hR62	hF139
fD45	fG121	fA220	hD65	hV140
fV48	fQ123	fV246	hE66	hH141
fF49	fA124	fN247	hC84	hY143
fL50	fI129		hA97	hK181
fD51	fD130		hW123	hS182
fP76	fL153		hN124	hG186
fT78	fW156		hE125	
fG96	fE159		hV126	

h. SPM 2100-2700 ns

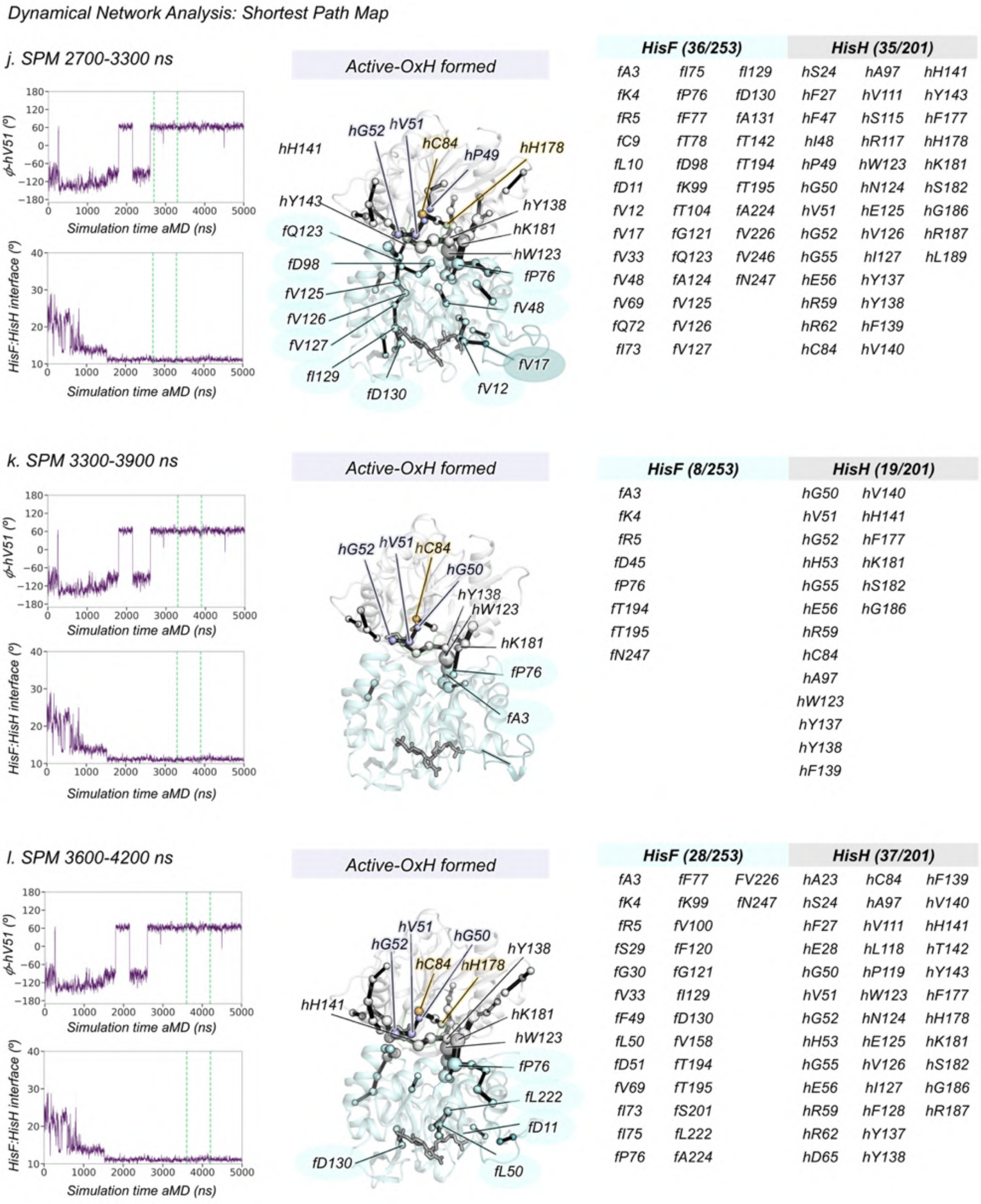


HisF (32/253)			HisH (29/201)		
fA3	fT78	fI129	hV8	hR117	hK181
fK4	fL94	fD130	hG9	hW123	hS182
fR5	fR95	fA224	hP10	hN124	hG186
fV12	fA97	fV226	hG11	hF125	
fV17	fK99	fV246	hE56	hV126	
fV33	fT114	fN247	hR59	hY137	
fV48	fA117		hR62	hY138	
fF49	fG121		hD65	hF139	
fL50	fQ123		hL66	hV140	
fD51	fA124		hG82	hH141	
fE67	fV125		hV83	hY143	
fV69	fV126		hC84	hF177	
fP76	fV127		hS115	hH178	

i. SPM 2400-3000 ns



HisF (45/253)			HisH (28/201)	
fA3	fQ72	fV125	hG9	hY137
fK4	fI73	fV126	hP10	hY138
fR5	fI75	fV127	hG55	hF139
fI7	fP76	fI129	hE56	hV140
fA8	fT78	fD130	hR59	hH141
fC9	fA97	fA131	hR62	hY143
fL10	fK99	fT142	hD65	hF177
fD11	fS101	fS201	hL66	hH178
fV12	fI102	fL222	hG82	hK181
fV17	fN103	fA224	hV83	hS182
fV33	fT104	fV226	hC84	hG186
fV48	fL112	fI232	hA97	hL189
fF49	fQ115	fD233	hW123	
fL50	fI116		hN124	
fD41	fQ123		hE125	
fV69	fA124		hV126	



HisF (36/253)			HisH (35/201)		
fA3	fI75	fI129	hS24	hA97	hH141
fK4	fP76	fD130	hF27	hV111	hY143
fR5	fF77	fA131	hF47	hS115	hF177
fC9	fT78	fT142	hI48	hR117	hH178
fL10	fD98	fT194	hP49	hW123	hK181
fD11	fK99	fT195	hG50	hN124	hS182
fV12	fT104	fA224	hV51	hE125	hG186
fV17	fG121	fV226	hG52	hV126	hR187
fV33	fQ123	fV246	hG55	hI127	hL189
fV48	fA124	fN247	hE56	hY137	
fV69	fV125		hR59	hY138	
fQ72	fV126		hR62	hF139	
fI73	fV127		hC84	hV140	

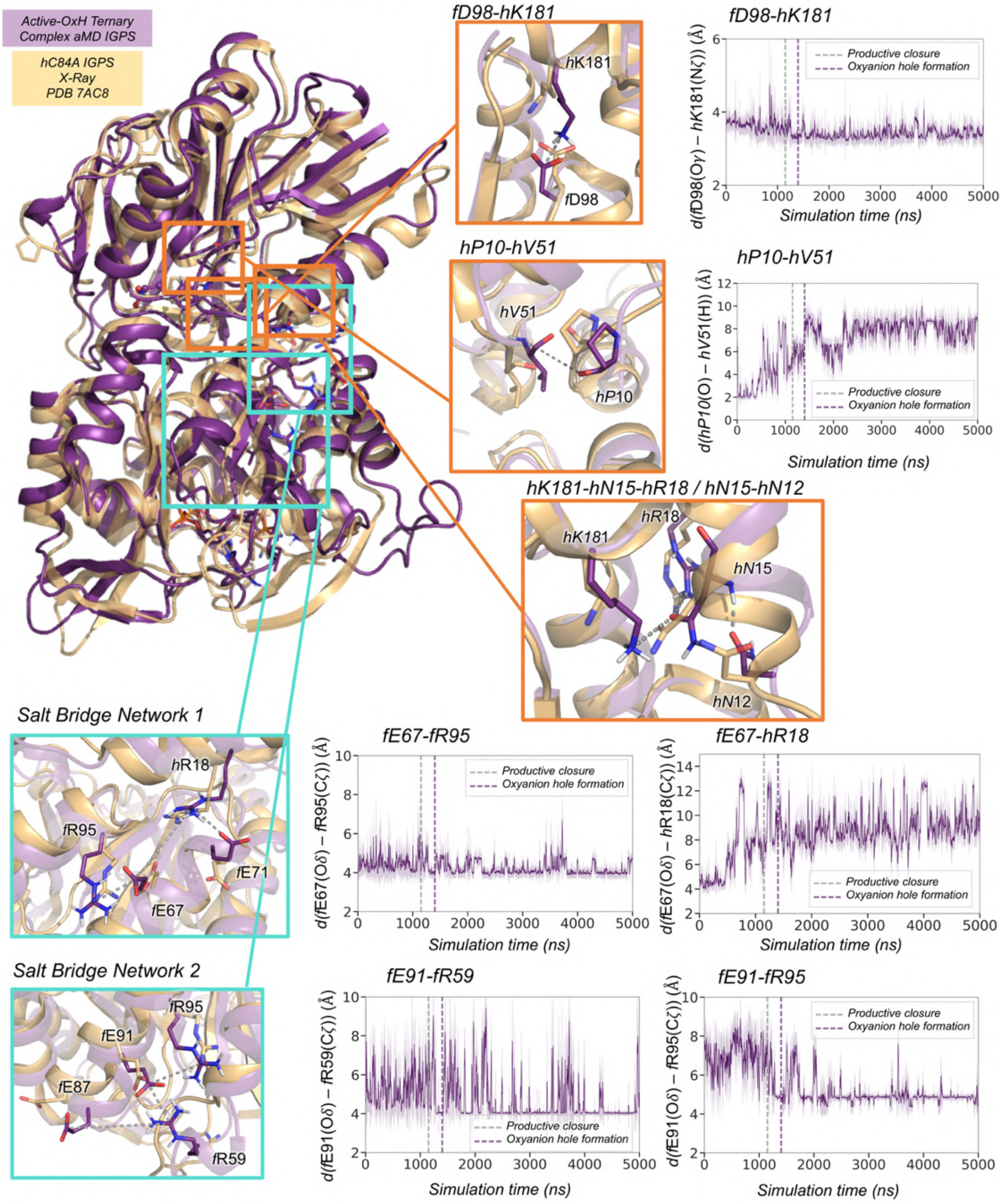
HisF (8/253)	HisH (19/201)	
fA3	hG50	hV140
fK4	hV51	hH141
fR5	hG52	hF177
fD45	hH53	hK181
fP76	hG55	hS182
fT194	hE56	hG186
fT195	hR59	
fN247	hC84	
	hA97	
	hW123	
	hY137	
	hY138	
	hF139	

HisF (28/253)			HisH (37/201)		
fA3	fF77	FV226	hA23	hC84	hF139
fK4	fK99	fN247	hS24	hA97	hV140
fR5	fV100		hF27	hV111	hH141
fS29	fF120		hE28	hL118	hT142
fG30	fG121		hG50	hP119	hY143
fV33	fI129		hV51	hW123	hF177
fF49	fD130		hG52	hN124	hH178
fL50	fV158		hH53	hE125	hK181
fD51	fT194		hG55	hV126	hS182
fV69	fT195		hE56	hI127	hG186
fI73	fS201		hR59	hF128	hR187
fI75	fL222		hR62	hY137	
fP76	fA224		hD65	hY138	

Figure A36. Detailed time-evolution Shortest-Path Map (SPM) analysis along a representative 5-μs aMD trajectory. Plot of the HisF:HisH interface angle along the simulation time. Plot of the ϕ dihedral angle of *h*V51. Vertical dashed green lines indicate the range of simulation time where the SPM was calculated. Structural representation of the PRFAR-IGPS SPM analysis with the most relevant residues highlighted. Residues belonging to HisF and HisH subunits are highlighted in cyan and white, respectively. oxyanion strand, Ω-loop, catalytic residues, and Loop1 are colored in purple, green, orange, and teal, respectively. List of all residues included in the SPM.

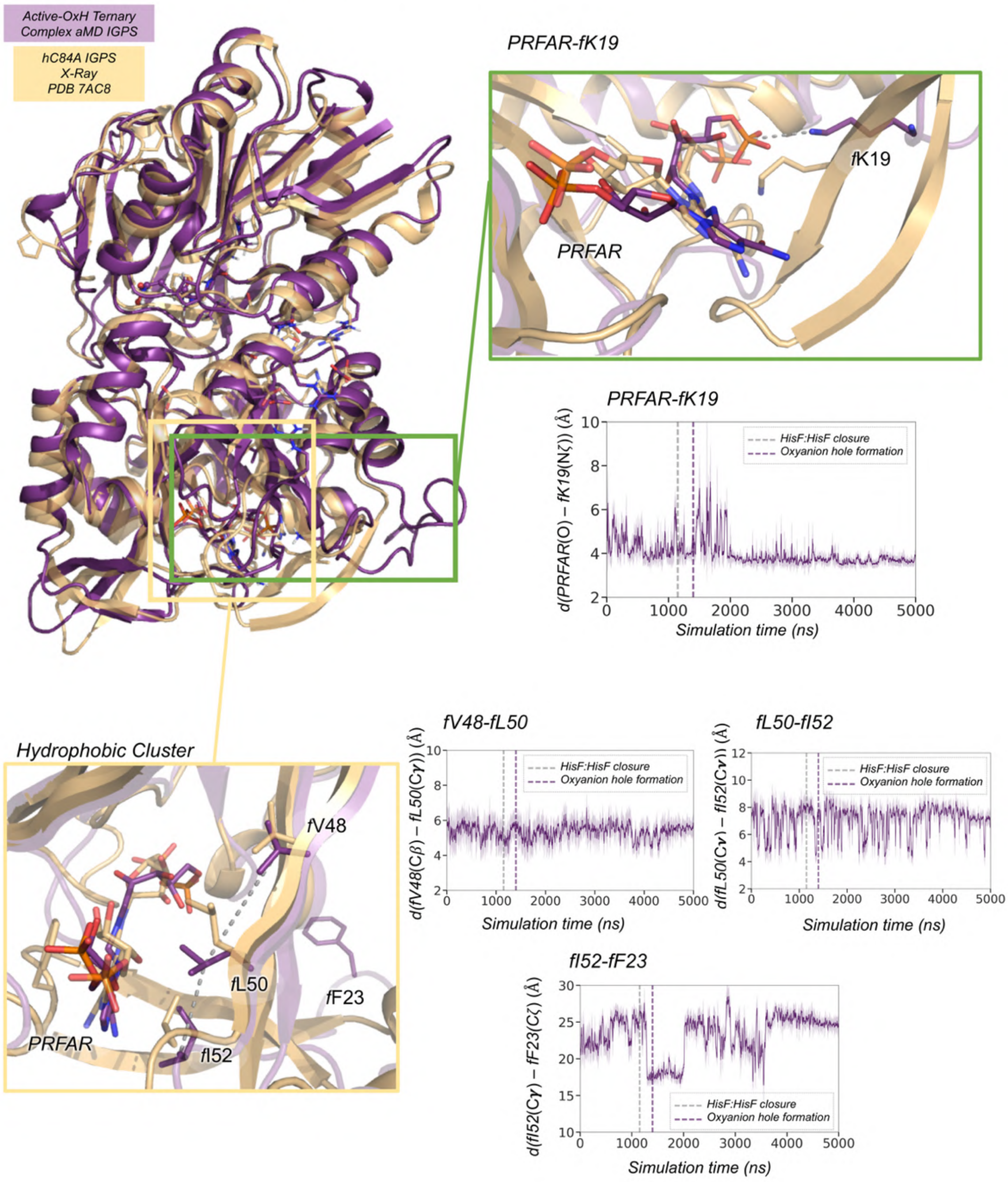
Accelerated MD ternary complex: HisF conformational dynamics along allosteric activation

a. HisF:HisH interface and HisF salt bridge network interactions



Accelerated MD: HisF conformational dynamics (continuation)

b. HisF hydrophobic cluster and fK19-PRFAR interactions



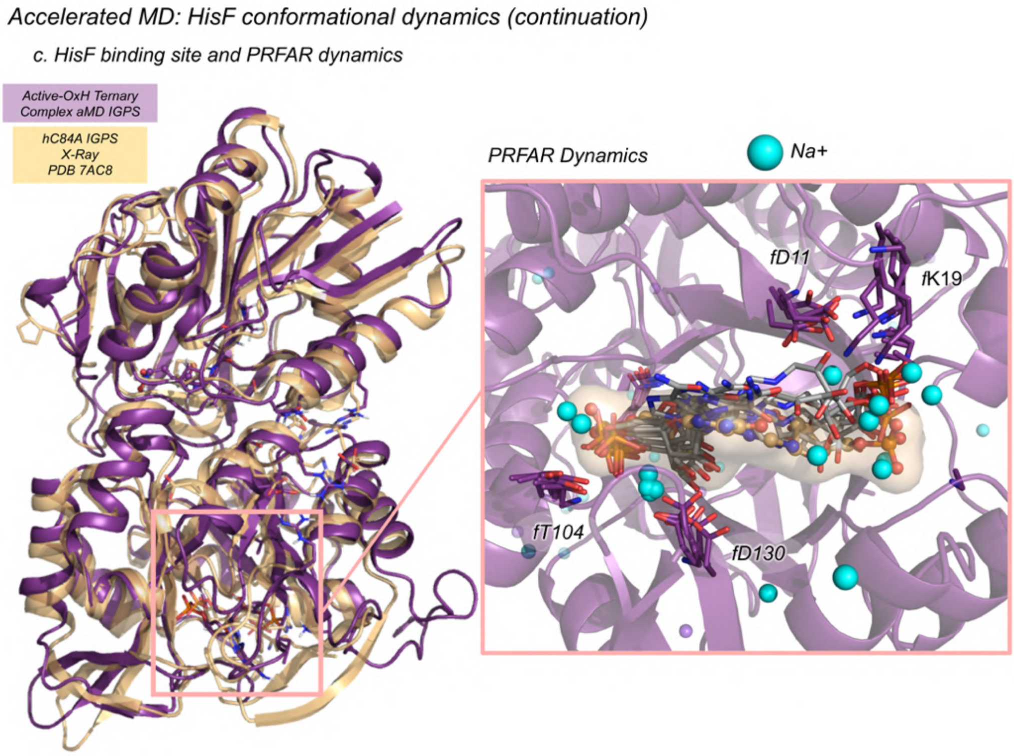

Accelerated MD: HisF conformational dynamics (continuation)
c. HisF binding site and PRFAR dynamics
Active-OxH Ternary Complex aMD IGPS
hC84A IGPS X-Ray PDB 7AC8
PRFAR Dynamics
Na+
fD11
fK19
fT104
fD130

Figure A37. HisF conformational dynamics in aMD simulations in the Ternary complex. Analysis of the most relevant interactions in HisF subunit in the ternary complex along the aMD simulations. Computationally predicted active state in purple and *h*C84A IGPS structure in orange. The selected distances were described by Rivalta et al.[8] and shown to be relevant for analyzing the effects of PRFAR. Overlay of the side chains of representative conformations in the computational prediction and *X-ray* structure are shown in purple and orange, respectively. **a)** Molecular representation of HisF salt bridge network (highlighted in cyan squares) and *f*D98-*h*K181 HisF:HisH interaction (highlighted in an orange square). Plot of the most relevant distances of the salt bridge network, *h*P10-*h*V51 and *f*D98-*h*K181 distance along the 5 μs aMD simulations where the allosteric activation occurs. The distances of the salt bridge network are calculated between the carbon atom of the carboxylate group of the glutamate side chain and the carbon atom of the guanidinium group of the arginine residues. The distance of the *f*D98-*h*K181 interaction is calculated between the carbon atom of the carboxylate group of *f*D98 side chain and the nitrogen of the side chain of *h*K181. The distance of the *h*P10-*h*V51 interaction is calculated between the carbonyl oxygen of *h*P10 and the amide hydrogen of *h*V51. **b)** Molecular representation of HisF hydrophobic cluster (highlighted in a yellow square) and PRFAR-*f*K19 HisF:HisH interaction (highlighted in a green square). Plot of the distance for the most relevant distances of the hydrophobic cluster and PRFAR-*f*K19 distance. The distances between the residues forming the hydrophobic cluster are monitored between the β-carbon of *f*V48, the γ-carbon of *f*L50, the γ-carbon of *f*I52, and the ζ-carbon of *f*F23. The distance of the PRFAR-*f*K19 interaction is calculated between the phosphorus atom of PRFAR and the nitrogen of the side chain of *f*K19. All distances are in Å. See Appendix A Extended text section D

for a complete description of the results. **c)** Overlay of six relevant IGPS conformations (in purple) of PRFAR (in gray) in the HisF active site extracted from aMD simulations of the active ternary complex. The conformation of ProFAR (PRFAR precursor) co-crystallized in PDB 7AC8 (chain E) is shown in orange using a sphere and surface representation. PRFAR displays significant flexibility in the HisF active site. However, the position of the phosphate groups remains similar throughout the aMD simulation and resembles the position of ProFAR phosphate groups. Sodium ions are depicted as cyan spheres and accumulate near phosphate groups of PRFAR and carboxylate groups of catalytic *f*D11 and *f*D130.

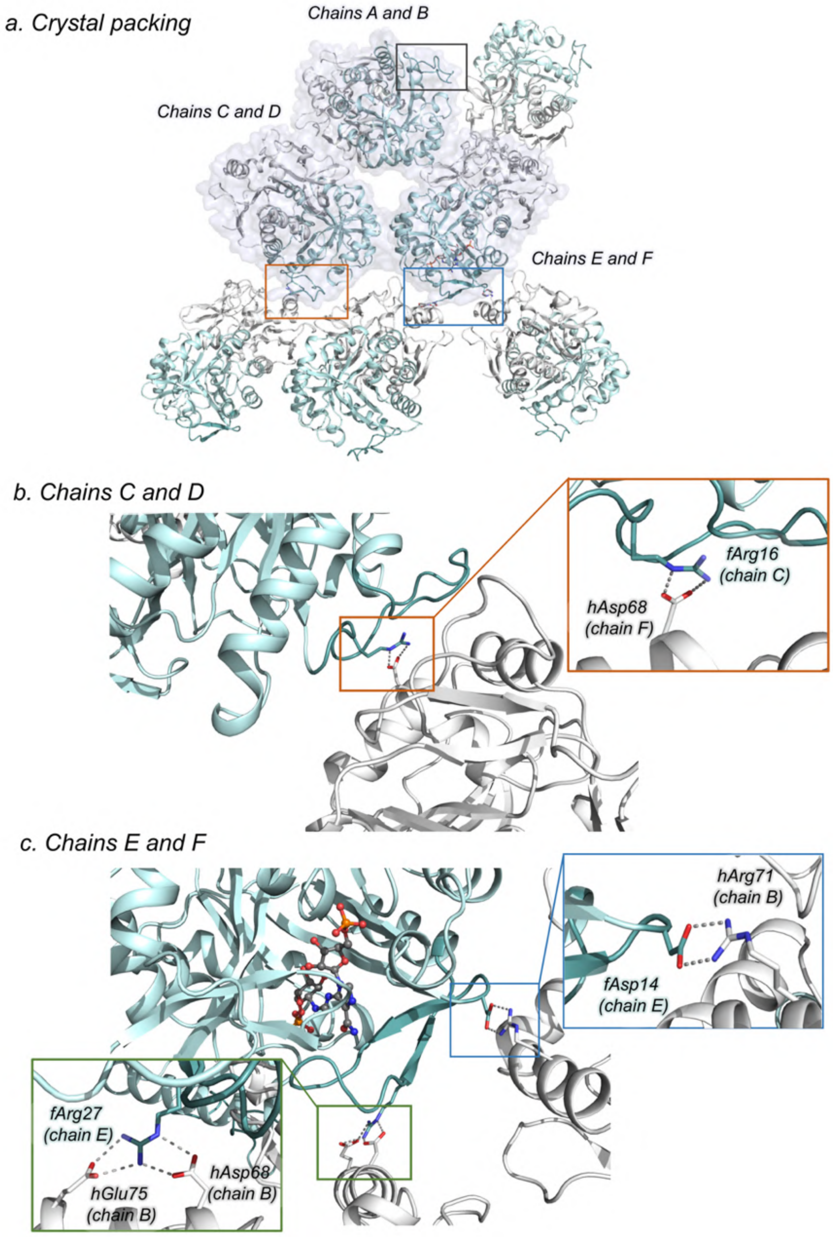
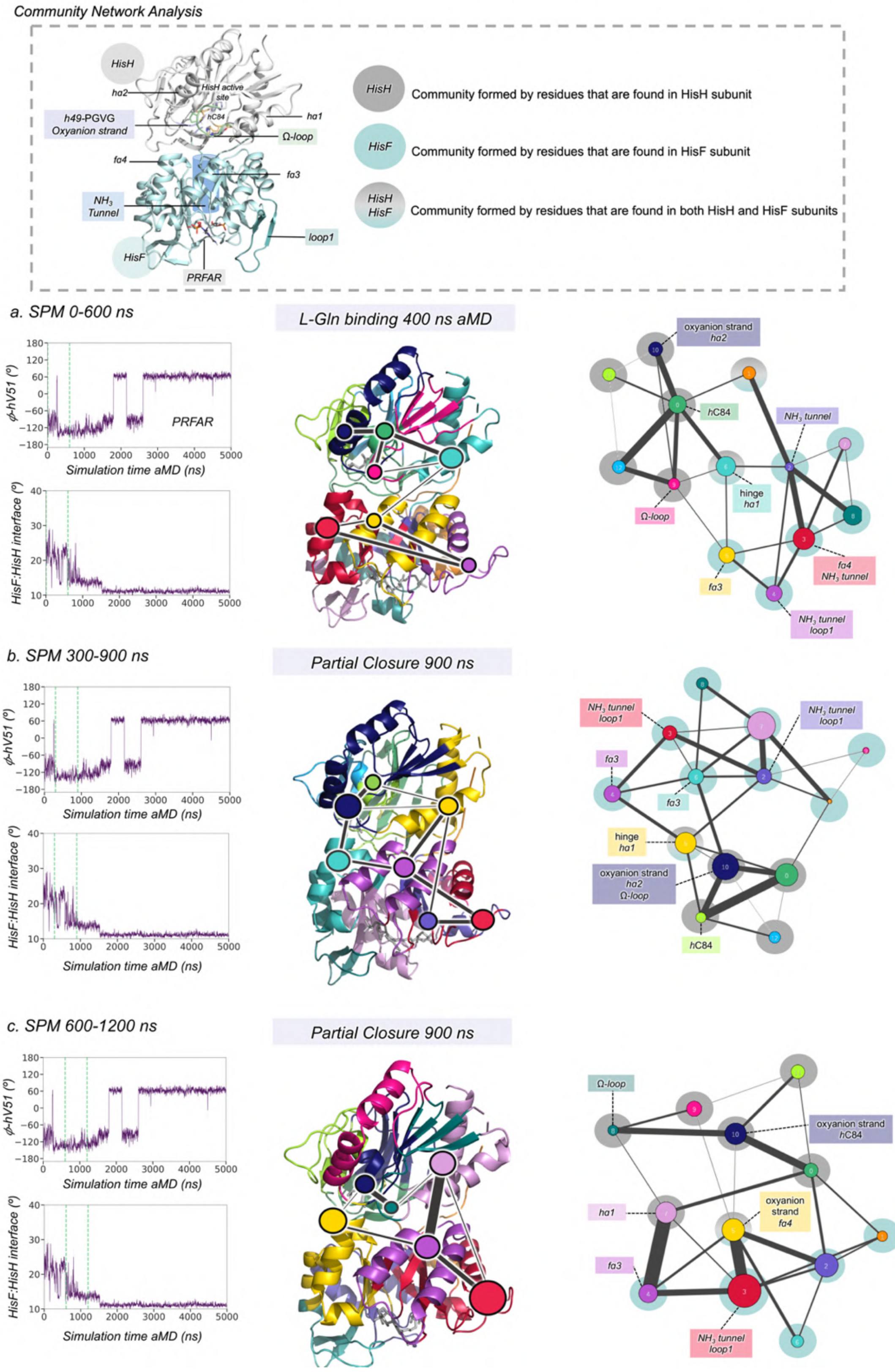
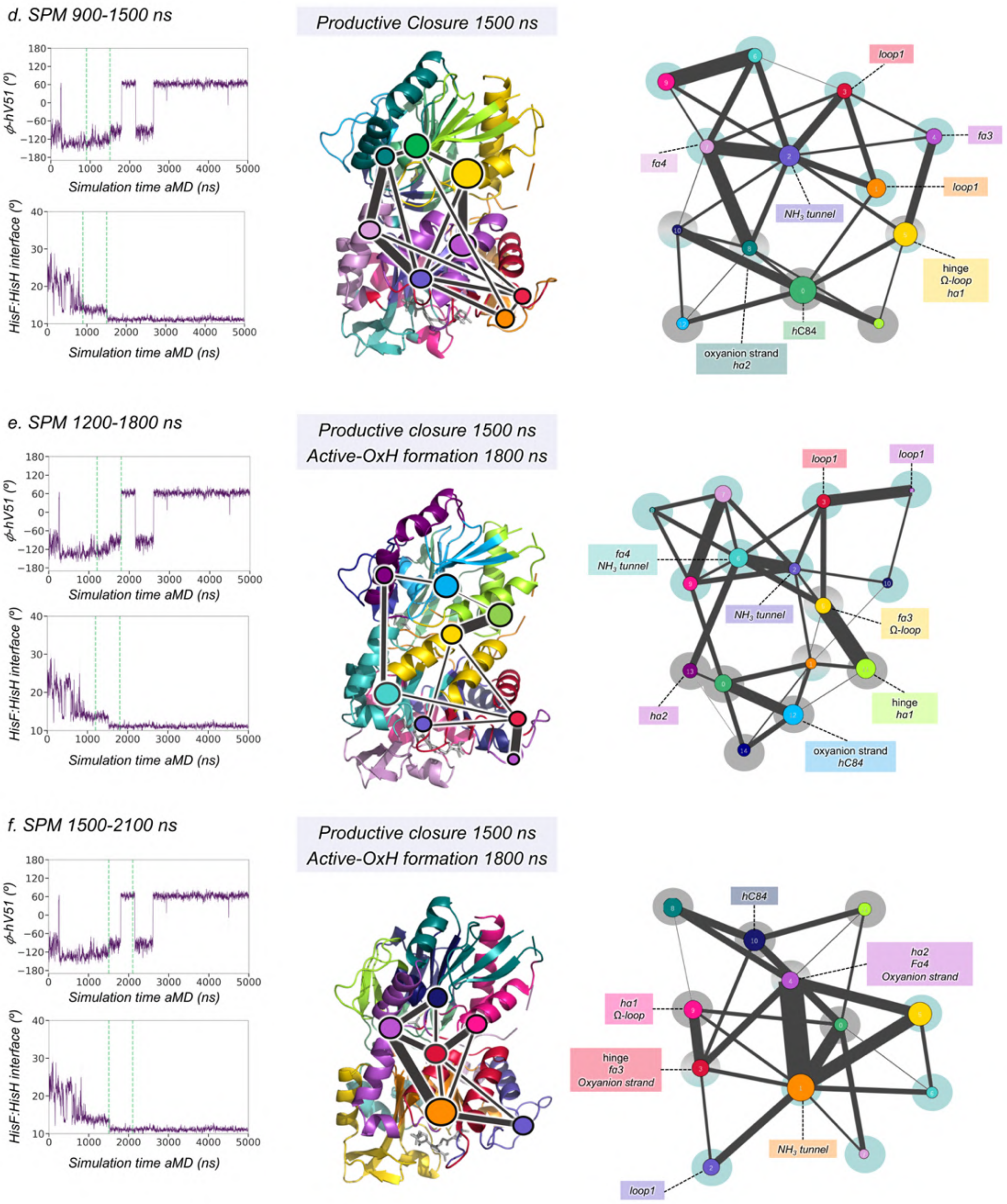


Figure A38. *f*Loop1 crystal packing of *h*C84A IGPS in PDB 7AC8. a) Crystal packing of *h*C84A IGPS mutant. The three IGPS units of the crystal structure together with the closed structures in the crystal of 7AC8 are shown. *f*Loop1 of chain A is not establishing interactions in the crystal packing and show an open disordered structure **b)** *f*Loop1 of chain C is interacting with *h*D68 of another chain F. The loop shows an open disordered conformation. **c)** *f*Loop1 of chain F is interacting with *h*R71, *h*D68, and *h*E75 of other IGPS chains of the crystal. The loops displays a closed ordered structure.


Community Network Analysis
HisH
hα2
HisH active site
hC84
hα1
h49-PGVG Oxyanion strand
Ω-loop
fα4
fα3
NH3 Tunnel
loop1
HisF
PRFAR
HisH Community formed by residues that are found in HisH subunit
HisF Community formed by residues that are found in HisF subunit
HisH HisF Community formed by residues that are found in both HisH and HisF subunits
a. SPM 0-600 ns
L-Gln binding 400 ns aMD
ϕ-hV51 (º)
PRFAR
Simulation time aMD (ns)
HisF:HisH interface (º)
oxyanion strand hα2
hC84
NH3 tunnel
hinge hα1
Ω-loop
fα4 NH3 tunnel
fα3
NH3 tunnel loop1
b. SPM 300-900 ns
Partial Closure 900 ns
ϕ-hV51 (º)
Simulation time aMD (ns)
HisF:HisH interface (º)
Simulation time aMD (ns)
NH3 tunnel loop1
NH3 tunnel loop1
fα3
fα3
hinge hα1
oxyanion strand hα2 Ω-loop
hC84
c. SPM 600-1200 ns
Partial Closure 900 ns
ϕ-hV51 (º)
Simulation time aMD (ns)
HisF:HisH interface (º)
Simulation time aMD (ns)
Ω-loop
oxyanion strand hC84
hα1
oxyanion strand fα4
fα3
NH3 tunnel loop1

Community Network Analysis



d. SPM 900-1500 ns

Productive Closure 1500 ns

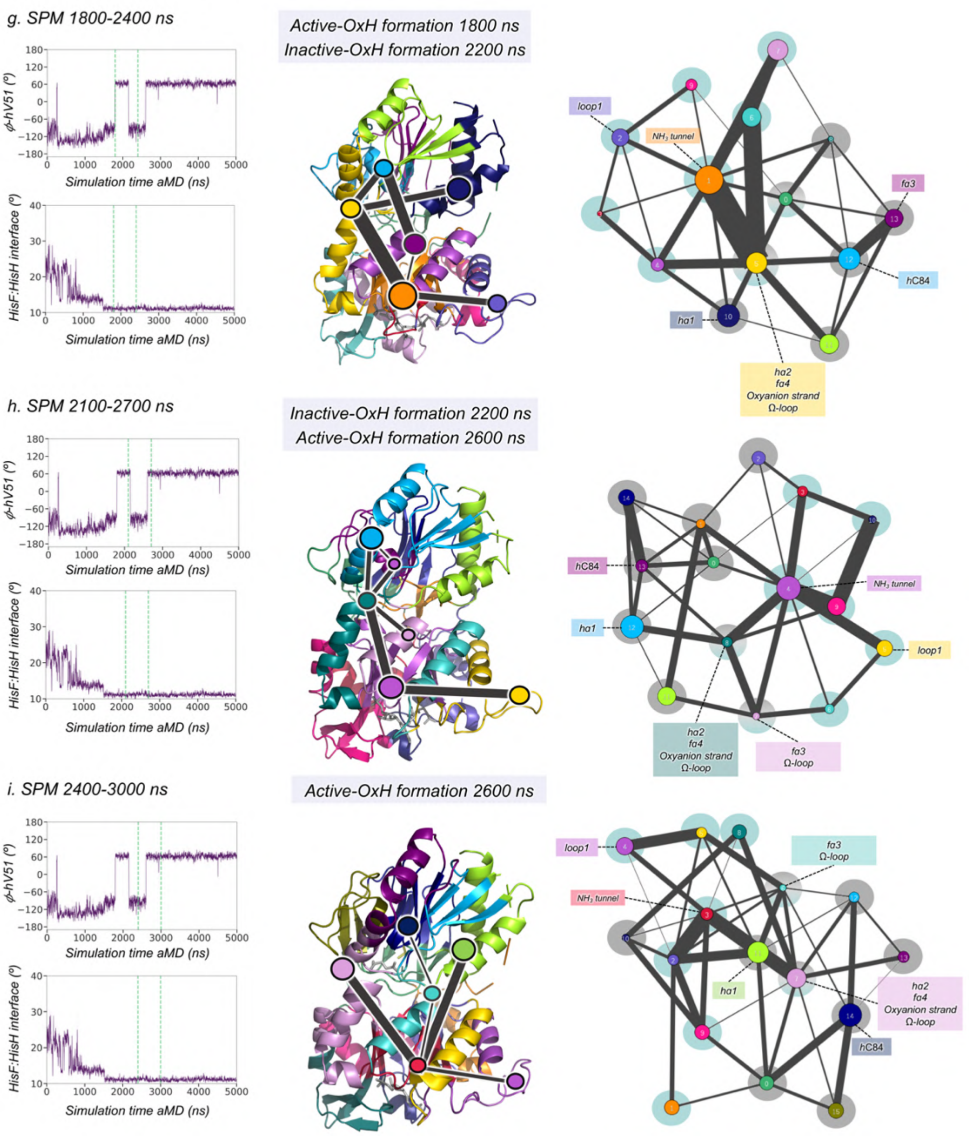
e. SPM 1200-1800 ns

Productive closure 1500 ns
Active-OxH formation 1800 ns

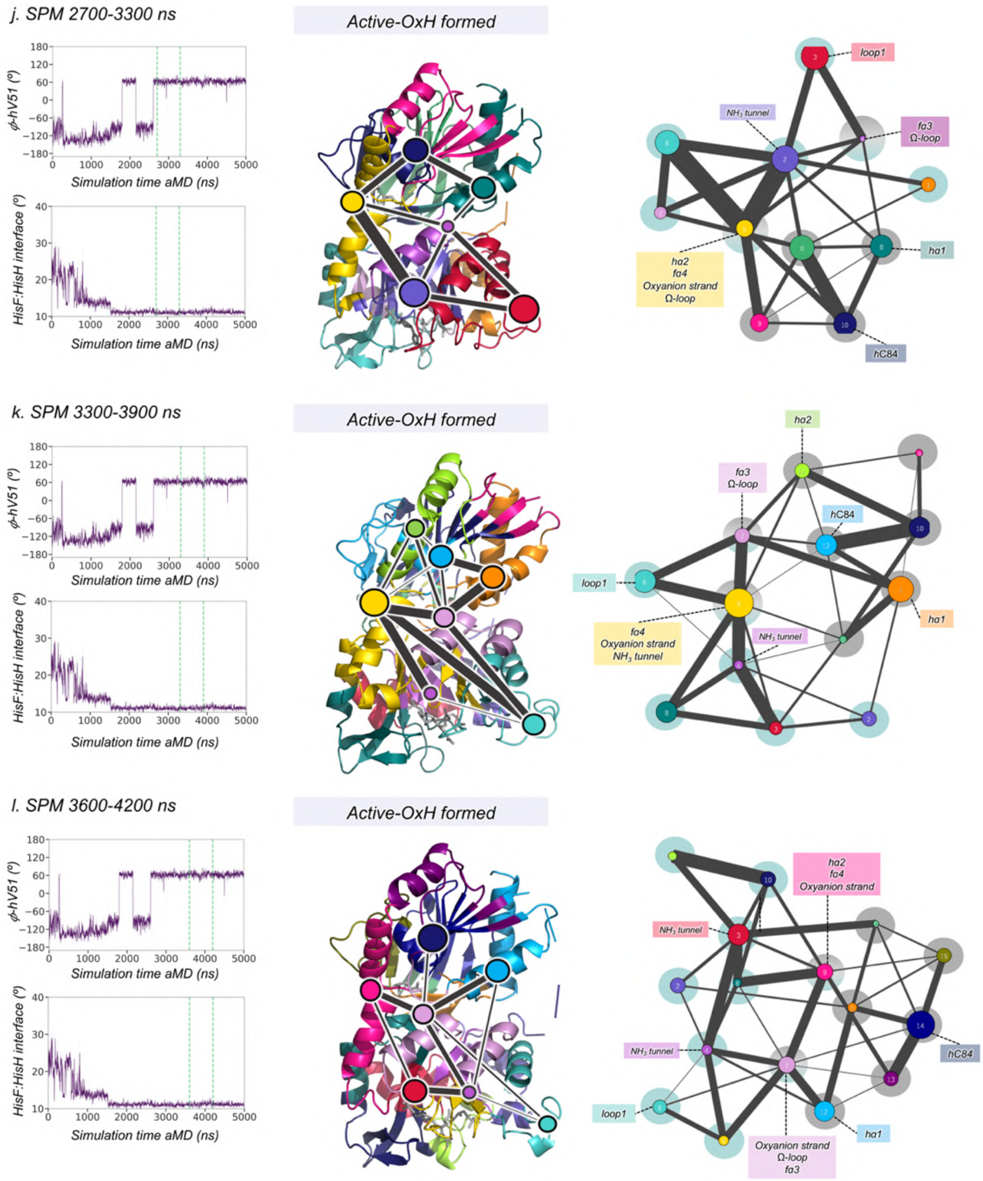
f. SPM 1500-2100 ns

Productive closure 1500 ns
Active-OxH formation 1800 ns

Community Network Analysis



Community Network Analysis



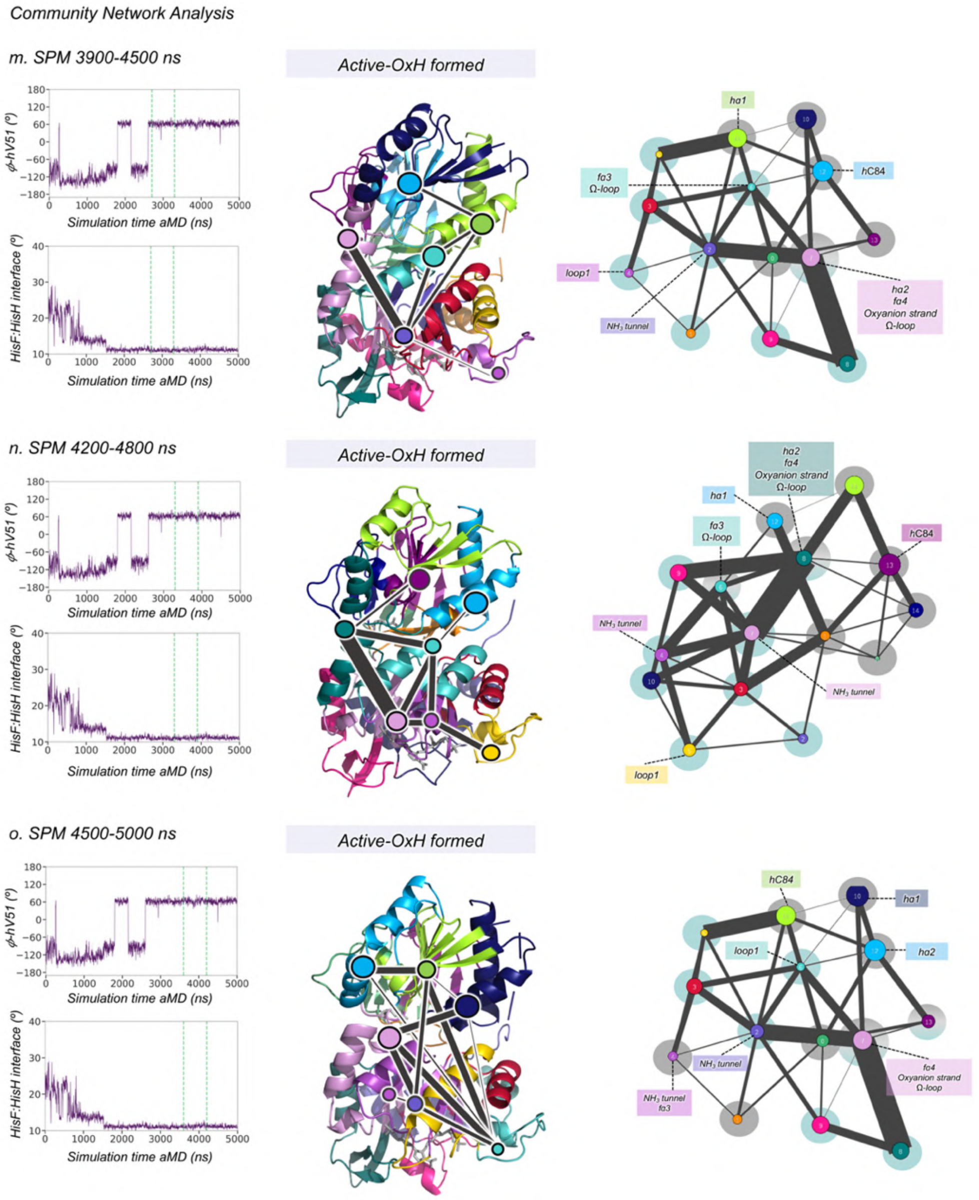


Figure A39. Detailed time-evolution community network analysis along a representative 5-μs aMD trajectory. Plot of the HisF:HisH interface angle along the simulation time. Plot of the ϕ dihedral angle of *h*V51. Vertical dashed green lines indicate the range of simulation time where the SPM was calculated. Structural representation of the PRFAR-IGPS community network analysis with the different communities shown in different colors and labeled accordingly. Graph of the community network with the most important structural elements of IGPS highlighted. Communities made of residues belonging to either HisF or HisH subunits are highlighted with a cyan and gray sphere, respectively. Communities made of residues from both HisF and HisH subunits are highlighted with a mixed cyan and gray circle.

Appendix A Movies

Movies can be found at: https://pubs.acs.org/doi/10.1021/jacs.1c12629

Movie A1. Conventional molecular dynamics simulations: transient *hV*51 oxyanion hole formation in substrate free PRFAR-IGPS. The movie shows the time-evolution of the HisH active site (PRFAR-IGPS) along a 4 μs conventional molecular dynamics simulation. This simulation started with the oxyanion strand in the Inactive-OxH conformation and evolves toward the Active-OxH state. The transient *h*V51 oxyanion hole formation occurs after 1 μs of simulation time, remaining formed for around 1 μs, and subsequently evolving to unblocked-OxH conformation. See Figure A7 for complete details. The *h*49-PGVG oxyanion strand residues are shown in purple. Catalytic (*h*C84, *h*H178, and *h*E180) and Ω-loop (*h*P10) residues are highlighted in orange and green, respectively.

Movie A2. Accelerated molecular dynamics simulations: spontaneous *L-Gln* substrate binding in the HisH active site. The movie shows the spontaneous substrate binding of *L-Gln* into the HisH active site for PRFAR-IGPS along a 600 ns accelerated molecular dynamics simulation. This simulation started with the oxyanion strand in the Active-OxH conformation and a single *L-Gln* molecule situated *ca.* 25 Å away from the HisH active site. Substrate recognition takes place when the oxyanion strand is in the Active-OxH state. Subsequently, the oxyanion strand readily transitions from the Active-OxH to the Unblocked-OxH and Inactive-OxH orientations. The population of the Inactive-OxH state allows the reorientation of the substrate in the HisH active site. When *L-Gln* eventually binds the HisH active site in the inactive-OxH state (step 4), the carbonyl of *L-Gln* is stabilized by the H^N backbone of *h*G52. The *h*49-PGVG oxyanion strand residues are shown in purple. Catalytic (*h*C84, *h*H178, and *h*E180) and Ω-loop (*h*P10) residues are highlighted in gray and green, respectively. *L-Gln* is depicted in gray spheres.

Movie A3. Accelerated molecular dynamics simulations: spontaneous *L-Gln* substrate binding in IGPS (global view). The movie shows the spontaneous substrate binding of *L-Gln* into the HisH active site for PRFAR-IGPS along a 600 ns accelerated molecular dynamics simulation. The *h*49-PGVG oxyanion strand residues are shown in purple. Catalytic (*h*C84, *h*H178, and *h*E180) and Ω-loop (*h*P10) residues are highlighted in gray and green, respectively. *L-Gln* is depicted in gray spheres.

Movie A4. Accelerated molecular dynamics simulations: allosteric activation of IGPS in the ternary complex. The movie shows the complete allosteric activation of IGPS in the ternary complex along a 10 μs accelerated molecular dynamics simulation. The movie starts with the spontaneous substrate binding process and follows with the subsequent allosteric activation. Without using *a priori* information of the active state, this simulation uncovers how IGPS, with the allosteric effector bound in HisF, spontaneously captures glutamine in a catalytically Inactive-OxH conformation, subsequently attains a closed HisF:HisH interface, and finally forms the *h*V51 oxyanion hole in HisH for efficient glutamine hydrolysis. The formation of the *h*V51 oxyanion hole takes place multiple times along the same simulation. The *h*49-PGVG oxyanion strand residues are shown in purple. Catalytic (*h*C84, *h*H178, and *h*E180) and Ω-loop (*h*P10) residues are highlighted in gray and green, respectively. *L-Gln* is depicted in gray spheres.

Appendix A References

(1) Kneuttinger, A. C.; Rajendran, C.; Simeth, N. A.; Bruckmann, A.; König, B.; Sterner, R. Significance of the Protein Interface Configuration for Allostery in Imidazole Glycerol Phosphate Synthase. *Biochemistry* **2020**, *59* (29), 2729–2742. https://doi.org/10.1021/acs.biochem.0c00332.

(2) Beismann-Driemeyer, S.; Sterner, R. Imidazole Glycerol Phosphate Synthase from Thermotoga Maritima. Quaternary Structure, Steady-State Kinetics, and Reaction Mechanism of the Bienzyme Complex. *J. Biol. Chem.* **2001**, *276* (23), 20387–20396. https://doi.org/10.1074/jbc.M102012200.

(3) Lisi, G. P.; East, K. W.; Batista, V. S.; Loria, J. P. Altering the Allosteric Pathway in IGPS Suppresses Millisecond Motions and Catalytic Activity. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114* (17), E3414--E3423. https://doi.org/10.1073/pnas.1700448114.

(4) Jiménez-Osés, G.; Osuna, S.; Gao, X.; Sawaya, M. R.; Gilson, L.; Collier, S. J.; Huisman, G. W.; Yeates, T. O.; Tang, Y.; Houk, K. N. The Role of Distant Mutations and Allosteric Regulation on LovD Active Site Dynamics. *Nat. Chem. Biol. 2014 106* **2014**, *10* (6), 431–436. https://doi.org/10.1038/nchembio.1503.

(5) Kuzmanic, A.; Sutto, L.; Saladino, G.; Nebreda, A. R.; Gervasio, F. L.; Orozco, M. Changes in the Free-Energy Landscape of P38\$α\$ MAP Kinase through Its Canonical Activation and Binding Events as Studied by Enhanced Molecular Dynamics Simulations. *Elife* **2017**, *6*. https://doi.org/10.7554/eLife.22175.

(6) Chaudhuri, B. N.; Lange, S. C.; Myers, R. S.; Davisson, V. J.; Smith, J. L. Toward Understanding the Mechanism of the Complex Cyclization Reaction Catalyzed by Imidazole Glycerolphosphate Synthase: Crystal Structures of a Ternary Complex and the Free Enzyme. *Biochemistry* **2003**, *42* (23), 7003–7012. https://doi.org/10.1021/bi034320h.

(7) Lipchock, J. M.; Loria, J. P. Nanometer Propagation of Millisecond Motions in V-Type Allostery. *Structure* **2010**, *18* (12), 1596–1607. https://doi.org/10.1016/j.str.2010.09.020.

(8) Rivalta, I.; Sultan, M. M.; Lee, N. S.; Manley, G. A.; Loria, J. P.; Batista, V. S. Allosteric Pathways in Imidazole Glycerol Phosphate Synthase. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109* (22). https://doi.org/10.1073/pnas.1120536109.

(9) Boto, R. A.; Peccati, F.; Laplaza, R.; Quan, C.; Carbone, A.; Piquemal, J.-P.; Maday, Y.; Contreras-García, J. NCIPLOT4: Fast, Robust, and Quantitative Analysis of Noncovalent Interactions. *J. Chem. Theory Comput.* **2020**, *16* (7), 4150–4158. https://doi.org/10.1021/ACS.JCTC.0C00063.

Appendix B

Supplementary information Chapter 5. Material from: Liu, Z., Calvó-Tusell, C., Zhou, A.Z. et al. Dual-function enzyme catalysis for enantioselective carbon–nitrogen bond formation. *Nat. Chem.* 13, 1166–1172 (**2021**) published 2021, Springer Nature.

Computational Methods

Molecular dynamics (MD) simulations

Molecular Dynamics simulations were performed using the GPU code (*pmemd*)[1] of the AMBER 18 package.[2] Parameters for the lactone carbene covalently bound to Fe-haem Ser ligated cofactor, ylides and amine substrates, and Fe-haem Ser-ligated cofactor were generated within the *antechamber* and MCPB.py[3] modules in AMBER18 package using the general AMBER force field (GAFF),[4] with partial charges set to fit the electrostatic potential generated at the B3LYP/6-31G(d) level by the RESP model.[5] The charges were calculated according to the Merz-Singh-Kollman scheme[6,7] using the Gaussian 09 package.[8]

Protonation states of protein residues at pH 7.4 were predicted using H++ server. Each protein was immersed in a pre-equilibrated truncated cuboid box with a 10 Å buffer of TIP3P[9] water molecules using the *leap* module, resulting in the addition of around 15,300 solvent molecules. The systems were neutralized by addition of explicit counter ions (Na^+ and Cl^-). All subsequent calculations were done using the widely tested Stony Brook modification of the Amber14 force field (*ff14sb*).[10] A two-stage geometry optimization approach was performed. The first stage minimizes the positions of solvent molecules and ions imposing positional restraints on the solute by a harmonic potential with a force constant of 500 $kcal\cdot mol^{-1}\cdot Å^{-2}$ and the second stage minimizes all the atoms in the simulation cell except those involved in the harmonic distance restraint. The systems were gently heated using six 50 ps steps, incrementing the temperature by 50 K for each step (0–300 K) under constant-volume and periodic-boundary conditions. Water molecules were treated with the SHAKE algorithm such that the angle between the hydrogen atoms was kept fixed. Long-range electrostatic effects were modelled using the particle-mesh-Ewald method.[11] An 8 Å cutoff was applied to Lennard–Jones and electrostatic interactions. Harmonic restraints of 30 $kcal\cdot mol^{-1}$ were applied to the solute and the Andersen equilibration scheme was used to control and equalize the temperature. The time step was kept at 2 fs during the heating stages, allowing potential inhomogeneities to self-adjust. Each system was then equilibrated for 2 ns with a 2-fs time step at a constant pressure of 1 atm. Finally, conventional MD trajectories at constant volume and temperature (300 K) were propagated.

Constrained-MD simulations included a restrained distance between the substrate amine N atom and the lactone carbene central C atom (up to 3.1–3.6 Å) or between the ylide lactone central C atom and Fe (up to 3.7–4.2 Å), that was defined by adding a harmonic potential with k = 50 and 200 $\mathrm{mol}^{-1}\cdot\text{Å}^{-2}$ to the respective coordinate during the respective equilibrations and production runs. Since the distance of TS1 for the amide insertion, calculated using DFT, is 2.04 Å (see Figure B10), we considered that 3.6 Å as a maximum value is representative of the reactant complex. The distance restrains applied for the ylide complex was also selected based on the DFT calculations shown in Figure B10. Trajectories were processed and analyzed using the *cpptraj*[12] module from AmberTools utilities.

The total simulation time accumulated for each system is detailed below:

- P411-**L1** variant in the *apo* state: 5 replicas of 1000 ns each (5,000 ns in total).
- P411 variants (**L5**, **L6**, and **L7**) in the *apo* state: 3 independent replicas of 500 ns each (1,500 ns in total).
- Lactone carbene bound into P411 studied variants (**L5**, **L6**, and **L7**): 5 independent replicas of 500 ns each (2500 ns in total).
- Amine substrate bound in a near attack conformation respect to the lactone carbene bound into P411 studied variants (**L5**, **L6**, and **L7**), through constrained-MD simulations: 2 sets of 3 independent replicas of 250 ns each (2 × 750 = 1,500 ns in total, for each system).
- Ylide bound into P411 studied variants (**L5**, **L6**, and **L7**) in a "just dissociated" conformation, through constrained-MD simulations: 3 independent replicas of 100 ns each (300 ns in total, for each system).
- P411-**L5-B3** variant in the *apo* state: 3 replicas of 500 ns each (1,500 ns in total).
- Lactone carbene bound into P411-**L5-B3** variant: 5 independent replicas of 500 ns each (2500 ns in total).
- Amine substrate **1a** bound in a near attack conformation respect to the lactone carbene bound into P411-**L5-B3** variant, through constrained-MD simulations: 2 sets of 3 independent replicas of 250 ns each (2 × 750 ns = 1,500 ns in total).
- Ylide bound into P411-**L5-B3** in a "just dissociated" conformation, through constrained-MD simulations: 5 independent replicas of 100 ns each (500 ns in total).
- Spontaneous substrate binding reconstruction of substrate **1a** in P411 variants (**L5-B3** and **L6**): 10 spontaneous binding simulations of 250 ns each for the selected variants (10 x 250 ns = 2500 ns each). Simulations where effective substrate binding was observed, were further extended up to 1,000 ns each.

Quantum Mechanics Density Functional Theory (DFT) calculations

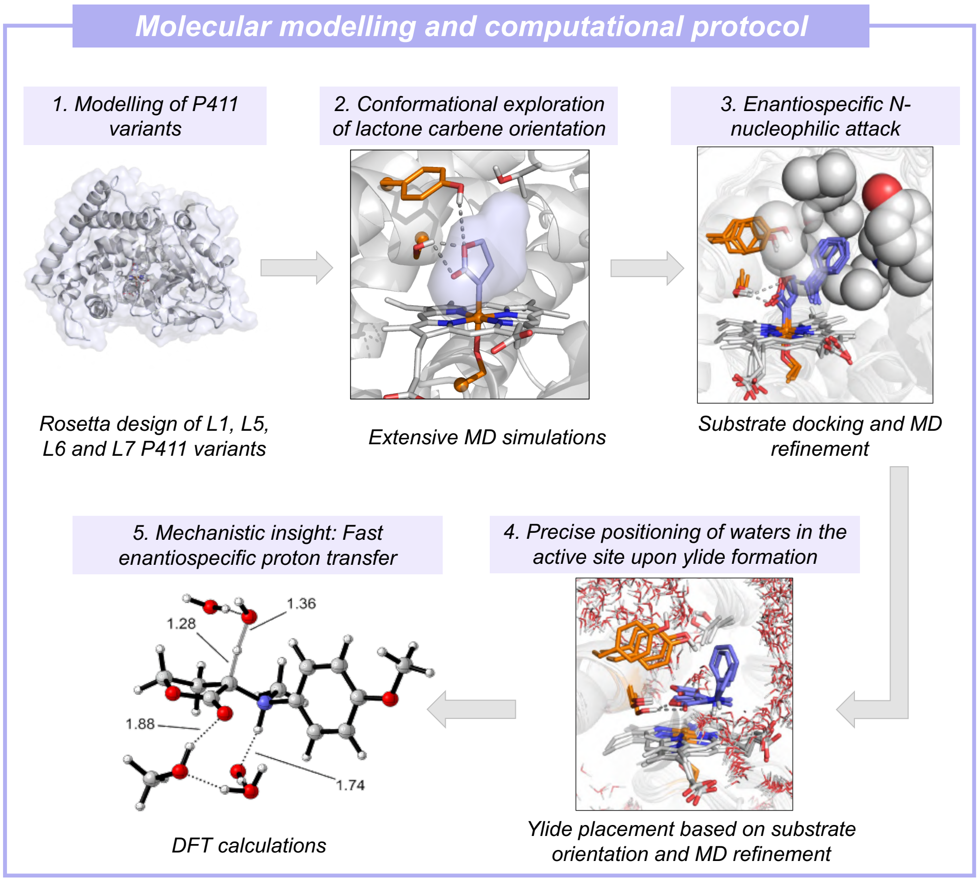
Density Functional Theory (DFT) calculations were carried out using Gaussian09.[8] A truncated computational model containing the porphyrin pyrrole core, Fe center and a methoxy group to mimic serine as Fe-axial ligand was used. Geometry optimizations and frequency calculations were performed using (U)B3LYP[13-15] functional with the SDD basis set for iron and 6-31G(d) on all other atoms. Transition states had one negative force constant corresponding to the desired reaction coordinate. All stationary points were verified as minima or first-order saddle points by a vibrational frequency analysis. Intrinsic reaction coordinate (IRC) calculations were performed to ensure that the optimized transition states connect the corresponding desired reactants and products. Enthalpies and entropies were calculated at 1 atm and 298.15 K. Single point (SP) energy calculations were performed using the dispersion-corrected functional (U)B3LYP-D3(BJ)[16,17] with the Def2TZVP basis set on all atoms. The CPCM polarizable conductor model (diethyl ether, $\varepsilon = 4$)[18,19] to have an estimation of the dielectric permittivity in the enzyme active site was included during the optimizations and SP calculations. The use of a dielectric constant $\varepsilon = 4$ has been proved to be a good and general model to account for electronic polarization and small backbone fluctuations in enzyme active sites.[20,21]

The methodology employed in this study, based on the use of (U)B3LYP density functional, is very similar to the previously used by us and other groups for the study of haem-iron carbene transfer reaction mechanisms.[22-26] Independent benchmark studies by Prof. Shaik[22] and Prof. Liu[25] groups demonstrated that this method performs very well in the computational modelling of these carbene transfer reactions.

The modeling of the open-shell electronic state was done by using a Gaussian09 “stable = opt” calculation[27-29] to generate a singlet open-shell orbital guess from the triplet optimized geometry, followed by a full optimization of the system starting from this guess. Using this approach, we could successfully determine the open-shell singlet pathway for the studied nucleophilic attack of the amine to the lactone carbene.

Optimized DFT structures are illustrated with CYLView.[30]

Computational protocols



Molecular modelling and computational protocol

1. Modelling of P411 variants

Rosetta design of L1, L5, L6 and L7 P411 variants

2. Conformational exploration of lactone carbene orientation

Extensive MD simulations

3. Enantiospecific N-nucleophilic attack

Substrate docking and MD refinement

4. Precise positioning of waters in the active site upon ylide formation

Ylide placement based on substrate orientation and MD refinement

5. Mechanistic insight: Fast enantiospecific proton transfer

DFT calculations

1. Starting from the available crystal structure of P411-**E10** variant (PDB: 5UCW.pdb),[31] the P411-**C10** variant (P411-**L1**) was prepared using RosettaDesign.[32] P411-**L1** contains 13 mutations with respect to P411-**E10** (N70E, A74G, V78L, M118S, F162L, M177L, L263Y, H266V, A330Y, I401L, T436L, L437Q, S438T).

The P411-**L1** structure obtained from Rosetta was used as starting point for MD simulations. A total of 5 independent replicas of 1,000 ns each were carried out, accumulating a total of 5,000 ns (5 µs). The conformations visited by the enzyme along all this simulation time were clustered based on protein backbone RMSD, and the most populated cluster was selected as a representative structure of P411-**L1** *apo* state.

The representative structure obtained for P411-**L1** was used to prepare P411 variants **L5**, **L6** and **L7** using the mutagenesis tool from Pymol.[33] Variant **L5** includes 4 additional mutations with respect to **L1** (T327V, Q437L, S332A, A87P), **L6** has 1 additional mutation respect **L5** (A264S), and **L7** has 1 mutation respect to **L6** (V327P).

These structures generated for **L5**, **L6**, and **L7** were then used as starting points for MD simulations. A total of 3 independent replicas of 500 ns each were propagated, accumulating a total of 1,500 ns (1.5 µs) for each of these variants. These were clustered considering protein backbone RMSD, and the most populated cluster was selected as representative structure of P411-**L5**, **L6** and **L7** variants in the *apo* state, respectively.

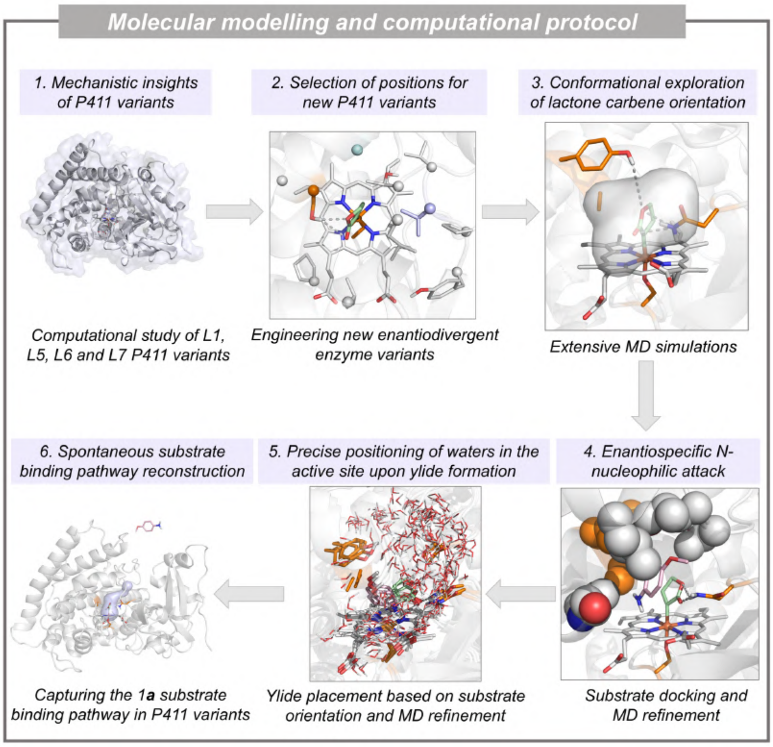
2. The lactone carbene covalently bound to the iron was manually docked in the representative structures for P411 variants **L5**, **L6** and **L7** characterized previously in their *apo* state. For each system, 5 independent replicas of 500 ns each and starting from different orientations of the lactone carbene were carried out, accumulating a total of 2,500 ns (2.5 μs). The conformations explored by the lactone carbene and the interactions that it established with active site residues were analyzed. The most representative conformations visited by the enzyme along the accumulated simulation time were clustered based on protein backbone RMSD.

3. The amine substrate was docked in the lactone carbene-bound P411-**L5**, **L6**, and **L7** variants. Two different representative structures (the two most populated clusters) previously characterized for each system (in step 2) were used. Docking calculations were carried out using Autodock Vina.[34,35] Docking predictions were refined by performing constrained-MD simulations, where the distance between the amine and the carbene central C atom was kept restrained up to 3.6 Å. For each system 2 different structures were considered as starting points, and 3 independent replicas of 250 ns each were carried out, accumulating a total of 2 × 750 ns = 1,500 ns of sampling. The most representative conformations visited along the trajectories were characterized by clustering, considering the protein backbone RMSD. These simulations provided good descriptions of catalytically relevant binding poses explored by the amine substrate, in which it is in a near attack conformation to perform the nucleophilic attack to the carbene.

4. Next, the ylide-bound complex was prepared. The ylide was manually docked in P411 **L5**, **L6**, and **L7** variants, starting from the most populated clusters obtained from the amine and lactone carbene bound simulations (in step 3). The characterized amine substrate binding pose was used as a template, and the ylide was placed by superimposing it with the amine and the lactone carbene. Starting from these ylide bound structures constrained-MD simulations were performed, in which the distance between the ylide central C atom and the Fe was kept restrained up to 4.1 Å. With this geometric constraint it was aimed to characterize the dissociated ylide complex in the enzyme active site, mimicking the ylide complex characterized from model DFT calculations (see Figure B9). For each system, total of three replicas of 100 ns each were carried out, accumulating a total of 300 ns of simulation time.

5. Finally, visual inspections of the constrained-MD simulations of P411 **L6** and **L7** variants with ylide intermediate bound (from step 4) were carried out. This led to prepare different truncated active site models that included the ylide intermediate, S264 residue, and key active site water molecules, that were used to explore the possible ylide proton transfer pathways by DFT calculations.

The mechanistic information obtained from the computational modelling of variants **L1**, **L5**, **L6** and **L7** is used to design enzyme variants that revert the enantioselectivity toward the other enantiomer. To do so, the following computational protocol is applied:


Molecular modelling and computational protocol
1. Mechanistic insights of P411 variants
2. Selection of positions for new P411 variants
3. Conformational exploration of lactone carbene orientation
Computational study of L1, L5, L6 and L7 P411 variants
Engineering new enantiodivergent enzyme variants
Extensive MD simulations
6. Spontaneous substrate binding pathway reconstruction
5. Precise positioning of waters in the active site upon ylide formation
4. Enantiospecific N-nucleophilic attack
Capturing the 1a substrate binding pathway in P411 variants
Ylide placement based on substrate orientation and MD refinement
Substrate docking and MD refinement

1. In the first project, we reported a series of P411 enzymes (engineered P450 enzymes substituted with serine as the heme-ligating residue) that perform efficient carbene N–H insertion with enantioselectivity over 95:5 *er* toward the *(S)*-enantiomer. In the evolutionary trajectory, a dramatic change in enantioselectivity (from -21% ee to 92% ee for *N*-methyl aniline) was observed after the introduction of a single mutation at position A264S when going from **L5** to **L6**. Molecular Dynamics (MD) simulations revealed that the orientation of the carbene-lactone in the active site is key to determine the enantioselectivity of the variant. In **L5**, which showed poor enantioselectivity, the lactone explores multiple conformations. However, when a serine is introduced at position 264 (**L6**), hydrogen-bond interactions between this residue and the lactone-carbene ester group keep the lactone in a single preferred orientation.

Mechanistic insights indicated that the orientation of the lactone-carbene formed in the enzyme active site plays a key role for the enantioselective formation of the ylide intermediate, which undergoes then a stereoselective protonation step to yield the final product.

2. Based on the mechanistic insights gathered by applying the previous protocol, we aimed to rationally engineer new enantiodivergent enzyme variants. To achieve this, mutational positions were rationally selected based on their spatial distribution in the active site around the lactone-carbene intermediate, based on the computational models generated for the lactone-carbene bound in **L5**, **L6** and **L7** systems. The selection of the positions to be mutated was made with the final aim of reshaping the active site, in order to control the conformations accessible for the lactone-carbene biocatalytic intermediate. The goal was to invert the major orientation the lactone-carbene explores in **L6** and **L7** *S*-selective variants, to favor the amine *N*-nucleophilic attack from the opposite face of the lactone ring (*re* face) and access to the opposite product enantiomer.

In order to modulate the orientation of the lactone-carbene, we hypothesized that it would be required to (1) replace the serine at position 264 for a non-polar residue to disrupt this H-bond interaction occurring in **L6** and **L7**; and (2) introduce a new H-bond donor residue at the opposite side of the active site that could act as a new anchoring point for the lactone-carbene and invert its orientation in the enzyme active site (see Figure 5.5, Chapter 5).

By analyzing the structural arrangement of the active site of the computational models generated for **L5** and **L6** variants with the lactone-carbene bound, we identified two positions that could be mutated to act as new anchoring points from the opposite side. The selected positions for site-saturation mutagenesis were 268 and 328. S264 polar residue in *S*-selective **L6** and **L7** was proposed to be mutated to a similar sized but non-polar alanine residue. Accordingly, **L6** variant with S264A mutation corresponds to the parent **L5**, which was used as starting point for the new evolution campaign.

Protein engineering based on site-saturation mutagenesis (SSM) and screening at the 328 and 268 positions, led to the identification of two *R*-selective variants: **L5_FL-B2** and **L5_FL-B3**, where residue V328 is mutated to a glutamine (Q) and an asparagine (N), respectively.

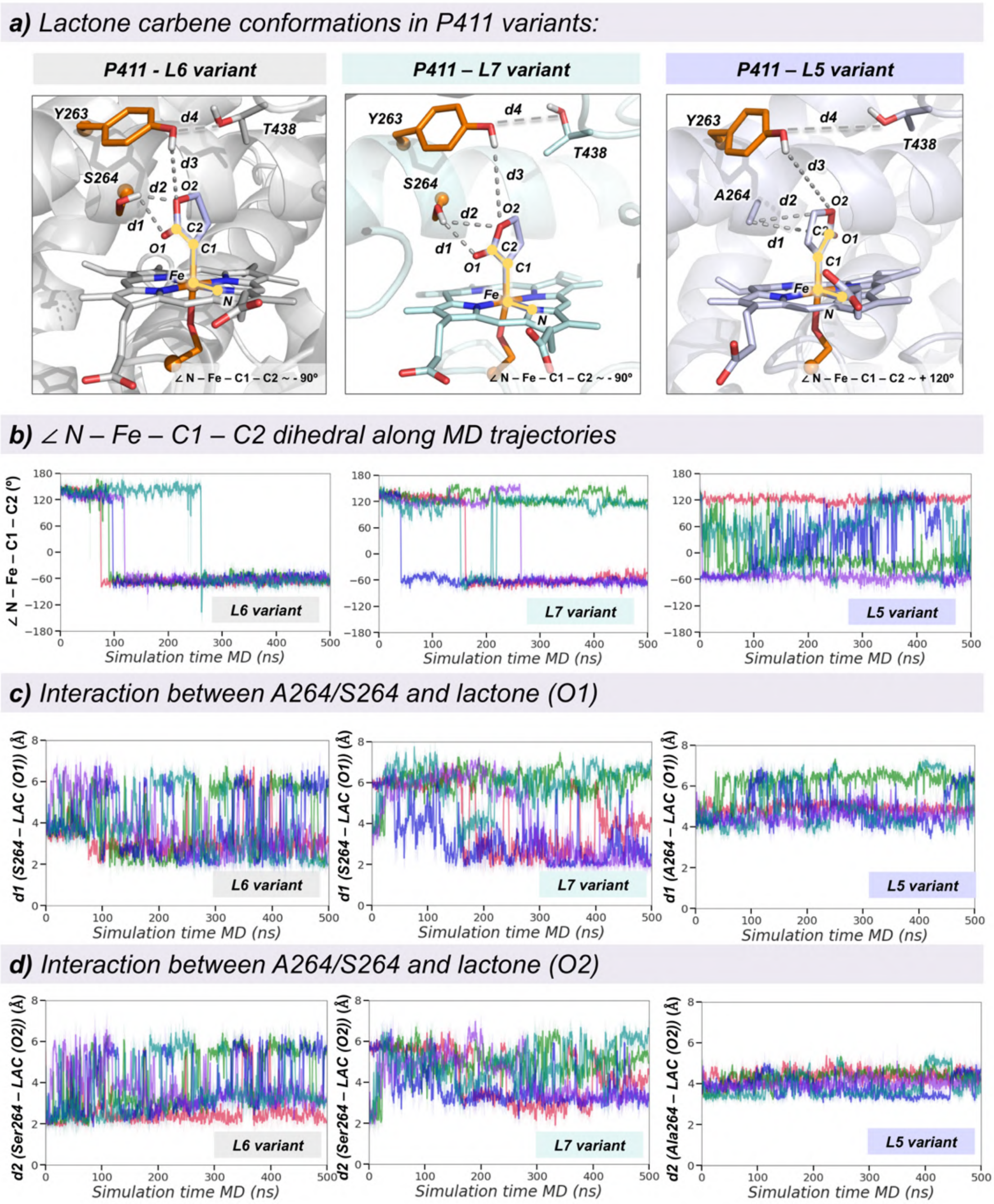
Computationally, we modelled the variant **L5-B3** in the *holo* state, using the previously generated model for **L5** variant (see computational methods). A total of 3 independent replicas of 500 ns were performed, accumulating a total of 1500 ns. These were clustered considering protein backbone RMSD, and the most populated cluster was selected as representative structure of P411-**L5-B3** in the *apo* state.

3. The lactone-carbene covalently bound to the iron was manually docked in a representative structure of the most populated conformational state of P411-**L5-B3**, as characterized previously in its *holo* state. Then, 5 independent replicas of 500 ns each and starting from different orientations of the lactone-carbene were carried out, accumulating a total of 2,500 ns (2.5 μs). The conformations explored by the lactone-carbene and the interactions that it establishes with active site residues were analyzed. The most populated conformational state visited by the enzyme along the accumulated simulation time was characterized based on clustering analysis considering the protein backbone RMSD.

4. The amine substrate 1a was docked in the active site of the lactone-carbene bound P411-**L5-B3** modelled structure. Two different representative structures (the two most populated clusters) previously characterized in step 3 for **L5-B3** with lactone-carbene bound were used. Docking calculations were carried out using Autodock Vina. Docking predictions were refined by performing restrained-MD simulations, where the distance between the amine and the lactone-carbene central C atom was kept restrained up to 3.6 Å (see computational methods section). As a starting point, 2 different structures were considered, and 3 independent replicas of 250 ns each were carried out for each system, accumulating a total of 2 × [3 × 250 ns] = 1,500 ns of sampling. The most representative conformations visited along the trajectories were characterized by clustering, considering the protein backbone RMSD. These simulations provided good descriptions of catalytically relevant binding poses explored by the amine substrate with respect to the carbene, in which the amine is in a near attack conformation to perform the *N*-nucleophilic attack to the carbene.

5. Next, the ylide-bound complex was studied. The ylide was manually docked in P411-**L5-B3** variant starting from the most populated clusters obtained from the amine and lactone-carbene bound simulations (step 4). The characterized amine substrate binding pose was used as a template, and the ylide was placed by superimposing it with the amine and the lactone-carbene. Starting from these ylide-bound structures restrained-MD simulations were performed, in which the distance between the ylide central C atom and the Fe was kept restrained up to 4.1 Å. With this geometric constraint it was aimed to characterize the ylide once it is formed in the enzyme active site, mimicking the ylide-heme no-covalent complex characterized from model DFT calculations. A total of 5 independent replicas of 100 ns each were carried out, accumulating a total of 500 ns of simulation time.

6. Finally, spontaneous substrate binding simulations to reconstruct the substrate binding pathway in P411-**L5-B3** and **L6** variants were performed. For each variant, 4 molecules of substrate 1a were placed outside the protein in the bulk solvent (*ca.* 20 Å away from the active site), and then trajectories were propagated to capture the spontaneous substrate binding pathway without predefining any reaction coordinate or including any bias. For each system, 10 spontaneous binding simulations of 250 ns each for the selected variants (**L5-B3** and **L6**) were carried out. For **L5-B3**, a binding event was observed in one out of 10 simulations. For **L6**, binding events were observed in two out of 10 simulations. The simulations where substrate binding was observed were extended up to 1,000 ns.


a) Lactone carbene conformations in P411 variants:
P411 - L6 variant
P411 – L7 variant
P411 – L5 variant
Y263
d4
T438
d3
S264
d2
O2
d1
C2
O1
C1
Fe
N
A264
∠ N – Fe – C1 – C2 ~ - 90°
∠ N – Fe – C1 – C2 ~ - 90°
∠ N – Fe – C1 – C2 ~ + 120°
b) ∠ N – Fe – C1 – C2 dihedral along MD trajectories
∠ N – Fe – C1 – C2 (°)
180
120
60
0
−60
−120
−180
L6 variant
L7 variant
L5 variant
0
100
200
300
400
500
Simulation time MD (ns)
c) Interaction between A264/S264 and lactone (O1)
d1 (S264 – LAC (O1)) (Å)
d1 (A264 – LAC (O1)) (Å)
8
6
4
2
0
d) Interaction between A264/S264 and lactone (O2)
d2 (Ser264 – LAC (O2)) (Å)
d2 (Ala264 – LAC (O2)) (Å)

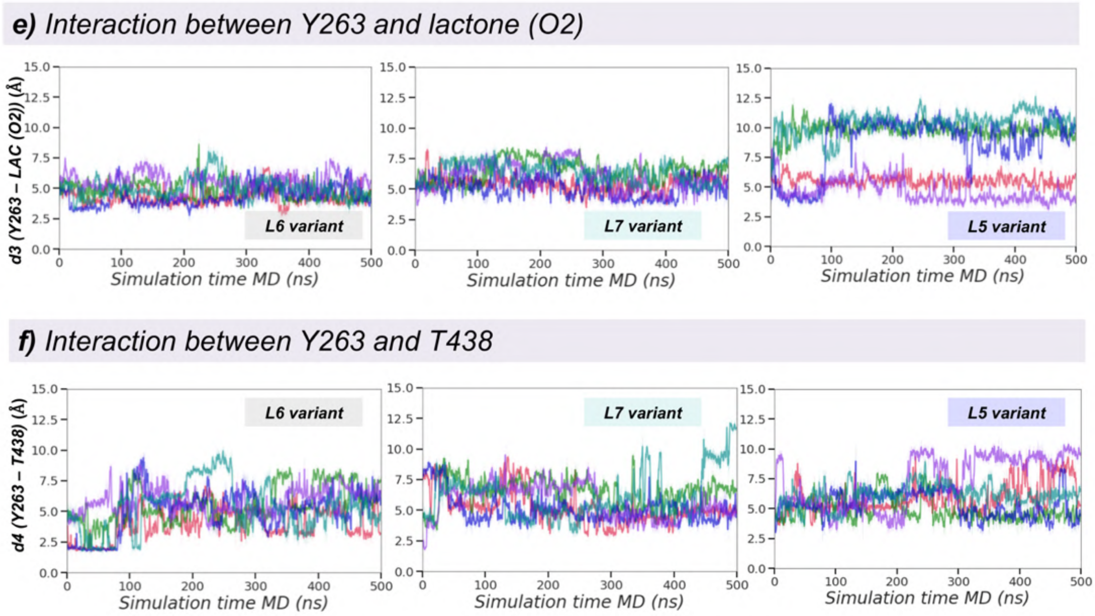


Figure B1. Conformational control of lactone carbene formed in P411 variants. a, Representative snapshot corresponding to the most populated cluster extracted from Molecular Dynamics (MD) simulations describing the preferred orientation of the lactone carbene when formed in P411-**L6** (gray), P411-**L7** (cyan) and P411-**L5** (purple) variants. **b,** The ∠(N – Fe – C1 – C2) dihedral angle measured along independent MD trajectories (5 replicas, 500 ns each) describes the relative orientation explored by the carbene. In P411-**L6** and P411-**L7**, simulations show that the lactone preferentially explores a single conformation (dihedral angle ca. –90º), which is stabilized by H-bond interactions established between the carbene ester group and S264 and Y263 (see below). In P411-**L5**, simulations show that the lactone explores multiple orientations without a clear preference. **c,** Distance vs. time plot describing the H-bond interaction between the S264 hydroxyl group and the O1-oxygen of the lactone in P411-**L6** and P411-**L7**, and the A264 C_β and the O1-oxygen of the lactone in P411-**L5** (*d1*, as shown in Figure B1a). **d,** Distance vs. time plot describing the H-bond interaction between the S264 hydroxyl group and the O2-oxygen of the lactone in P411-**L6** and P411-**L7**, and the A264 C_β and the O2-oxygen of the lactone in P411-**L5** (*d2*, as shown in Figure B1a). **e,** Distance vs. time plot describing the H-bond interaction between the Y263 hydroxyl group and the O2-oxygen of the lactone in P411-**L6**, P411-**L7** and P411-**L5** (*d3*, as shown in Figure B1a). **f,** Distance vs. time plot describing the interaction between the Y263 hydroxyl group and the T438 hydroxyl group in P411-**L6**, P411-**L7** and P411-**L5** (*d4*, as shown in Figure B1a).

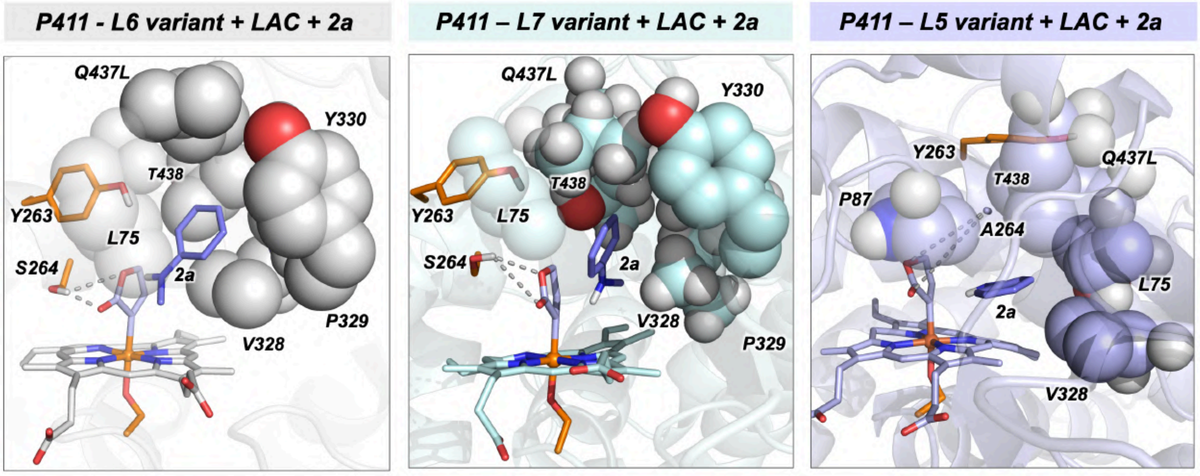
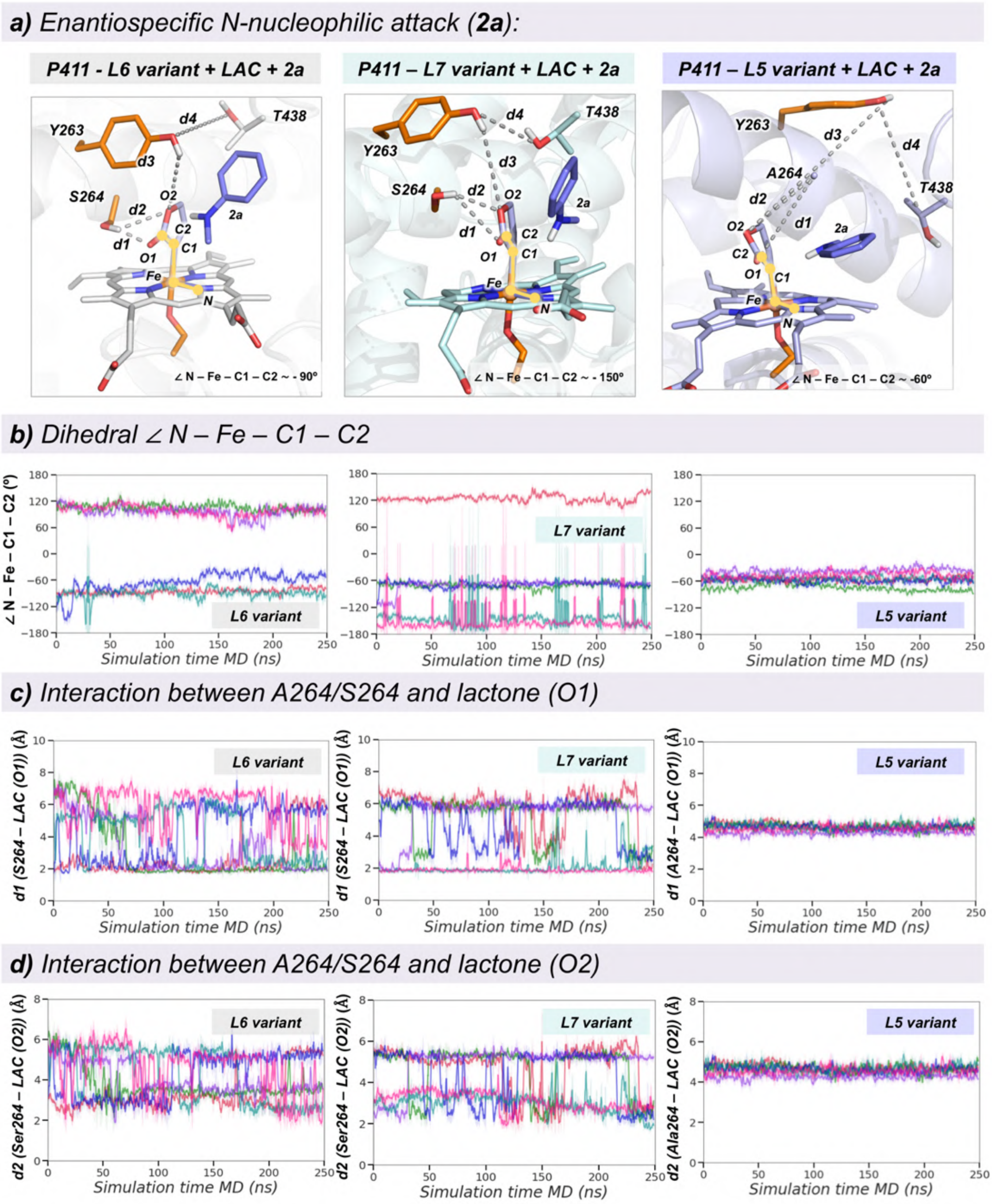


Figure B2. Binding of substrate 2a in a near attack conformation for N-nucleophilic attack. Representative snapshot corresponding to the most populated cluster extracted from constrained-MD simulations describing the binding of **2a** in a near attack conformation for the N-nucleophilic attack to the lactone carbene bound in variants P411-**L6** (gray), P411-**L7** (cyan) and P411-**L5** (purple). Substrate **2a** is shown in slate purple. The residues that are directly establishing hydrophobic interactions with the substrate are displayed as spheres. In P411-**L6** and P411-**L7**, hydrophobic interactions occurring between the aromatic ring of the aniline derivative and active site residues (L75, V328, L437, Y330, P329) stabilize this binding mode, while H-bond interactions between the carbene ester group and S264 are maintained (see Figure B3). On the other hand, in P411-**L5** variant the substrate is bound in a slightly different position in the active site, and interacts with residues P87, L75, V328, L437 and P329.



a) *Enantiospecific N-nucleophilic attack (**2a**):*

b) *Dihedral ∠ N – Fe – C1 – C2*

c) *Interaction between A264/S264 and lactone (O1)*

d) *Interaction between A264/S264 and lactone (O2)*

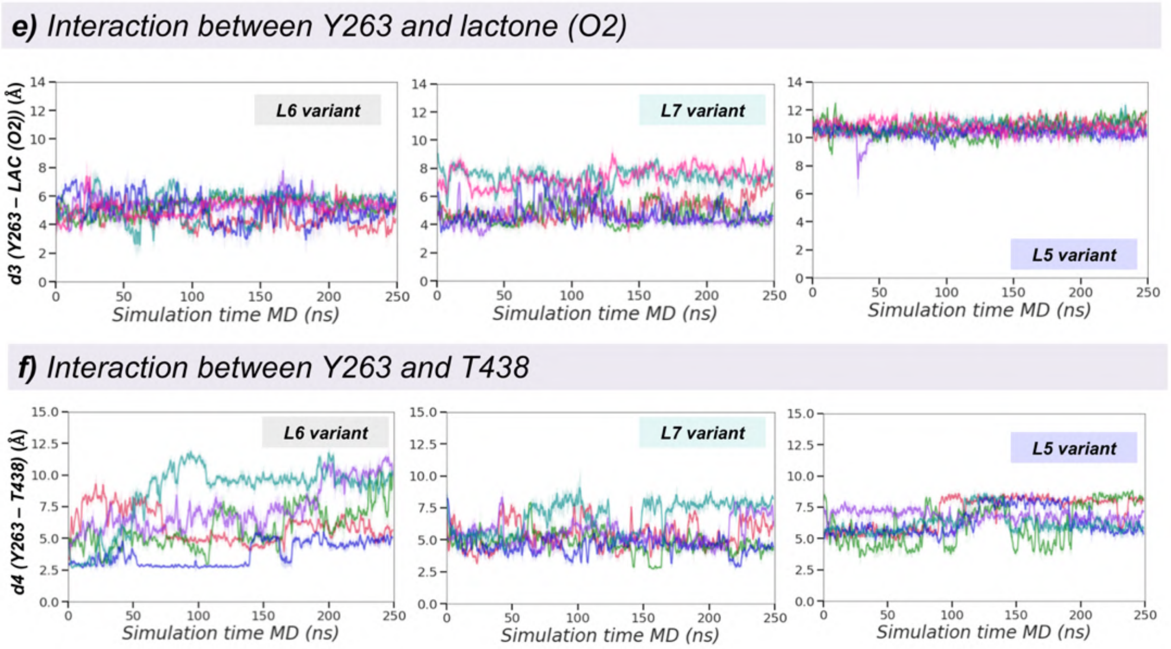


Figure B3. Analysis of substrate 2a binding in a near attack conformation for N-nucleophilic attack. a) Representative snapshot corresponding to the most populated cluster extracted from constrained-MD simulations describing the near attack conformations for the N-nucleophilic attack of **2a** to the lactone carbene in variants P411-**L6** (gray), P411-**L7** (cyan) and P411-**L5** (purple). **b)** The ∠(N – Fe – C1 – C2) dihedral angle measured along independent MD trajectories (2 different starting poses, 3 replicas of 250 ns for each) describes the relative orientation explored by the carbene in the presence of substrate **2a** in the active site. Lactone carbene does not reorient when the amine substrate is bound, and its initial conformation determines which face of the carbene will be exposed for the N-nucleophilic attack. **c)** Distance vs. time plot describing the H-bond interaction between the S264 hydroxyl group and the O1-oxygen of the lactone in P411-**L6** and P411-**L7**, and the A264 C_β and the O1-oxygen of the lactone in P411-**L5** (*d1*, as shown in Figure B3a). **d)** Distance vs. time plot describing the H-bond interaction between the S264 hydroxyl group and the O2-oxygen of the lactone in P411-**L6** and P411-**L7**, and the A264 C_β and the O2-oxygen of the lactone in P411-**L5** (*d2*, as shown in Figure B3a). **e)** Distance vs. time plot describing the H-bond interaction between the Y263 hydroxyl group and the O2-oxygen of the lactone in P411-**L6**, P411-**L7** and P411-**L5** (*d3*, as shown in Figure B3a). In P411-**L5**, the distance between Y263 and the lactone is significantly longer than in P411-**L6** and P411-**L7**. This affects to the positioning of specific water molecules in the active site (see also Figure B4) **f)** Distance vs. time plot describing the interaction between the Y263 hydroxyl group and the T438 hydroxyl group in P411-**L6**, P411-**L7** and P411-**L5** (*d4*, as shown in Figure B3a).

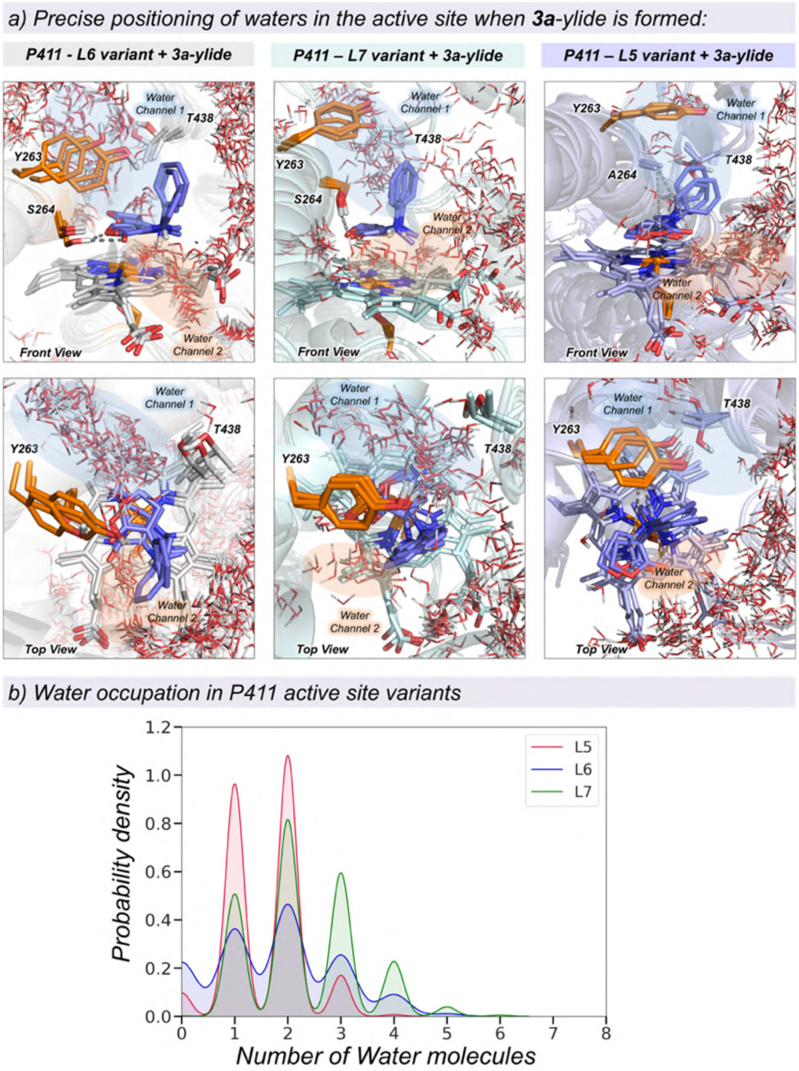
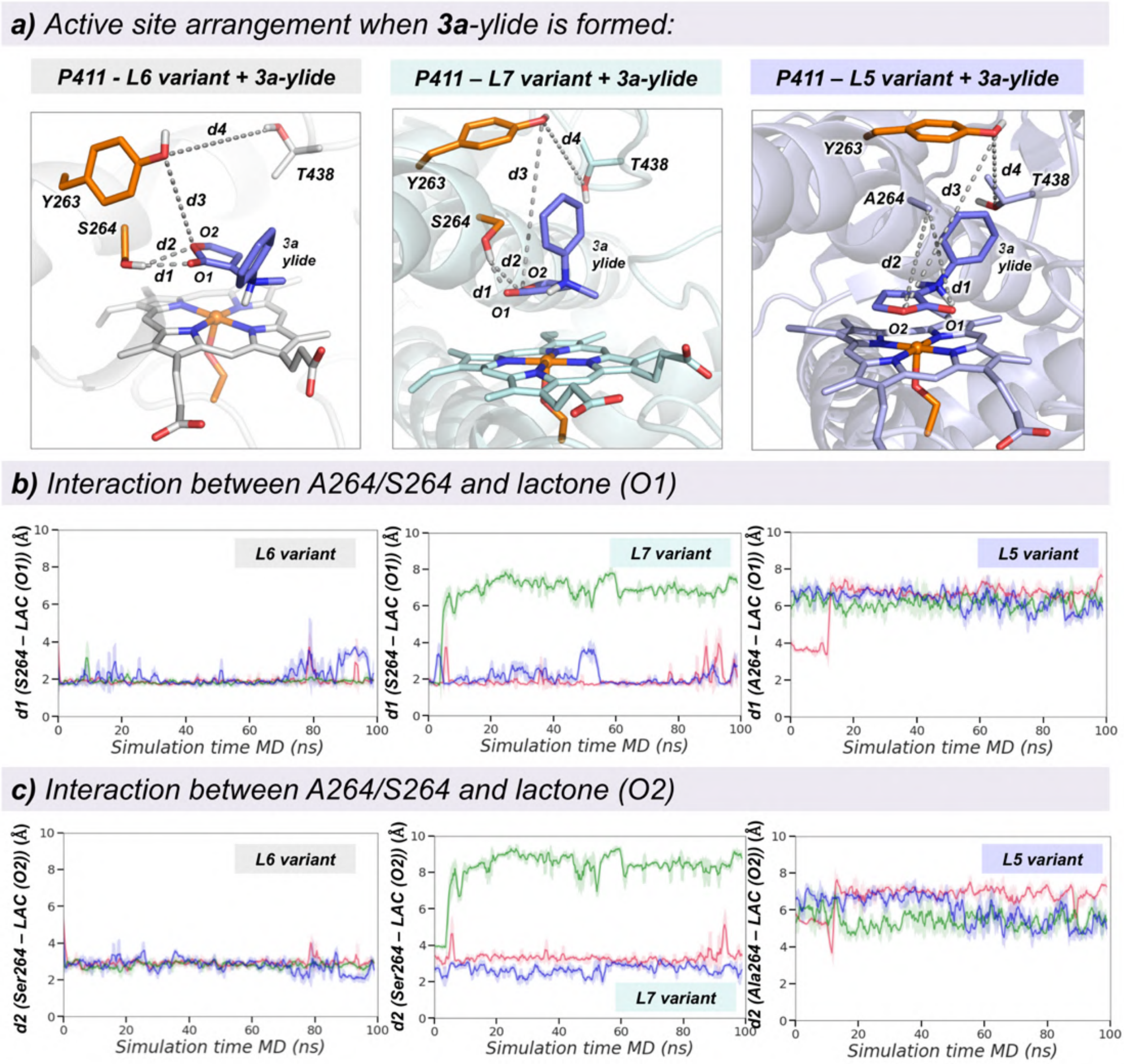

a) Precise positioning of waters in the active site when 3a-ylide is formed:
P411 - L6 variant + 3a-ylide
P411 – L7 variant + 3a-ylide
P411 – L5 variant + 3a-ylide
Water Channel 1
T438
Y263
S264
A264
Water Channel 2
Front View
Top View
b) Water occupation in P411 active site variants
Probability density
Number of Water molecules
L5
L6
L7
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
1
2
3
4
5
6
7
8

Figure B4. Precise positioning of waters in the active site upon 3a-ylide formation. a, Overlay of 3 representative snapshots from constrained-MD simulations exploring the active site arrangement in P411 variants when **3a**-ylide is formed. In P411-**L6** and P411-**L7**, water molecules are precisely positioned on the top-face of the lactone ring through *water channel 1* driven by Y263 and T438, and nearby the protonated ylide amine group through *water channel 2*. These water molecules are able to stereoselectively protonate the ylide intermediate from the pro-*S* face (see DFT model calculations in Figure B14). In P411-**L5** variant, water molecules cannot effectively access the top-face of lactone ring. Displayed water molecules are drawn from 25 random structures across the 100 ns MD trajectory. **b,** Representation of the normalized kernel density plot of the number of water molecules in the active site of P411 variants nearby the **3a**-ylide, considering its first solvation shell (using a distance cut-off of 3.4 Å). The average number of water molecules in the active site is 1.8 ± 1.2 (P411-**L6**), 2.3 ± 1.0 (P411-**L7**) and 1.6 ± 0.7 (P411-**L5**), respectively. The presence of water molecules is monitored through visual inspection of MD trajectories and using the *watershell* function of cpptraj.[6]



a) *Active site arrangement when* ***3a****-ylide is formed:*

P411 - L6 variant + 3a-ylide

P411 – L7 variant + 3a-ylide

P411 – L5 variant + 3a-ylide

b) *Interaction between A264/S264 and lactone (O1)*

c) *Interaction between A264/S264 and lactone (O2)*

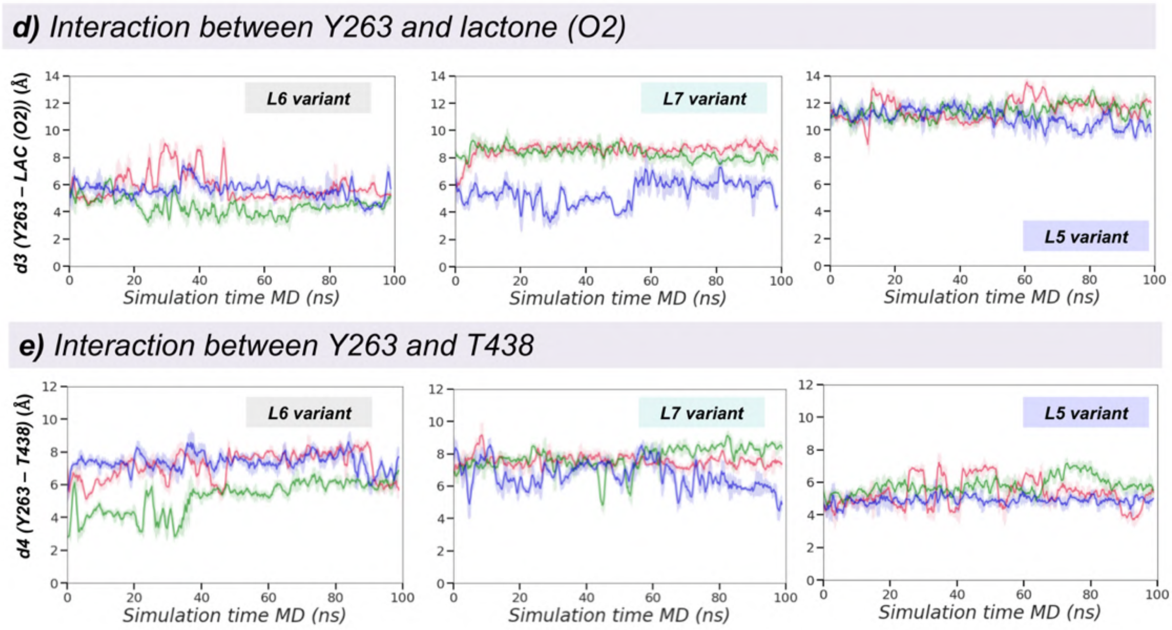


Figure B5. Analysis of active site arrangement when 3a-ylide is formed. a, Representative snapshot corresponding to the most populated cluster extracted from constrained-MD simulations describing the active site arrangement when **3a**-ylide is formed (right after dissociation from Fe) in variants P411-**L6** (gray), P411-**L7** (cyan) and P411-**L5** (purple). **b,** Distance vs. time plot describing the H-bond interaction between the S264 hydroxyl group and the O1-oxygen of the lactone in P411-**L6** and P411-**L7**, and the A264 C_β and the O1-oxygen of the lactone in P411-**L5** (*d1*, as shown in Figure B5a). **c,** Distance vs. time plot describing the H-bond interaction between the S264 hydroxyl group and the O2-oxygen of the lactone in P411-**L6** and P411-**L7**, and the A264 C_β and the O2-oxygen of the lactone in P411-**L5** (*d2*, as shown in Figure B5a). **d,** Distance vs. time plot describing the H-bond interaction between the hydrogen of the Y263 hydroxyl group and the O2-oxygen of the lactone in P411-**L6**, P411-**L7** and P411-**L5** (*d3*, as shown in Figure B5a). In P411-**L5**, the distance between Y263 and the lactone is significantly longer than in P411-**L6** and P411-**L7**. This prevents the precise positioning of water molecules in the active site. **e,** Distance vs. time plot describing the interaction between the oxygen of the Y263 hydroxyl group and the T438 hydroxyl group in P411-**L6**, P411-**L7** and P411-**L5** (*d4*, as shown in Figure B5a).

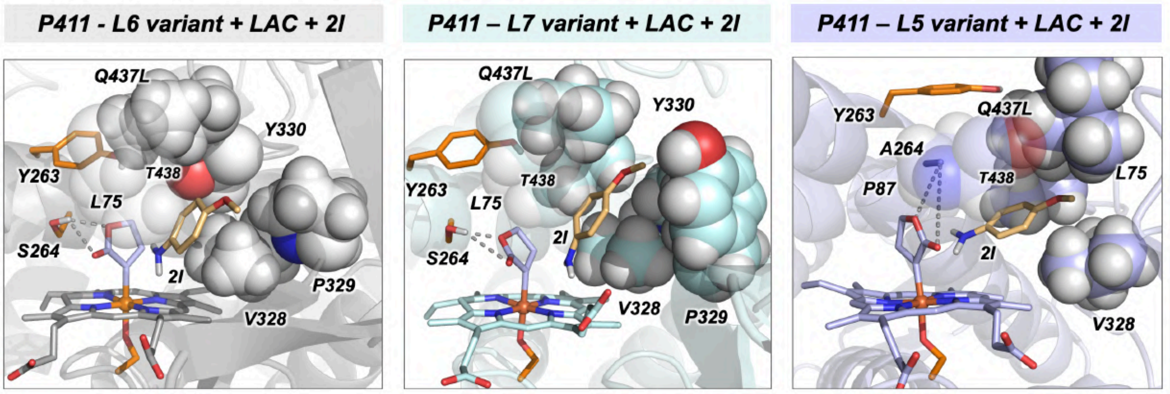
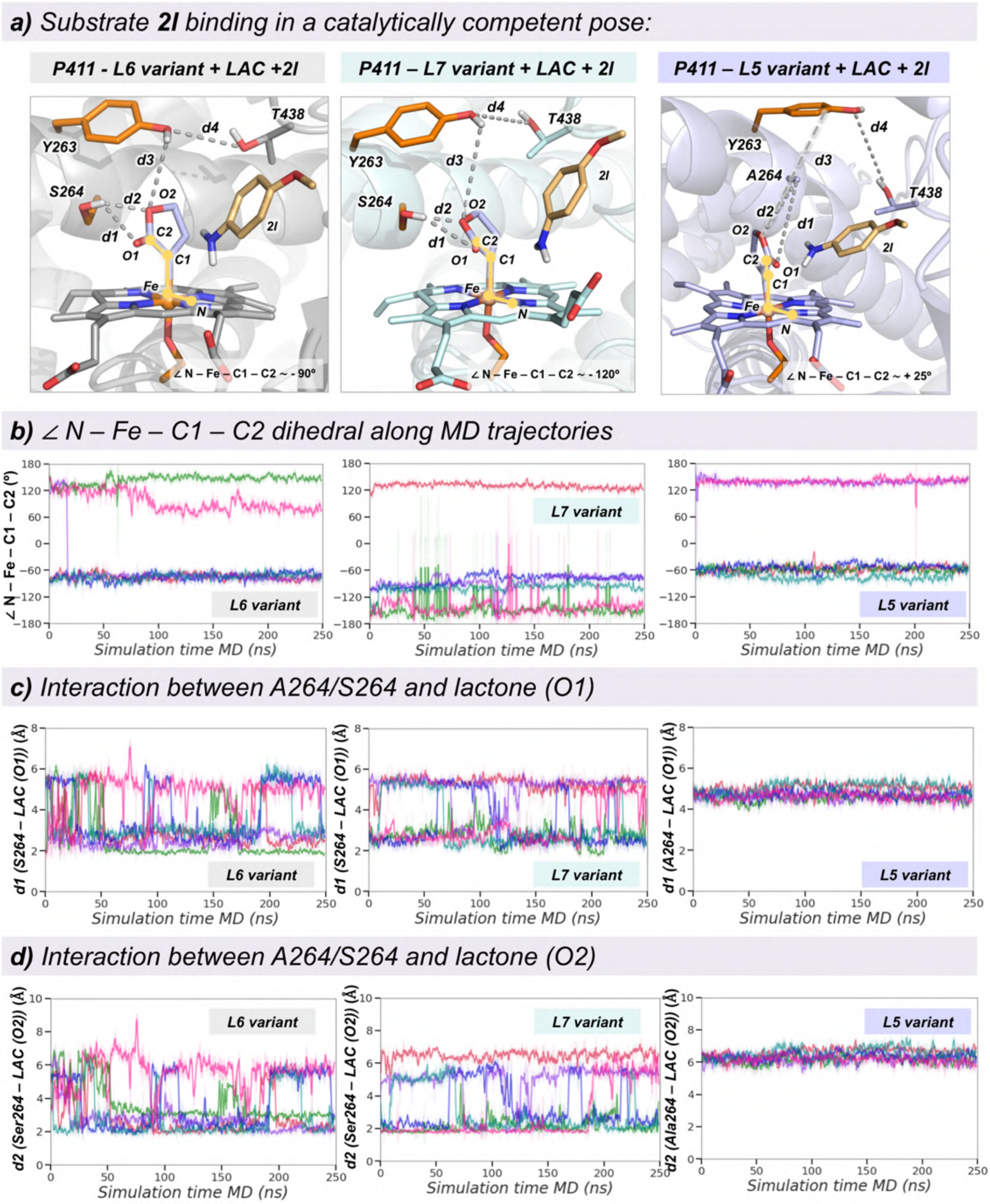


Figure B6. Binding of substrate 2l in a near attack conformation for N-nucleophilic attack. Representative snapshot corresponding to the most populated cluster extracted from constrained-MD simulations describing the binding of **2l** in a near attack conformation for the N-nucleophilic attack to the lactone carbene bound in variants P411-**L6** (gray), P411-**L7** (cyan) and P411-**L5** (purple). Substrate **2l** is colored in light orange. The residues that are directly establishing hydrophobic interactions with the substrate are displayed as spheres. In P411-**L6** and P411-**L7**, hydrophobic interactions occurring between the aromatic ring of the aniline derivative and active site residues (L75, V328, L437, Y330, P329) stabilize this binding mode, while H-bond interactions between the carbene ester group and S264 are maintained (see Figure B7). On the other hand, in P411-**L5** the substrate is bound in a slightly different position in the active site, and interacts with residues P87, L75, V328, L437 and P329.



a) Substrate **2l** binding in a catalytically competent pose:

b) ∠ N – Fe – C1 – C2 dihedral along MD trajectories

c) Interaction between A264/S264 and lactone (O1)

d) Interaction between A264/S264 and lactone (O2)

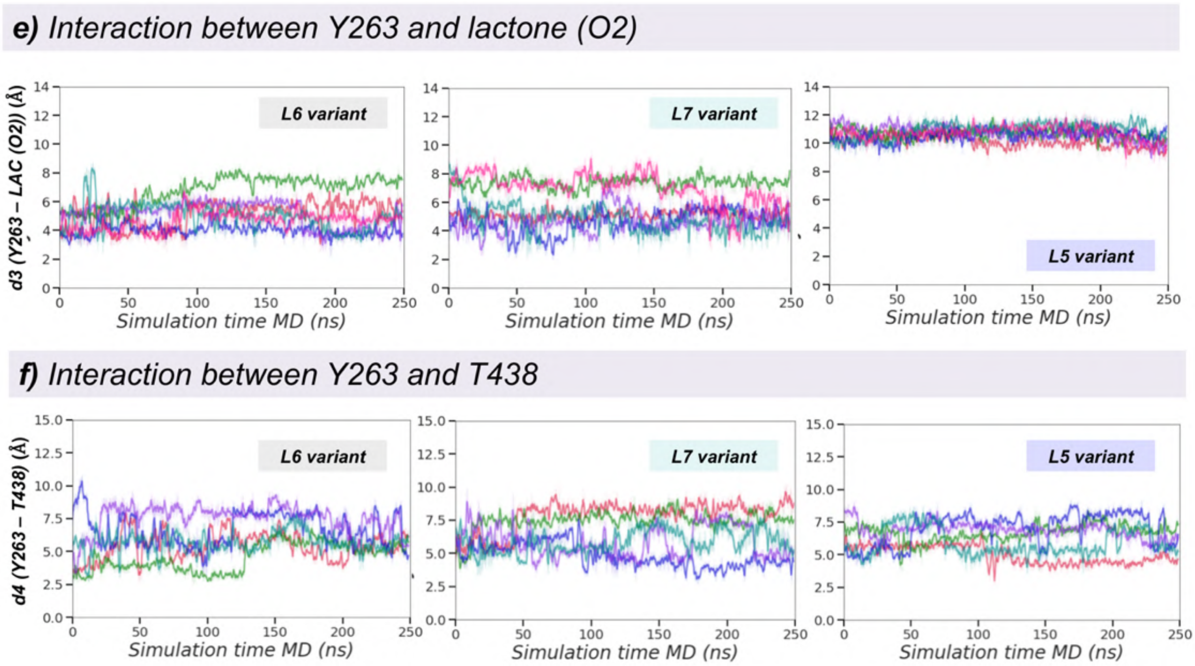


Figure B7. Analysis of substrate 2l binding in a near attack conformation for N-nucleophilic attack. a, Representative snapshot corresponding to the most populated cluster extracted from constrained-MD simulations describing the near attack conformations for the N-nucleophilic attack of **2l** to the lactone carbene in variants P411-**L6** (gray), P411-**L7** (cyan) and P411-**L5** (purple). **b,** The ∠(N – Fe – C1 – C2) dihedral angle measured along independent MD trajectories (2 different starting poses, 3 replicas of 250 ns for each) describes the relative orientation explored by the carbene in the presence of substrate **2l** in the active site. Lactone carbene does not reorient when the amine substrate is bound, and its initial conformation determines which face of the carbene will be exposed for the N-nucleophilic attack. **c,** Distance vs. time plot describing the H-bond interaction between the S264 hydroxyl group and the O1-oxygen of the lactone in P411-**L6** and P411-**L7**, and the A264 C_β and the O1-oxygen of the lactone in P411-**L5** (*d1*, as shown in Figure B7a). **d,** Distance vs. time plot describing the H-bond interaction between the S264 hydroxyl group and the O2-oxygen of the lactone in P411-**L6** and P411-**L7**, and the A264 C_β and the O2-oxygen of the lactone in P411-**L5** (*d2*, as shown in Figure B7a). **e,** Distance vs. time plot describing the H-bond interaction between the Y263 hydroxyl group and the O2-oxygen of the lactone in P411-**L6**, P411-**L7** and P411-**L5** (*d3*, as shown in Figure B7a). In P411-**L5**, the distance between Y263 and the lactone is significantly longer than in P411-**L6** and P411-**L7**. This affects to the positioning of specific water molecules in the active site (see also Figure B8) **f,** Distance vs. time plot describing the interaction between the Y263 hydroxyl group and the T438 hydroxyl group in P411-**L6**, P411-**L7** and P411-**L5** (*d4*, as shown in Figure B7a).

These observations for substrate **2l** are equivalent to those reported for substrate **2a** (Figure B3).

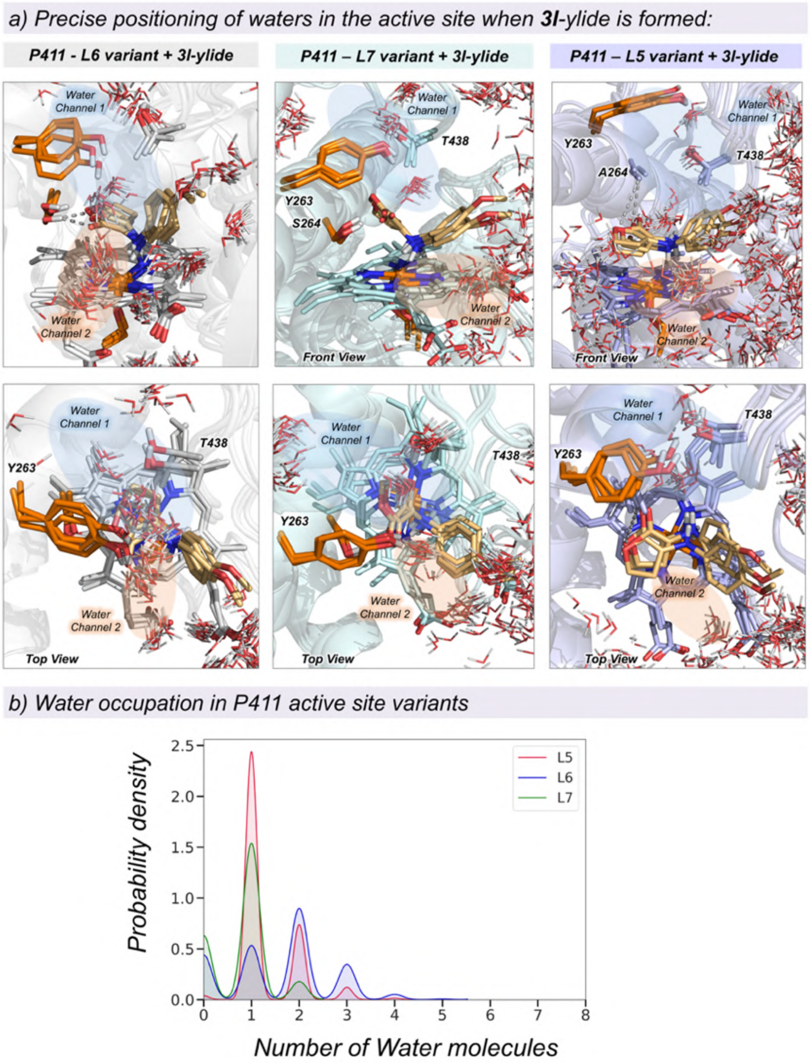
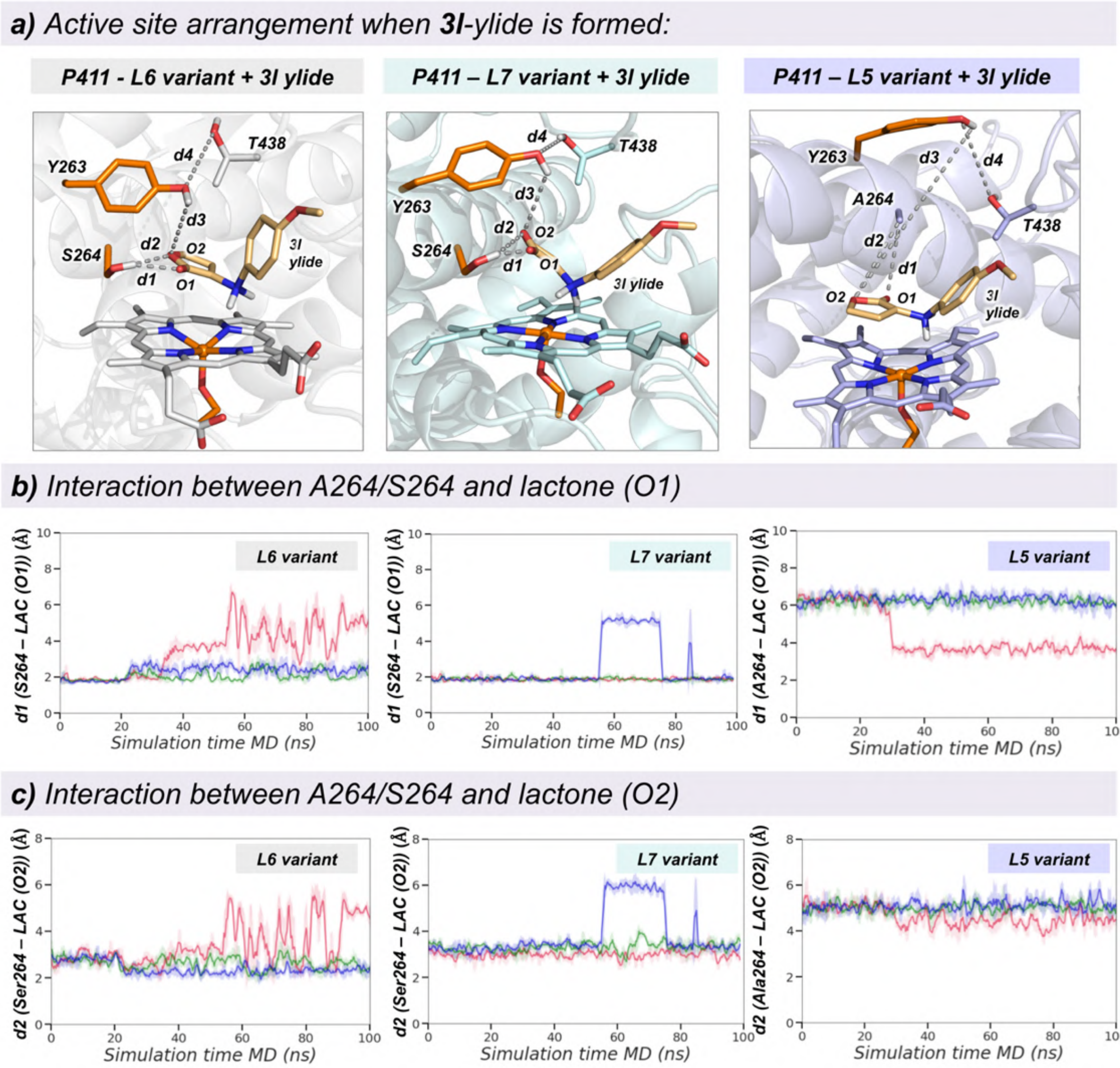


Figure B8. Precise positioning of waters in the active site upon 3l-ylide formation. a, Overlay of 3 representative snapshots from constrained-MD simulations exploring the active site arrangement in P411 variants when **3l**-ylide is formed. In P411-**L6** and P411-**L7**, water molecules are precisely positioned on the top-face of the lactone ring through *water channel 1* driven by Y263 and T438, and nearby the protonated ylide amine group through *water channel 2*. These water molecules are able to stereoselectively protonate the ylide intermediate from the pro-*S* face (see DFT model calculations in Figure B15). In P411-**L5** variant, water molecules cannot effectively access the top-face of lactone ring. Displayed water molecules are drawn from 25 random structures across the 100 ns MD trajectory. **b,** Representation of the normalized kernel density plot of the number of water molecules in the active site of P411 variants nearby the **3l**-ylide, considering its first solvation shell (using a distance cut-off of 3.4 Å). The average number of water molecules in the active site is 1.6 ± 1.1 (P411-**L6**), 0.8 ± 0.5 (P411-**L7**) and 1.3 ± 0.6 (P411-**L5**), respectively. The presence of water molecules is monitored through visual inspection of MD trajectories and using the *watershell* function of cpptraj.[6]

The results obtained for **3l**-ylide are very similar to those observed for **3a**-ylide (Figure B4).



***a)** Active site arrangement when **3l**-ylide is formed:*

***b)** Interaction between A264/S264 and lactone (O1)*

***c)** Interaction between A264/S264 and lactone (O2)*

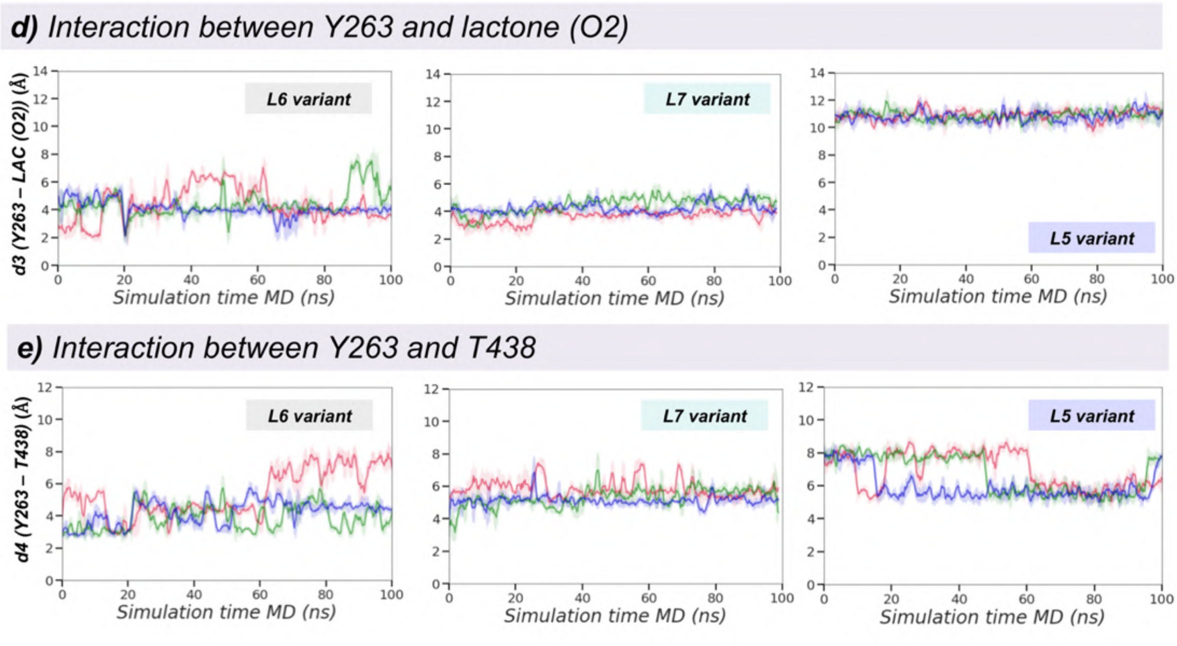


Figure B9. Analysis of active site arrangement when 3l-ylide is formed. a, Representative snapshot corresponding to the most populated cluster extracted from constrained-MD simulations describing the active site arrangement when **3l**-ylide is formed (right after dissociation from Fe) in variants P411-**L6** (gray), P411-**L7** (cyan) and P411-**L5** (purple). **b,** Distance vs. time plot describing the H-bond interaction between the S264 hydroxyl group and the O1-oxygen of the lactone in P411-**L6** and P411-**L7**, and the A264 C_β and the O1-oxygen of the lactone in P411-**L5** (*d1*, as shown in Figure B9a). **c,** Distance vs. time plot describing the H-bond interaction between the S264 hydroxyl group and the O2-oxygen of the lactone in P411-**L6** and P411-**L7**, and the A264 C_β and the O2-oxygen of the lactone in P411-**L5** (*d2*, as shown in Figure B9a). **d,** Distance vs. time plot describing the H-bond interaction between the Y263 hydroxyl group and the O2-oxygen of the lactone in P411-**L6**, P411-**L7** and P411-**L5** (*d3*, as shown in Figure B9a). In P411-**L5**, the distance between Y263 and the lactone is significantly longer than in P411-**L6** and P411-**L7**. This prevents the precise positioning of water molecules in this region of the active site. **e,** Distance vs. time plot describing the interaction between the Y263 hydroxyl group and the T438 hydroxyl group in P411-**L6**, P411-**L7** and P411-**L5** (*d4*, as shown in Figure B9a).

The conclusions obtained for **3l**-ylide are very similar to those for **3a**-ylide (Figure B5).

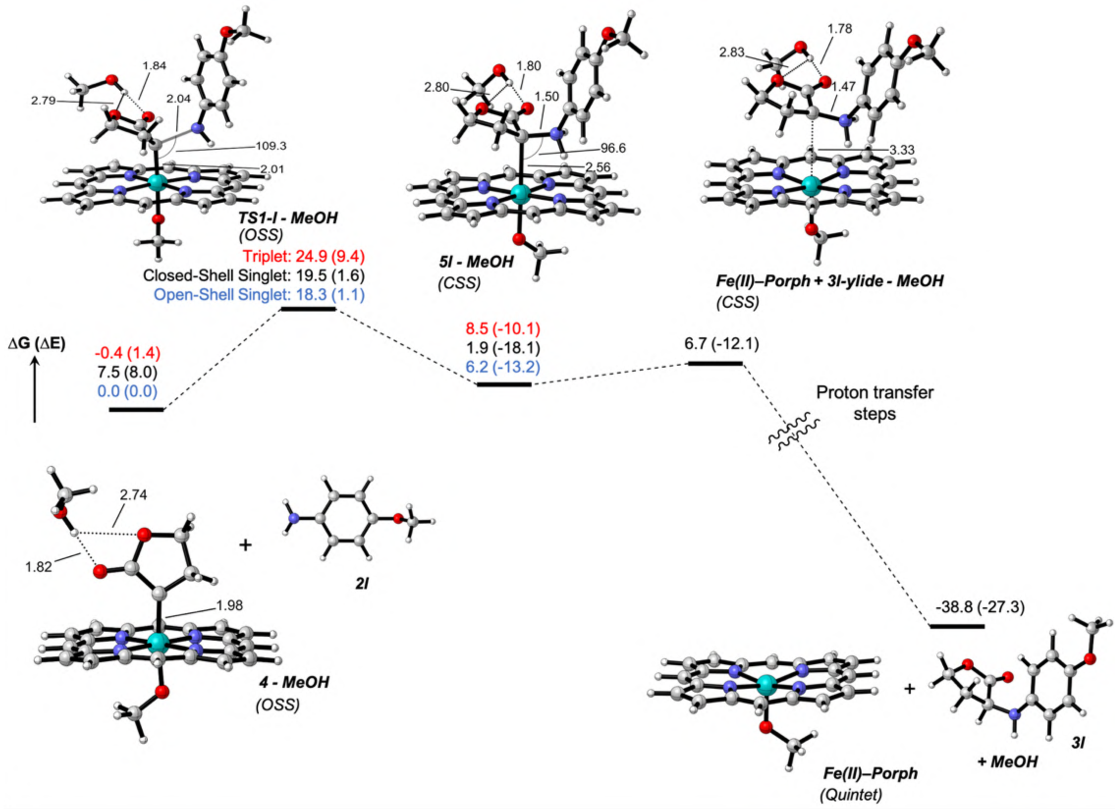


Figure B10. Energy profile for lactone carbene N–H insertion involving model substrate **2l**. A truncated model that includes a methanol molecule to mimic P411-**L6** active site S264 has been used. Results obtained for different spin states are reported and the lowest in energy optimized geometries for each stationary point are shown. **3l**-ylide dissociation from **5l** – MeOH is found to be barrierless (see Figure B12). Proton transfer steps to yield the final product **3l** are studied in Figure B13 and B14. Electronic and Gibbs free energies are obtained at B3LYP-D3BJ/def2-TZVP//B3LYP/6-31G(d)-SDD. Energy values are given in kcal·mol^{-1}. Key distances and angles are given in Å and deg., respectively.

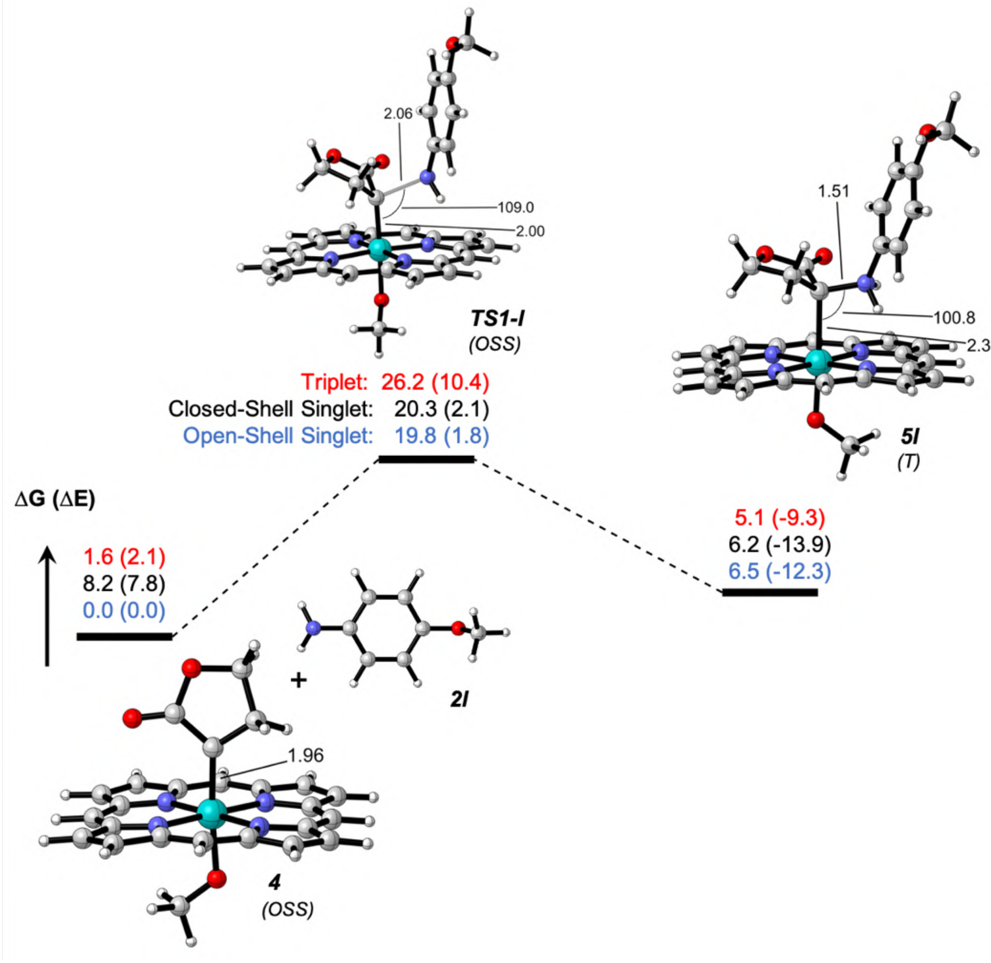


Figure B11. N-nucleophilic attack step for carbene N–H insertion reaction involving substrate **2l** in the absence of H-bond interaction with methanol molecule mimicking P411-**L6** active site S264. Results obtained for different spin states are reported. The lowest in energy optimized geometries for each stationary point are shown. Electronic and Gibbs free energies are obtained at B3LYP-D3BJ/def2-TZVP// B3LYP/6-31G(d)-SDD. Energy values are given in kcal·mol^{-1}. Key distances and angles are given in Å and deg., respectively.

The presence of MeOH (to mimic S264 H-bonding to the lactone ester group) decreases the N-nucleophilic attack barrier by ca. 1.5 kcal·mol^{-1} ($\Delta G^{\ddagger}_{(\text{MeOH-model})}$ = 18.3 kcal·mol^{-1} from Figure B9, vs $\Delta G^{\ddagger}$ = 19.8 kcal·mol^{-1}). The presence of a H-bond interaction with the carbene ester group activating it is in line with observations from previous works.[26,36]

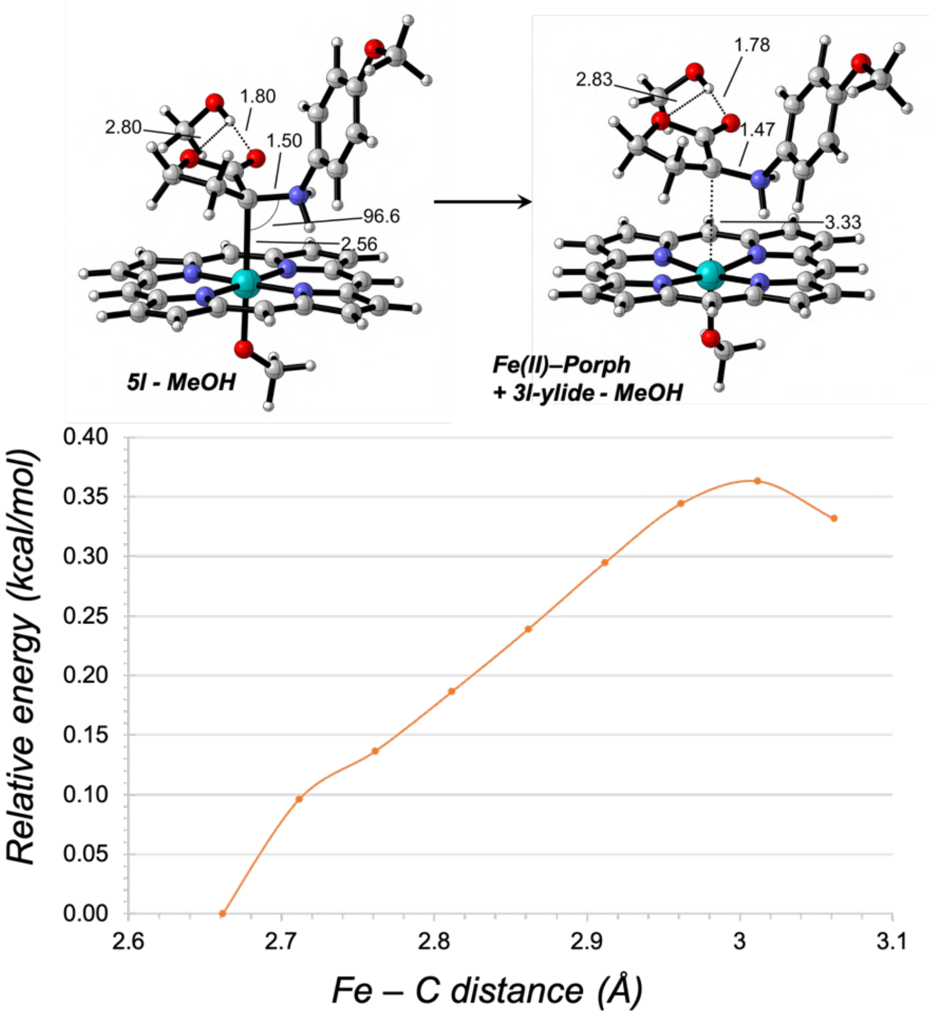
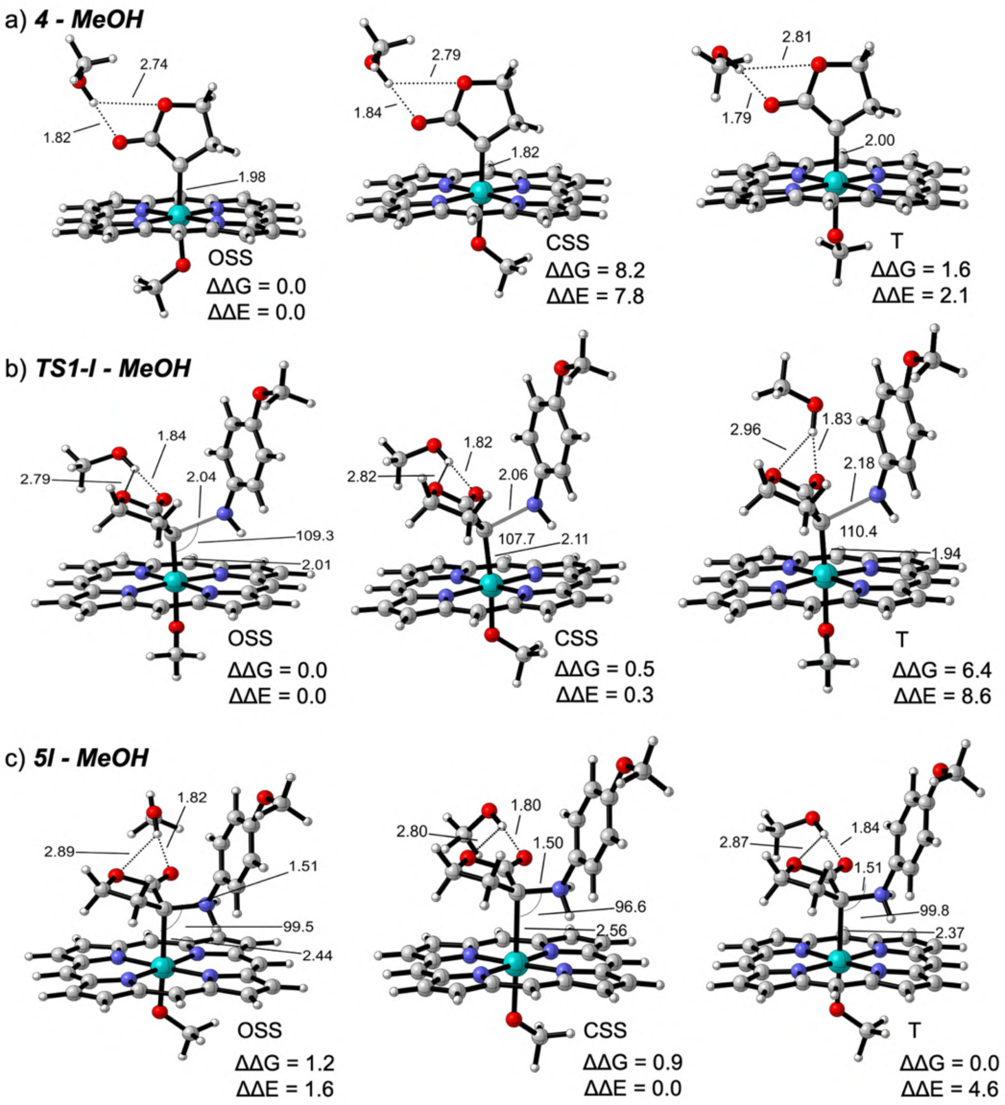


Figure B12. Potential energy scan (electronic energy) along Fe – C coordinate that describes **3l**-ylide dissociation, in the lowest energy closed-shell singlet electronic state, calculated at B3LYP/6-31G(d)-SDD level. The energies obtained at the lower computational level and do not include ZPE, Gibbs, or dispersion corrections. Relative energies are given in kcal·mol^{-1} and distances in Å.

The potential energy scan indicates a barrierless dissociation of the covalent **5l** intermediate to generate the ylide. All the attempts to optimize a dissociation transition state were unsuccessful. This is in line with previous work on a similar reaction.[22] No spontaneous proton transfer to form the enol is observed during dissociation (see also Figure B16).


a) 4 - MeOH
2.74
1.82
1.98
OSS
ΔΔG = 0.0
ΔΔE = 0.0
2.79
1.84
1.82
CSS
ΔΔG = 8.2
ΔΔE = 7.8
2.81
1.79
2.00
T
ΔΔG = 1.6
ΔΔE = 2.1
b) TS1-I - MeOH
1.84
2.79
2.04
109.3
2.01
OSS
ΔΔG = 0.0
ΔΔE = 0.0
1.82
2.82
2.06
107.7
2.11
CSS
ΔΔG = 0.5
ΔΔE = 0.3
2.96
1.83
2.18
110.4
1.94
T
ΔΔG = 6.4
ΔΔE = 8.6
c) 5I - MeOH
1.82
2.89
1.51
99.5
2.44
OSS
ΔΔG = 1.2
ΔΔE = 1.6
1.80
2.80
1.50
96.6
2.56
CSS
ΔΔG = 0.9
ΔΔE = 0.0
2.87
1.84
1.51
99.8
2.37
T
ΔΔG = 0.0
ΔΔE = 4.6

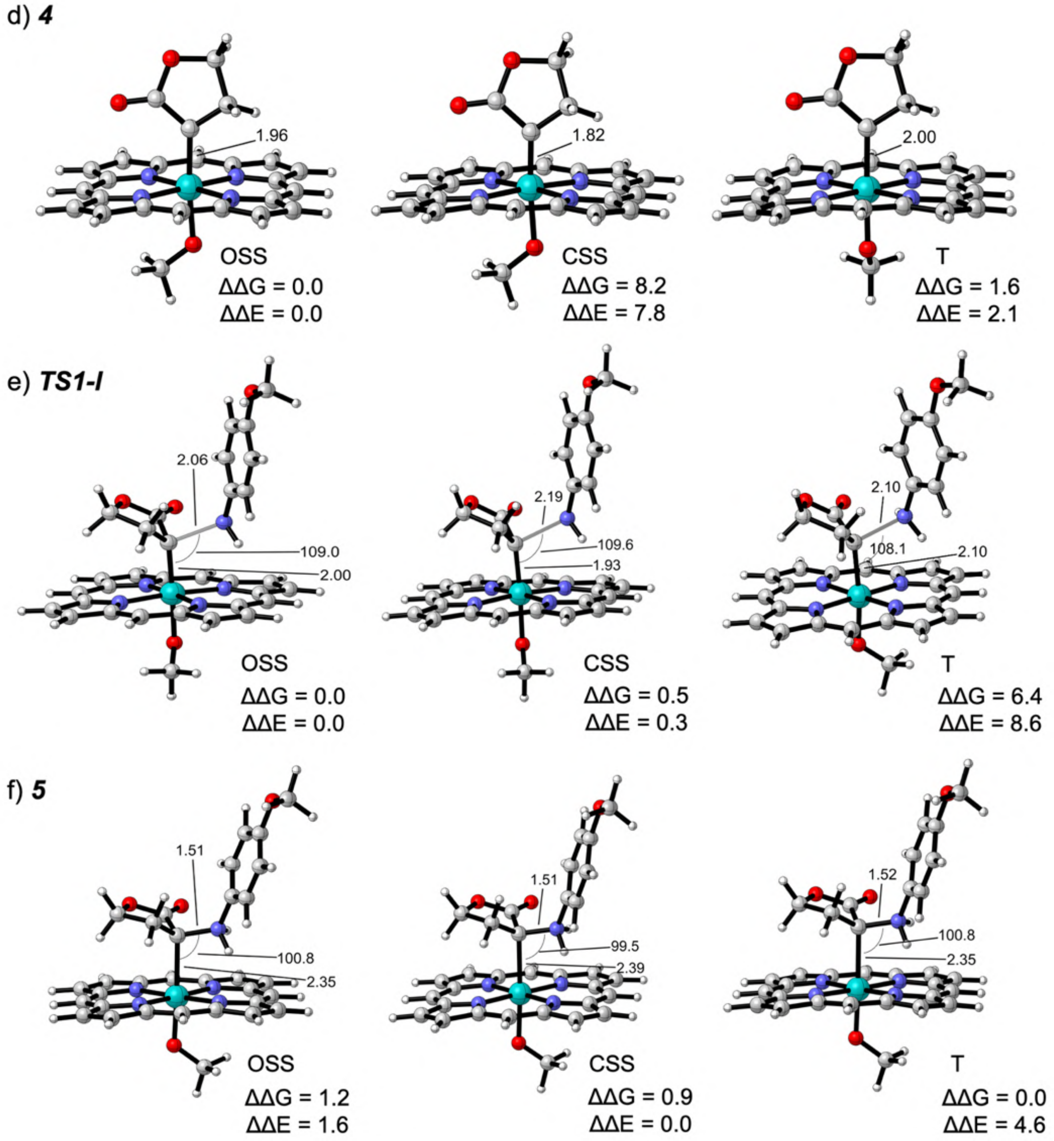


Figure B13. Optimized geometries for the different species reported in Figure B9 and B10 in their different electronic states. Relative electronic and Gibbs free energies are obtained at B3LYP-D3BJ/def2-TZVP// B3LYP/6-31G(d)-SDD. Energy values are given in kcal·mol^{-1}. Key distances and angles are given in Å and deg., respectively.

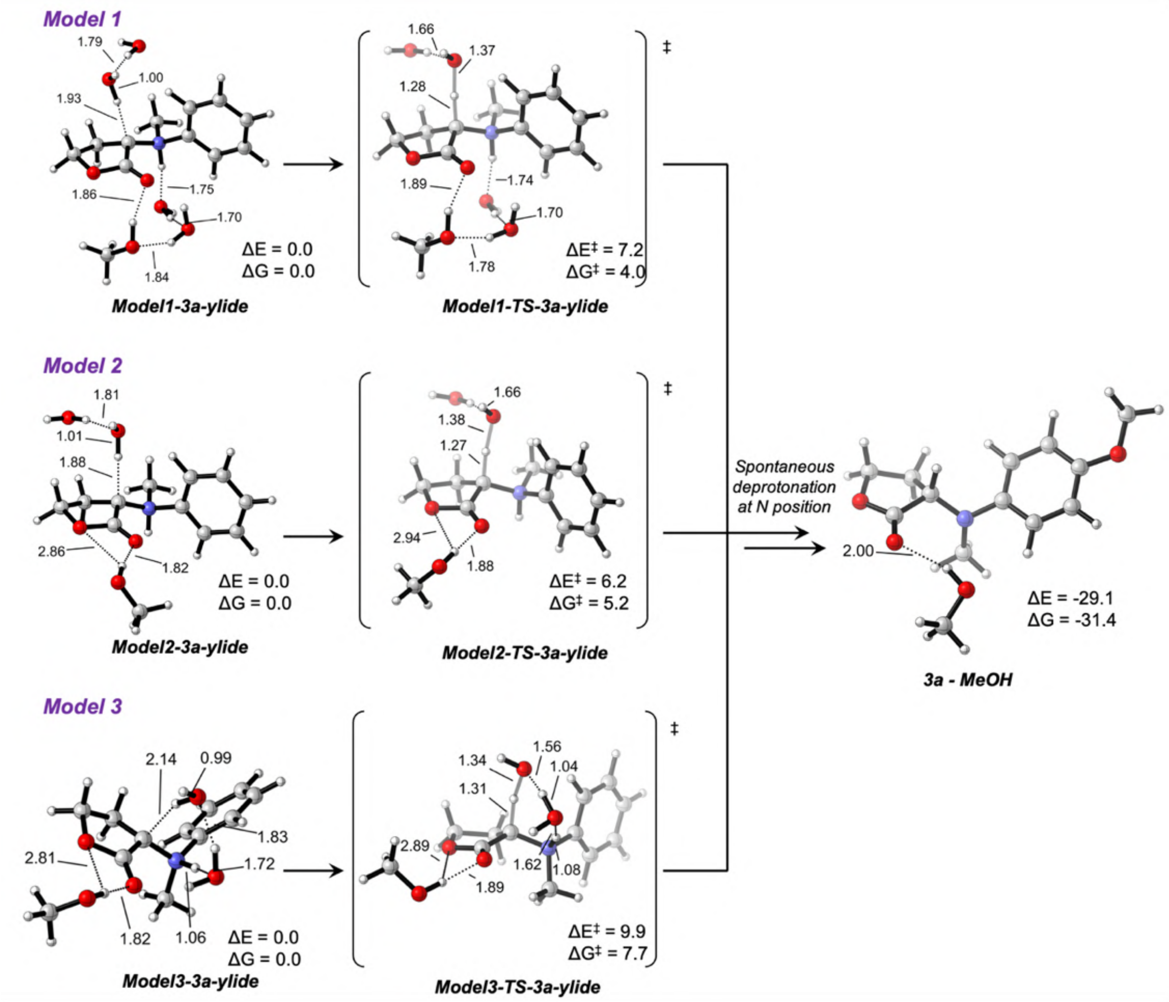


Figure B14. Optimized model transition states (TSs) for stereoselective 3a formation from 3a-ylide. Computational models are built based on the conformations explored by the **3a**-ylide when formed in P411-**L6** active site and the arrangement of water molecules around the ylide as observed from MD simulations (Figure B4). Electronic and Gibbs free energies are obtained at B3LYP-D3BJ/def2-TZVP// B3LYP/6-31G(d)-SDD. Energy values are given in kcal·mol^{-1}. Key distances and angles are given in Å and deg., respectively. Optimizations are carried out in the absence of any geometrical constraint.

Model 1 includes a methanol molecule (model for active site S264 residue) and four water molecules: two on the top face (pro-*S* face) of the lactone, and two near the protonated amine.

Model 2 includes a methanol molecule (model for active site S264 residue) and two water molecules on the top face (pro-*S* face) of the lactone ring.

Model 3 includes a methanol molecule (model for active site S264 residue) and two water molecules: one on the top face (pro-*S* face) of the lactone, and another one near the protonated amine. Although the two water molecules are not directly interacting in the initial structure, they establish a new H-bond interaction upon geometry optimization. Additionally, the N–C(lactone) bond rotates during optimization to establish a new H-bond network between these two waters and the amine. All these interactions could probably not occur in the enzyme active site due to geometric restraints imposed to the ylide.

The low activation barriers calculated specially for Model 1 and Model 2 are indicating that this proton transfer step can rapidly take place in the active site once the ylide dissociates from the iron.

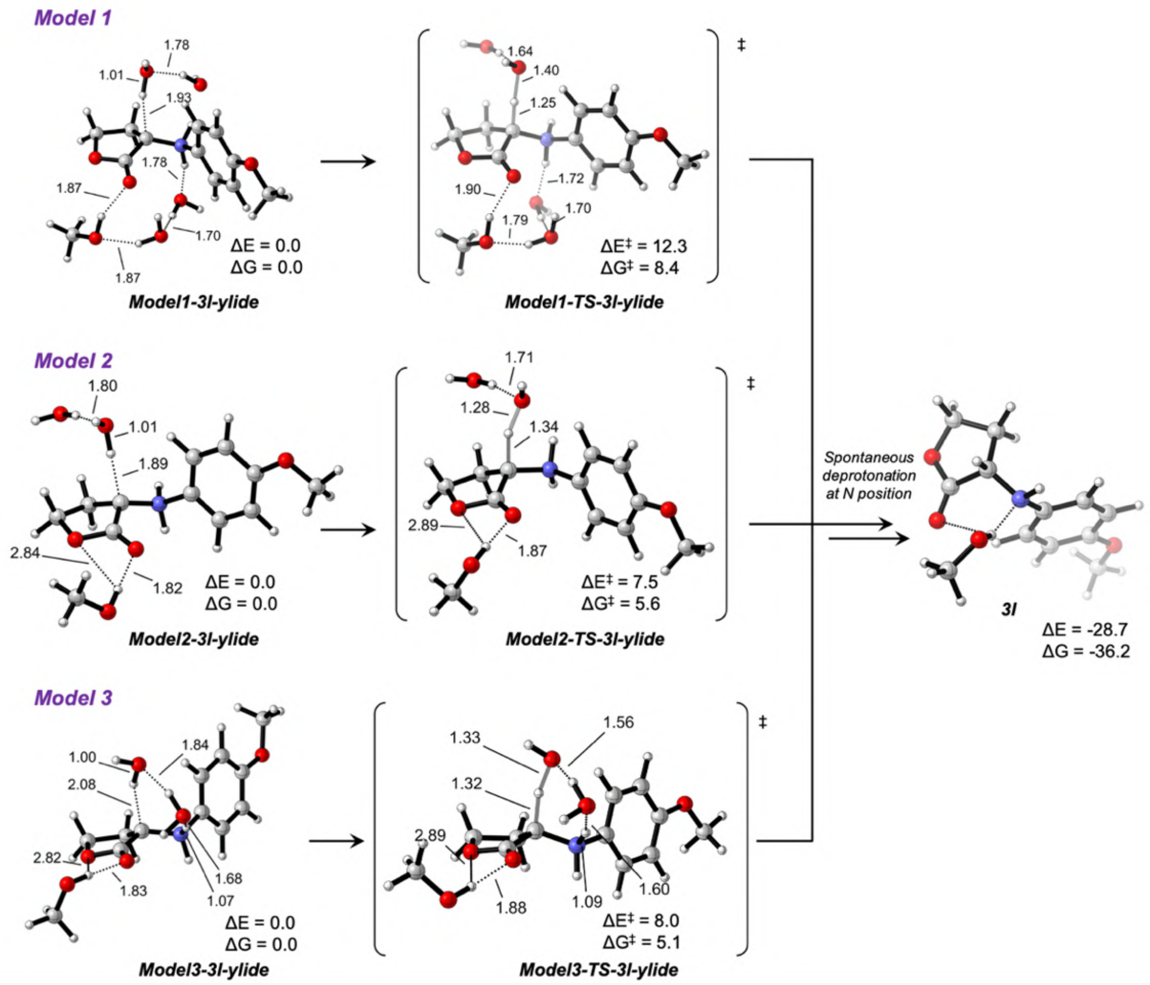


Figure B15. Optimized model transition states (TSs) for stereoselective 3l formation from 3l-ylide. Computational models are built based on the conformations explored by the **3l**-ylide when formed in P411-**L6** active site and the arrangement of water molecules around the ylide as observed from MD simulations (Figure B8). Electronic and Gibbs free energies are obtained at B3LYP-D3BJ/def2-TZVP// B3LYP/6-31G(d)-SDD. Energy values are given in kcal·mol^{-1}. Key distances and angles are given in Å and deg., respectively. Optimizations are carried out in the absence of any geometrical constraint. Models were constructed as described in Figure B14.

As found for *N*-methylated **3a**-ylide, model calculations for the enantioselective proton transfer involving **3l**-ylide describe low activation barriers calculated, indicating that this proton transfer step can rapidly take place in the active site once the ylide dissociates from the iron.

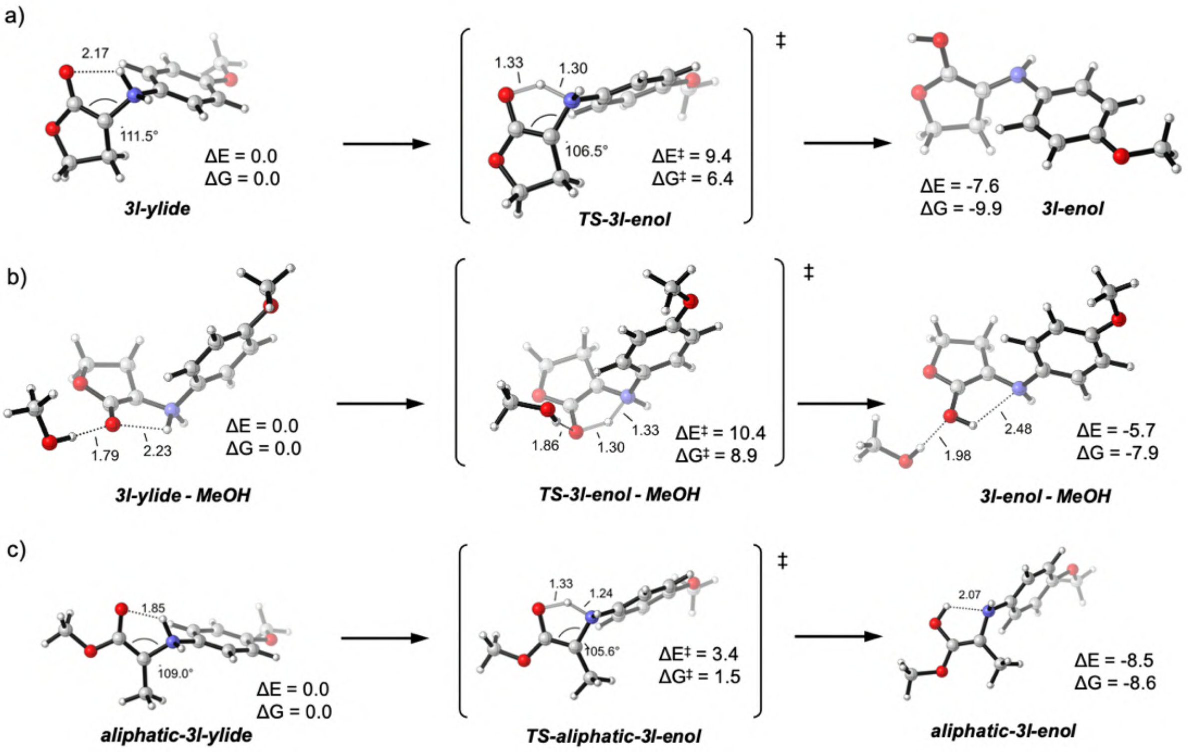


Figure B16. **Computational explorations of enol formation from ylide considering 3l-ylide as model substrate. a)** Intramolecular **3l**-enol formation from **3l**-ylide. **b)** Equivalent enol formation from **3l**-ylide but considering the H-bond interactions that the lactone carbonyl is establishing with a molecule of methanol that mimics the side chain of active site S264 residue in P411-**L6**. **c)** Intramolecular enol formation from a model aliphatic-**3l**-ylide. Electronic and Gibbs free energies are obtained at B3LYP-D3BJ/def2-TZVP// B3LYP/6-31G(d)-SDD. Energy values are given in kcal·mol^{-1}. Optimizations are carried out in the absence of any geometrical constraint. Key distances and angles are given in Å and deg., respectively.

In a previous computational study on similar N–H carbene insertion reactions catalyzed by Fe-porphyrin,[22] it was described that dissociation of the ylide form the Fe center could involve a spontaneous proton transfer to form the corresponding enol. This proton transfer could occur during the dissociation, or through a very low in energy H-transfer step. In that case, the studied carbene includes an aliphatic ester, which is more flexible than the current lactone used in this study.

In our current case, no spontaneous enol formation is observed upon ylide dissociation from Fe (Figure B12). To rationalize these differences, we studied the intramolecular ylide-enol rearrangement using the models described in Figure B16a-c.

Enol formation from **3l**-ylide (B16a) has a much higher barrier than enol formation when an aliphatic ester is present (B16c). This is due to the higher strain on the 5-membered ring formed during the H-transfer transition state in the lactone system, than in the much more flexible aliphatic ester system. In addition, the presence of an additional H-bond interaction involving the carbonyl group of the lactone (B16b), stabilizes the ylide form with respect to the enol and increases the intramolecular proton transfer barrier. Consequently, in this particular system the interaction established by the active site S264 side chain and the lactone disfavors even more the formation of the enol from the ylide when it is formed in the enzyme active site.

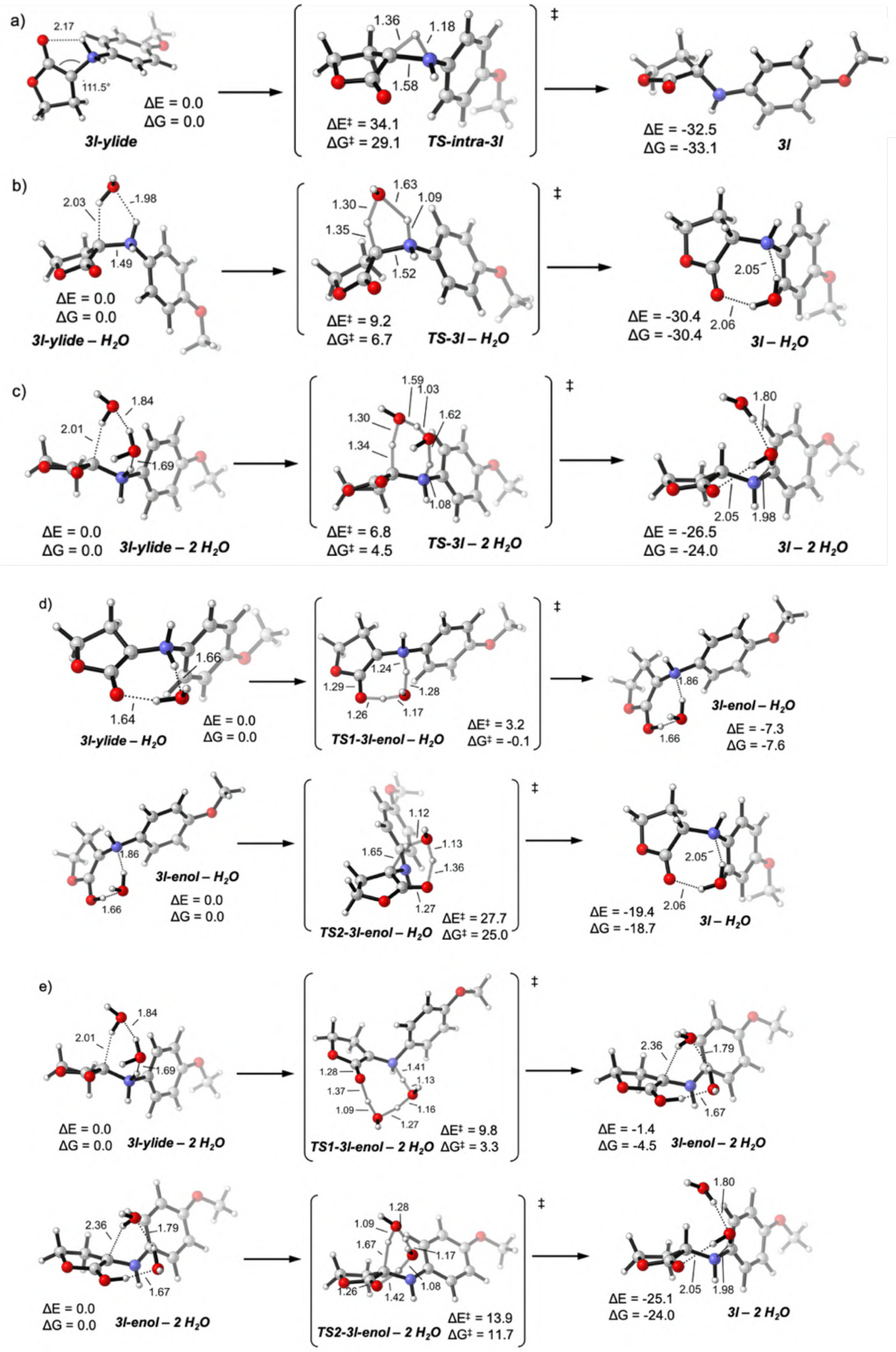


Figure B17. Alternative proton transfer rearrangements explored for model **3l**-ylide in the absence of geometrical restraints imposed by the enzyme active site (*i.e.,* considering any possible geometry and water assistance) to stereoselectively yield **3l** product. **a**) Direct intramolecular proton transfer; **b**) and **c**) Proton transfer assisted by one or two water molecules, respectively; **d**) and **e**) first enol formation, followed by subsequent product formation assisted by one or two water molecules, respectively. Electronic and Gibbs free energies are obtained at B3LYP-D3BJ/def2-TZVP//B3LYP/6-31G(d)-SDD. Energy values are given in kcal·mol^{-1}. Optimizations are carried out in the absence of any geometrical constraint. Key distances and angles are given in Å and deg., respectively.

- Direct intramolecular proton transfer to yield product **3l** has a very high barrier (B17a).
- Intramolecular proton transfer assisted by one or two water molecules from the ylide to yield product **3l** would be possible (B17b-c). However, the required conformations for the ylide and the specific positioning of the water molecules respect to the migrating proton from the amine and the lactone carbon are not possible in the enzyme active site (Figure B8).
- Although enol formation from **3l**-ylide via direct intramolecular proton transfer is very high in energy (Figure B16), this proton transfer assisted by a single water molecule would be feasible if the catalytic water molecule could interreact in a coplanar mode with a proton from the amine and the lactone carbonyl group (B17d). However, geometrical restraints in enzyme active site prevent the ylide to explore those conformations.
- The direct formation of **3l** from **3l**-ylide is energetically more favorable (lower reaction barriers) than the pathway involving first the formation of the enol.

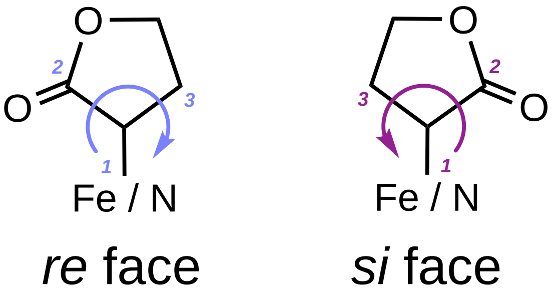


Figure B18. Prochiral faces of the lactone-carbene/ylide (Fe/N) intermediates.

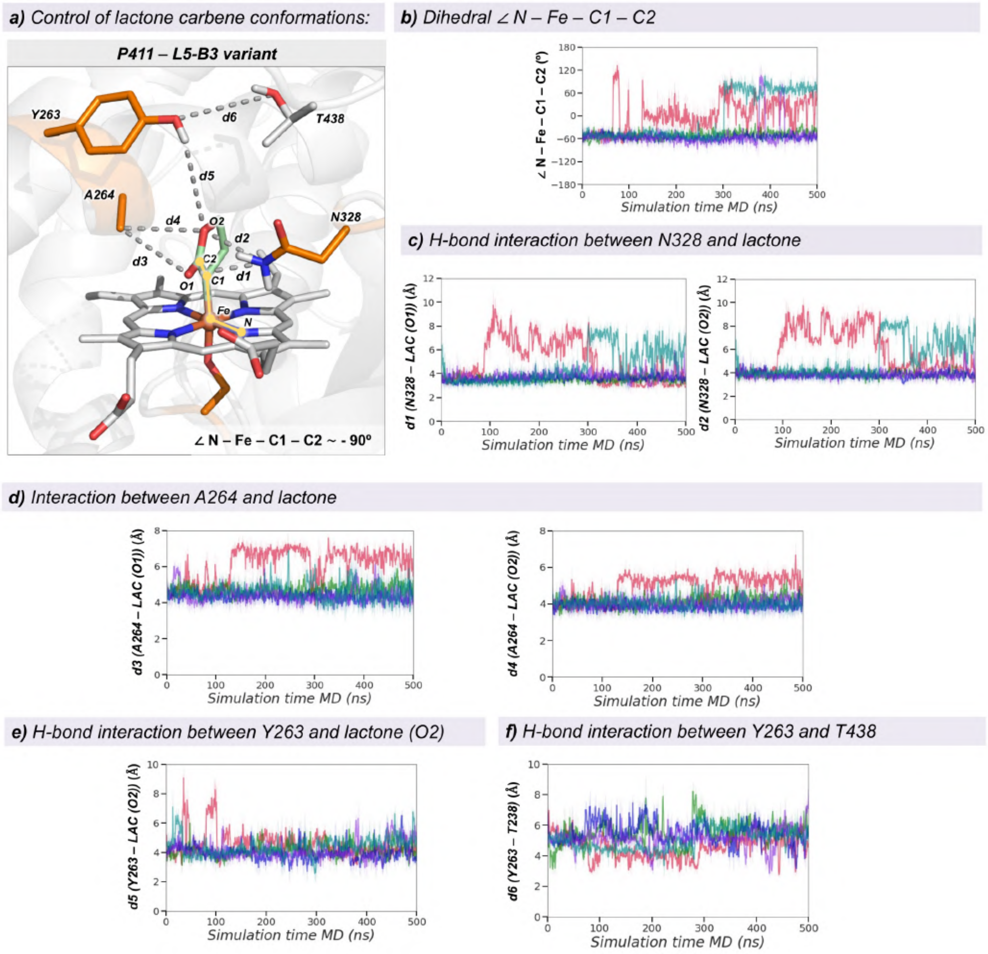
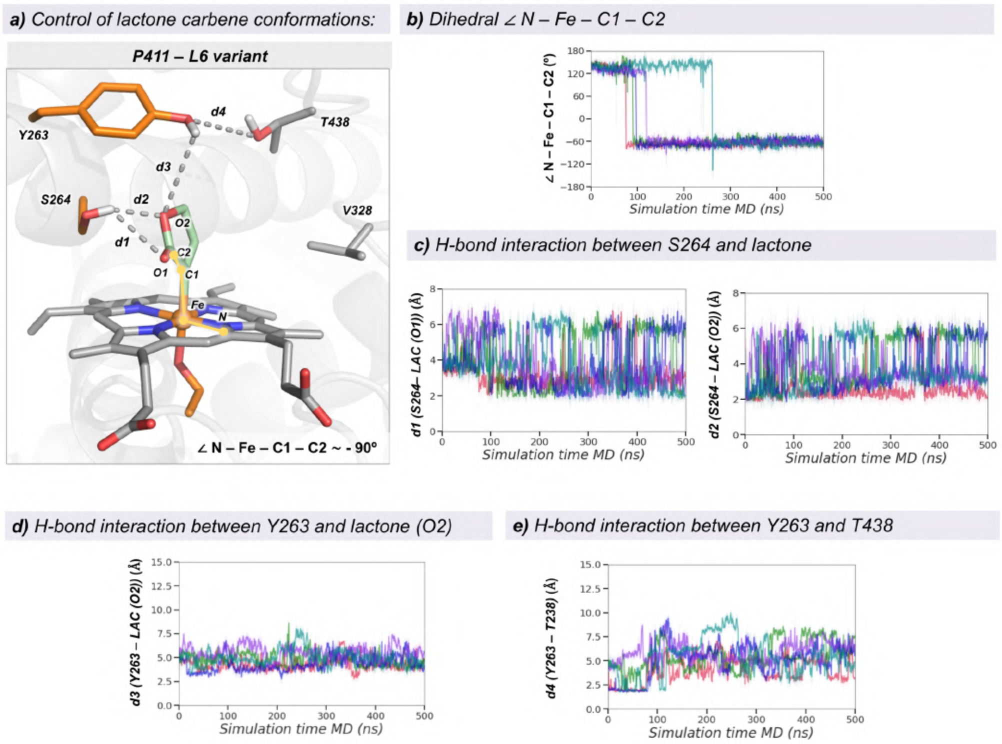


Figure B19. Conformational control of lactone carbene formed in P411 L5-B3 variant. a, Representative snapshot corresponding to the most populated cluster extracted from Molecular Dynamics (MD) simulations describing the preferred orientation of the lactone carbene when formed in P411-**L5-B3** variant. **b,** The ∠(N – Fe – C1 – C2) dihedral angle measured along independent MD trajectories (5 replicas, 500 ns each) describes the relative orientation explored by the carbene. In P411-**L5-B3** simulations show that the lactone preferentially explores a single conformation (dihedral angle ca. –50º), which is stabilized by H-bond interactions established between the carbene ester group and N328. **c,** Distance vs. time plot describing the H-bond interaction between the N328 amide group and the O1-oxygen of the lactone (left) and the N328 amide group and the O2-oxygen of the lactone (right) in P411-**L5-B3** variant (*d1* and *d2* respectively, as shown in Figure B19a). **d,** Distance vs. time plot describing the interaction between the A264 Cβ and the O1-oxygen of the lactone (left) and the A264 Cβ and the O2-oxygen of the lactone (right) in P411-**L5-B3** variant (*d3* and *d4* respectively, as shown in Figure B19a). **e,** Distance vs. time plot describing the H-bond interaction between the Y263 hydroxyl group and the O2-oxygen of the lactone in P411-**L5-B3** (*d5*, as shown in Figure B19a). **f,** Distance vs. time plot describing the interaction between the Y263 hydroxyl group and the T438 hydroxyl group in P411-**L5-B3** (*d6*, as shown in Figure B19a).



a) Control of lactone carbene conformations:

b) Dihedral ∠ N – Fe – C1 – C2

c) H-bond interaction between S264 and lactone

d) H-bond interaction between Y263 and lactone (O2)

e) H-bond interaction between Y263 and T438

Figure B20. Conformational control of lactone carbene formed in P411-L6 variant. a, Representative snapshot corresponding to the most populated cluster extracted from Molecular Dynamics (MD) simulations describing the preferred orientation of the lactone carbene when formed in P411-**L6** variant. **b,** The ∠(N – Fe – C1 – C2) dihedral angle measured along independent MD trajectories (5 replicas, 500 ns each) describes the relative orientation explored by the carbene. In P411-**L6** simulations show that the lactone preferentially explores a single conformation (dihedral angle ca. –90º), which is stabilized by H-bond interactions established between the carbene ester group and S264. **c,** Distance vs. time plot describing the H-bond interaction between the S264 hydroxyl group and the O1-oxygen of the lactone (left) and the S264 hydroxyl group and the O2-oxygen of the lactone (right) in P411-**L6** variant (*d1* and *d2* respectively, as shown in Figure B20a). **d,** Distance vs. time plot describing the H-bond interaction between the Y263 hydroxyl group and the O2-oxygen of the lactone in P411-**L6** (*d5*, as shown in Figure B20a). **e,** Distance vs. time plot describing the interaction between the Y263 hydroxyl group and the T438 hydroxyl group in P411-**L6** (*d6*, as shown in Figure B20a).

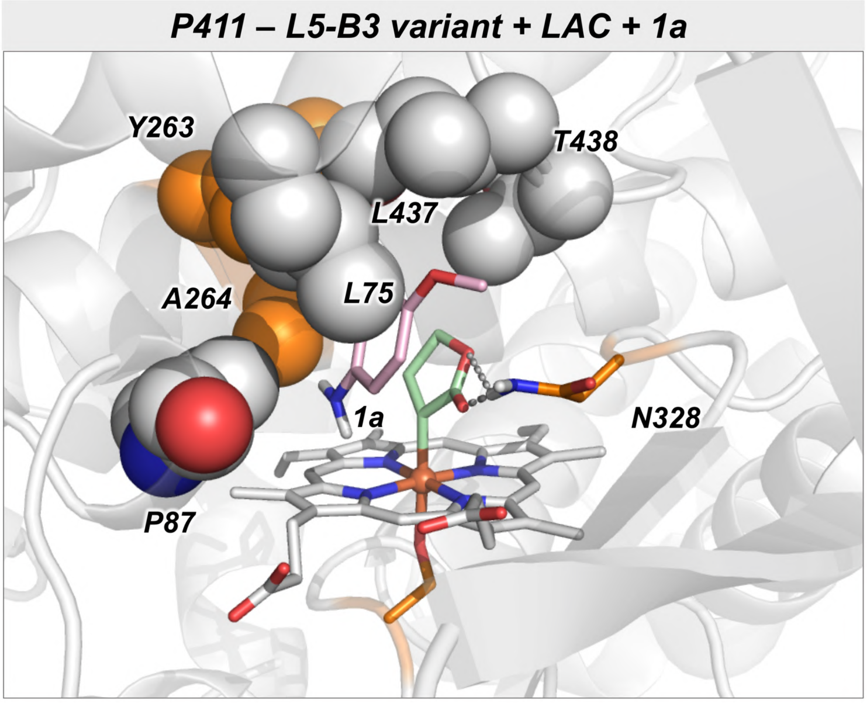


Figure B21. Binding of substrate 1a in a near attack conformation for *N*-nucleophilic attack for L5-B3 variant. Representative snapshot corresponding to the most populated cluster extracted from constrained MD simulations describing the binding of **1a** in a near attack conformation for the *N*-nucleophilic attack to the lactone carbene bound in P411-**L5-B3** variant. Substrate **1a** is shown in light pink. The residues that are directly establishing hydrophobic interactions with the substrate are displayed as spheres. Hydrophobic interactions occurring between the aromatic ring of the aniline derivative and active site residues (L75, P87, Y263, L437, T438) stabilize this binding mode, while H-bond interactions between the carbene ester group and N328 are maintained (see Figure B23).

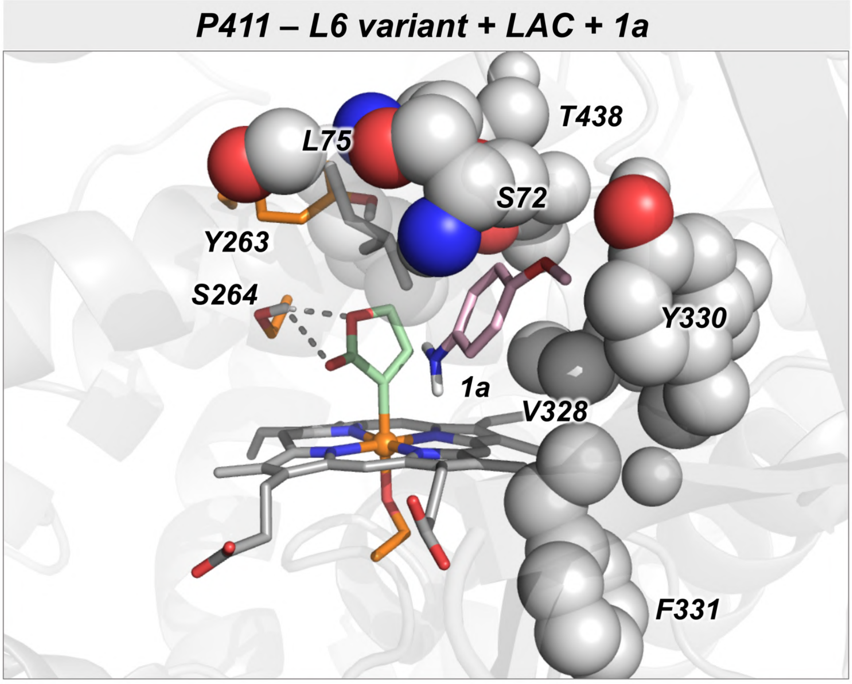


Figure B22. Binding of substrate 1a in a near attack conformation for *N*-nucleophilic attack for L6 variant. Representative snapshot corresponding to the most populated cluster extracted from constrained MD simulations describing the binding of **1a** in a near attack conformation for the *N*-nucleophilic attack to the lactone carbene bound in P411-**L6** variant. Substrate **1a** is shown in light pink. The residues that are directly establishing hydrophobic interactions with the substrate are displayed as spheres. Hydrophobic interactions occurring between the aromatic ring of the aniline derivative and active site residues (L75, S72, Y263, T438, T330 and V328) stabilize this binding mode, while H-bond interactions between the carbene ester group and S264 are maintained (see Figure B24).

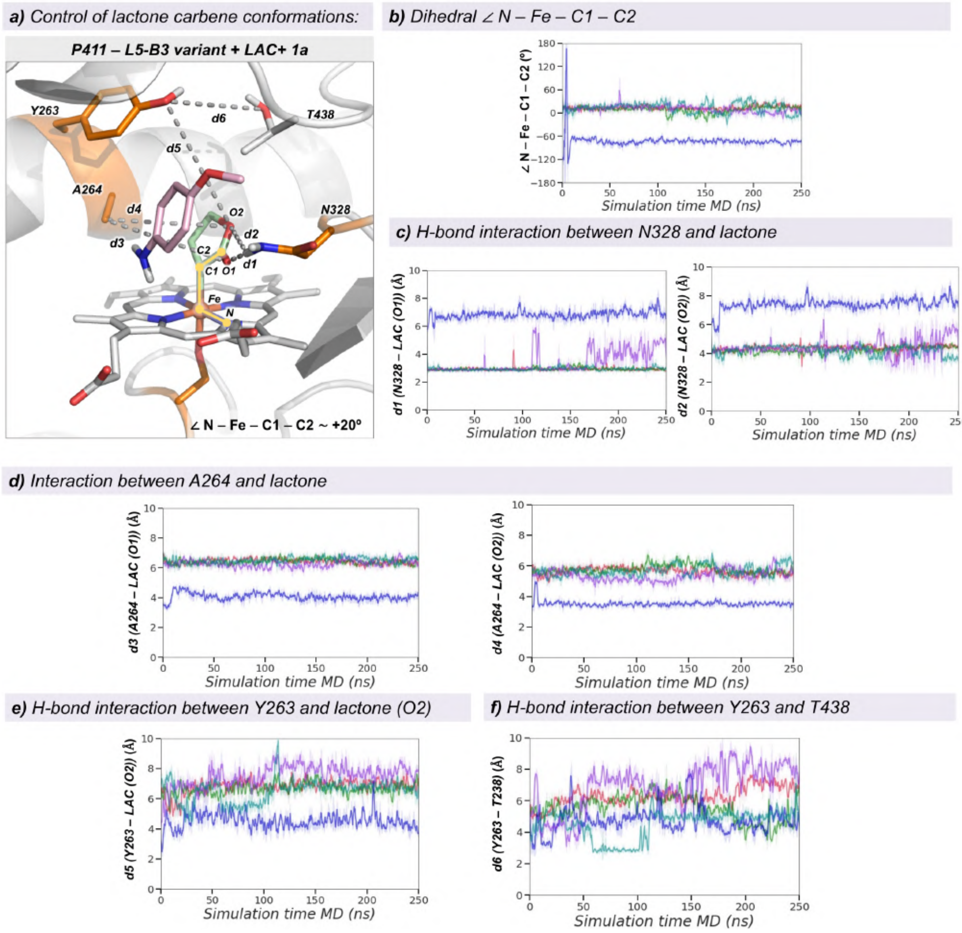


Figure B23. Analysis of substrate 1a binding in a near attack conformation for *N*-nucleophilic attack for L5-B3 variant. a, Representative snapshot corresponding to the most populated cluster extracted from constrained-MD simulations describing the near attack conformations for the N-nucleophilic attack of **1a** to the lactone carbene in variant P411-**L5-B3**. **b,** The ∠(N – Fe – C1 – C2) dihedral angle measured along independent MD trajectories (5 replicas, 500 ns each) describes the relative orientation explored by the carbene. In P411-**L5-B3** simulations show that the lactone preferentially explores a single conformation (dihedral angle ca. +20º), which is stabilized by H-bond interactions established between the carbene ester group and N328. **c,** Distance vs. time plot describing the H-bond interaction between the N328 amide group and the O1-oxygen of the lactone (left) and the N328 amide group and the O2-oxygen of the lactone (right) in P411-**L5-B3** variant (*d1* and *d2* respectively, as shown in Figure B23a). **d,** Distance vs. time plot describing the interaction between the A264 Cβ and the O1-oxygen of the lactone (left) and the A264 Cβ and the O2-oxygen of the lactone (right) in P411-**L5-B3** variant (*d3* and *d4* respectively, as shown in Figure B23a). **e,** Distance vs. time plot describing the H-bond interaction between the Y263 hydroxyl group and the O2-oxygen of the lactone in P411-**L5-B3** (*d5*, as shown in Figure B23a). **f,** Distance vs. time plot describing the interaction between the Y263 hydroxyl group and the T438 hydroxyl group in P411-**L5-B3** (*d6*, as shown in Figure B23a).

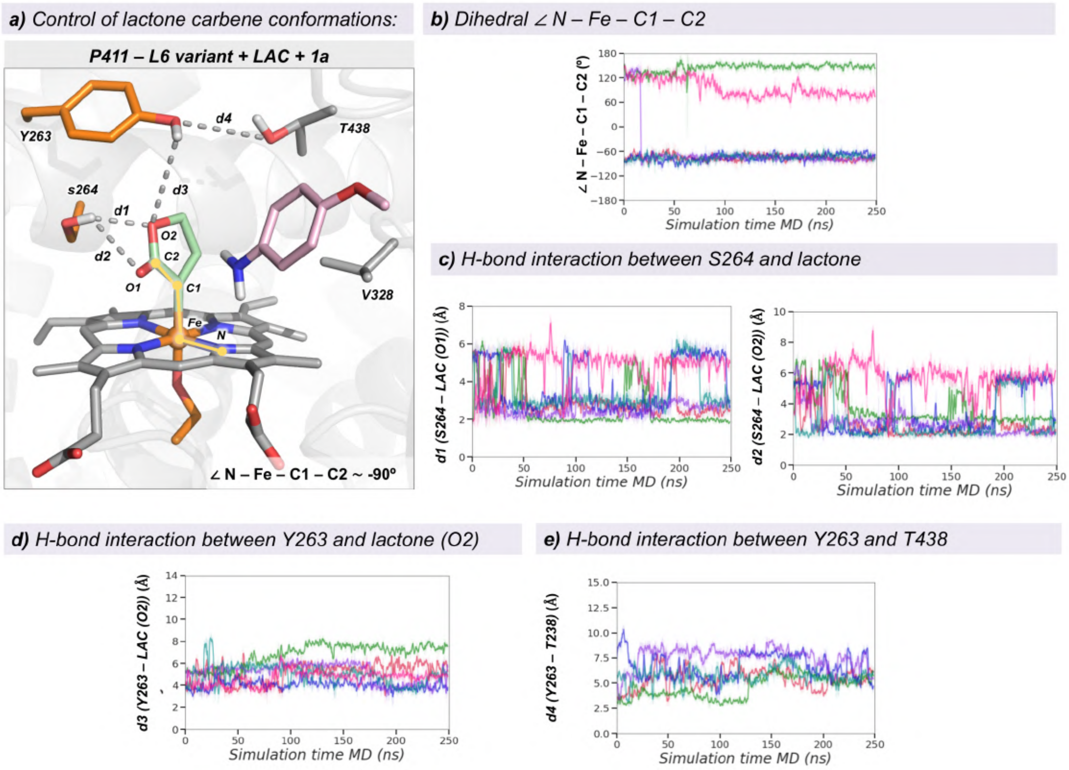
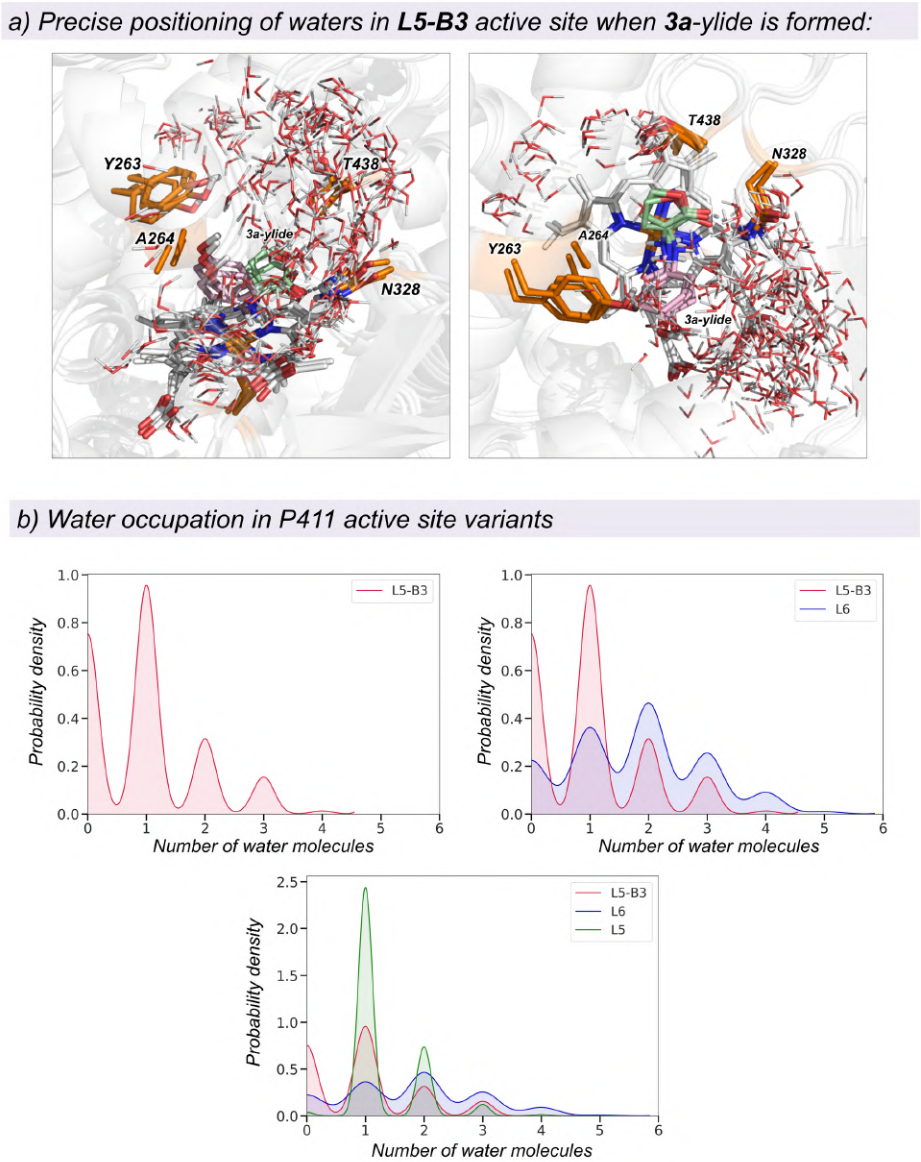


Figure B24. Analysis of substrate 1a binding in a near attack conformation for N-nucleophilic attack for L6 variant. a, Representative snapshot corresponding to the most populated cluster extracted from constrained-MD simulations describing the near attack conformations for the *N*-nucleophilic attack of **1a** to the lactone carbene in variant P411-**L6**. **b,** The ∠(N – Fe – C1 – C2) dihedral angle measured along independent MD trajectories (5 replicas, 500 ns each) describes the relative orientation explored by the carbene. In P411-**L6** simulations show that the lactone preferentially explores a single conformation (dihedral angle ca. -90º), which is stabilized by H-bond interactions established between the carbene ester group and S264. **c,** Distance vs. time plot describing the H-bond interaction between the S264 hydroxyl group and the O1-oxygen of the lactone (left) and the S264 hydroxyl group and the O2-oxygen of the lactone (right) in P411-**L6** variant (*d1* and *d2* respectively, as shown in Figure B24a). **d,** Distance vs. time plot describing the H-bond interaction between the Y263 hydroxyl group and the O2-oxygen of the lactone in P411-**L6** (*d3*, as shown in Figure B24a). **e,** Distance vs. time plot describing the interaction between the Y263 hydroxyl group and the T438 hydroxyl group in P411-**L6** (*d4*, as shown in Figure B24a).



a) Precise positioning of waters in **L5-B3** active site when **3a**-ylide is formed:

b) Water occupation in P411 active site variants

Figure B25. Precise positioning of waters in the active site upon 3a-ylide formation. a, Overlay of 3 representative snapshots from constrained-MD simulations exploring the active site arrangement in P411-**L5-B3** variant when **3a**-ylide is formed. Water molecules are precisely positioned on the top-face of the lactone ring through a water channel driven by Y263 and T438, and nearby the protonated ylide amine group through the water channel. These water molecules can stereoselectively protonate the ylide intermediate from the pro-*R* face. Displayed water molecules are drawn from 25 random structures across the 100 ns MD trajectory. **b,** Representation of the normalized kernel density plot of the number of water molecules in the active site of P411 variants nearby the **3a**-ylide, considering its first solvation shell (using a distance cut-off of 3.4 Å). The average number of water molecules in the active site is 1.0 ± 0.9 (P411-**L5-B3**), 1.8 ± 1.2 (P411-**L6**), and 1.6 ± 0.7 (P411-**L5**), respectively. The presence of water molecules is monitored through visual inspection of MD trajectories and using the *watershell* function of *cpptraj*.[14]

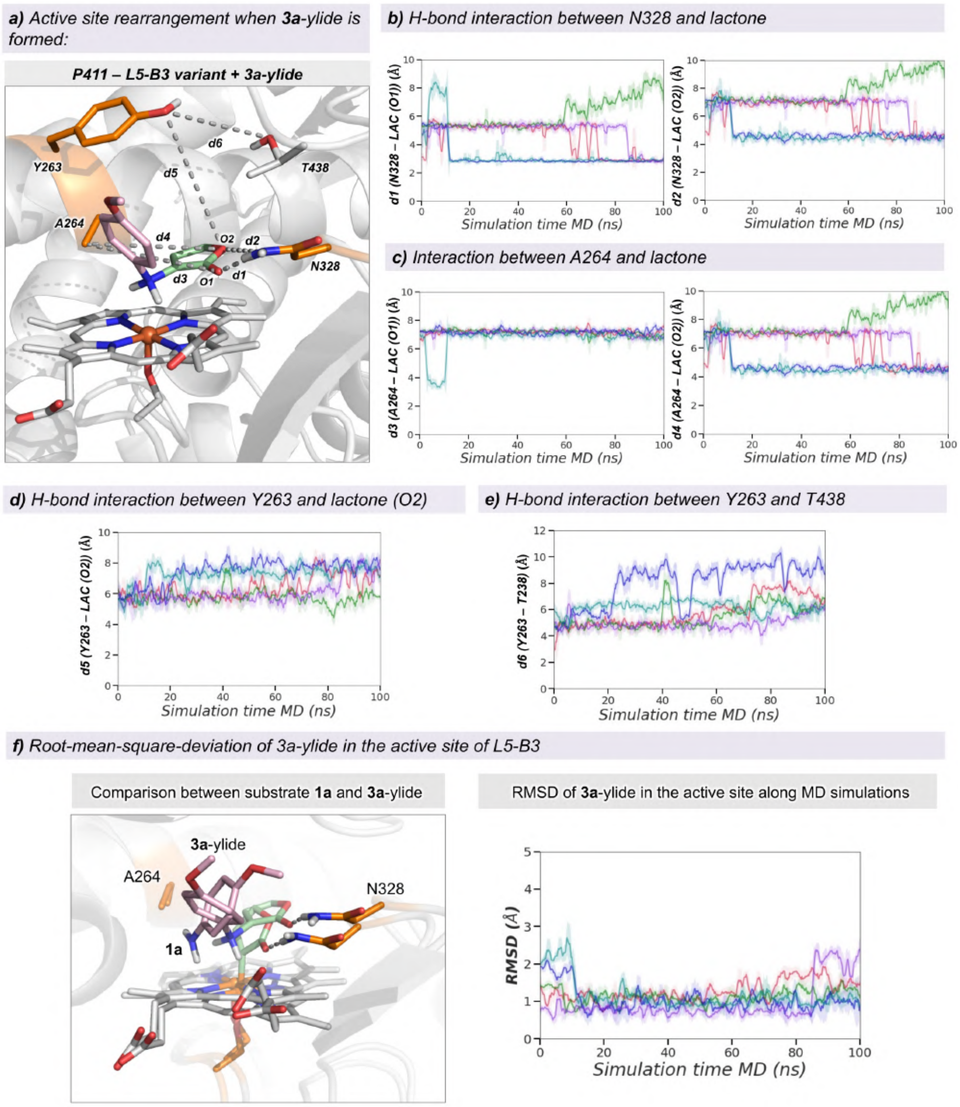


Figure B26. Analysis of active site arrangement when 3a-ylide is formed for L5-B3 variant. a, Representative snapshot corresponding to the most populated cluster extracted from constrained-MD simulations describing the active site arrangement when **3a**-ylide is formed (right after dissociation from Fe) in variant P411-**L5-B3**. **b,** Distance vs. time plot describing the H-bond interaction between the N328 amide group and the O1-oxygen of the lactone (left) and the N328 amide group and the O2-oxygen of the lactone (right) in P411-**L5-B3** variant (*d1* and *d2* respectively, as shown in Figure B26a). **c,** Distance vs. time plot describing the interaction between the A264 Cβ and the O1-oxygen of the lactone (left) and the A264 Cβ and the O2-oxygen of the lactone (right) in P411-**L5-B3** variant (*d3* and *d4* respectively, as shown in Figure B26a). **d,** Distance vs. time plot describing the H-bond interaction between the Y263 hydroxyl group and the O2-oxygen of the lactone in P411-**L5-B3** (*d5*, as shown in Figure B26a). **e,** Distance vs. time plot describing the interaction between the Y263 hydroxyl group and the T438 hydroxyl group in P411-**L5-B3** (*d6*, as shown in Figure B26a). **f,** Root-mean-square-deviation (RMSD) of 3a-ylide in the active site of P411-L5-B3. In the left, snapshots of the substrate **1a** and **3a**-ylide obtained from most populated clusters

of MD simulations are overlayed and depicted in light pink. Substrate **1a** and **3a**-ylide show a similar orientation in the P411-**L5-B3** active site. In the right, RMSD of **3a**-ylide in the active site of P411-**L5-B3** along five replicas of 100 ns MD simulations. RMSD is calculated with respect to the average structure of each replica. RMSD values indicate that ylide conformation remains stable along the MD simulations.

.

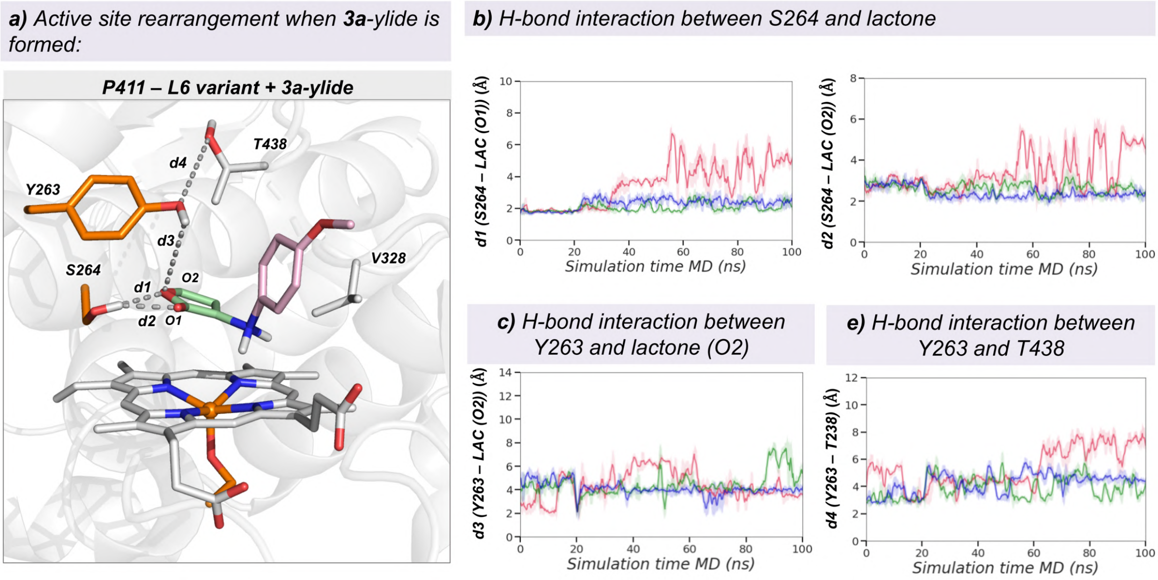


Figure B27. Analysis of active site arrangement when 3a-ylide is formed for L6 variant. a, Representative snapshot corresponding to the most populated cluster extracted from constrained-MD simulations describing the active site arrangement when **3a**-ylide is formed (right after dissociation from Fe) in variant P411-**L6**. **c,** Distance vs. time plot describing the H-bond interaction between the S264 hydroxyl group and the O1-oxygen of the lactone (left) and the N328 amide group and the O2-oxygen of the lactone (right) in P411-**L6** variant (*d1* and *d2* respectively, as shown in Figure B27a). **d,** Distance vs. time plot describing the H-bond interaction between the Y263 hydroxyl group and the O2-oxygen of the lactone in P411-**L6** (*d3*, as shown in Figure B27a). **f,** Distance vs. time plot describing the interaction between the Y263 hydroxyl group and the T438 hydroxyl group in P411-**L6** (*d4*, as shown in Figure B27a).

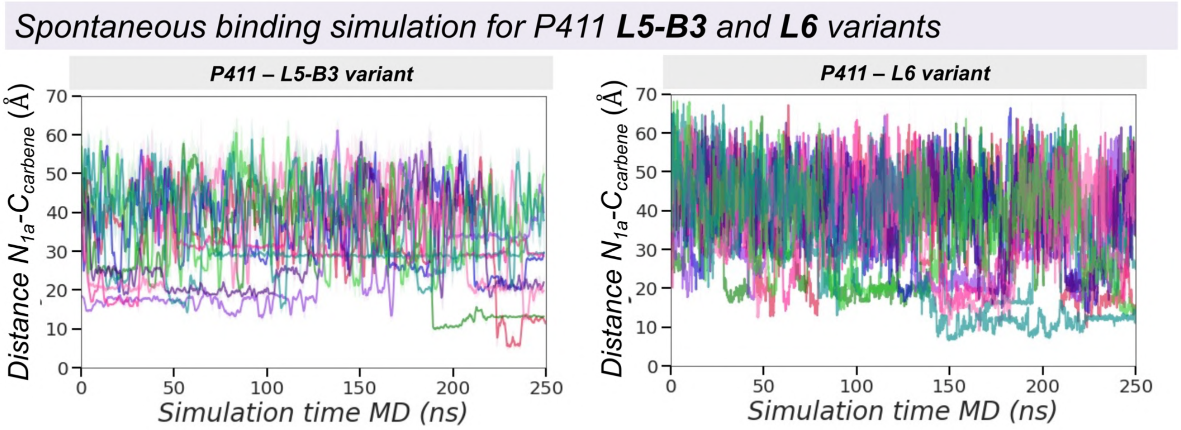


Figure B28. Spontaneous 1a substrate binding process in P411 L5-B3 and L6 variants. Plot of the distance (in Å) between the carbene carbon of the lactone and the nitrogen of the substrate for ten replicas of 250 ns cMD simulations for P411 variant **L5-B3** (left) and **L6** (right). Each replica is depicted in a different color. Spontaneous **1a** substrate binding occurs in 1/10 (red) and 2/10 (green and red) replicas of P411-**L5-B3** and **L6** variants, respectively. The simulations where substrate binding was observed were extended up to 1000 ns (see Figure 5.7 in Chapter 5 and Figure B29).

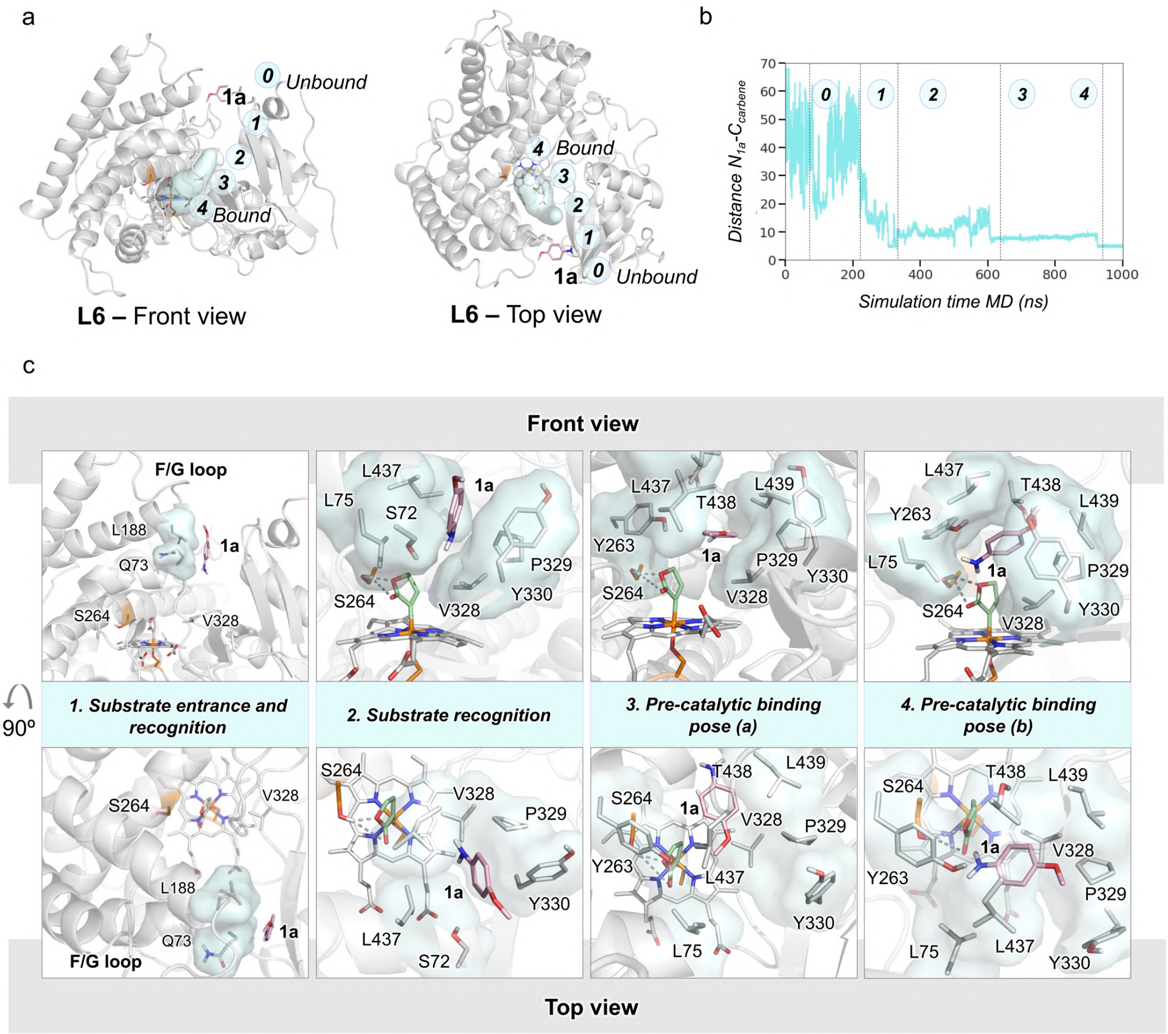


Figure B29. Molecular basis of 1a substrate binding in P411-L6 variant. a, General scheme of the spontaneous **1a** substrate binding process in P411-**L6** variant. The substrate entrance channel calculated with CAVER[24] is depicted as a blue surface. The numbers indicate the most relevant steps of the binding process (detailed in **c**). **b,** Plot of the distance (in Å) between the carbene carbon of the lactone and the nitrogen of the substrate for 1000 ns cMD simulations. The dashed lines indicate the distances where entrance, recognition and pre-catalytic steps take place, respectively. **c**, Molecular representation of selected key conformational states of the **1a** substrate binding pathway in P411-**L6** variant. The substrate is shown in light pink, key active site residue 264S is depicted in orange and key residues for each step (entrance, recognition, and pre-catalytic binding poses) are shown in grey sticks and surrounded with a cyan surface. In **L6** variant, substrate **1a** explores two pre-catalytic binding poses: *pre-catalytic binding pose (a)*, where the nitrogen of the substrate is not oriented towards the carbon of the carbene; and *pre-catalytic pose (b)*, where the nitrogen of **1a** is oriented towards the lactone. Unlike the catalytic step sampled in variant **L5-B3** (see Figure 5.7, Chapter 5), more simulation time would be required to explore catalytic shorter distances in variant **L6**.

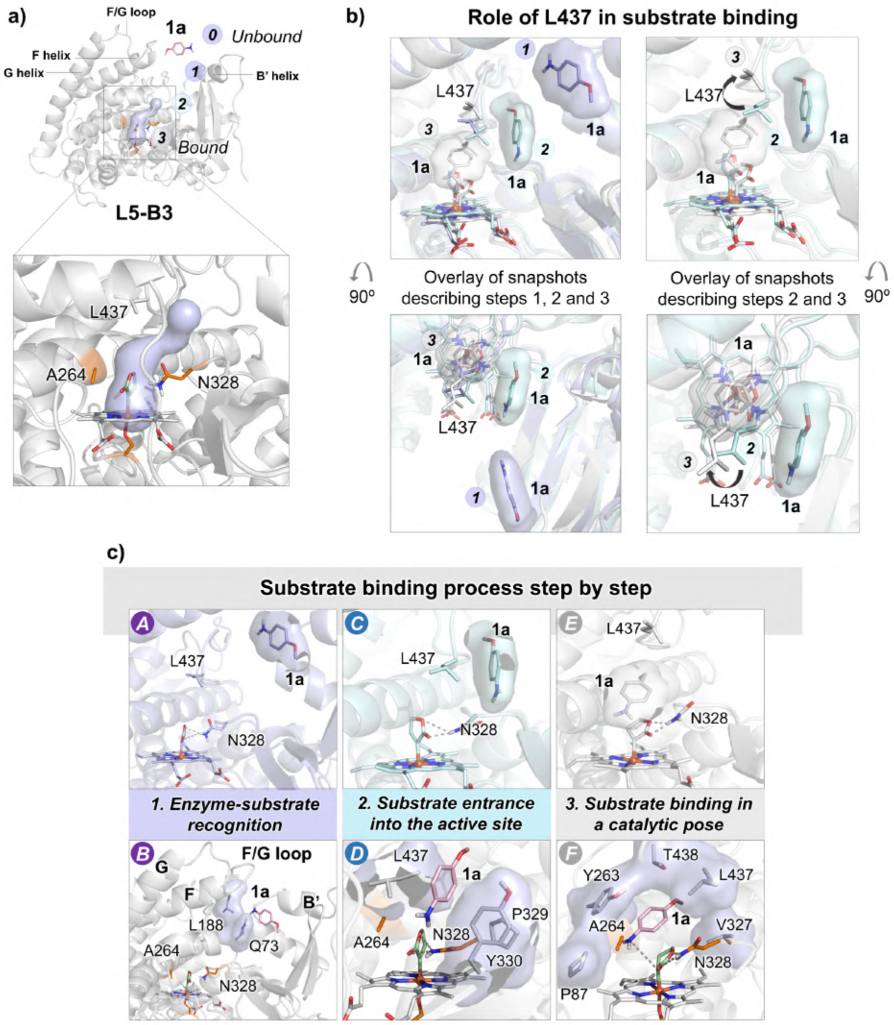

a)
F/G loop
1a
0 Unbound
F helix
G helix
1
B' helix
2
3
Bound
L5-B3
L437
A264
N328
b)
Role of L437 in substrate binding
1
L437
3
1a
2
1a
1a
3
L437
2
1a
1a
90°
Overlay of snapshots describing steps 1, 2 and 3
Overlay of snapshots describing steps 2 and 3
90°
3
1a
2
1a
L437
1
1a
1a
2
3
L437
1a
c)
Substrate binding process step by step
A
L437
1a
N328
C
1a
L437
N328
E
L437
1a
N328
1. Enzyme-substrate recognition
2. Substrate entrance into the active site
3. Substrate binding in a catalytic pose
B
G
F/G loop
F
1a
B'
L188
A264
Q73
N328
D
L437
1a
P329
A264
N328
Y330
F
T438
Y263
L437
A264
1a
V327
N328
P87

Figure B30. Role of L437 in 1a substrate binding in P411-L5-B3 variant. a, General scheme for the spontaneous substrate binding process of **1a** substrate in P411-**L5-B3** variant. The substrate entrance channel calculated with CAVER[18] software is depicted as a light blue surface. The numbers indicate the most relevant steps of the binding process (detailed in **c**). **b,** Front and top views of snapshots obtained from spontaneous binding MD simulations. Molecular representation of the role of L437 during the substrate binding process. Snapshots corresponding to steps 1) Enzyme-substrate recognition; 2) Substrate entrance into the active site and 3) Substrate binding in a catalytic pose are overlayed and depicted in light blue, cyan and grey, respectively. L437 acts as a gate after substrate recognition in the substrate entrance channel and entrance into the active site (steps 1 and 2) leading to effective substrate binding into the catalytic pocket (step 3). No large conformational changes of the enzyme structure are observed. **c**, Molecular representation of selected key conformational states of the characterized **1a** substrate binding pathway. A and B describe the step 1- Enzyme-substrate recognition; C and D represent step 2- Substrate entrance into the active site and E and F correspond to step 3- Substrate binding in a catalytic pose. In A, C and E steps the role of L437 is highlighted. In B, D and F, key residues for each step are depicted (shown as a surface).

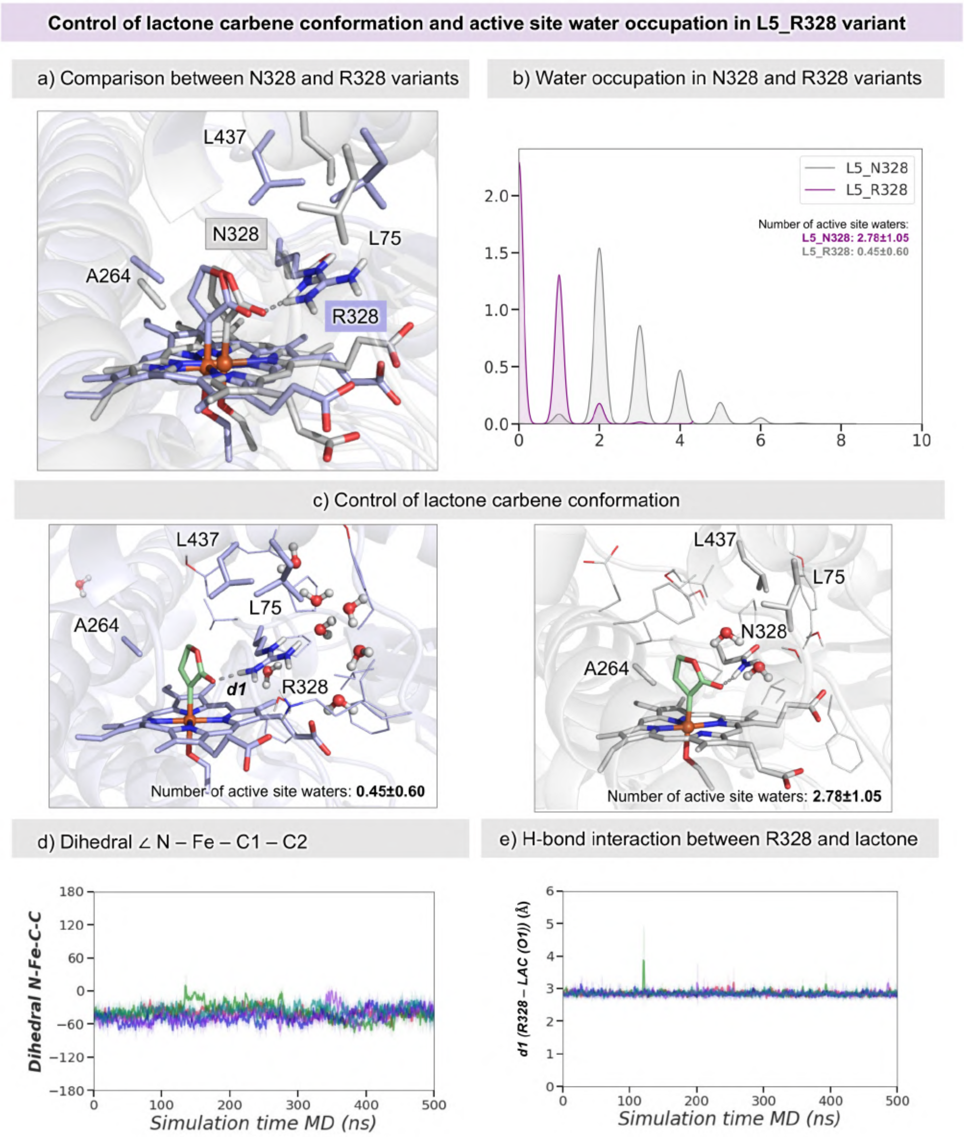


Figure B31. Comparison of P411-L5-B3 (N328) with P411-L5-R328 variant. a, Representative snapshot corresponding to the most populated cluster extracted from MD simulations describing the preferred orientation of the lactone carbene when formed in P411-**L5-B3** variant (in gray) and P411-**L5-R328** variant (in purple) respectively. **b,** Representation of the normalized kernel density plot of the number of water molecules in the active site of P411-**L5-B3** (in gray) and P411-**L5-R328** (in purple) variants nearby the lactone carbene, considering its first solvation shell (using a distance cut-off of 3.4 Å). The average number of water molecules in the active site is 0.45 ± 0.60 (P411-L5-B3), and 2.78 ± 1.05 (P411-L5-R328), respectively. The presence of water molecules is monitored through visual inspection of MD trajectories and using the *watershell* function of *cpptraj*. **c,** Water molecules present in the vicinity of the P411 active site in the most populated cluster extracted from MD simulations. Two water molecules are present nearby the lactone carbene of P411-**L5-B3** while zero water molecules are found in the P411-**L5-R328** variant. In the P411-**L5-R328** variant, L437 and L75 are directly interacting and preventing the access of water molecules in the active site. **d,**

The ∠(N – Fe – C1 – C2) dihedral angle measured along independent MD trajectories of **L5-R328** variant (5 replicas, 500 ns each) describes the relative orientation explored by the carbene. In P411-**L5-R328** simulations show that the lactone preferentially explores a single conformation (dihedral angle ca. –50º), which is stabilized by H-bond interactions established between the carbene ester group and R328. **e,** Distance vs. time plot describing the H-bond interaction between the R328 guanidinium group and the O1-oxygen of the lactone in P411-**L5-R328** variant.

Appendix B References

(1) Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6**, 343–345 (2009).

(2) Barr, I. & Guo, F. Pyridine hemochromagen assay for determining the concentration of heme in purified protein solutions. *Bio. Protoc.* **5**, e1594 (2015).

(3) Chen, K., Zhang, S.-Q., Brandenberg, O. F., Hong, X. & Arnold, F. H. Alternate heme ligation steers activity and selectivity in engineered cytochrome P450-catalyzed carbene-transfer reactions. *J. Am. Chem. Soc.* **140**, 16402–16407 (2018).

(4) Yang, Y. & Arnold, F. H. Navigating the unnatural reaction space: directed evolution of heme proteins for selective carbene and nitrene transfer. *Acc. Chem. Res.* **54**, 1209–1225 (2021).

(5) Gu, X.-S. *et al.* Enantioselective hydrogenation of racemic *α*-arylamino lactones to chiral amino diols with site-specifically modified chiral spiro iridium catalysts. *Org. Lett.* **21**, 4111–4115 (2019).

(6) Zhou, A. Z., Chen, K. & Arnold, F. H. Enzymatic lactone-carbene C–H insertion to build contiguous chiral centers. *ACS Catal.* **10**, 5393–5398 (2020).

(7) Salomon-Ferrer, R., Götz, A. W., Poole, D., Le Grand, S. & Walker, R. C. Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. Explicit solvent particle mesh Ewald. *J. Chem. Theory Comput.* **9**, 3878–3888 (2013).

(8) D.A. Case *et al.* Amber 2017, University of California, San Francisco. (2017) doi:10.13140/RG.2.2.36172.41606.

(9) Li, P. & Merz, K. M. MCPB.py: a Python based metal center parameter builder. *J. Chem. Inf. Model.* **56**, 599–604 (2016).

(10) Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. Development and testing of a general amber force field. *J. Comput. Chem.* **25**, 1157–1174 (2004).

(11) Bayly, C. I., Cieplak, P., Cornell, W. & Kollman, P. A. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *J. Phys. Chem.* **97**, 10269–10280 (1993).

(12) Besler, B. H., Merz, K. M. & Kollman, P. A. Atomic charges derived from semiempirical methods. *J. Comput. Chem.* **11**, 431–439 (1990).

(13) Singh, U. C. & Kollman, P. A. An approach to computing electrostatic charges for molecules. *J. Comput. Chem.* **5**, 129–145 (1984).

(14) Frisch, M. J. *et al.* Gaussian 09, Revision A. 02. Gaussian. *Inc.: Wallingford, CT* (2009).

(15) Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926–935 (1983).

(16) Maier, J. A. *et al.* ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *J. Chem. Theory Comput.* **11**, 3696–3713 (2015).

(17) Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: an *N* □log(*N*) method for Ewald sums in large systems. *J. Chem. Phys.* **98**, 10089–10092 (1993).

(18) Roe, D. R. & Cheatham, T. E. PTRAJ and CPPTRAJ: software for processing and analysis of molecular dynamics trajectory data. *J. Chem. Theory Comput.* **9**, 3084–3095 (2013).

(19) Becke, A. D. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A* **38**, 3098–3100 (1988).

(20) Becke, A. D. Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.* **98**, 5648–5652 (1993).

(21) Lee, C., Yang, W. & Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B* **37**, 785–789 (1988).

(22) Grimme, S., Ehrlich, S. & Goerigk, L. Effect of the damping function in dispersion corrected density functional theory. *J. Comput. Chem.* **32**, 1456–1465 (2011).

(23) Grimme, S., Antony, J., Ehrlich, S. & Krieg, H. A consistent and accurate *ab initio* parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.* **132**, 154104 (2010).

(24) Barone, V. & Cossi, M. Quantum calculation of molecular energies and energy gradients in solution by a conductor solvent model. *J. Phys. Chem. A* **102**, 1995–2001 (1998).

(25) Cossi, M., Rega, N., Scalmani, G. & Barone, V. Energies, structures, and electronic properties of molecules in solution with the C-PCM solvation model. *J. Comput. Chem.* **24**, 669–681 (2003).

(26) Schutz, C. N. & Warshel, A. What are the dielectric 'constants' of proteins and how to validate electrostatic models? *Proteins: Struct., Funct., Bioinf.* **44**, 400–417 (2001).

(27) Li, L., Li, C., Zhang, Z. & Alexov, E. On the dielectric "constant" of proteins: smooth dielectric function for macromolecular modeling and its implementation in DelPhi. *J. Chem. Theory Comput.* **9**, 2126–2136 (2013).

(28) Sharon, D. A., Mallick, D., Wang, B. & Shaik, S. Computation sheds insight into iron porphyrin carbenes' electronic structure, formation, and N–H insertion reactivity. *J. Am. Chem. Soc.* **138**, 9597–9610 (2016).

(29) Lewis, R. D. *et al.* Catalytic iron-carbene intermediate revealed in a cytochrome *c* carbene transferase. *Proc. Natl. Acad. Sci. USA* **115**, 7308–7313 (2018).

(30) Huang, X. *et al.* A biocatalytic platform for synthesis of chiral *α*-trifluoromethylated organoborons. *ACS Cent. Sci.* **5**, 270–276 (2019).

(31) Yang, Y., Cho, I., Qi, X., Liu, P. & Arnold, F. H. An enzymatic platform for the asymmetric amination of primary, secondary and tertiary C(sp^3)–H bonds. *Nat. Chem.* **11**, 987–993 (2019).

(32) Garcia-Borràs, M. *et al.* Origin and control of chemoselectivity in cytochrome *c*-catalyzed carbene transfer into Si–H and N–H bonds. *ChemRxiv* (2021) doi:10.26434/chemrxiv.14102363.v1.

(33) Seeger, R. & Pople, J. A. Self-consistent molecular orbital methods. XVIII. Constraints and stability in Hartree–Fock theory. *J. Chem. Phys.* **66**, 3045–3050 (1977).

(34) Bauernschmitt, R. & Ahlrichs, R. Stability analysis for solutions of the closed shell Kohn–Sham equation. *J. Chem. Phys.* **104**, 9047–9052 (1996).

(35) Schlegel, H. B. & McDouall, J. J. W. Do you have SCF stability and convergence problems? In *Computational Advances in Organic Chemistry: Molecular Structure and Reactivity* (eds. Ögretir, C. & Csizmadia, I. G.) 167–185 (Springer Netherlands, 1991). doi:10.1007/978-94-011-3262-6_2.

(36) Legault, C. CYLview, 1.0 b, Université de Sherbrooke, Sherbrooke, Québec, Canada (2009). *URL http://www.cylview.org (accessed February 1, 2016)*.

(37) Prier, C. K., Zhang, R. K., Buller, A. R., Brinkmann-Chen, S. & Arnold, F. H. Enantioselective, intermolecular benzylic C–H amination catalysed by an engineered iron-haem enzyme. *Nat. Chem.* **9**, 629–634 (2017).

(38) Richter, F., Leaver-Fay, A., Khare, S. D., Bjelic, S. & Baker, D. De novo enzyme design using Rosetta3. *PLoS ONE* **6**, e19230 (2011).

(39) Schrödinger, LLC *The PyMOL Molecular Graphics System*, Version 1.8. (2015).

(40) Morris, G. M. *et al.* AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. *J. Comput. Chem.* **30**, 2785–2791 (2009).

(41) Trott, O. & Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **31**, 455–461 (2010).

Appendix C

Supplementary information Chapter 6. Reprinted with permission from *ACS Catal.* **2020**, 10, 14, 7512–7525 (https://pubs.acs.org/doi/full/10.1021/acscatal.0c01487). Copyright 2020 American Chemical Society. Further permissions to the material shown should be directed to the ACS.

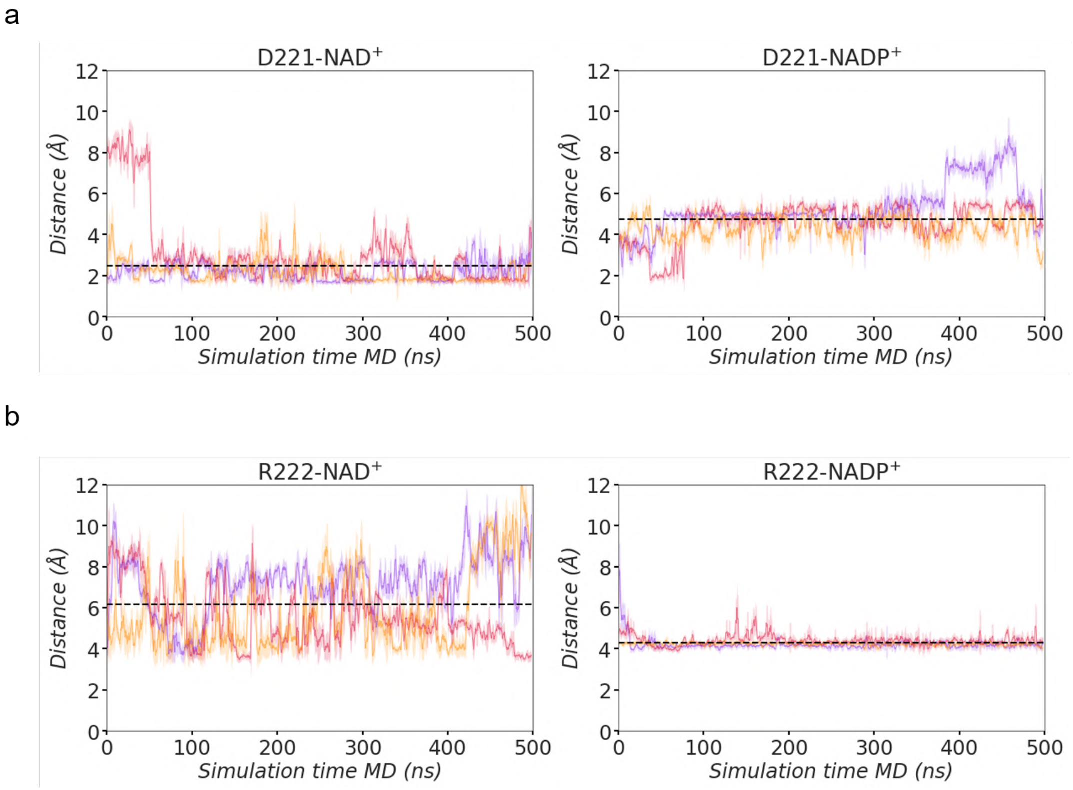


Figure C1. Comparison of the active site conformational dynamics of WT *Pse*FDH in the presence of NAD$^+$ or NADP$^+$ cofactor and formate. a) Plot of the distance between the carbon of the carboxylate group of D221 and 2'-OH group of NAD$^+$ (left) and the distance between the carbon of the carboxylate group of D221 and NADP$^+$ 2-' phosphate group (right) along 3 representative 500 ns replicas of MD simulations (shown in red, orange, and purple) for both WT-NAD$^+$ and WT-NADP$^+$ systems. Average distances from all replicas of 2.5 ± 1.2 Å and 4.7 ± 1.0 Å, respectively, are also shown with a dashed black line; and **b)** Plot of the distance between the carbon of the guanidinium group of R222 and 2'-OH group of NAD$^+$ (left) and the distance between the carbon of the guanidinium group of R222 and NADP$^+$ 2-' phosphate group (right) along 3 representative 500 ns replicas of MD simulations for both WT-NAD$^+$ and WT-NADP$^+$ systems. Average distances (dashed black line) of 6.2 ± 1.9 Å and 4.3 ± 0.4 Å, respectively, are also shown. All distances are represented in Å.

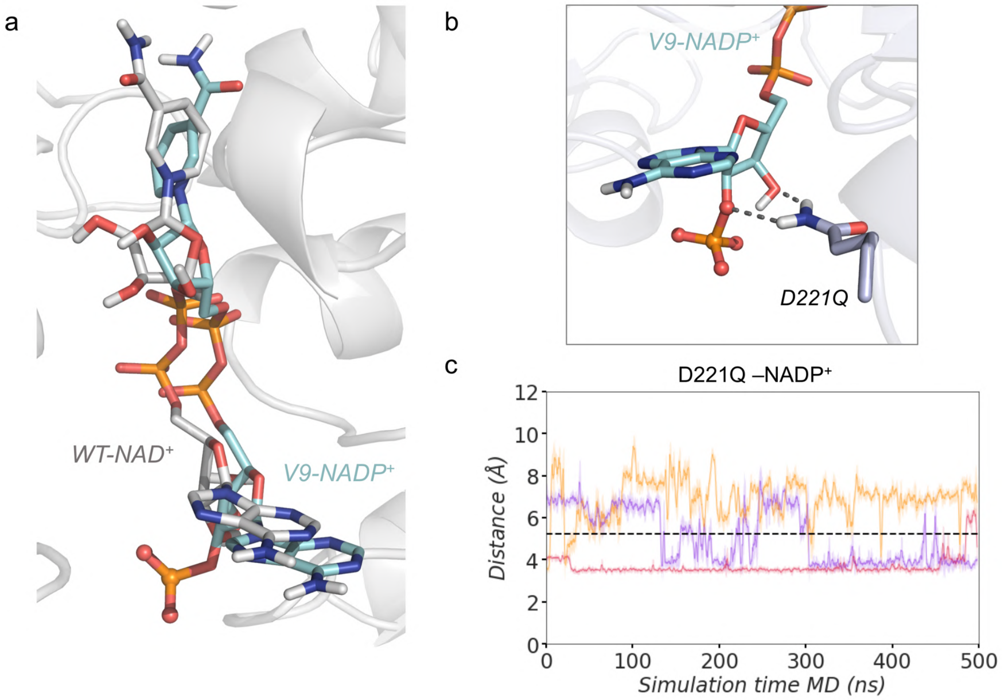
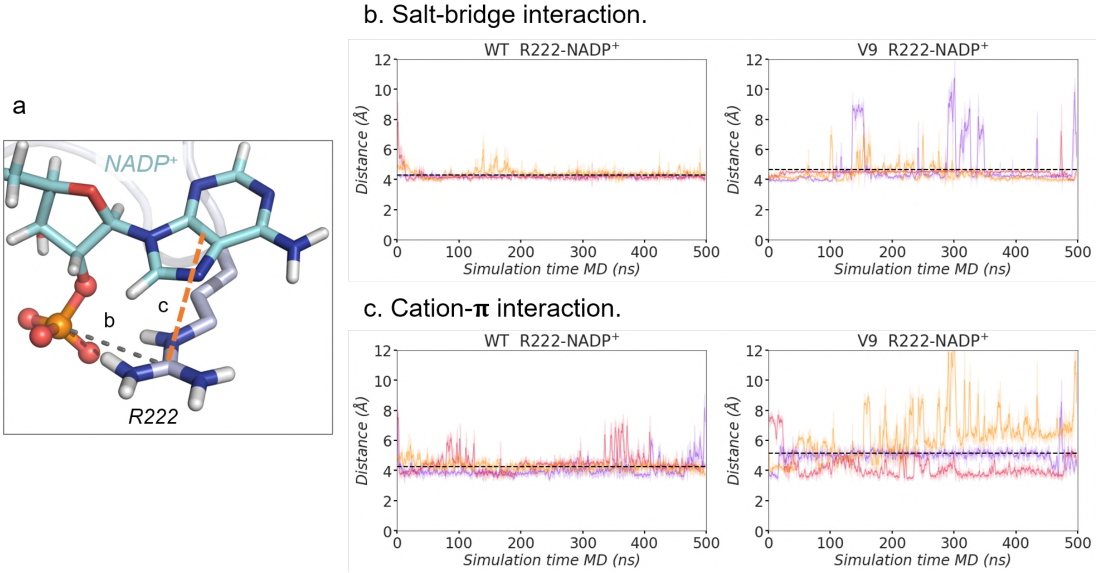


Figure C2. MD simulations analysis with focus on mutation D221Q. a) Overlay of *Pse*FDH WT-NAD$^+$ (grey) and variant V9-NADP$^+$ (cyan) representative conformations of the binding pocket. The figure shows that the adenine ring of both WT-NAD$^+$ and V9-NADP$^+$ are found in the same orientation. However, the nicotinamide ring of the V9-NADP$^+$ is rotated with respect to WT-NAD$^+$. The Root-Mean-Square-Deviation (RMSD) of the cofactor NADP$^+$ in the variant V9 with respect to the natural NAD$^+$ cofactor in the WT enzyme is 2.1 Å. **b)** Representative structure of the frequently observed hydrogen bonds established between D221Q and the 2'-phosphate and 3'-OH group of NADP$^+$. **c)** Plot of the distance of the hydrogen bond established between D221Q and the 2'-phosphate and 3'-OH group of NADP$^+$ along 3 replicas of 500 ns of MD simulations for the V9-NADP$^+$ (replicas are shown in purple, orange and red). Average distance from all replicas (dashed black line) of 5.2 ± 1.7 Å is also depicted. All distances are represented in Å.



b. Salt-bridge interaction.

c. Cation-π interaction.

Figure C3. MD simulations analysis with focus on residue R222. a) Representative structure of V9-NADP+ binding pocket with the salt-bridge interaction between the guanidinium group of R222 and the 2'-phosphate group of NADP+ and the cation-*π* between the guanidinium group of R222 and the adenine group of NADP+ highlighted. **b)** Plot of the distance of the salt-bridge interaction between the carbon of the guanidinium group of R222 and the 2'-phosphate group of NADP+ along 3 replicas of 500 ns of MD simulations for the WT-NADP+ and the V9-NADP+ (red, orange and purple lines). Average distances (dashed black line) of 4.3 ± 0.4 Å and 4.4 ± 1.2 Å, respectively, are also included. **c)** Plot of the distance of the cation-*π* interaction between the guanidinium group of R222 and the center of mass of the adenine group of NADP+ along 3 replicas of 500 ns of MD simulations for the WT-NADP+ and the V9-NADP+. Average distances (dashed black line) of 4.3 ± 0.7 Å and 5.2 ± 1.4 Å, respectively, are also shown. All distances are represented in Å.

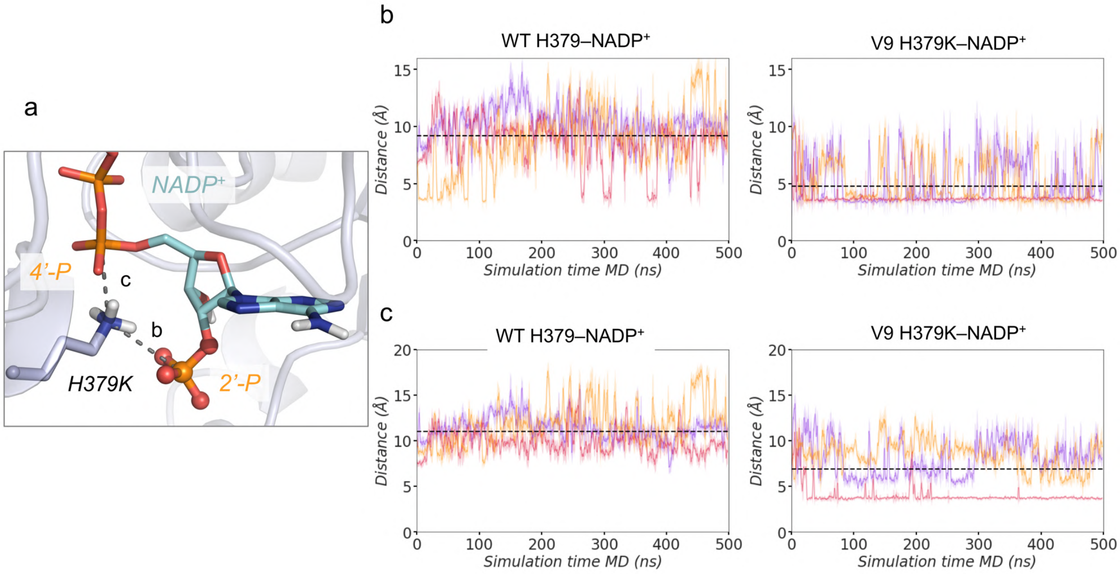


Figure C4. MD simulations analysis with focus on mutation H379K. a) Representative structure of V9-NADP$^+$ binding pocket with the salt-bridge interaction between the positively charged amino group of H379K and the 2'-phosphate group of NADP$^+$ and the salt-bridge interaction between the positively charged amino group of H379K and the linker 4'-phosphate group of NADP$^+$ highlighted. **b)** Plot of the distance between the ε-nitrogen of H379 and the 2'-phosphate group of NADP$^+$ along 3 replicas of 500 ns of MD simulations (shown in red, orange and purple) for the WT-NADP$^+$ (left) and plot of the distance of the salt-bridge interaction between the positively charged amino group of H379K and the 2'-phosphate group of NADP$^+$ along 3 replicas of 500 ns of MD simulations for the V9-NADP$^+$ (right). Average distances from all replicas (dashed black line) of 9.2±2.5 Å and 4.8±2.0 Å, respectively, are additionally included. **c)** Plot of the distance between the ε-nitrogen of H379 and the linker 4'-phosphate group of NADP$^+$ along 3 replicas of 500 ns of MD simulations for the WT-NADP$^+$ (left) and plot of the distance of the salt-bridge interaction between the positively charged amino group of H379K and the linker 4'-phosphate group of NADP$^+$ along 3 replicas of 500 ns of MD simulations for the V9-NADP$^+$ (right). Average distances (dashed black line) of 11.0±1.2 Å and 6.9±2.7 Å, respectively, are also shown. All distances are represented in Å.

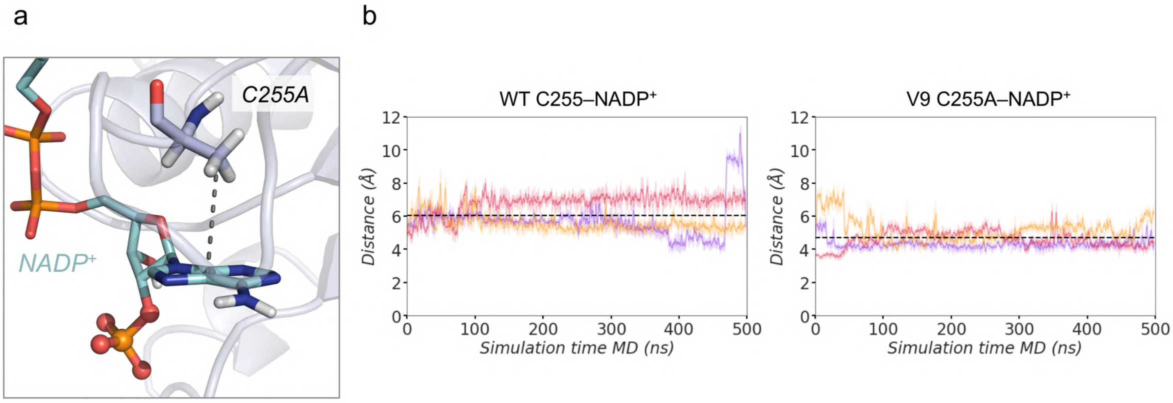


Figure C5. MD simulations analysis with focus on mutation C255A. a) Representative structure of V9-NADP$^+$ binding pocket with the CH··π interaction between the adenine ring of NADP$^+$ and the β-carbon of the side chain of C255A. **b)** Plot of the distance between the center of mass (COM) of the NADP$^+$ adenine ring and the β-carbon of the side chain of C255 in the case of WT (left) and the β-carbon of the side chain of C255A in the case of V9 (right) along 3 replicas of 500 ns of MD simulations (shown in red, orange, and purple) for the WT-NADP$^+$ and the V9-NADP$^+$. Average distances from all replicas of 6.0±1.0 Å and 4.7±0.7 Å, respectively, are also shown with a dashed black line. All distances are represented in Å.

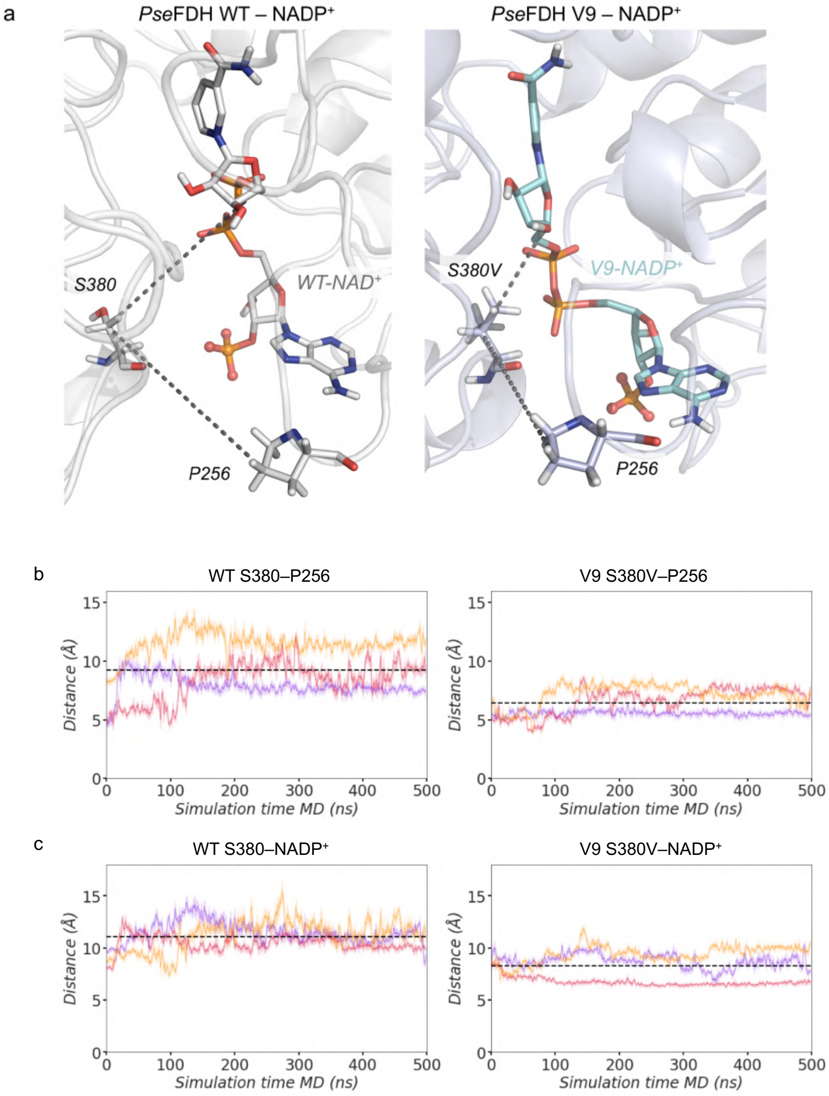


Figure C6. MD simulations analysis with focus on mutation S380V. a) Representative structures of WT-NADP+ (left) and V9-NADP+ (right) binding pocket with the interactions between the β-carbon of the side chain of S380(WT)/V380(V9) and the β-carbon of the side chain of P256 and the interactions between the β-carbon of the side chain of S380(WT)/V380(V9) and the nicotinamide ribose group of NADP+ highlighted. **b)** Plot of the distance between the β-carbon of the side chain of S380(WT)/V380(V9) and the β-carbon of the side chain of P256 along 3 replicas of 500 ns of MD simulations for the WT-NADP+ and the V9-NADP+. Average distances from all replicas (dashed black line) of 9.2±2.1 Å and 6.5±1.1 Å, respectively, are also included. **c)** Plot of the distance between the β-carbon of the side chain of S380(WT)/V380(V9) and the nicotinamide ribose group of NADP+ along 3 replicas of 500 ns of MD simulations for the WT-NADP+ and the V9-NADP+. Average distances of 11.1±1.3 Å and 8.3±1.4 Å, respectively, are shown with a dashed black line. All distances are represented in Å.

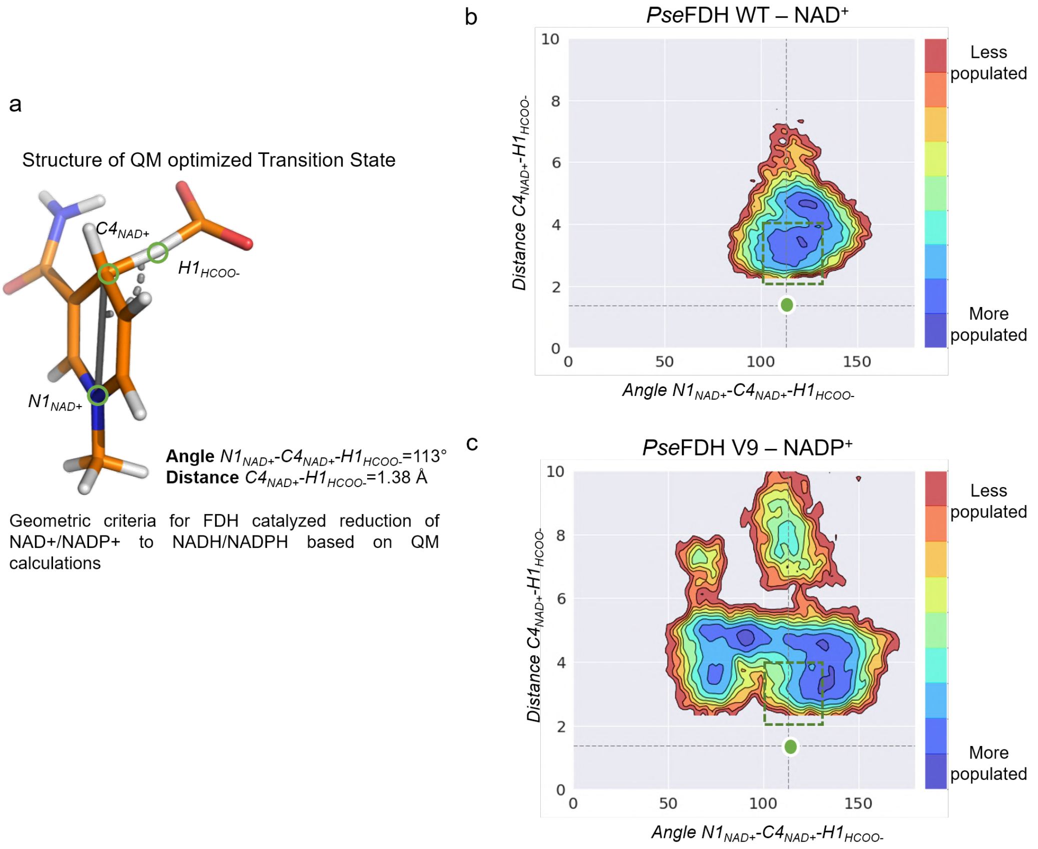


Figure C7. QM studies and conformational population analysis based on the QM-derived geometric criteria. a) Structure of the QM optimized Transition State for PseFDH catalyzed reduction of NAD$^+$/NADP$^+$ with the optimal angle and distance for hydride transfer reaction. A truncated computational model of the cofactor was used in the TS calculations (see computational details). **b)** Conformational population analysis based on the geometric criteria (hydride transfer distance versus angle) for PseFDH hydride transfer in the case of WT-NAD$^+$. **c)** Conformational population analysis based on the geometric criteria (hydride transfer distance versus angle) for PseFDH hydride transfer in the case of V9-NADP$^+$. The plots have been constructed using the angle $N1_{NAD+/NADP+}$-$C4_{NAD+/NADP+}$-$H1_{HCOO-}$ and the distance $C4_{NAD+/NADP+}$-$H1_{Hcoo-}$ sampled along 3 replicas of 500 ns MD simulations for WT-NAD$^+$ and V9-NADP$^+$. The catalytic distance (1.38 Å, represented by a horizontal dashed black line; value obtained from QM calculation) and the proper angle (*ca.* 113°, represented by a vertical dashed black line; value obtained from QM calculation) required for hydride transfer is represented by a green dot. The range of distances and angles considered as catalytically relevant in our MD simulations are those found within the green box (distances that range from 2 to 4 Å and angles from 100° to 130°).