COMPUTATIONAL EXPLORATION AND DESIGN OF HHDH VARIANTS WITH NOVEL SYNTHETICALLY USEFUL FUNCTIONALITIES

Miquel Estévez-Gay

Universitat
de Girona

DOCTORAL THESIS

Computational exploration and design of HHDH variants with novel synthetically useful functionalities

Miquel ESTÉVEZ-GAY

May 10th, 2023

Universitat de Girona

DOCTORAL THESIS

Computational exploration and design of HHDH variants with novel synthetically useful functionalities

Author:
Miquel ESTÉVEZ-GAY

Supervisor:
Prof. Dr. Sílvia OSUNA

A thesis submitted in fulfillment of the requirements for the degree of Doctor in Bioinformatics

in the

CompBioLab

May 10th, 2023

List of Publications

Chapters 4 to 6 of this thesis are respectively based on the following publications:

L. Wang, M. Marciello, M. Estévez-Gay, P. E. D. Soto Rodriguez, Y. Luengo Morato, J. Iglesias-Fernández, X. Huang, S. Osuna, M. Filice, S. Sánchez, *Angewandthe Chemie International Edition*. **2020**; *59*, 21080. DOI: 10.1002/anie.202008339

M. Estévez-Gay, J. Iglesias-Fernández, S. Osuna. Conformational Landscapes of Halohydrin Dehalogenases and Their Accessible Active Site Tunnels. *Catalysts*. **2020**; *10*(12):1403. DOI: 10.3390/catal10121403

J. Wessel, G. Petrillo,M. Estevez-Gay, S. Bosch, M. Seeger, W.P. Dijkman, J. Iglesias-Fernández, A. Hidalgo, I. Uson, S. Osuna, A. Schallmey. Insights into the molecular determinants of thermal stability in halohydrin dehalogenase HheD2. *The FEBS Journal*. **2021**; *288*, 4683-4701. DOI: 10.1111/febs.15777

Other publications not included in this thesis:

M. Floor, K. Li, M. Estévez-Gay, L. Agulló, P. M. Muñoz-Torres, J. K. Hwang, S. Osuna, J. Villà-Freixa. SBMOpenMM: A Builder of Structure-Based Models for OpenMM. *Journal of Chemical Information and Modeling*. **2021**; *61*(7), 3166-3171. DOI: 10.1021/acs.jcim.1c00122

G. Casadevall, C. Duran, M. Estévez-Gay, S. Osuna. Estimating conformational heterogeneity of tryptophan synthase with a template-based Alphafold2 approach. *Protein Science*. **2022**; *31*(10):e4426. DOI: 10.1002/pro.4426

List of Abbreviations

HHDH	**H**alo**H**ydrin **D**e**H**alogenase
MD	**M**olecular **D**ynamics
HheA	**H**alo**H**ydrin D**e**halogenase (type) **A**
HheB	**H**alo**H**ydrin D**e**halogenase (type) **B**
HheC	**H**alo**H**ydrin D**e**halogenase (type) **C**
HheD	**H**alo**H**ydrin D**e**halogenase (type) **D**
HheG	**H**alo**H**ydrin D**e**halogenase (type) **G**
WT	**W**ild **T**ype
R9	**R**ound **9**
R18	**R**ound **18**
RNA	**R**ibo**N**ucleic **A**cid
DNA	**D**eoxyribo**N**ucleic **A**cid
mRNA	**m**essenger **R**ibo**N**ucleic **A**cid
TS	**T**ransition **S**tate
TTS	**T**ransition **S**tate **S**tabilization
PKA	**P**rotein **K**inase (type) **A**
PKC	**P**rotein **K**inase (type) **C**
MAP	**M**itogen-**A**ctivated **P**rotein
MAPKK	**M**itogen-**A**ctivated **K**inase **K**inase
Vmax	maximum **V**elocity
$\mathbf{K}_{M}$	**M**ichaelis **K**inetics constant
$\mathbf{K}_{cat}$	**Cat**alytic **K**inetics constant
$\mathbf{K}_{50}$	half (**50%**) **K**inetics concentration constant
$\mathbf{n}_{H}$	**H**ills′ coefficient
PCR	**P**olimerase **C**hain **R**eaction
qPCR	**q**uantitative **P**olimerase **C**hain **R**eaction
DE	**D**irected **E**volution
QM	**Q**uantum **M**echanics
MM	**M**olecular **M**echanics

ff	**F**orce**F**ield
GAFF	**G**eneral **A**mber **F**orce**F**ield
AF2	**A**lpha **F**old **2**
PDB	**P**rotein **D**ata**B**ase
PBC	**P**eriodic **B**oundary **C**conditions
RMS	**R**oot **M**ean **S**quare
RMSF	**R**oot **M**ean **S**quare **F**luctuation
PCA	**P**rincipal **C**omponents **A**nalysis
PCA	**P**rincipal **C**omponent
ICA	**I**ndependent **C**omponents **A**nalysis
tICA	**T**ime-lagged **I**ndependent **C**omponents **A**nalysis
TIC	**T**ime-lagged **I**ndependent **C**omponent
t-SNE	**t**-distributed **S**tochastic **N**eighbour **E**mbedding
FEL	**F**ree **E**nergy **L**andscape
CV	**C**ollective **V**ariable
LB	**L**ysogeny **B**roth
HDBSCAN	**H**ierarchical **D**ensity-**B**ased **S**patial **C**lustering and **A**pplications with **N**oise
BR	**B**ottleneck **R**adius
ΔG	**G**ibbs energy
MSM	**M**arkov **S**tate **M**odel

Universitat de Girona

Prof. Dr. Sílvia Osuna of Universitat de Girona,

I DECLARE:

That this thesis entitled "Computational exploration and design of HHDH variants with novel synthetically useful functionalities", presented to obtain a doctoral degree. has been carried and completed by Mr. Miquel Estévez Gay under my supervision and that meets the requirements to opt for an International Doctorate.

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

Signature

Prof. Dr. Sílvia Osuna Oliveras

Girona, May 9th, 2023

a la famila,
als amics,
als que ja no hi son,
als que ho han fet possible.

“It is those who possess wisdom who are the greatest fools. History has shown us this. You could say that this is the final warning from God to those who resist.”

Rintaro Okabe

Acknowledgements

A tothom qui ha fet aquesta tesi possible. Des de la família que sempre ha estat quan es necessitava, els amics que han fet que tot sigui molt més senzill, companys de grup que han fet tot possible. També per descomptat a la **Sílvia** que va veure a un pobre biòleg amb ganes d'apendre aquest nou món i em va donar tots els recursos, opcions i paciència per fer-ho possible.

Contents

List of Figures

List of Tables

Summary of the thesis

Enzymes are the best catalysts. They are the main catalysts in cells and have been exposed to millions of years of natural evolution by including random mutations in their sequence and posterior selection. Some enzymes show extreme catalytic rates, selectivity, or stability, but not all are suitable for industrial or pharmaceutical applications. Their use in industrial contents would be very advantageous, thanks to the fact that enzymes are naturally biodegradable molecules, work in water-based solvents, and are non-toxic. These properties are critical for a suitable future for the new generations. It is crucial to design enzymes for catalyzing industrially relevant reactions.

One enzyme family with significant usage in the pharmaceutical industry is Halohydrin Dehalogenasses (HHDH). These enzymes convert (S)-4-chloro-3-hydroxybutyrate into ethyl (R)-4-cyano-3-hydroxybutyrate, a precursor of statin drugs that need to be enantiomerically pure. Not all HHDHs can perform this catalysis due to their inability to accept the substrate (they have a limited substrate scope), and insufficient enantioselectivity, stability, or activity. The design of new enzymes that display good properties in the selected industrial environment is nowadays possible, thanks to the experimental Directed Evolution technique. Still, this protocol mutates residues randomly. The effect of the mutations is not rationalized and usually requires the production and screening of multiple (thousands) variants, which has a high cost associated.

Computational protocols, based, for instance, on Molecular Dynamics (MD) simulations, allow for rationalizing the effect of the introduced mutations onto the ensemble of conformations that the enzymes can explore. Still, analyzing MD simulation outputs can be challenging, and there is no gold standard that gives good results and is computationally feasible. From these MD simulations, the identification of which amino acid positions need to be changed to enhance a given property is also not straightforward.

In this thesis, a novel pipeline for analyzing the variance obtained during the MD simulations and the accessible tunnels has been developed and is described in **Chapter 4**. This new protocol was applied to explore the tunnels in all HHDH families, compare the results with the reported features of each HHDH, and propose new mutagenesis sites (**Chapter 5**). **Chapter 6** describes the newly discovered D-family HHDHs and some proposed variants from Prof. Anett Schallmey's group, and the thermal stability mechanism is unveiled. Finally, in **Chapter 7**, the most evolved variant generated via Directed Evolution, i.e., HheC R18, is experimentally and computationally characterized to rationalize the effect of the randomly introduced mutations and how these affect the stability, oligomerization, cooperativity, and catalytic parameters of the HHDH enzyme.

Resum de la tesi

Els enzims són els millors catalitzadors que hi ha. Són els principals catalitzadors a les cèl·lules i han estat exposats a milions d'anys d'evolució natural, incloent-hi mutacions aleatòries en la seqüència i posterior selecció. Alguns enzims mostren velocitats catalítiques, selectivitat o estabilitat molt elevades, però no tots són adequats per a aplicacions industrials o farmacèutiques. El seu ús industrial seria molt avantatjós, gràcies al fet que els enzims són molècules naturalment biodegradables, funcionen en dissolvents de base aquosa i no són tòxiques. Aquestes propietats són crítiques per a un futur adequat per a les noves generacions. És crucial dissenyar enzims per catalitzar reaccions industrialment rellevants.

Una família d'enzims amb un ús significatiu a la indústria farmacèutica són les Deshalogenases d'Halohidrines (HHDH). Aquests enzims converteixen el (S)-4-clor-3-hidroxibutirat a (R)-4-cià-3-hidroxibutirat d'etil, un precursor de les estatines que han de ser enantiomèricament pur. No totes les HHDH poden realitzar aquesta catàlisi a causa de la seva incapacitat per acceptar el substrat (tenen un ventall de substrats limitat), enantioselectivitat, estabilitat o activitat insuficients. El disseny de nous enzims que mostrin bones propietats a l'entorn industrial seleccionat és avui en dia possible gràcies a la tècnica experimental d'Evolució Dirigida. Tot i això, aquest protocol muta residus aleatòriament. L'efecte de les mutacions no està racionalitzat i normalment requereix la producció i selecció de múltiples (milers) variants, cosa que té un alt cost associat.

Els protocols computacionals, basats, per exemple, en simulacions de Dinàmica Molecular (MD), permeten racionalitzar l'efecte de les mutacions introduïdes al conjunt de conformacions que els enzims poden explorar. Així i tot, analitzar els resultats de la simulació de MD pot ser un desafiament, i no hi ha un estàndard que sempre brindi bons resultats i sigui computacionalment factible. A partir d'aquestes simulacions de MD, la identificació de les posicions d'aminoàcids que s'han de canviar per millorar una propietat determinada no és senzill.

En aquesta tesi, s'ha desenvolupat un protocol nou per analitzar la variància obtinguda durant les simulacions MD i els túnels accessibles i es descriu al **Capítol 4**. Aquest nou protocol es va aplicar per explorar els túnels a totes les famílies HHDH, comparar els resultats amb les característiques reportades de cada HHDH, i proposar nous llocs de mutagènesi (**Capítol 5**). El **Capítol 6** descriu els HHDH de la família D acabats de descobrir i algunes variants proposades pel grup de la Prof. Anett Schallmey, i es revela el mecanisme d'estabilitat tèrmica. Finalment, al **Capítol 7**, la variant més evolucionada generada a través de l'Evolució Dirigida, és a dir, HheC R18, es caracteritza experimentalment i computacionalment per racionalitzar l'efecte de les mutacions introduïdes aleatòriament i com aquestes afecten l'estabilitat, l'oligomerització, la cooperativitat i paràmetres catalítics de l'enzim HHDH.

Resumen de la tesis

Las enzimas son los mejores catalizadores que existen. Son los principales catalizadores en las células y han estado expuestos a millones de años de evolución natural, incluyendo mutaciones aleatorias en su secuencia y posterior selección. Algunas enzimas muestran velocidades catalíticas, selectividad o estabilidad muy elevadas, pero no todas son adecuadas para aplicaciones industriales o farmacéuticas. Su uso industrial sería muy ventajoso, gracias a que las enzimas son moléculas naturalmente biodegradables, funcionan en disolventes de base acuosa y no son tóxicas. Estas propiedades son críticas para un futuro adecuado para las nuevas generaciones. Es crucial diseñar enzimas para catalizar reacciones industrialmente relevantes.

Una familia de enzimas con un uso significativo en la industria farmacéutica son las Deshalogenasas de Halohidrinas (HHDH). Estas enzimas convierten el (S)-4-cloro-3-hidroxibutirato en (R)-4-ciano-3-hidroxibutirato de etilo, un precursor de las estatinas que deben ser enantioméricamente puras. No todos los HHDH pueden realizar esta catálisis debido a su incapacidad para aceptar el sustrato (tienen un abanico de sustratos limitado), enantioselectividad, estabilidad o actividad insuficientes. El diseño de nuevas enzimas que muestren buenas propiedades en el entorno industrial seleccionado es hoy posible gracias a la técnica experimental de Evolución Dirigida. Aún así, este protocolo muta residuos aleatoriamente. El efecto de las mutaciones no está racionalizado y normalmente requiere la producción y selección de múltiples (miles) variantes, lo que tiene un alto costo asociado.

Los protocolos computacionales, basados, por ejemplo, en simulaciones de Dinámica Molecular (MD), permiten racionalizar el efecto de las mutaciones introducidas en el conjunto de conformaciones que las enzimas pueden explorar. Aún así, analizar los resultados de la simulación MD puede ser un desafío, y no existe un estándar que siempre brinde buenos resultados y sea computacionalmente factible. A partir de estas simulaciones MD, la

identificación de qué posiciones de aminoácidos deben cambiarse para mejorar una propiedad determinada tampoco es sencilla.

En esta tesis, se ha desarrollado un protocolo novedoso para analizar la varianza obtenida durante las simulaciones MD y los túneles accesibles y se describe en el **Capítulo 4**. Este nuevo protocolo se aplicó para explorar los túneles en todas las familias HHDH, comparar los resultados con las características reportadas de cada HHDH, y proponer nuevos sitios de mutagénesis (**Capítulo 5**). El **Capítulo 6** describe los HHDH de la familia D recién descubiertos y algunas variantes propuestas por el grupo de la Prof. Anett Schallmey, y se revela el mecanismo de estabilidad térmica. Finalmente, en el **Capítulo 7**, la variante más evolucionada generada a través de la Evolución Dirigida, es decir, HheC R18, se caracteriza experimental y computacionalmente para racionalizar el efecto de las mutaciones introducidas aleatoriamente y cómo estas afectan la estabilidad, la oligomerización, la cooperatividad y los parámetros catalíticos del enzima HHDH.

Chapter 1

Introduction

1.1 Preface

The most prominent product obtained by the pharmaceutical industry is organic solvent waste, even with all the recycling and recovery processes in place (Slater et al., 2010). Most organic solvent waste is toxic, and its disposal could be more straightforward. Also, accidents and spills are not uncommon events[1], even with all the security protocols in place (Shareefdeen, 2022). There is, therefore, a need to reduce the amount of organic solvent used in the pharmaceutical industry. However, to tackle this problem, we need to understand why organic solvents are used in the first place.

Organic solvents are used for extracting, solubilizing, and purifying the targeted compound. Usually, the compound must be modified to get the final product. A chemical reaction has to occur with a high reaction rate and selectivity, often accomplished thanks to the use of a catalyst. A **catalyst**[2] is a molecule that speeds up a reaction rate without being consumed in the process and without modifying the overall standard Gibbs energy change in the reaction, so it can be reused and can speed up (i.e., catalyze) the reaction again. Many catalysts (and many substrates) are soluble in an organic solvent, and thus the reaction occurs in the organic solvent.

Catalysts are essential to the chemical reactions that sustain life in all living organisms but also drive many industrial processes (van Santen et al., 1999). **Enzymes**, in particular, are highly efficient catalysts indispensable for the proper functioning of all living organisms. Enzymes are proteins capable of catalyzing a wide variety of reactions, including those central to metabolism, such as the breakdown of sugars and the synthesis of nucleic acids and proteins. Compared to traditional catalysts, such as metal-based catalysts, enzymes have several advantages. They are highly specific and able to catalyze a single reaction or a small set of closely related reactions (thus, they present high **selectivity**). They are also much more environmentally friendly, can operate under milder

[1] from: http://www.factsonline.nl/accidents/%205405/91993_INFLAMMABLE%20SOLVENT/chemical-accidents-with-inflammable-solvent

[2] IUPAC definition: https://goldbook.iupac.org/terms/view/C00876

greener conditions, and are water-soluble, making them suitable for various applications.

Obtaining enzymes for all reactions used in industry would be a giant step towards a greener world. Despite the importance of enzymes, their biological catalytic activity is often insufficient for many industrial and biomedical applications. Therefore, there is a growing interest in enzyme design, which involves the modification or engineering of enzymes to enhance or alter their activity, stability, selectivity, and tolerance to organic solvents.

This work might impact how we can study and make statistically relevant analyses of the computational simulations of enzymes, inferring in the importance of the conformational dynamics of enzymes in order to rationalize the mutation-induced changes in activity with the final long-term goal of designing new enzyme variants. To that end, the synergistic collaboration between computational and experimental work is essential for an efficient design of industrially-relevant.

1.1.1 Enzymes: Nature's biocatalysts

Enzymes are the catalysts that Nature and evolution have selected for performing the catalysis needed for all living creatures (Cooper, 2000). Enzymes are only one of many biocatalyst types, but the major ones. Other types of biocatalysts present in nature are RNA and metal organo-complexes, but those are usually coupled to a proteic part. Cells have mechanisms for regulating the activity of the biocatalysts through the amino acid chain. This is a process called allostery, which will be discussed later in this introduction.

Enzymes are excellent biocatalysts thanks to their high activity, water solubility, and selectivity. But the main feature that makes enzymes the best catalyst for living organisms is the central dogma of molecular biology [Figure 1.1]: DNA can duplicate via replication, can be transformed into RNA via transcription, and the latter one can be converted into a protein via translation. This makes it easy to modify the structure of proteins by changing the DNA sequence, and the next generation can carry out these modifications if they are good. That confers enzymes with an enormous evolutive advantage over other catalysts. Nevertheless, RNA can also do this, but the structure of RNA is limited to 4 canonic bases or building blocks, compared to twenty canonic amino acids (AA) that form proteins (Crick, 1970).

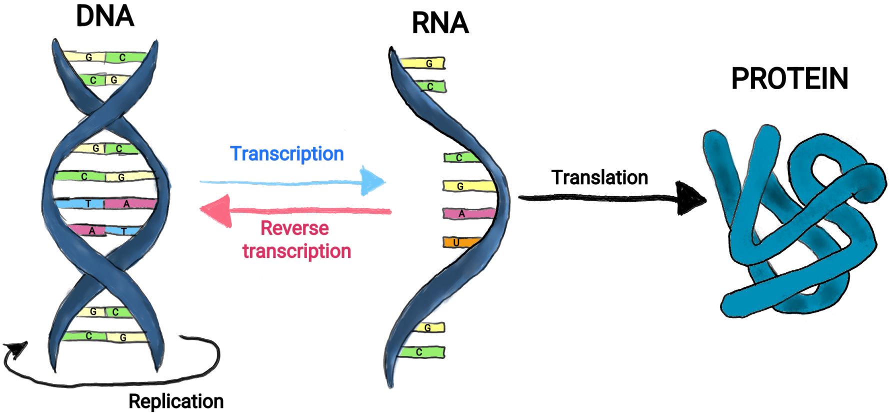

FIGURE 1.1: Scheme of the central dogma of molecular biology. DNA is transcribed into mRNA that it gets translated into a protein. In the Scheme it is also depicted the replication of the DNA and the Reverse transcription from RNA to DNA.

Enzymes are proteins with catalytic features. Proteins are one of the most diverse and important biomolecules that exist and are essential for life. It is important to know what proteins are, how they are made, and what structural properties have. Proteins are made up of long chains of amino acids. Each amino acid has a central carbon atom, called the alpha-carbon, which is bonded to four different groups: a hydrogen atom, an amino group, a carboxyl group, and a side chain. The side chain is unique to each amino acid and gives it unique chemical properties. Twenty different amino (canonic) acids can be chained through a peptide bond to make a protein. A protein's sequence of amino acids determines its unique three-dimensional structure and specific function (Brändén and Tooze, 1999).

Proteins are synthesized by cells through a process called protein synthesis, which occurs in the ribosomes. Protein synthesis involves two main steps: transcription and translation. During transcription, the information stored in DNA is transferred to a complementary RNA molecule. This RNA molecule is called messenger RNA (mRNA) (Alberts et al., 2014).

During translation, the mRNA molecule is used as a template to guide protein synthesis. The process of translation occurs in the ribosomes, which are tiny organelles found in the cytoplasm of cells. Translation involves using transfer RNA (tRNA) molecules, which bring the correct amino acids to the ribosomes and link them together in the right order to form a protein (Alberts et al., 2014).

The primary structure of a protein is the specific sequence of amino acids in the protein. The secondary structure of a protein refers to the local geometric arrangements of the amino acid residues within the protein, such as the alpha helices and beta sheets. The tertiary structure of a protein is the three-dimensional structure of the protein as a whole, which is determined by the interactions between the amino acid residues and their side chains. The quaternary structure of a protein refers to the arrangement of multiple protein subunits in a protein complex (oligomerization) and the presence of non-proteic parts like metals, RNA, or other organic molecules (Alberts et al., 2014).

1.1.2 Enzymes: lowering the activation energy

Enzymes transform substrates into products by reducing the energy barrier (named **activation energy**) that needs to be overcome for the reaction to happen. The maximum energy point along the reaction coordinate is actualy the **transition state (TS)**. The higher the activation energy is, more difficult it is for the reaction to happen. If the barrier is high, other parameters like increasing the pressure or temperature might add energy to the system, therefore increasing the probability of overcoming the TS.

Apart from that, the enzyme can stabilize the TS making it lower in energy compared to in solution, thereby reducing the activation energy. This concept central to enzyme catalysis is usually known as **transition state stabilization (TSS)**(Szefczyk et al., 2004).

Another way to decrease the activation energy is by **Ground State Destabilization**(Anderson, 2001). Ground state destabilization theory proposes that a catalyst lowers the activation energy required for a reaction to proceed by increasing the energy of the reactants, making it easier for them to overcome the energy barrier and reach the transition state. This is achieved by providing an alternative reaction pathway that involves the formation of a new intermediate or transition state that has a lower energy barrier than the original pathway. This is still a topic of debate(Rindfleisch et al., 2022).

For achieving this TSS, enzymes present a highly preorganized active site pocket, which contains all catalytic and binding residues precisely positioned for catalysis. As it will be discussed in the next section, enzymes apart from this preorganized active site have the ability to change conformation, which is also crucial for enhanced function.

1.1.3 Protein Dynamics

An extra layer of complexity about anzymes and proteins is that they are not static structures, unlike the impression one can get from the X-ray structures or the one obtained by homology modeling. In solution, proteins are flexible and can adopt an ensemble of conformations key for their function. The protein may display different properties in all the different conformations they can adopt. Thanks to this, while studying proteins, it is vital to understand how a mutation (i.e., amino acid change) may affect the 3D structure and also the relative stabilities of all different conformations, thus changing the properties of the protein (Kirby, 1996).

More specifically, in the case of enzymes, it is known that the different conformations they can adopt have a high impact on their catalytic properties. For instance a change in conformation might be needed for the enzyme-substrate formation. There are three models that describe the enzyme-substrate binding process: the lock-and-key model, the induced fit model, and the conformational selection model (Orosz and Vértessy, 2021).

The **lock-and-key model** of enzyme-substrate binding is one of the oldest and most widely recognized. According to this model, the enzyme and substrate are compared to a lock and key, respectively, where the enzyme's active site is shaped to fit precisely the substrate. This specific shape complementarity allows the enzyme to bind the substrate and catalyze the reaction. This model can explain some enzymes' good selectivity but fails, for instance, to explain substrate promiscuity (Orosz and Vértessy, 2021).

The **induced fit** model is a modification of the lock-and-key model. It is based on the idea that the active site of an enzyme is not static. Still, it changes slightly upon substrate binding. This change in conformation helps the enzyme to bind the substrate more tightly and leads to a more optimal orientation of the substrate for catalysis. In this model, the protein conformation shift is mandatory and sometimes is the rate-determining step of the reaction (Orosz and Vértessy, 2021).

The **conformational selection** model describes the enzyme as an ensemble of conformations, which can be shifted due to binding of substrate, inhibitor, cofactor or effector. In fact, along the catalytic cycle, the enzyme changes conformation for allowing substrate binding or product release. Apart from that, enzymes can change conformations due to alterations in the physical conditions (such as temperature and preasure). In this mechanism, the ligand seems to "select" and stabilize a higher-energy conformation for binding (Orosz and Vértessy, 2021).

It is also important to note that these models are not mutually exclusive. In reality, enzymes use a combination of these mechanisms to increase their specificity and efficiency of substrate binding and catalysis (Orosz and Vértessy, 2021).

1.1.4 Allostery

Allostery (Motlagh et al., 2014) refers to the ability of a protein, such as an enzyme, to change its conformation in response to the binding of a molecule at a specific site on the protein that is distinct from the active site. This change in conformation can lead to changes in the activity of the protein, including changes in the rate at which it catalyzes a reaction, its specificity or affinity for a particular substrate. Allosteric regulation can be either positive or negative. Positive regulation will increase the protein's activity, whereas negative regulation will lead to a decrease in the protein's activity.

Allostery is an extremely useful tool for the cells to downregulate or upregulate certain metabolic routes in response to an external stimulus, such as a hormone. In human cells, almost all proteins involved in metabolic routes need an allosteric effect to have the desired effect. Some examples are the membrane receptors, G-coupled proteins, the adenylate cyclase, all kinds of kinases (PKA, PKC, MAP, MAPKK, etc.), ion channels, transcription factors, and many others.

However, additional allosteric effects often take place and are essential. If we focus on the G-protein example, it has different activity depending on whether it is in a monomeric or multimeric state. It is known that the presence of other amino acid chains (i.e., protein partners or even peptides) can regulate the enzyme's activity. Thus not only a small molecule but also the oligomerization state may have a considerable effect on enzyme activity (this regulation is also called **cooperativity**). Cooperativity can be negative, like in the previous example (the other chain inactivates the enzyme), or positive, when the presence of another subunit makes the enzyme more efficient.

It is proposed that the allosteric effect travels from the binding site to the catalytic site via a series of internal interactions that the protein residues have (H-bonds, stacking, van der Waals and electrostatic)(Hu et al., 2009), inducing a change in the active site. This alteration of the active site makes the protein more or less active or efficient. This concept is known as **intrinsic allostery** (Gunasekaran, Ma, and Nussinov, 2004; Boehr, Nussinov, and Wright, 2009). Any change in the amino acid interactions of a protein may have a significant effect on the enzyme activity, even if such modification is far away from the active site. This means that any mutation in the protein's amino acid sequence can affect activity thanks to the intrinsic allosteric concept.

1.1.5 Enzyme Kinetics

Enzyme kinetics studies the reaction rates for an enzyme-catalyzed chemical reaction. As described previously, an enzyme speeds up the chemical reactions by decreasing the activation energy. This affects the reaction kinetics and the rates of the reactions by increasing the concentration of reactive species, either by breaking down reactants into more reactive forms or by forming intermediates that can react more readily.

The reaction rate is the value that describes the reaction velocity only by monitoring the substrate consumption or product release. To explore the enzyme's effect on the reaction rate, multiple models have been developed to monitor different catalysts and elucidate how they work.

It is known that (Segel, 2013) usually adding more substrate into the catalyzed reaction media increases the reaction velocity, but there is a limit called Vmax. Once the rate is close to Vmax, adding more substrate will have almost no effect, creating the hyperbolic curve that has an asymptote at the Vmax value.

In 1913, the German biochemist **Leonor Michaelis** and the Canadian physician **Maud Menten** published the widely known model nowadays (**Michaelis-Menten**). This model assumes that the reaction follows the next formula:

$$E + S \rightleftharpoons [ES] \rightarrow E + P \tag{1.1}$$

where E is the enzyme, S is the substrate(s), ES is the complex formed by the enzyme and substrate(s), and finally, P is the product(s) of the reaction. In this representation, reactions have different rate constants, where the **catalytic rate constant (**k_{cat}**)** is the rate of the non-reversible product formation reaction (Johnson and Goody, 2011).

Assuming the equilibrium, the steady-state and that the enzyme is at a much lower concentration than the substrate, we can calculate the reaction rate in the following way:

$$v = \frac{d[P]}{dx} = V_{max}\frac{[S]}{K_M + [S]} = k_{cat}[E]_0\frac{[S]}{K_M + [S]} \tag{1.2}$$

where $[E]_0$ is the initial enzyme concentration, and K_M is the value of $[S]$ when the rate is at its half maximum. K_M measures the enzyme's affinity for the substrate.

This model has become the standard way of getting the kinetic values of catalysts and makes it easy to compare multiple enzymes' performance for the same reaction by comparing k_{cat}, K_M, or even the catalytic efficiency (k_{cat}/K_M).

To get the parameters using experimental values, the amount of substrate consumed and/or product formed needs to be measured at different times for a set of initial substrate concentrations. Then, the data has to be fitted in a non-linear regression, and the parameters can be extracted from the regression curve.

Even with that, not all enzyme-catalyzed reactions follow a Michaelis-Menten distribution. Some enzymes display a sigmoid saturation curve, which often indicates a cooperative binding in the catalyst. This means the enzyme has multiple active sites, but they do not work entirely independently. Positive cooperativity implies that the active site has a higher affinity for the substrate (K_M) if another active site has a substrate in the active site. To model this behavior, the **Hill–Langmuir (Hill)** equation is used (Srinivasan, 2021):

$$v_0 = V_{max}\frac{[S]^{n_H}}{K_{50} + [S]^{n_H}} \tag{1.3}$$

The only difference between the Michaelis-Menten and Hill equation is the introduction of the n parameter (Hill coefficient). The Hill coefficient is the parameter coefficient, where $n_H = 1$ means

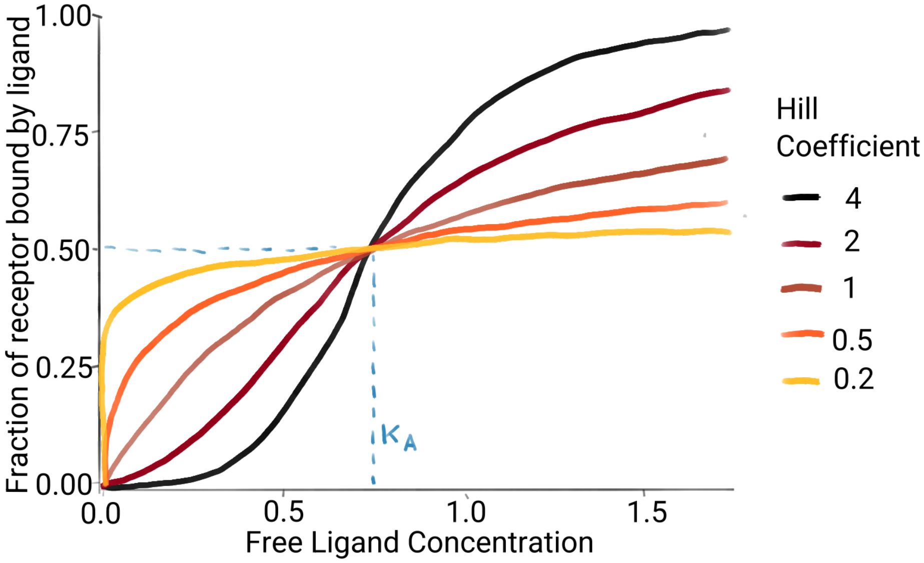

FIGURE 1.2: Plot that shows how the Hill distribution changes by changing the n_H value. In the plot it is depicted the concentration of ligand where the substrate concentration is at half in a Michaelis-Menten. This is useful to see that K_M and K_{50} are not comparable parameters.

no coefficient (and thus the equation equals the Michaelis-Menten), and a higher value denotes larger cooperativity between active sites. Also, because the formula is slightly different, the K_M parameter does not exist, but rather K_{50} can also be used to describe affinity[see Figure 1.2].

1.1.6 Enzyme design: an overview of strategies

Various protocols have been developed, given the importance of enzyme engineering to improve their efficiency (Leveson-Gower, Mayer, and Roelfes, 2019; Vaissier Welborn and Head-Gordon, 2019). These enzyme design strategies can be classified into rational and non-rational. Generally speaking, the non-rational techniques provide excellent results in terms of enzyme optimization but do not provide knowledge about the role of the added mutations. This knowledge is crucial for developing more robust and economically viable enzyme design strategies. On the contrary, rational methods have the advantage of being potentially cheaper and able to provide detailed information on the effect of mutations. Despite some success in the field, rationally designed enzymes still perform relatively poor compared to experimentally-engineered enzymes.

Non-rational enzyme design techniques are based on experimental methods, thanks to the ability to produce and test multiple enzymes using screening techniques. Screening refers to the testing of the new designs towards the target reaction pursued to be improved and then selecting those designs that showed an improvement. It can be done in multiple ways, limited to the ability to test each reaction and the necessary conditions. Thanks to this, in non-rational strategies, thousands and even millions of constructs can be tested.

There are multiple techniques that allow the introduction of random mutations into the enzyme, but almost all of them are based on modifying the DNA in a mutagenic Polymerase Chain Reaction (PCR). The PCR is the laboratory equivalent of the DNA replication that occurs in the cells, but a mutagenic PCR is the one that introduces variations in the DNA on purpose. It is usually done by introducing designed fragments of DNA (mutagenic Primers) that will give a modified protein or by changing the PCR conditions to force errors in the replication process so that random mutations will be introduced (error-prone PCR). This process creates diversity in the gene, and coupled with the proper screening method and repeated several times in order to have a Darwinian selection, we obtain the **Directed Evolution (DE)** (Wang et al., 2021) method, the most known technique to design new enzymes and the one that shows better results. Thanks to this, the Nobel Prize of Chemistry 2018 was given to Frances H. Arnold, George P. Smith, and Sir Gregory P. Winter for the development of Directed Evolution and Phage display screening methods[3].

Rational enzyme design involves experimental or computational techniques (or a combination of both) to determine which residues

[3] https://www.nobelprize.org/prizes/chemistry/2018

need to be mutated and to which specific amino acid (Lassila et al., 2006). Information on the reaction and the enzyme's structure is needed first. With this, the function of specific amino acids in the enzyme activity and structure can be understood, and therefore mutations or even new enzyme scaffolds can be proposed. One of the most known examples of rational design approaches is the **Inside-Out protocol** (Zanghellini et al., 2006).

The Inside-Out protocol is based on the concept of theozyme (short for ¨theoretical enzyme¨) to model the transition state(s) of interest and apropiate it into a protein scaffold. Using Quantum Mechanics (QM) calculations, the transition state of the reaction of interest is modeled (Romero-Rivera, Garcia-Borràs, and Osuna, 2017). Additional amino acids can be added to the active site of the theoretical enzyme to have a stabilization effect. This protocol usually creates multiple theozymes for the desired reaction, which are then incorporated into an enzyme scaffold with Rosetta (or related software). Subsequent steps can introduce new variations in the scaffold, and the new variants are sometimes further evaluated by employing Molecular Dynamics (MD) in an explicit solvent before the experimental validation. Unfortunately, despite the initial successes, enzymes developed using this protocol are usually not active enough for practical usage, and Directed evolution needs to be applied to enhance the activity of the constructs. Since the first development of the inside-out protocol, many other approaches have been proposed as discussed in different reviews (Osuna, 2021), but none of them can design highly efficient enzymes rivaling natural ones or those obtained with DE.

Rational and non-rational techniques are not mutually exclusive. Enzyme structure, dynamics, and mechanism can be studied, revealing several critical positions in the amino acid sequence that can be selected for iterative mutagenesis. The techniques that use knowledge of the enzyme to propose positions and create ¨Small but smart¨(Jochens and Bornscheuer, 2010) mutation libraries are described as **Semi-Rational** enzyme design approaches.

1.2 Halohydrin dehalogenases (HHDHs)

This thesis focuses on the work done in understanding a family of enzymes called **Halohydrin dehalogenases (HHDHs)** (Schallmey and Schallmey, 2016). HHDHs are enzymes that catalyze the **dehalogenation** reaction of some toxic compounds for the host organisms, creating an epoxide and releasing the halide. The generated halide is then held in the so-called **halide binding site** and is released after the product. The active form of HHDH is reported to be tetrameric, but the presence of the dimeric form has also been found for some HHDHs(Wijngaard, Reuvekamp, and Janssen, 1991). Each active site has a conserved catalytic triad composed of Serine 132, Tyrosine 145, and Arginine 149 (numeration from HheC).

The mechanism of the dehalogenation reaction works as follows. The catalytic Serine and Tyrosine side chains are oriented towards the oxygen of the alcohol group of the substrate. The serine-alcohol hydrogen bond stabilizes the substrate in place, while the interaction between tyrosine and arginine polarizes tyrosine's oxygen, lowering its pK_a. After the deprotonation of the tyrosine, it can act as a base and deprotonate the alcohol group of the substrate, thus promoting the intramolecular nucleophilic attack of the alcoxide to the adjacent carbon releasing the halide and generating an epoxide [see Figure 1.3].

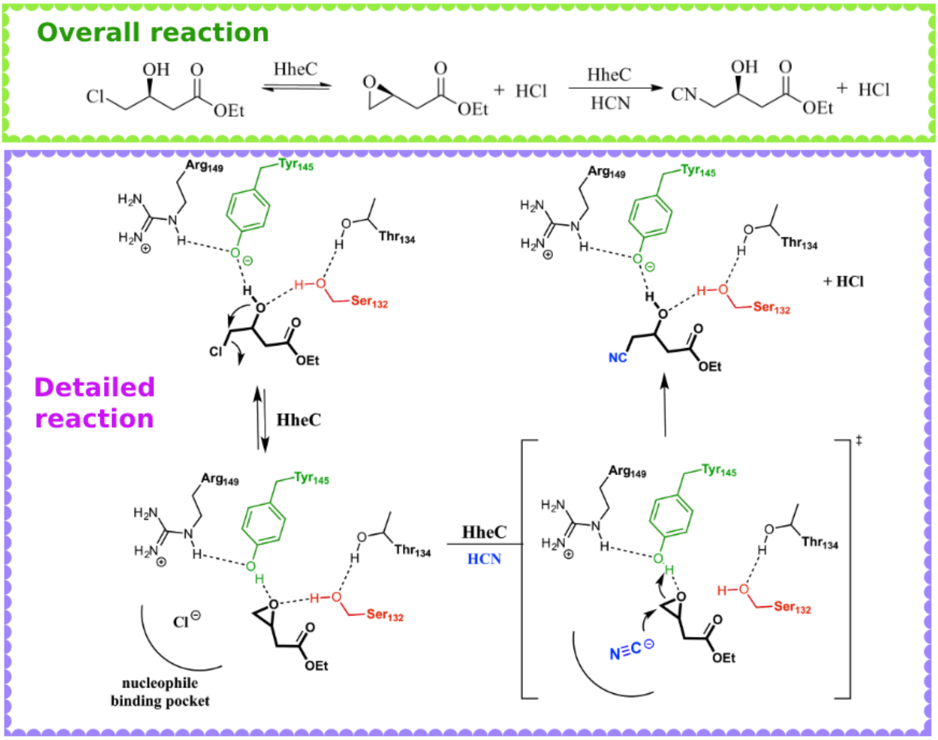

FIGURE 1.3: Chem-draw scheme displaying the mechanisms of HHDHs (residue numbers from HheC). On the left, the catalytic tyrosine acts as a base and abstracts the proton from the oxygen of the halohydrin, thus promoting the intramolecular nucleophylic attack on the carbon, releasing the halide (Cl^{-}) and creating the epoxide. Afterwards, a non-natural nucleophile can be positioned in the halide-binding site, and perform a nucleophilic attack at the (usually) less substituted carbon of the epoxide. In this step, Tyr acts as an acid.

This reaction is, however, not the only reason why HHDHs are interesting and important for the pharmaceutical and agrochemical industry. HHDHs can also catalyze the reverse reaction using other small and anionic nucleophiles, the **epoxide-ring opening reaction**. HHDHs can catalyze this secondary promiscuous reaction, where the epoxide is held in the active site, the chloride is released, and another small and charged nucleophile can occupy the halide binding site. With this new configuration, the catalytic serine and tyrosine interact with the oxygen of the epoxide. The new nucleophile attacks the carbon of the epoxide ring (usually the terminal carbon), and the tyrosine acts as an acid, thus protonating the oxygen and regenerating the alcohol [see Figure 1.3]. This second reaction is more important for the pharmaceutical industry due to the high enantio- and regio- selectivity that some HHDHs can show towards specific epoxides, making HHDHs an excellent choice for obtaining laboratory-evolved catalysts for obtaining precursors of drugs such as atorvastatin.

These enzymes were discovered in 1968 (Castro and Bartnicki, 1968), but HHDHs demonstrated their ability to catalyze epoxide ring-opening with different nucleophiles in 1991 (Nagasawa et al., 1992). It was found that they accept a wide range of charged and small nucleophiles for the epoxide ring-opening reactions, which give access to the corresponding unnatural alcohols like β-nitro and β-azido, as well as β-hydroxynitriles (Hasnaoui-Dijoux et al., 2008). From those nucleophiles, cyanide is one of the most used and studied, thanks to offering further transformation of compounds into an amino, amide, or carboxy group-containing molecules by the action of nitrile-amide converting enzymes (Elenkov, Hauer, and Janssen, 2006).

HHDHs show different substrate scope, stereoselectivity, enantioselectivity, and thermal stability. Thanks to the ability of HHDHs to clean toxic compounds and create important molecules for drug synthesis, these enzymes have been heavily studied and multiple types of HHDHs have been reported and grouped in families that go from A to G up to this date (Koopmeiners et al., 2016). The most studied and used HHDH is HheC, one of the first HHDHs reported and showing high affinity for aromatic substrates, high β-regioselectivity, and (R)-enantio-selectivity (Jong et al., 2005) as a wild type. Thanks to that, numerous mutations have been designed on top of the HheC scaffold to obtain higher activity, change the selectivity, and enhance the substrate scope and stability (Schallmey and Schallmey, 2016). Also, DE of HheC has been performed (Fox et al., 2007) to design the best HHDH to date for catalyzing (Ma et al., 2010) the conversion of halohydrin (ethyl (S)-4-chloro-3-hydroxybutyrate) into the corresponding hydroxynitrile with high selectivity fulfilling the requested industrial demands.

This thesis reports the knowledge generated from the computational and experimental study of this enzyme family and generated mutants. This knowledge is key for the generation of new designs for obtaining HHDHs with better functionalities. This hybrid computational and experimental thesis shows how a strong collaboration between both sides might lead to, but also highlights that theoretical work can successfully be used to predict and propose new mutations and not only to explain the observables.

Chapter 2

Methods

In this Chapter, the methods and the basic fundamentals for understanding the results reported in this thesis are provided. Molecular dynamics (MD) simulations, computational analysis methods, and laboratory techniques used to test, design, and validate the multiple Halohydrin dehalogenases (HHDHs) will be described in a theoretical and practical way.

2.1 Computational Methods for studying enzymatic systems

2.1.1 Classical Mechanics and Molecular Dynamics simulations

The big impact of enzymes' dynamism on catalysis, stability, substrate scope, and selectivity makes it the main point that will be studied in this thesis to understand and design better catalysts. In order to evaluate the enzyme conformational dynamics, the principal methodology used is Molecular Dynamics (MD) simulations. With MD simulations, one can explore the position of the atoms that compose the enzyme over time, but it is crucial to know the methodology used and the limitations of the methods (Dror et al., 2012).

For simulating the movements of the atoms composing an enzyme, multiple methods can be used. However, in contrast to more precise methods based on quantum mechanics (QM) that can be used to study bond breaking/forming but are computationally to expensive for evaluating the enzyme's conformational dynamics, molecular mechanics (MM) allow exploring the motions of all the atoms for an extended period of time (picoseconds up to some cases milliseconds). In MM, it is assumed that all atoms are charged spheres with a given radius and mass connected through previously defined permanent bonds; thus, there are no bond-forming or bond-breaking events. This approach based on studying the conformational dynamics with MM is also called classical MD simulations. In classical MD simulations, the potential energy of the system is computed by means of the sum of bonded and non-bonded terms, as shown in:

$$U_{tot} = U_{bonded} + U_{non-bonded} \tag{2.1}$$

The multiple parts or formulas that describe all the forces that take place at a given time in one atom are determined by the Force-Field (FF) (Ponder and Case, 2003a). In most FFs used in protein MD simulations, the bonded energies are divided into three terms: bond, torsions, and angles. Otherwise, the non-bonded terms are the van der Waals and electrostatic terms (Salomon-Ferrer et al., 2013).

$$U_{\vec{r}} = \sum_{i}^{bonds} k_{r,i}(r_i - r_{eq,i})^2 + \sum_{i}^{angles} (\theta - \theta_{eq,i})^2 + \\ + \sum_{i}^{dihedrals} \sum_{n} \frac{v_{n,i}}{2}[1 + cos(n\phi_i - \gamma_i)] + \sum_{i<j}^{atoms} (\frac{A_{j,i}}{R_{j,i}^{12}}) - \frac{B_{j,i}}{R_{j,i}^{6}}) + \sum_{i<j}^{atoms} \frac{q_i q_j}{4\pi\varepsilon_0 R_{j,i}} \tag{2.2}$$

- The **first term** (U_{bond}) contains the sum of potential energies that describe the bonds. These energies are described as a typical elastic or harmonic potential energy. This means that it is only determined by the square of the distance between atoms (x) and a constant (k). The constant depends on the type of bond we describe and are defined in the FF. This means that we can understand the bonds in the MD simulations as springs that join the atoms.

- The **second term** (U_{angle}) describes the sum over all angles formed by two consecutive bonds. With that, the potential energy caused by the displacement of the angles is defined. In this case, all angles are treated the same way, so no constant is used, and there is no equivalent "type of angles."

- The **third term** ($U_{torsions}$), the last one from the bonded terms, is the one that describes the bond torsions or twists. It is defined by a Fourier expansion of a constant that defines the type of bond and bond order and the cosine of the torsion angle.

This term describes the force that captures the steric and electrostatic interactions that might impede the twist of a bond.

- The **fourth term** is the van der Waals (U_{vdW}) interactions. This term is the sum of all interactions done by the simple fact that all atoms have electrons or "electron clouds," and those induce some short-lived multipoles into other atoms. These interactions are mainly attractive but are strongly repulsive over short distances. This is the typical part of the forcefield that exerts a higher computational cost, so a distance threshold is set in place (8-12 Å) where this interaction potential is zero. Also, a less resource-hungry descriptor is used, such as Lennard-Jones 12-6 function. This is described by the two parameters called well depth and collision diameter. The collision diameter is the parameter that defines when the vdW interaction between two atoms is zero, and the well depth is the parameter that describes the maximum attraction vdW potential energy that the two atoms can have. After this is defined, the only variable that changes this is the distance between the two atoms.

- The **fifth** and last **therm** is the electrostatic term, which describes the Coulombic (electrostatic) interactions between charged atoms and is computed using Coulomb's law, where q_1 and q_2 are the charges on the particles, r is the distance between the particles and (epsilon) the electric constant.

Then, the FF requires not only the formulas for the multiple energetic terms needed to be computed but also all parameters that the particles under investigation will need. Furthermore, in Amber forcefields, these parameters are created to replicate the characteristics of the molecules. For instance, several amber forcefields are made to simulate proteins, enzymes, and other biomolecules like DNA or RNA as accurately as possible. Other force fields are used to simulate water, solvents, and salts. The General Amber force field (GAFF) is widely used for parametrizing non-canonical amino acids, cofactors, or general organic molecules (Wang et al., 2004; Ponder and Case, 2003b).

With all of that, one can compute the interactions and energies in most scenarios one can encounter in simulating enzymes and biomolecules. However, to generate and propagate trajectories over time and evaluate how the system behaves, additional steps need to be taken. This is done in MD by following Newton's second law:

$$F_i = m_i * a_i \tag{2.3}$$

As the forces that each particle in the system is affected by have already been described, one can split the calculations into three components and get the acceleration of the particle after computing the total potential energy described by the force field:

$$-\frac{dU}{dr_i} = m_i a_i(t) \tag{2.4}$$

In MD simulations, the forces acting on the atoms in a molecule are calculated at each time step(Δ_T). The accelerations are then calculated using Newton's second law. The positions and velocities of the atoms are then updated using the accelerations, and the process is repeated at the next time step. This allows the motion of the atoms to be simulated over time, allowing the properties of the molecule to be studied. The time step used can vary and be modified depending on the situation, but it is often in the femtosecond range. In this thesis, and typically done in MD simulations of big systems, the time step is set at two femtoseconds. In this way, and because the vibration of C-H bonds is a much faster event, we can set the distance between H and a heavy atom as constant (Susskind and Hrabovsky, 2014).

Another important concept in MD simulations is the different available thermodynamic ensembles. Each ensemble represents a different set of conditions that the system is subjected to, and the choice of ensemble depends on the specific goals of the simulation; the conditions are the number of atoms (N), pressure (P), volume (V), and energy (E). These variables are used to define the thermodynamic state of the system. Several different types of ensembles are commonly used in MD simulations, including the microcanonical ensemble (NVE), the canonical ensemble (NVT), and the isothermal-isobaric ensemble (NPT) (Frenkel and Smit, 2002).

The microcanonical ensemble is used to study isolated systems in which the total energy is fixed. This type of ensemble helps study the thermodynamic properties of a system, such as its temperature, entropy, and specific heat.

The canonical ensemble is used to study systems with a thermostat so that the temperature is fixed. This type of ensemble helps study the behavior of a system at a constant temperature and predict how the system will respond to changes in temperature. This is the ensemble in which the production simulations of this thesis are made due to the mentioned control in the temperature and because the volume is also kept constant.

The isothermal-isobaric ensemble is used to study systems in which the temperature and pressure are fixed. This type of ensemble helps study the behavior of a system at a constant temperature and pressure and for predicting how the system will respond to changes in these variables.

2.1.2 Running MD simulations

The first thing needed to perform an MD simulation is the molecule or set of molecules in a given starting conformation. In this thesis, these molecules are mainly enzymes and other molecules that are substrates (if applicable). To get the initial protein structure, the large Protein Database (PDB) can be explored to check if crystal structures of the system in good quality and resolution are available. If this is the case, one must check if the protein structure is complete. If not, the missing parts need to be reconstructed with programs like the recent *alphafold2*(AF2) (Jumper et al., 2021), *Swiss-Model*[1], or others. Also, the X-ray will not contain the protons due to the resolution limit of the technique, so these need to be added to the system, setting a proper protonation state for all the residues. To do so, programs like *H++* (Gordon et al., 2005) or *PropKa* (Olsson et al., 2011) can be used to estimate the pKa of the residues in the

[1] https://swissmodel.expasy.org/

protein environment, at the pH at which the structure was obtained. Some exceptions are the histidine, aspartic and lysine residues because multiple protonation states are available. One has to look at the interactions that the residue can make to guess the most probable place for the protons to be, as well as check whether they have a catalytic role and assign the proper protonation state for the modeled reaction state.

Regarding the mutants, there are no X-ray resolved structures of most of the mutants and variants accessed in this thesis. To get the structure of the mutant, the Wild-Type (WT) protein can be modified using programs like PyMOL (Schrödinger, LLC, 2015a; Schrödinger, LLC, 2015b; Schrödinger, LLC, 2015c), ChimeraX(Pettersen et al., 2021), rosetta (Rohl et al., 2004), or more recently with AF2. For a few modifications, using the mutagenesis tool from PyMOL can be suitable if one carefully chooses the best rotamer and the minimization steps are increased further. For mutants with many modifications, such as HheC R18 (37 positions mutated), the program rosetta was used so that the mutations are done automatically (or AF2 could be used, which was only recently made available). The rotamers are selected based on an internal force field-based field minimization and scoring. The obtained protein is further verified and compared with the result from AF2.

Another important and applicable modification in this thesis is to access the proteins' oligomerization state. Proteins can, and normally do, form oligomers. We need to obtain the proper oligomerization state for the system. To do so, we can use programs like PISA (Krissinel and Henrick, 2007), Alphafold2 multimer, or PyMOL.

Once the quaternary structure is accessed, the cofactor (if the protein require it) needs to be included in the active site if we want to simulate the HOLO state or remove it (in the case that the substrate was in the X-ray structure) for the APO state.

The last step that needs to be considered before running the MD simulation is that the system needs to be solvated. To do so, and because enzymes are usually solvated in water, explicit solvents (water models) are added[see Figure 2.1]. This can be done by adding the protein to a box of pre-equilibrated water that comprehensively covers the protein in all directions. The solvent box must be sufficiently big to accommodate the protein and extra space from the protein surface until we reached the threshold used for the Lennard-Jones potential. To avoid edge effects, copies of the box are placed around the original's limits, creating the so-called Periodic Boundary Conditions (PBC). These new boxes are copies of the original one, and the new atoms are not part of the potential energy, but the presence of the PBC makes the effects of the box disappear.

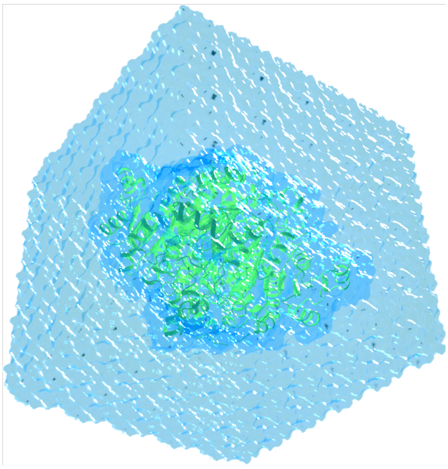

FIGURE 2.1: 3D render of a HHDH in the Solvent box. This system is prepared for being simulated.

The solvation box is added using the program tleap from the Ambertools (Case et al., 2005). With this, the box's dimensions are selected, and the ff and the water force field are set. In this thesis, TIP3P water (MacKerell Jr. et al., 1998) model is used. Counterions (Cl^- and Na^+) are usually added to set the charge of the whole system to zero.

The process of running an MD simulation typically involves the following steps:

1. **Minimization**: Once the system is prepared, the energy of

the system must be minimized by relaxing the positions of the atoms or molecules. This is done using an optimization algorithm that iteratively adjusts the positions of the atoms until the energy of the system is minimized. In the first minimization step, the protein is kept rigid by using restraints or external forces, minimizing the solvent box. Water molecules set by the tleap program are arranged in a pattern, which should be minimized to a more stable conformation. In the second minimization, the restraints are removed, and the whole system is minimized. This is especially relevant if we have modified the crystal a lot (including mutations or substrates)

2. **Heating**: After the system has been minimized, the next step is to heat the system to the desired temperature. This is usually done by gradually increasing the temperature over a series of time steps until the desired temperature is reached.

3. **Equilibration**: After the system has been heated, it is vital to ensure it has reached thermodynamic equilibrium. This is done by allowing the system to evolve for a sufficient number of time steps, during which the temperature, pressure, and other physical parameters are monitored to ensure they are stable. This step is typically executed in an NPT ensemble. This is done because the system and solvent need more volume after the heating process, so the system is allowed to expand and keep it at 1atm. The final volume will be fixed (NVT) during the production run.

4. **Production runs**: Once the system has been equilibrated, the actual production run of the MD simulation can begin. This involves simulating a specified number of time steps, during which the atoms' or molecules' positions and velocities are updated according to the laws of classical mechanics.

2.1.3 MD Analysis techniques

After the production run has been completed, the simulation results can be analyzed to study the system's properties and predict its behavior. The type of analysis executed depends on the system and the property we want to analyze.

- **Visual Inspection**. It is a good idea to check and visualize the frames extracted from the MD simulations with a visualization program such as *PyMOL*, or *ChimeraX*. This visual inspection might be more relevant after we determined key frames of positions in the aminoacid chain that might be more relevant.

- **RMSF of Atomic Coordinates**. The Root-Mean-Square Fluctuation (RMSF) measures the average deviation of the positions of atoms in a molecule from their mean positions. It is often used to analyze MD simulations to evaluate the flexibility of different molecule regions. The RMSF is calculated as the fluctuation of the Root Mean Square Deviation (RMSD) of the atomic positions from their mean positions over a given time period. For example, having an MD simulation of a protein with N atoms, the RMSD for atom i can be calculated as follows:

  $$RMSD = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(p_{i,j} - \bar{p}_i)^2} \qquad (2.5)$$

  Where $p_{i,j}$ is the position of atom i at time t and p_i is the position of atom *i* at the start of the simulation. The RMSF can be calculated for each atom in the molecule, allowing us to identify which regions are more flexible than others. The RMSF can help understand a molecule's conformational changes during an MD simulation and identify potential sites of protein-ligand binding or other functional groups.

- **Computing measurables during the MD**. A great technique is to compute relevant distances between atoms in the system, angles, or dihedrals. With that, we can evaluate how some parts of the system move concerning other parts; we can define substrate binding distances, attack angles, important torsions necessary for an event to happen, and much more. We can also evaluate cavities and spaces in the molecule, such as the catalytic pocket and the tunnels that connect this with the solvent. To do that, we can use specific softwares such as CAVER (Pavelka et al., 2015). With this software, we can compute the presence and absence of tunnels, the narrowest regions of the tunnel, i.e. bottleneck radius, and the more critical residues that shape the tunnels.

- **Shortest Path Map**. The Shortest Path Map (SPM)(Romero-Rivera, Garcia-Borràs, and Osuna, 2017a; Osuna, 2021) is a correlation-based tool that explores residue-by-residue correlated movements and inter-residue distances. This generates a complex graph based on proximity and correlation. The latter graph is further evaluated by making use of the Dijkstra algorithm to identify the shortest path lengths. The algorithm goes through all nodes of the graph and identifies which is the shortest length path to go from the first until the last protein residue. The method therefore identifies which edges of the graph are shorter, i.e. more correlated, and which are more central for the communication pathway. Only those edges having a higher contribution are represented, and they are weighted according to their contribution.

Due to the system's naturally high amount of variables, we cannot analyze all distances, coordinates, and angles during the MD simulations. For example, if one is interested in computing the coordinates of all alpha-carbons of a monomer of HheC during the MD simulation, the final data array will have the dimensions: 252 atoms, times three coordinates each, and times the number of frames. This will exceed the 10 million features if we consider an MD with 10.000 frames. In order to get the essential features of the system and be able to analyze it, we need to use techniques that reduce the dimensionality of the array, and these techniques are **Dimensionality Reduction techniques**.

- **Non-parametric tests**. If one is interested in comparing how the interactions and distances differ between two different conditions (or systems), we can use non-parametric tests to determine which contacts or distances differ the most to focus the analysis on these key distances. It is essential to use a non-parametric test because we do not know the distribution of the data, and the data often does not tend to follow a normal distribution. The **Smirnov-Kolmogorov** test("Kolmogorov–Smirnov Test" 2008) is a nonparametric statistical test used to evaluate whether a sample comes from a particular population. It is based on the difference between the empirical cumulative distribution function of the sample and the theoretical cumulative distribution function of the population. The test determines whether the sample is drawn from a specific distribution. The test can be used to determine whether the two datasets come from the same population or different populations. It is a powerful technique for big sample size datasets.

- **Machine Learning**. Other statistical techniques that are extremely useful for recognizing patterns in the data are machine learning, and for this thesis, in specific **Decision Trees** and the **Random Forest** were used (Ho, 1995; Ho, 1998). The random forest algorithm was employed to analyze which distances or contacts are the most important and correlate the best with a target distance that we know is important for function. With that, the parts of the proteins that make possible the selected event can be elucidated. Random forest is an ensemble method that trains multiple decision trees on random subsets of the data and aggregates their predictions to make a final prediction. The idea behind using multiple decision trees is that they can learn from different parts of the data, and the final prediction will be more accurate than any individual decision tree. Using a random forest regressor, many decision trees are trained on random subsets of the data. For each tree, a random subset of the features is chosen as the split points at each node. This helps to decorrelate the trees, which can improve the model's overall performance. During training, for each tree, the algorithm selects a random sample of the data with replacement (i.e., it allows some data points to be chosen multiple times) and uses this sample to fit the decision tree. The rest of the data is used to estimate the error of the tree. This process is repeated for each tree in the forest. The algorithm feeds the input data through each decision tree in the forest to make a prediction using the random forest regressor. For each tree, the output of the tree is recorded, and the final prediction is the average of all the trees' predictions. Random forest is a robust algorithm that can handle high-dimensional data with many features and provide accurate predictions. It is also relatively faster to train and predict than other machine learning algorithms. However, it can be prone to overfitting if the number of trees in the forest is too large.

- **PCA, tICA, and t-SNE**. Another statistical approach to reduce dimensionality is by using techniques that can group variables or features in a new space. The most common and known technique is **Principal Component Analysis** (*PCA*) (Jolliffe and Cadima, 2016). *PCA* is a statistical technique used to reduce a dataset's dimensionality. It makes this reduction by projecting the data onto a lower-dimensional space, known as the principal components while retaining as much variation in the data as possible. *PCA* is a linear technique, which assumes that the data is linear and the relationships between the variables are linear. It is often used to reduce the complexity of the data, to make it easier to visualize or to remove noise or outliers from the data. The following steps are typically taken to perform *PCA*:

  1. The data is centered, so each feature has a mean of zero.
  2. The covariance matrix of the data is calculated.
  3. The eigenvectors and eigenvalues of the covariance matrix are calculated.
  4. The eigenvectors are ranked by their corresponding eigenvalues in descending order.

5. The eigenvectors with the highest eigenvalues are selected as the principal components.
6. The data is projected onto the principal components to obtain the reduced-dimensional representation of the data.

With this, we obtain the new space where the variables are grouped so that the ones in the same Principal Component (*PC*) reduce the covariance of the data correlatedly.

A similar technique is the **Time-Lagged Independent Component Analysis** (*tICA*) (Molgedey and Schuster, 1994). *tICA* is a variant of Independent Components Analysis (*ICA*) ("What is Independent Component Analysis?" 2001) that considers the temporal dependencies between the variables in the data. tICA works by shifting the variables in the time series data by a certain number of time steps and then applying dimensionality reduction to the shifted data. The resulting independent components are called time-lagged independent components (*TIC*). These components capture the relationships between the variables at different time lags. They can identify patterns in the data that may not be apparent when looking at the data in its original form and represent the slowest movement in a dimension [see comparison between *PCA* and *tICA* in Figure 2.2].

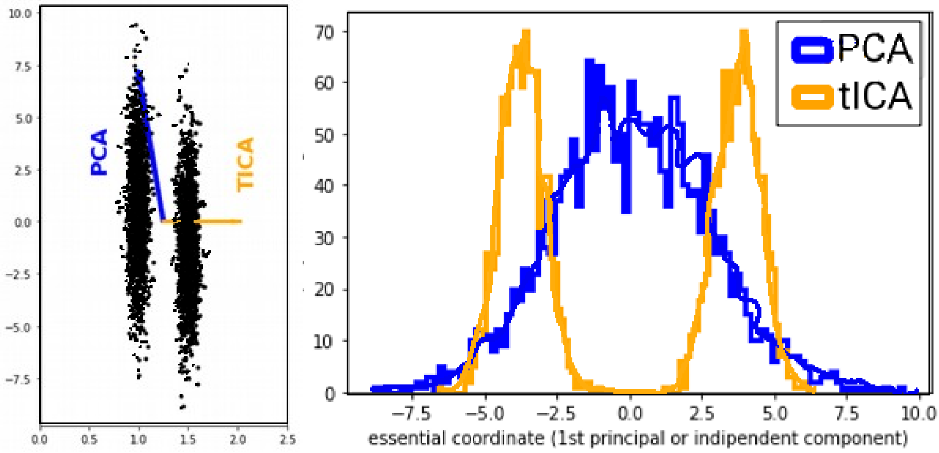

FIGURE 2.2: Comparison between the dimensionality reduction techniques *PCA* and *tICA* on a set of dummy data. Using this data, one can see how *tICA* finds the slowest movement in the left plot, and how separates the conformations on the right.

The last dimensionality reduction technique explained in this thesis will be the **t-distributed Stochastic Neighbor Embedding** (*t-SNE*) (Maaten and Hinton, 2008). *t-SNE* is a nonlinear dimensionality reduction technique that maps the data points from a high-dimensional space to a lower-dimensional space, typically a two- or three-dimensional space while preserving the relationships between the data points as much as possible.

t-SNE works by measuring the similarities between the data points in the high-dimensional space and constructing a probability distribution over the pairs of points. It then optimizes a low-dimensional representation of the data such that the similarities between the points in the low-dimensional space match the probabilities in the high-dimensional space as closely as possible. This creates a low-dimensional space where the divergence between the probability distributions between pairs of points is minimized. One big downside is that because the new space is not created from transforming the original data, the new dimensions have no functional or physical meaning; the data is simply grouped in similar data points. *t-SNE* is a valuable tool for data exploration and visualization. It has been widely used to visualize high-dimensional data in various applications, including natural language processing, image classification, and gene expression analysis. However, in this thesis (in particular chapter 4 and 5) it will be explained how *t-SNE* was applied in MD analysis for biocatalytic aplications for the first time.

Finally, once we have the data that we want to analyze further selected, the **Free Energy Landscape (FEL)** can be reconstructed from the new variables created (i.e., Collective Variables (CVs)). The CVs selected may be distances, angles, or other measurables selected by visual inspection and RMSF analysis, or even components from tICA or PCA (TICs and PCs). By reconstructing the FEL, we can compute the barriers by assigning energy values to the different conformations explored based on the density of points of the 2D histogram of the data. This is done by computing the thermodynamic integration ($-log(p)$, where p is the probability computed from the histogram count).

2.2 Experimental Methods for studying enzymatic systems

Ideally, to check if the theoretical predictions and claims are accurate, experimental assays can be performed. This means that we will test in the lab computational predictions and measure the effects of mutations on kinetics, activity, and thermal stability. However, before setting up the experiments and testing the parameters of the selected enzymes, these must be produced and purified. It is vital to understand the basics of creating recombinant proteins, cloning, and purification.

Recombinant proteins are those produced using genetic engineering techniques (Rosano and Ceccarelli, 2014). There are several steps involved in producing recombinant proteins in the laboratory. In a nutshell the following steps need to be followed:

1. **Construct Design**: The first step is to design the DNA construct that will be used to produce the protein. This involves selecting the appropriate expression vector (a piece of DNA that can replicate in a host cell) and inserting the gene for the protein of interest into the vector.

2. **Construct Cloning**: The next step is to clone the construct, which involves making multiple copies of the DNA construct in a host cell, such as *E. coli* or yeast. This is usually done using a process called transformation, in which the construct is introduced into the host cell using a plasmid (a small, circular piece of DNA) or a virus.

3. **Protein Expression**: Once the construct has been cloned, the host cells are grown in a culture medium under conditions that favor protein production. The protein of interest is then expressed in the host cells and secreted into the culture medium.

4. **Protein Purification**: The next step is to purify the protein from the culture medium (Wingfield, 2015). This step typically involves several steps, including centrifugation to remove cell debris, filtration to remove contaminants, and chromatography to isolate the protein.

5. **Protein Analysis**: Finally, the purified protein is analyzed and sequenced to confirm that it is the correct protein and to determine its purity and activity.

All these steps will be detailed for the HHDH systems studied in this thesis.

- **Construct Design**. Most of the proteins produced and constructed were done on top of the WT and available genes in the lab, except for the HheC R18 variant. This is because it is much better to buy the synthetic gene than to include the 37 mutations manually. For the other systems that need to be mutated, the DNA of the WT is subjected to site-directed mutagenesis using Polymerase Chain Reaction (PCR) and mutagenic plasmids.

To get the mutagenic PCR, we need to design and prepare the primers that will change the bases that transcribe into the selected mutagenesis. This is carried out using the PrimerX program and ensuring that the melting temperature stays similar.

Once the primers are created, and the mutagenic PCR is performed, the genes are transferred into a plasmid (*pET22*, *pET28*, or *psKBAD*). This is done using the same restriction enzymes for the empty plasmid and the gene fragment. Thanks to this, the ligation reaction is straightforward, and the proper ligation can be tested via agarose-gel electrophoresis. Once done, each construct is cloned into competent *E.coli DH5α(gold)* by heat shock and grown overnight at 37°C in agar plates with an antibiotic. Thanks to this, the ligation reaction is straightforward, and the proper ligation can be tested via agarose-gel electrophoresis.

The following day, the colonies can be collected and resuspended individually in lysogeny broth (LB) supplemented with antibiotics. Once the culture is grown, we can extract the plasmids cloned by the *E.coli DH5α(gold)* using a miniprep kit and store it.

- **Construct Cloning**. In the last step, we cloned the gene in *E.coli DH5α(gold)*, but that strain is used because it greatly multiplies the number of plasmids and is unsuitable for protein production. To do so, *E.coli BL21(DE3)* is used. This strain creates lots of protein once the required inductor (depending on the plasmid used) is added to the media.

  To do so, the plasmid is cloned into competent *E.coli BL21(DE3)* and cultured overnight in agar plates. After that, some LB is inoculated with a colony from the plate with the corresponding antibiotic corresponding tho the construct. The amount of LB should be at least 1/100 of the final culture. The following day, the broth is inoculated in the final culture in 0.5L Erlenmeyer (not more than 40% of the flask needs to be full, *E.coli* needs oxygen and agitation). The final culture is set into the incubator with intense agitation for up to 4 h approx. The culture needs to grow exponentially when the inductor is inoculated.

- **Protein Expression**. Once the *E.coli BL21(DE3)* culture is at exponential growth, a sample of the culture must be collected and then the inductor needs to be added to the media. This will make the production of the protein start inside the cells. Approximately four hours later, having the culture in agitation in the incubator at 37ºC or overnight at approximately 30ºC, the culture can be extracted and centrifuged to separate the cells from the culture media. After that, the cells are resuspended in the minimum amount of buffer and the cell membranes need to be broken. Multiple methods can be used, but sonicator and french press were used with similar results in this thesis. After that, the protein should be in the buffer, and the purification process can start. However, to know if the protein was induced properly, an SDS-page acrylamide-gel electrophoresis [Figure 2.3] should be done to see if the expected gel band has appeared after induction compared to the sample pre-induction.

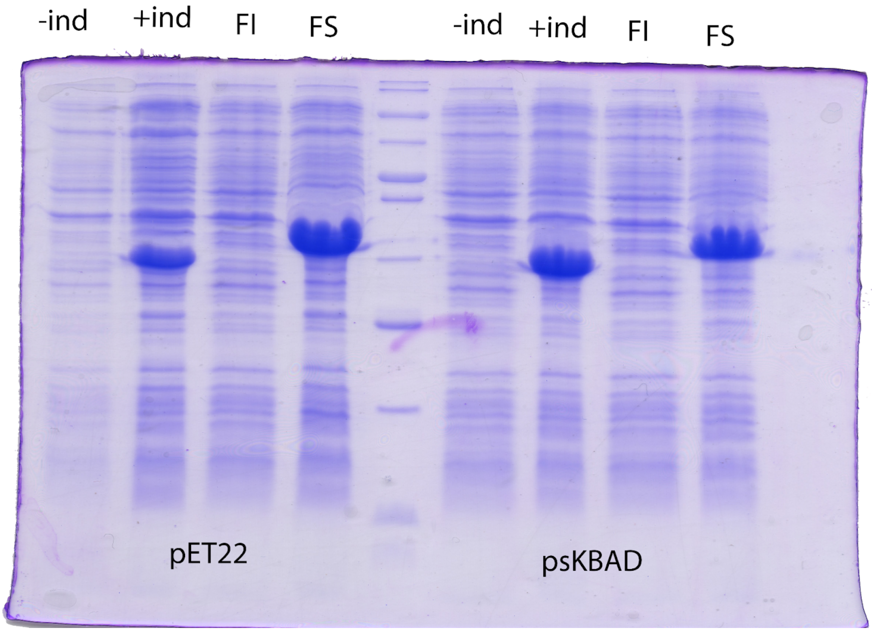

FIGURE 2.3: Photo of an Acrilamide-Gel showing all the proteins pre-induction (-ind), post-induction (+ind), in the insoluble phase (FI) and in the soluble phase (FS) for both constructs build in pET22 and psKBAD plasmids for HheC constructs. Soluble and insoluble phases were separated via ultracentrifugation

- **Protein Purification**. The first purification step is centrifuging the sample at high speed to separate the soluble and the insoluble proteins. The first time this is done, samples of the soluble and insoluble phases are sampled and analyzed via SDS-page.

  Further protein purification was done in two different ways, depending on the presence or not of His-Tag in the construct. The presence of the His-tag at the end of the protein makes it easier to purify, thanks to $Ni2^+$ affinity columns. These columns have a high affinity for the histidine side chains and can retain the proteins that display the His-tag. Later, the column is washed with a plain buffer to remove all proteins that are not binding in the column. After the column is clean, a buffer containing imidazole (histidine side chain) is passed through the column, displacing the protein and eluting it with high purity. After this, the sample needs to be cleaned using a desalting column to remove all excess of imidazole.

For the proteins without His-tag, the HHDHs were purified employing ion-exchange chromatography, being more exact, anion-exchange chromatography. This chromatography technique uses a column with fixed cations in the stationary phase. Then the sample is added with all soluble proteins in the column and eluded with a buffer. This will make the proteins with positive and neutral charge elute, and the negatively charged ones will be attached in the stationary phase. After this, a gradient of the salted buffer will start to elute in the column, and the concentration of salts will increase gradually, thanks to the usage of the FPLC machine (*ÄKTA pure*). Once each protein reaches the isoelectric point, it will be released into the mobile phase. All samples collected are tested for activity and also molecular weight on the SDS page to check when the HHDH is released.

Later, size exclusion chromatography is used with multiple finalities: to get the sample even purer, to desalt the sample, and to verify the oligomerization state of the proteins in solution (see more in the Results part). Size exclusion chromatography uses a resin with pores ranging from multiple sizes, making smaller proteins take more time to elute than bigger ones. With this, we can separate proteins with different particle sizes and check whether changes on the oligomerization state induced by mutations were introduced.

- **Protein Analysis**. After pure samples for the different HHDHs are obtained, multiple analyses can be performed to study the desired properties for each variant.

  As mentioned above, the oligomerization state of the different variants can be analyzed by running size exclusion chromatography. This method separates the proteins by particle size. For every peak, activity is tested, and SDS-page chromatography is used to verify the molecular weight of the eluted protein.

Another property we want to measure is the **thermal resistance** by getting the melting temperature (Semisotnov et al., 1991). In this thesis, two different methods were used. In the first one, the absorbance at 280 *nm* was monitored using a spectrophotometer while a water bath was heating the sample. This is based on the principle that tryptophan (a residue with a high absorbance at 280nm) is a hydrophobic residue and is mainly located inside the protein and not exposed to the solvent. Once the protein starts denaturing, the tryptophans are exposed, and the absorbance increases. With this, the melting curve can be generated, and the melting temperature estimated.

Another method is the thermofluor assay (Huynh and Partch, 2015). This assay uses a fluorometer with temperature control (qPCR machine) and a hydrophobic die like *SYPRO orange*. This dye emits light at 570 nm after receiving light at 470 *nm*, but only if it is attached to a hydrophobic area because water inhibits its fluorescence. The die is added to the sample and into the qPCR machine. Then the temperature gradually rises, and when the protein unfolds and more hydrophobic residues start to get exposed, the light emitted at 570 *nm* increases. The melting temperature is the temperature at which the fluorescence stops increasing its slope; in other words, it is the maximum value of the first derivative.

Another key property is the **catalytic activity** of the different enzymes, which is obtained through the kinetic parameters. **Steady-state kinetic** assays (Cleland, 1990) must be done to fit the data into a kinetic model, such as Michaeli-Menten or Hill's, to get k_{cat} and K_M or K_{50} (in Hill's kinetic model). Steady-state kinetics refers to the study of enzyme-catalyzed reactions under conditions in which the concentration of the enzyme and its cofactor(s) remain constant over time. One advantage of using steady-state kinetics to study enzyme-catalyzed reactions is that it allows for the determination of the kinetic parameters of the enzyme, such as the maximum velocity (V_{max}) and the substrate concentration at which V_{max} is achieved (K_M).

Different assays must be done depending on the reaction we want to monitor. For the case of the epoxide-ring forming reaction, the **halide-release assay** is used. The halide-release assay monitors the amount of halide (Cl^-) released in the media. To do so, the reaction is stopped multiple times, adding large amounts of acid with iron sulfate salts and mercuric thiocyanate. The chlorine forms a salt with the mercury, and ferric thiocyanate is formed in identical stoichiometric amounts as the chlorine is consumed. Ferric thiocyanate has a peak in absorbance at 480 nm and can be easily monitored using a spectrophotometer, and the concentration can be determined using a calibrated curve (Najib and Hayder, 2011).

$$Cl^- + Hg(SCN)_2 \rightarrow HgCl_2 + 2Fe(SCN)^{2+} \qquad (2.6)$$

For the kinetic parameters of the epoxide-ring opening reaction using cyanide, we used a nickel chloride that, in the presence of cyanide, makes a tetracyanonickelate with absorption at 267 nm. The salt is added and solved in 1M ammonia, killing the enzyme in the process and stopping the reaction (Schallmey et al., 2015).

Chapter 3

Objectives

In this thesis, we aim to understand and rationalize the conformational differences in natural and laboratory-evolved HHDHs and evaluate how these affect their catalytic activity and available tunnels. Computations and experiments are combined to elucidate how natural and laboratory evolution has yielded synthetically useful HHDHs.

Computational methods like Molecular Dynamics (MD) simulations and novel dimensionality reduction and feature selection techniques for this field were used and developed to reconstruct the associated Free-Energy Landscapes (FELs). To properly do this, multiple objectives were targeted:

- Develop and test a new computational pipeline that describes most of the variance sampled along the MD simulations in a human-readable two-dimensional space.
- Apply the new computational pipeline coupled with tunnel-analysis techniques to explore the effect of the conformational changes into the tunnels in naturally-occurring HHDHs families. The goal is to better understand the tunnels' effect in the substrate scope displayed among families and pinpoint critical residues.
- Decipher the flexibility and conformations that provide the D-family HHDHs with more thermal stability and often lower catalytic activity. This was based on the HheD2 designs provided by Prof. Anett Schmalley's group.
- Computationally and experimentally study the effect of the mutations randomly introduced via Directed Evolution (DE) in HheC to synthesize statin precursors. Experimentally study the thermal stability, oligomerization, and kinetic parameters and rationalize the obtained data using long timescale MD simulations.

Achieving these objectives will bring to the table an effective and statistically valid computational methodology that can potencially help researchers to understand, propose, and predict mutations in HHDHs and by extension to unrelated other proteins.

Chapter 4

Towards developing new computational pipelines for estimating the conformational landscape and its impact on the available tunnels

4.1 Precedents in CRL in nanomotors: State-of-the-Art

In recent years, there has been a remarkable growth in the development of micro- and nanomotors, particularly in the biomedical and environmental sectors(Abdelmohsen et al., 2013; Wang et al., 2019a; Gao and Wang, 2014). The operation of catalytic micro/nanomotors relies on converting chemical energy into mechanical force, leading to active motion(Luo et al., 2018). Enzymes, which are well-known for their biocompatibility and highly efficient catalysis in biosystems, have emerged as strong candidates for powering micro/nanomotors. In this regard, various enzymes such as catalase(Ma et al., 2015), urease(Dey et al., 2015), glucose oxidase(Ji et al., 2019), acetylcholinesterase(Arqué et al., 2019), trypsin(Schattling et al., 2017), lipase(Wang et al., 2019b), and multi-enzyme combinations(Gao et al., 2019) have been explored for the fabrication of micro/nanomotors. Despite significant progress in the development of enzyme-powered micro/nanomotors for in vitro/vivo applications, the impact of catalytic processes on the motion performance of nanomotors remains an area of concern(Arqué et al., 2019). Therefore, this part of the thesis aims to explore the effect of the enzyme dynamics for the development of catalytic micro/nanomotors and assess the effect of catalytic processes on their motion performance.

In this chapter of the thesis, *Candida rugosa* lipase (CRL) has been studied and used and nanomotor in diferent conditions. CRL is an enzyme that can catalyze the transformation of triacetin in glycerol in a highly efficient manner (María et al., 2005). A trait shared between lipases is that all have a lid domain covering the active site. This is usually blocking the entrance of the substrate, and the opening of the lid domain can be a slow event that hampers the catalysis, and it can be the rate-determining step(Khan et al., 2017). The CRL enzyme can be inmobilized using different techniques. The most hydrophobic technique used in this thesis is called OTES (hydrophobic interactions using trimethoxy-(octyl)silane), and the hydrophilic technique is called APTES (ionic absortion using

3-aminopropyltriethoxysilane). Based on the results obtained at Sanchez's lab, we were aware that if the hydrophobic method is used, the enzyme is fixed in an open conformation because most of the residues in the inner part of the lid domain are hydrophobic. Otherwise, by using the hydrophilic technique, the enzyme is trapped in a closed conformation. The awareness of the existing open and closed conformations within this enzyme makes this system an ideal reference for the purpose of validating our theoretical approach. In this thesis, the change of CRL lid domain, and its importance in catalysis, will be investigated. To that aim, we developen a computational pipeline for studying the effect of the lid conformation in the binding and unbinding events.

4.2 Computational details

4.2.1 System preparation

Open and closed CRL X-ray (PDB code *1CRL* and *1TRH*) structures were used as starting points for independent molecular dynamics (MD) simulations. Protonation states of enzyme residues were assigned based on *pKa* values provided by the *H*++ server (Gordon et al., 2005) and detailed information on the catalytic mechanism was used in order to select the proper protonation state of catalytic residues. The enzyme was then solvated in a pre-equilibrated cubic box with a 10 Å buffer of transferable intermolecular potential with 3 points (TIP3P) (Jorgensen et al., 1983) water molecules, adding 1620 solvent molecules. The systems were neutralized by adding 16 explicit counter ions (*Na*+). A two-stage geometry optimization was performed, first minimizing the positions of solvent molecules and ions by imposing harmonic positional restraints of 500 $kcal\ mol^{-1}$Å $^{-2}$ on solute molecules, followed by an unrestrained minimization. Afterward, the gradual heating of the systems was performed by increasing the temperature by 50 K along six 20 ps sequential MD simulations (0 - 300 K) under constant volume and periodic boundary conditions. Harmonic restraints of 10 $kcal\ mol^{-1}$ were applied to the solute, and the Langevin equilibration scheme was used to control and equalize the temperature. The time step was kept at 1 fs during the heating stages, allowing potential inhomogeneities to self-adjust. Each system was then equilibrated without restraints for 2 ns with a 2 fs time step at a constant pressure of 1 atm and temperature of 300 K.

4.2.2 Molecular dynamics

After equilibration in the isothermal isobaric ensemble (*NPT*), 61 replicas of 100 ns were run for each system in the canonical ensemble (*NVT*) for a total simulation time of 6.1 μs. More simulations including the substrate (triacetin) in the solvent (10%) were prepared using the packmol(Martínez et al., 2009) software and 5 replicas of 200 ns were simulated. All simulations were performed using the Amber 14SB force field (*ff14SB*) (Maier et al., 2015) using the AMBER16 software in the group-owned GPU cluster GALATEA.

4.2.3 Dimensionality reduction and clustering

The pyemma2 software package(Scherer et al., 2015) was used to extract data from all C-alpha coordinates along the MD simulations as representative features of the enzyme's conformational dynamics. A total of 1602 features were extracted for each system. Time-lagged Independent Component Analysis (*tICA*) was applied to reduce the dimensionality of our data. 390 *tIC* components were obtained, which account for the 95 % of kinetic variance. No interconversion between open and closed CRL structures was found in the reduced subspace. Later, a second dimensionality reduction technique was applied, the t- distributed stochastic neighbor embedding (*t-SNE*)(Maaten and Hinton, 2008). This technique was used to reduce the most informative first 20 *tICA* dimensions into a new 2D space and analyze the data in a more feasible way. The Hierarchical density based clustering (*HDBSCAN*) algorithm(McInnes, Healy, and Astels, 2017), with a *minimum_cluster_size* of 200 and other default parameters, was used to split the *t-SNE* space into 183 and 198 clusters for the open and closed systems, respectively.

4.2.4 Tunnels analysis

The tunnels in a given enzyme structure was determined using the CaverAnalyst(Jurcik et al., 2018) software . Additionally, the (un)binding process through the tunnels can be estimated by docking the substrate and product along different points of the tunnel, also taking into account the previous conformation and rotation of the substrate/product and the residues that shape the tunnel. This was done using the CaverDock software (Vavra et al., 2019). A 4 Å shell depth, 2.5 Å shell radius, clustering threshold value of 10 and a 0.9 Å minimum probe radius were used as tunnel search parameters. Transport energy profiles were computed for the CRL reaction substrate (triacetin).

4.3 Results and discussion

One of the most important and successful ways of understanding how the dynamism of an enzyme affects the catalytic activity is through exploring the cavities and tunnels present in the enzyme. The enzyme requires a proper pocket and tunnels to accommodate the substrate, the transition state (TS), and the product. Also, this shape complementary between the active site and TS is key for lowering the energy barrier(Szefczyk et al., 2004). For the substrate to enter the active site pocket, the enzyme needs at least one tunnel connecting the active site and the solvent to let the substrate(s) enter and undergo catalysis. This event, also known as binding (and in some cases the unbinding of the product), might be the rate-limiting step in some cases because the tunnels are not wide enough (usually depending on the LID domain conformation)(Khan et al., 2017), the residues that shape the tunnel and pocket are not precisely positioned for the proper interaction with the substrate, or the orientation that the substrate enters is not suitable for catalysis.

However, the analysis of the available tunnels on the crystal structure or generated models does not explain the bigger picture because, as mentioned in the introduction, the protein moves and adopts different conformations, which of course, impact the tunnels. It is, therefore, desirable to analyze the conformations and the presence (or absence) of the different tunnels from the multiple structures sampled in the MD simulations. The tunnels can be clustered in similar ones, making it possible to identify those that are more frequently observed.

The strategy used is to explore the different tunnels in the different conformations sampled from a specific protein is to first group/cluster together those similar conformations the protein can adopt. There are multiple ways to do so but to make a rational clustering taking into account most of the variance that the protein explores, in this thesis, we decided to use as geometrical features the information of all coordinates of the alpha-carbons that compose the protein, in this way we can identify distal regions from the tunnel that can play a role. Given the high dimensionality of the data, some dimensionality reduction techniques need to be applied. In particular, the data is transformed in another space using a dimensionality reduction technique called *tICA*. With this, we can explore the slow movements of the alpha-carbons that took place during the MD simulations.

The number of dimensions obtained using this method is still vast, and clustering in this high space may be extremely difficult and computationally costly. What is typically done is to sample using only the dimensions involved in the movements affecting the variables that we think are defining the tunnels. But following this strategy contradicts the fact that we are interested in determining if other regions far from active sites or tunnels are affecting the presence/absence of some of them. Also, some tunnels not present in the initial structure may not be sampled because of the bias that this decision will introduce.

To make the analysis as unbiased as possible, we chose to run a new dimensionality reduction technique which has become popular in the field of machine learning, the *t-SNE*. *t-SNE* creates a new space where all the variance can be displayed in the new 2 or 3 dimensions at maximum [Scheme for the dimensionality reduction: Figure 4.1].. By the end of this dimensionality reduction, we obtained a new 2D space where all the conformations explored are represented, and only the slow movements are present, thanks to its coupling with *tICA*. Thanks to this approach, the most frequently visited conformations can be identified, and the Caver analysis can be performed. With that, one can explore the conformations present in the MD simulations and all the tunnels with low bias [Scheme Tunnel analysis: Figure 4.2].

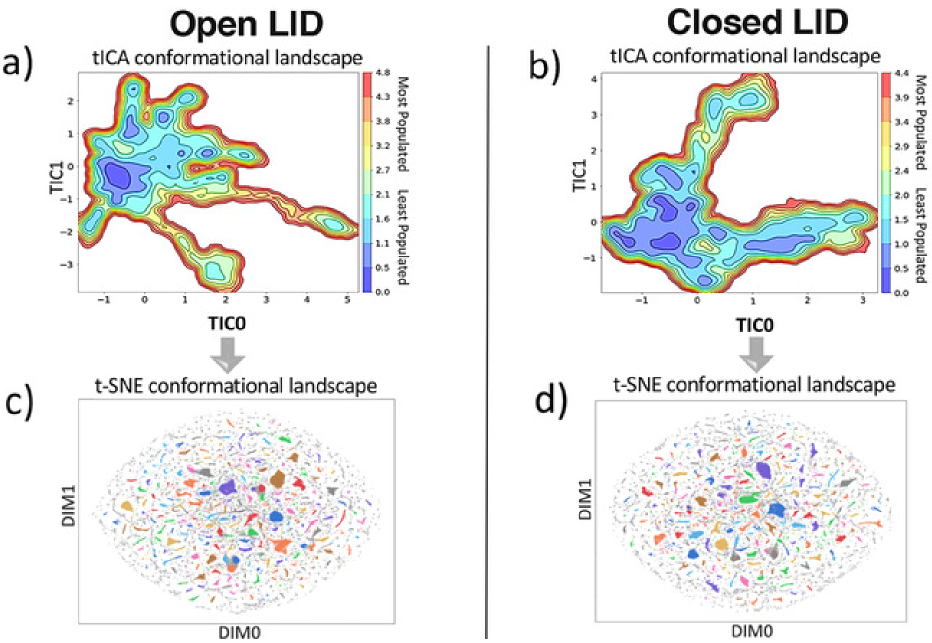

FIGURE 4.1: Scheme of the MD analysis process. The dataset obtained from multiple MD simulations for open and closed CRL was reduced using a two-step process: first, the dimensionality reduction technique *tICA* shown in a) and b) was applied for open and closed conformations of CRL, respectively. It was further reduced into a 2D dimension with *t-SNE* (c,d) and clustered with the HDBSCAN method. Each cluster in the *t-SNE* plots is colored differently. The four most populated clusters were subjected to tunnel and ligand binding analysis.

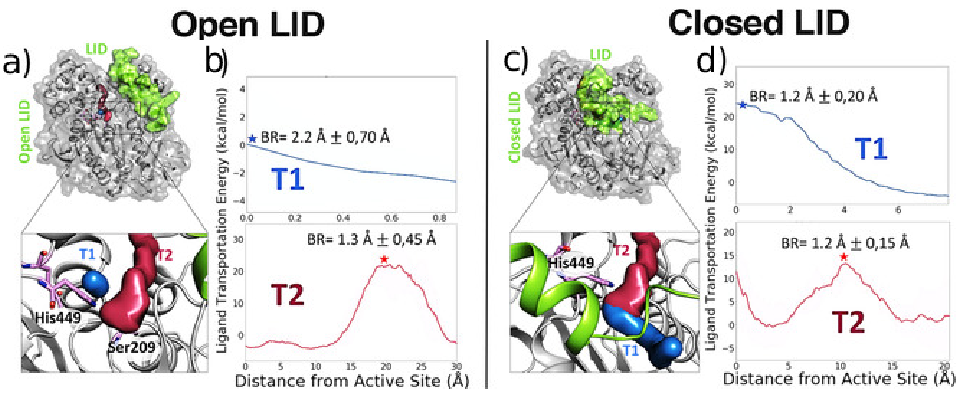

FIGURE 4.2: Scheme of the substrate binding accessibility profiles. Representative open (a) and closed (c) CRL conformations adopted during the MD simulations are represented as cartoon. LID domain is shown in green, active site residues in purple sticks, and tunnel T1 and T2 in blue and raspberry solid surfaces, respectively. Ligand transportation energy profiles (b,d) computed from the most populated clusters after *t-SNE* reduction. The mean energy profile for substrate accessibility to the active site computed on multiple MD snapshots is shown with a solid line and the standard deviation using a shaded region. Mean energy profiles for substrate accessibility through T1 and T2 tunnels are shown in blue and raspberry, respectively. The mean tunnel Bottleneck Radius (BR, marked with colored stars in (b,d), in Angstroms) together with the standard deviation is also shown.

We first tested this new methodology in collaboration with Prof. Dr. Samuel Sanchez and applied it to study the lipase protein from *Candida rugosa* (CRL) used as a motor for propelling nanoparticles. Our motivation is to test the effect of the different fixation techniques used, how this affected the catalysis, and, with that, how the product was released and the particle propelled.

The obtained specific activities of the nanoparticles in both OTES and APTES showed, as expected that the activity of CRL in OTES is much higher than in APTES. To rationalize how the fixation method affects the activity and evaluate whether this change in the conformation was affecting the tunnels, we run conventional MD simulations of CRL starting from the open state and closed state. Five replicas of 200 ns including 10% triacetin were simulated to explore the different open and closed conformations. We monitored the distance between residue P74 and P443 to monitor the opening and closing of the lid domain and saw that without restraints in the MD, the opening and closing event of CRL is slow (*ms* to *s* time-scale). We did not see any transition from closed to open or vice-versa[Figure 4.3]. To explore if the substrate affects the open-closed conformational change, we also executed five replicas from open and closed conformations, including 100 mM of triacetin in the solvent. The results showed that in the presence of triacetin, enhanced flexibility of the lid domain was observed, but not enough to sample an opening event in the five 200 ns MD replicas.

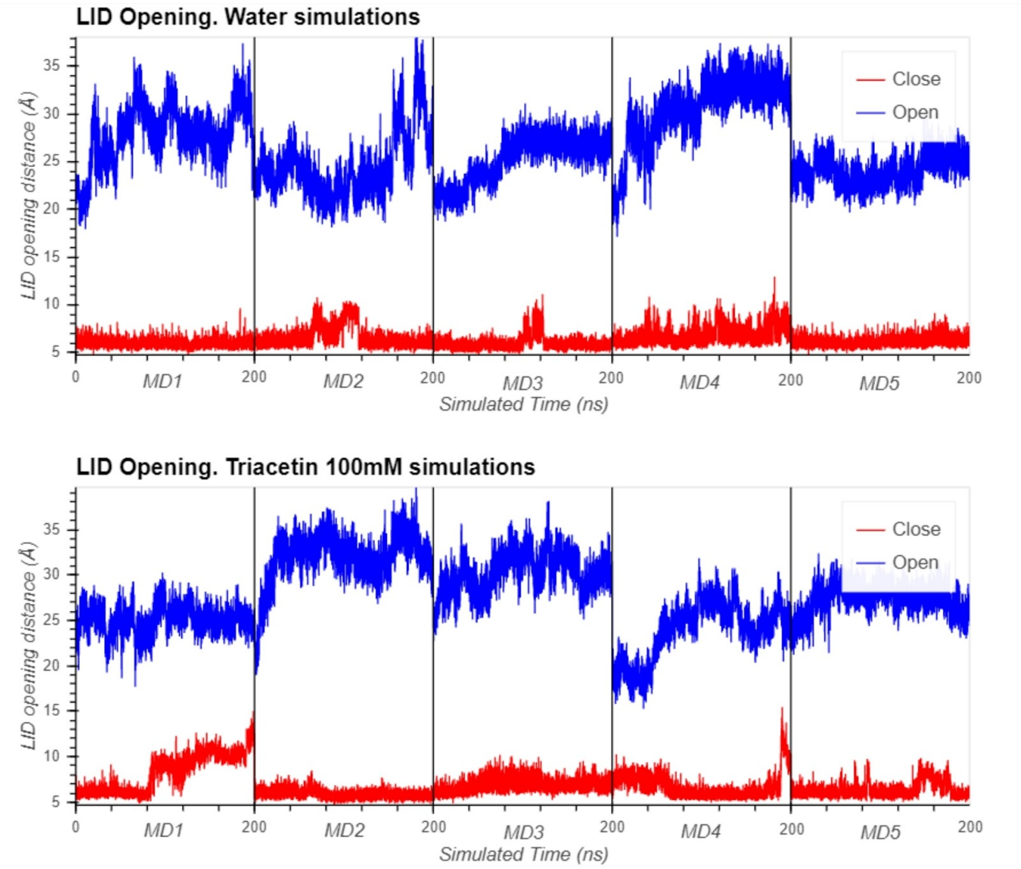

FIGURE 4.3: A plot of the distance between the alpha-carbons of residue P74 and P443 (that describe the state of the lid open-closed) along the five replicas of 200 ns MD simulations performed starting from either the open (shown with a blue line) or closed (red line) conformation of the CRL lid in water (top panel) and in 100 mM triacetin (down). This distance is ca. 10 Å in the closed conformation, and ca. 30 Å in the open lid conformation. In water, the closed conformation of the CRL lid is highly stable as the monitored distance stays at ca. 10 Å the whole simulation time. In contrast to what is observed in water, the closed state of the lid in 100 mM triacetin is substantially more flexible and explores partially open conformations of the lid (the distance is elongated from ca. 10 Å to 21 Å in some of the closed MD replicas).

With this in mind, 61 replicas of 100 ns were computed, starting from the open or closed conformation. We also explored the different conformations sampled and categorized all the open or closed conformations based on the distance between residues P74 and P443. Still, no opening or closing event was sampled, as observed before. All conformations that started from the open conformation remain open, the same for the closed state. All alpha-carbon coordinates were extracted, and the tICA dimensionality reduction was applied. Then, the 20 first dimensions were collected, and t-SNE was applied, thus obtaining the final 2D subspace where the data was clusterized.

The clusterization was done by using the HDBSCAN algorithm (Campello, Moulavi, and Sander, 2013; McInnes, Healy, and Astels, 2017). This algorithm clusters by density and hierarchy between points. With that, we can sample all the clusters in the t-SNE space by applying a minimum cluster size of 200 for the HDBSCAN. 183 and 198 clusters[Figure 4.1, c and d] were obtained from the simulations in open and closed conformations, respectively. To obtain all the tunnels, the CaverAnalyst software was used to explore the tunnels of the most representative structure of each cluster. This frame is obtained from the MD frames, the one closer to the geometrical centroid of the cluster. With this, we obtained three meaningful tunnels that can be sampled in the CRL: T1, T2, and T3.

Finally, to further analyze the tunnels and to obtain statistically meaningful results, 20 random structures of the most populated clusters of the open and closed conformations of CRL were further analyzed with CaverAnalyst. T1 is widely observed in the open conformations (94% of the structures showed the tunnel), with a bottleneck radius (BR) of 1.28 ± 0.24Å and an almost barrierless binding (ca. 2 $kcal \cdot mol^{-1}$). On the other hand, tunnel T1 is much less present in the closed conformation (48% of the conformations) and showed a narrower 1.10 ± 0.20Å BR and higher energetic barrier of ca. 13.3 $kcal \cdot mol^{-1}$ [Figure 4.2, f and h, In blue. Figure 4.4].

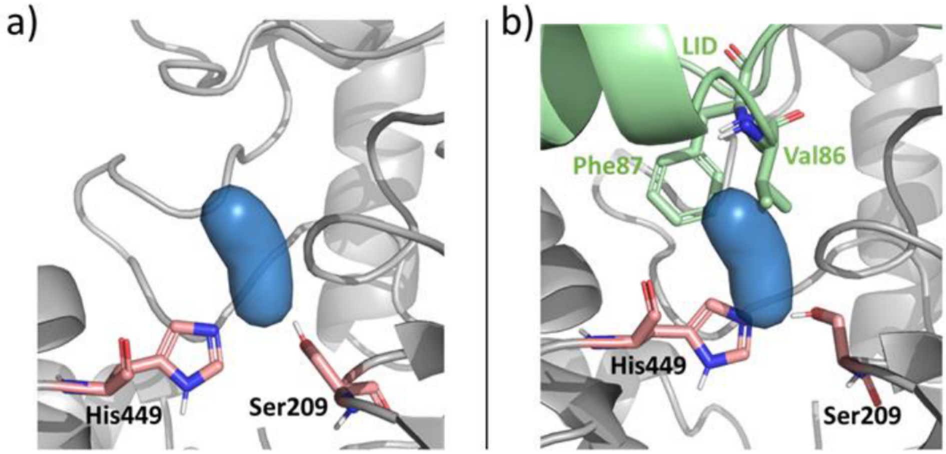

FIGURE 4.4: Active site comparison between open (a) and closed (b) CRL conformations. Tunnel T1, in blue, connects the catalytic residues (His449 and Ser209) to the solvent in the open conformation. In the closed conformation, the hydrophobic Val86 and Phe87 residues from the LID domain approach the active site reducing the dynamic site pocket volume and disfavoring the binding of the ligand in the *Candida rugosa* lipase (CRL) active site.

Tunnel T2 is always present when T1 is formed but showed a much narrower 1.04 ± 0.12Å BR for the open conformations. On the other hand, tunnel T2 is present in 39% of the closed structures sampled and shows a similar BR of 1.01 ± 0.11Å. For this tunnel, the binding energy barrier is estimated to be higher than in T1, showing a ca. 26.3 $kcal \cdot mol^{-1}$ in the open conformations and 28.8 $kcal \cdot mol^{-1}$ for closed conformations [Figure 4.2, f and h. In red]. By using the CaverAnalyst analysis tools, we could decipher that the residues that primarily contribute to the higher energy barrier are valine 86 and phenylalanine 87. These residues in the lid domain restrict the active site accessibility in the closed states.

The T3 tunnel is defined as the T1 tunnel that can escape between the chains that conform to the lid domain when in a closed conformation, taking a different path. The presence and BR of this tunnel are small, so they can be considered negligible.

In this study, we not only demonstrate the role of the dynamics of CRL for the affective immobilization in nanomotors and the role of the lid domain in catalysis, but we also developed and showed, for the first time, a novel and smart way to analyze the multiple conformations that are sampled during the MD simulations. This, coupled with an efficient computational cost analysis method, like tunnel exploration, leads to analyzing all the conformational space on this feature with great statistical sampling. This developed pipeline can be generally applied in other conformational studies and, concerning this thesis, for deciphering the tunnels on the available HHDHs families.

Chapter 5

Conformational Landscapes of Halohydrin Dehalogenases and Their Accessible Active Site Tunnels

5.1 Precedents in halohydrin dehalogenases: State-of-the-Art

As mentioned in the introduction, HHDHs share the same catalytic triad composed of Ser-Tyr-Arg, a halide binding site, and the active form is mostly the homotetrameric conformation. They catalyze the dehalogenation reaction and have some promiscuous activity towards the nucleophile-based epoxide ring-opening reaction, showing different activity, selectivity, substrate scope, and stability.

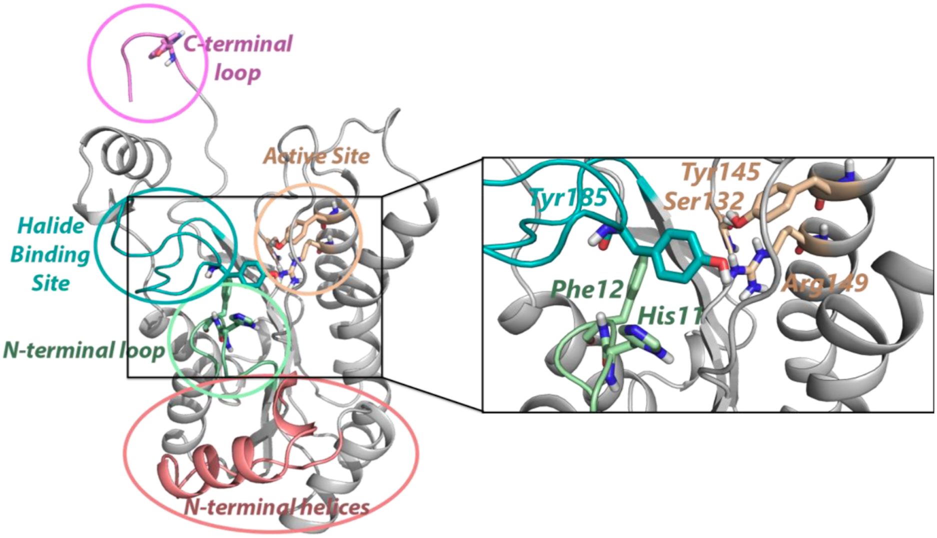

FIGURE 5.1: Distinct halohydrin dehalogenase (HHDH) structural elements and zoom of the active site and halide binding pockets based on the HheC structure. Active site residues are highlighted in wheat color, halide-binding site in teal, N-terminal loop in light green, C-terminal loop in purple, and N-terminal 6–7 helices in salmon. In the active site zoom, potential residues blocking the accessible active site tunnels are depicted using the same color scheme.

Different classes of HHDHs have been described up to date: A, B, C, D, E, F, and G (Koopmeiners et al., 2016). In particular, we focused our study on some representatives of each HHDH family, which were selected based on available structures. The selected structures did not contain any ligand bound, no mutations, and had the best resolution available. With that, the structures from the PDB considered were: 1ZMO for HheA2 (from *Arthrobacter sp. AD2*), 4ZD6 for HheB (from *Corynebacterium sp. N-1074*), 1PWZ for HheC (from *Agrobacterium radiobacter AD1*), 7B73 for HheD2 (from *gamma proteobacterium HTCC2207*) and 5O30 for HheG (from *Ilumatobacter coccineus YM16-304*).

Hereafter more details of the selected examples of HHDHs under study are provided.

HheA2 is 97.1% similar to HheA isolated from *Arthobacter sp.*, although another HheA has also been crystalized from Corynebacterium sp. HheA enzymes showed low enantioselectivity and preferred long-chain aliphatic molecules (C4-C5) as substrates. This is proposed to be due to a bigger substrate-binding pocket (Jong et al., 2006). The crystalized HheA enzymes show a 97.1% sequence identity but only 34% compared to the widely studied HheC. The preference for aliphatic substrates and the low selectivity made the HheA family of enzymes unsuitable for industrial applications.

The HheB studied in this thesis is from Corynebacterium sp. N-1074, and it is the only one reported in the PDB. The other HHDH B-type that has been studied in more detail is the HheB2 from *Mycobacterium sp. GP1*. These enzymes have 95% sequence identity and only show four differences in the sequence, but despite that, HheB shows much higher R-enantioselectivity than HheB2. In previous studies (Watanabe et al., 2015) three out of these four differences have been linked with the changes in the enantioselectivity of HheB. From a structural point of view, the 2nd and 3rd alpha helix (using HheC nomenclature, see Figure 5.1) are disordered structures in HheB.

HheC is the most studied enzyme from the HHDH family. The enzymes of the C-family show high S-enantioselectivity and prefer aliphatic and short substrates (C2-C3). They show only 20-30% of sequence identity with the other mentioned HHDH families and have many structural differences, mainly the N-terminal helix and C-terminal that cover the opposite active site [see Figure 5.1]. This

C-terminal part from the monomer in the diagonal is located on the top of the catalytic active site and includes bulky residues, like tryptophan 249, described as regulators of the entrance tunnels and substrate scope (Schallmey et al., 2015). It has been discussed that this terminal part may play a role in describing the selectivity, crucial role in the substrate binding and thermal stability. With that, mutations in the entrance near this region have been done to make HheC accept more significant and aromatic substrates and alter the selectivity (i.e., W249F mutation)(Tang et al., 2003).

For some time, these three families of HHDHs were the only ones known, but thanks to adding new motifs in the PHI-BLAST algorithm, new sequences were described, and new HHDHs had been and are being described (Schallmey and Schallmey, 2016; Schallmey et al., 2014; Koopmeiners et al., 2017). From these, the D, E, F, and G have been discovered, but only HheD2 and HheG have been crystallized(Koopmeiners et al., 2017).

The D subclass of HHDHs contains many enzymes, and several of them show remarkable thermal stability and also high activity. HheD2 is the only one that has been crystallized. As explained for HheB, from a structural point of view, the HheD2 presents the 2nd and 3rd alpha helix (using HheC nomenclature, see Figure 5.1) disordered. HheD2 shows high activity but low thermal stability as compared to the other members of the same family. The molecular basis of this stability/activity trade-off will be explained in the next chapter of this thesis.

Finally, the last family of HHDHs for which a crystal structure is available is HheG. This enzyme is more distantly related to the other HHDHs in the phylogenetic tree and shows low identity with other HHDHs, but is structurally similar. It is similar to HheC but has no C-terminal part covering the active site, contains a disordered 2nd alpha-helix, and an extra helix in the halide-binding site. HheG from Ilumatobacter coccineus shows high activity towards sterically demanding cyclic epoxides compared to all other HHDHs tested (Koopmeiners et al., 2017). This broader substrate scope is accompanied by a rather good enantiomeric excess, especially with azide. This enzyme has lower activity using cyanide as a nucleophile and, in this case, shows low stability and enantioselectivity. These features can be explained thanks to the prominent and solvent-exposed active site, as shown also below.

5.2 Computational details

5.2.1 System preparation

The selected HHDHs were prepared and simulated without any ligand and in tetrameric conformation. Protonation states of enzyme residues were assigned based on *pKa* values provided by the *H*++ server (Gordon et al., 2005). The enzymes were then solvated in a pre-equilibrated cubic box with a 10 Å buffer of transferable intermolecular potential with 3 points (*TIP3P*) water molecules(Jorgensen et al., 1983), resulting in the addition of approximately 27,000 solvent molecules per protein. The systems were neutralized by the addition of approximately 32 explicit counter ions (*Na*+).

A two-stage geometry optimization was performed, first minimizing the positions of solvent molecules and ions, by imposing harmonic positional restraints of 500 $kcal\ mol^{-1}$ Å $^{-2}$ on solute molecules, followed by an unrestrained minimization. Afterwards, a gradual heating of the systems was performed by increasing the temperature 50 K along six 20 ps sequential MD simulations (0-300 K) under constant volume and periodic boundary conditions. Harmonic restraints of 10 $kcal\ mol^{-1}$ were applied to the solute, and the Langevin equilibration scheme was used to control and equalize the temperature. The time step was kept at 1 fs during the heating stages, allowing potential inhomogeneities to self-adjust. Each system was then equilibrated without restraints for 2 ns with a 2 fs time step at a constant pressure of 1 atm and temperature of 300 K.

5.2.2 Molecular dynamics

All simulations were done using the Amber 99SB force field (*ff99SB-ildn*)(Lindorff-Larsen et al., 2010) After equilibration in the isothermal-isobaric ensemble (*NPT*), 5 replicas of 250 *ns* were run for each system (i.e., 1.25 μs per HHDH subclass) in the canonical ensemble (*NVT*) yielding a total MD simulation time for all systems of 6.25 μs. The graphics processing unit (GPU) version of pmemd in Amber16 was used for the MD simulations, which were executed on the in-house GPU cluster GALATEA.

5.2.3 Dimensionality reduction and clustering

The MD simulations were analyzed as monomers in order to make it more feasible. To do so, all MD simulations were separated into four different simulations for each monomer and were aligned, thus multiplying the simulated time analyzed by four (5 μs for each system and 25 μs in total). MD simulation trajectories were post-processed with the pyemma2 software package(Scherer et al., 2015). Alpha-carbon coordinates of the aligned protein subclasses at each nanosecond of MD simulation were used as initial features, resulting in 182,250,000, 168,000,000, 189,000,000, 168,000,000, 192,750,000 extracted values (features x frames x replicas) for the A2, B, C, D2, and G HHDH subclasses, making the statistical analysis unfeasible. Subsequently, the time-lagged Independent Component Analysis (*tICA*)(Molgedey and Schuster, 1994), with a lag time τ set to obtain the minimum number of reduced dimensions, was applied to reduce the dimensionality of the initial MD features. After applying *tICA*, we further reduced the dimensionality of the data by applying the t-Distributed Stochastic Neighbor Embedding (*t-SNE*)(Maaten and Hinton, 2008) method to the 20 most informative *tICA* dimensions. These 20 most informative *tICA* dimensions describe the 25% of the total variance. The resulting 2D *t-SNE* space was clustered with the hierarchical density based clustering (*HDBSCAN*) algorithm(McInnes, Healy, and Astels, 2017), with a minimum cluster size of 200 and other default parameters, resulting in 133, 126, 134, 124, 119 clusters for the A2, B, C, D2, and G variants, respectively. By applying the *t-SNE* dimensionality reduction, less than 75% of the variance was lost.

5.2.4 Tunnels analysis

CaverAnalyst(Jurcik et al., 2018) was used to compute substrate entry channels for the 10 most populated *HDBSCAN* clusters of each HHDH variant. For each *t-SNE* cluster, the nearest MD snapshot was extracted with the mdtraj software(McGibbon et al., 2015) for the analysis of accession tunnels, thus spanning the whole dynamical space of the enzyme. The parameters used for the tunnel search were 4 Å *shell depth*, 2.5 Å *shell radius*, *clustering threshold* value of 3.5 and a 1 Å *minimum probe radius* were used as tunnel search parameters.

5.2.5 Decision trees

All possible minimum distances between residues were defined as input features, defining the shape of the corresponding tunnel and the presence/absence of the studied tunnel as a target feature. We used a Python pipeline to standardize the input data and select the best Random Forest(Breiman, 2001) parameters for the classification. MD data was randomly split into a training set (80%) and test set (20%). We used Python packages Numpy (Harris et al., 2020), Pandas (pandas_dev_team, 2020), Scikit-Learn (Pedregosa et al., 2011), and Matplotlib (Hunter, 2007) for data manipulation, machine-learning, and visualization.

5.3 Results and discusion

As mentioned in the previous chapter, a new computational protocol to study the conformational space of active site tunnels in enzymes was developed, which we also applied in HHDHs. With this, we aim to get insights into the different classes of HHDHs known and understand the role of structural and dynamic differences of the enzymes for their application in the rational design of HHDHs.

The X-ray structure of HheC was analyzed in the tetrameric state to see the effect of the C-terminal part on the tunnels. For that, CaverAnalyst was used in the system in the tetrameric and monomeric state. In contrast to previously suggested, no significant effect of the C-terminal part of the residue W249 on the tunnels was observed. With this information in hand, the subsequent analysis only considered the monomers (although all the simulations were run in the tetrameric state, as mentioned before).

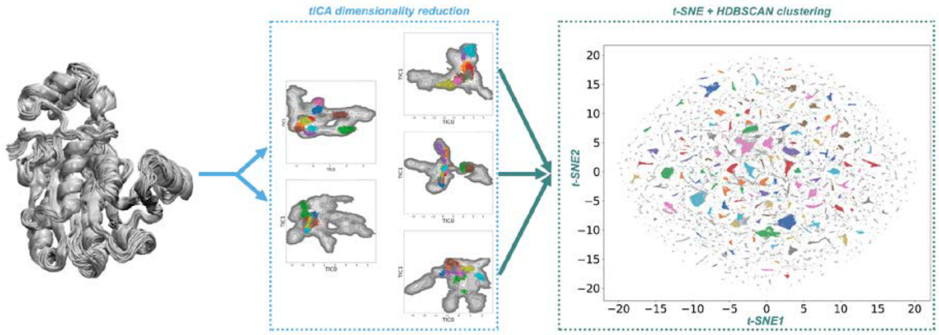

FIGURE 5.2: Computational protocol used to reconstruct the conformational landscapes of the different HHDH subclasses. It is based on a two-step process consisting of first applying to the MD dataset the linear time-Independent Component Analysis (*tICA*), followed by the application of the non-linear t-distributed Stochastic Neighbor Embedding (*t-SNE*) method. In this fashion, the high dimensional MD dataset is reduced into a 2D space that is subsequently clustered using HDBSCAN.

As mentioned in the previous project with CRL (See Chapter 4), the data from the MD simulations was collected, more specifically, the alpha-carbon coordinates. Then, *tICA* was performed to get the slowest movements sampled during the MD simulations, and then *t-SNE* and *HDBSCAN* clustering was applied [Figure 5.2].

Evaluation of the dynamical differences between the selected subclasses of HHDHs.

The first point to evaluate is the most flexible areas of the different HHDHs and understand how these changes in flexibility might affect activity, substrate scope, and stability. To that end, we first analyzed the conformational differences obtained using the *tICA* dimensionality reduction technique. As we can see in the representative overlays in the Figure 5.3 and Figure 5.4, the most flexible HHDHs are the ones that have the widest active site, i.e., HheA2 and HheG. The regions separated by the active-site pocket contain the catalytic residues (active site) and the halide binding site.

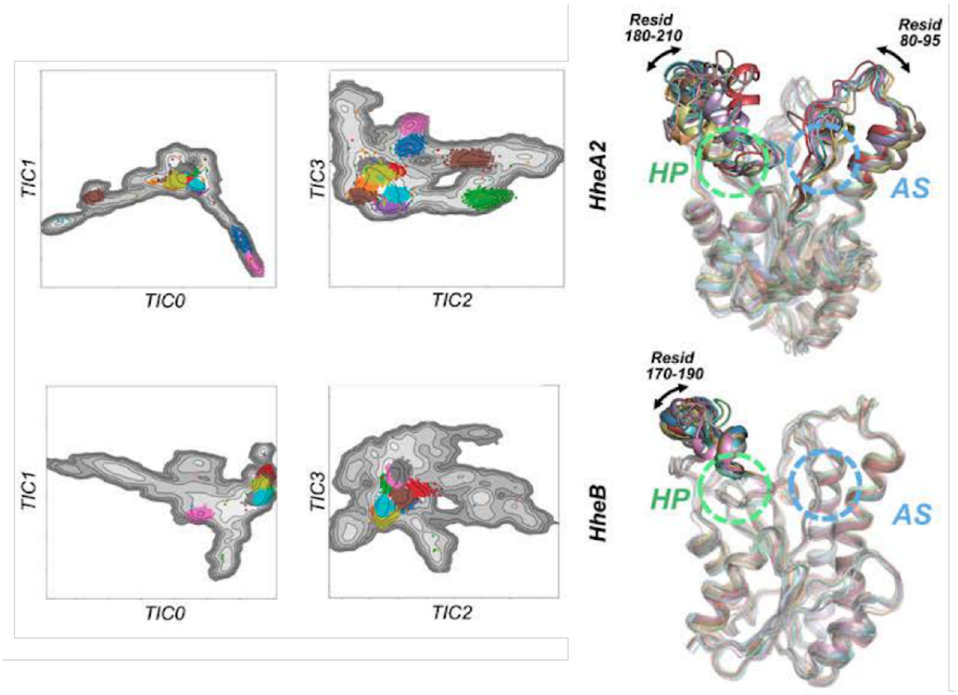

FIGURE 5.3: Representation of the 10 most populated MD conformations as described by the *t-SNE* technique for the different HHDH subclasses analyzed: HheA2 and HheB. The 10 different conformations (each one colored differently) are projected on the *tICA* conformational landscapes. The most flexible parts of the enzymes are marked and numbered accordingly. The active site (AS) and halide binding pocket (HP) locations are marked with a green and blue discontinuous circle, respectively.

For all the HHDHs studied in this chapter, the slowest movement sampled is the so-called "breathing" or open-closed conformational change of the halide binding site. This change implies that the halide binding site is positioned farther away from the active site (i.e., open state) or closer/collapsing on top of the active site region (i.e., closed state).

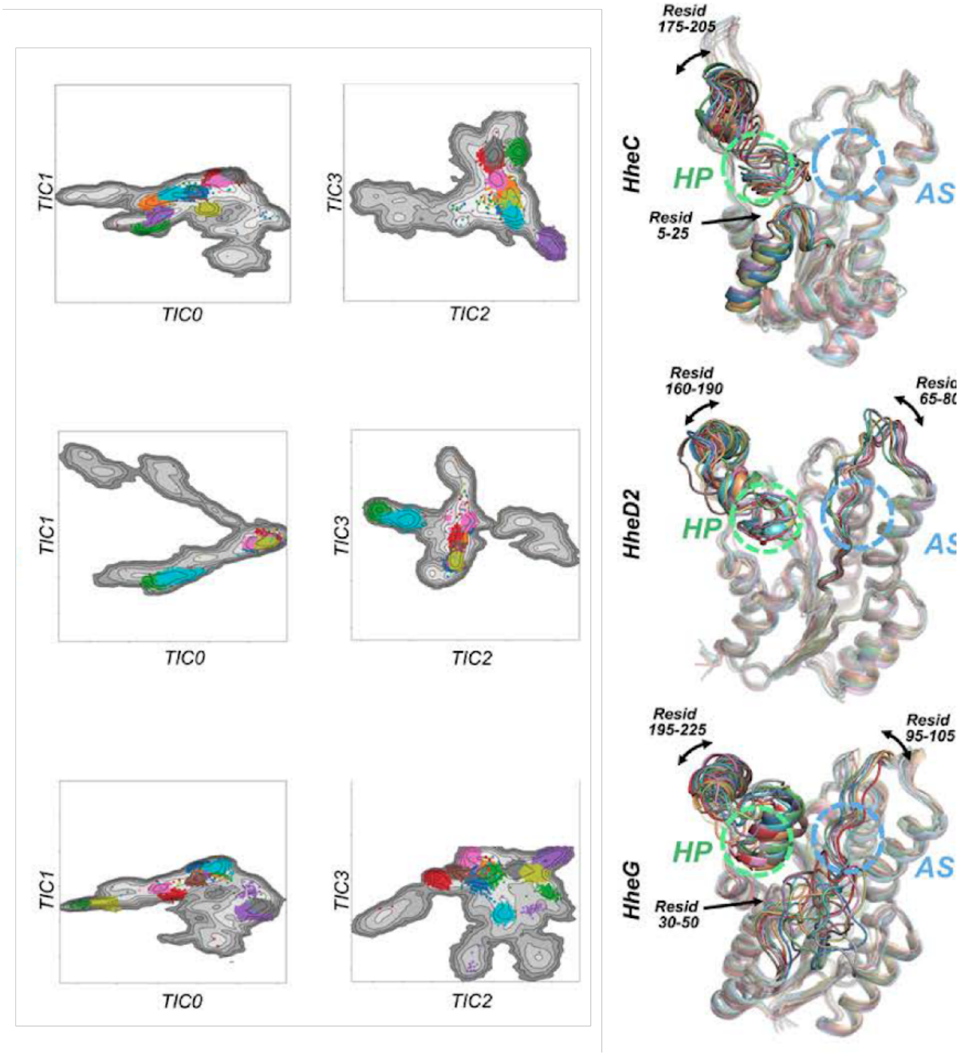

FIGURE 5.4: Representation of the 10 most populated MD conformations as described by the *t-SNE* technique for the different HHDH subclasses analyzed: HheC, HheD2, and HheG. The 10 different conformations (each one colored differently) are projected on the *tICA* conformational landscapes. The most flexible parts of the enzymes are marked and numbered accordingly. The active site (AS) and halide binding pocket (HP) locations are marked with a green and blue discontinuous circle, respectively.

For the HHDHs that have the extra helices (2nd and 3rd) in the N-terminal part (i.e., HheC and HheG), high flexibility in this area is also observed. Still, the most flexible enzyme explored is HheG, which presents a highly disorganized N-terminal region.

On the other side of the equation, we have HheB. This HHDH has a relatively limited conformational heterogeneity as it only exhibits a moderate "breathing" of the halide-binding site region, but but a rigid binding site region. These changes in conformational dynamics may be relevant to explain the high catalytic proficiency that HheB has.

Finally after the *tICA* was performed and explored, the 20 most contributing *TIC* dimensions were used as input for the *t-SNE* algorithm and clusterized using the HDBSCAN. The ten most populated clusters of the *t-SNE* were transformed into the tICA spaces again to evaluate if the *t-SNE* and subsequent clusterization successfully represented the explored flexibility. All *t-SNE* clusters appear to be in well-defined energy minima in the *tICA* space (see Figure 5.3 and Figure 5.4), thus confirming that the *t-SNE* dimensionality reduction faithfully represents the protein dynamics (colored clusters in Figure 5.2).

By analyzing the most populated clusters in the *t-SNE* space, we can extract that most of the variance included for HheA2 comes from the halide-binding site (residues 170-210) and the loop located close to the catalytic Tyr146 (Residues 80-95). HheD2 shows similar behavior for the halide-binding site residues (170-190) and catalytic region (130-150) but it is not as flexible as HheA2.

For HheC, the most populated conformations mainly involve motions from the N-terminal region (residues 32-36) mentioned before and the halide-binding site region. On the other side, for HheG, this N-terminal part is much more disorganized (residues 30-50), showing more flexibility, interacting with the residue Tyr13. This disorganization is the slowest relevant movement in this system. This residue is close to the active and halide-binding sites and may play an essential role in catalysis. HheG also presents a "breathing" movement like HheA2 and HheD2.

Contrary to the previously mentioned HHDHs, HheB displays an entirely different conformational behavior. The protein is tightly packed, and only minor rearrangements on the halide-binding site residues are observed (residues 170-190). The most populated clusters from the *t-SNE* fall into similar conformations in the *tICA* spaces, thus explaining the observed conformational rigidity of this protein.

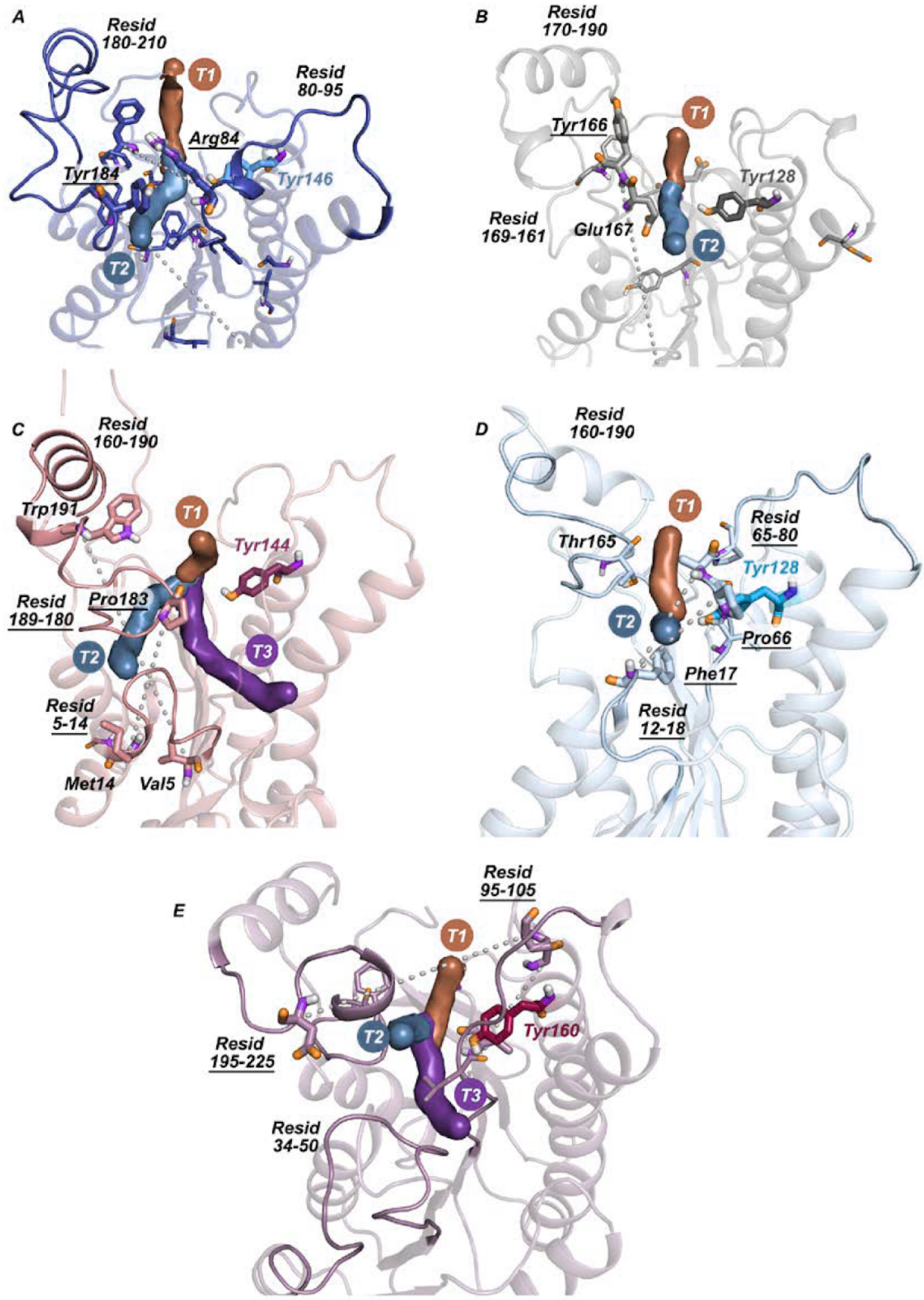

FIGURE 5.5: Representation of the three major tunnels that exist in (A) HheA2, (B) HheB, (C) HheC, (D) HheD2, and (E) HheG: T1 shown in brown, T2 in blue, and T3 in dark purple. The key elements that determine T2 formation in the different subclasses are highlighted.

TABLE 5.1: Mean tunnel bottleneck radius (BR, in Å) for each HHDH system computed on a representative structure of each cluster center. n.d. = not defined

HHDH	Tunnel T1	Tunnel T2	Tunnel T3
HheA2	1.8 ± 0.4	1.6 ± 0.6	n.d.
HheB	1.9 ± 0.6	1.8 ± 0.8	n.d.
HheC	2.0 ± 0.3	1.3 ± 0.2	1.0 ± 0.02
HheD2	1.8 ± 0.5	1.7 ± 0.4	n.d.
HheG	2.2 ± 0.4	1.9 ± 0.5	1.8 ± 0.5

Evaluation of the impact of conformational heterogeneity on the available tunnels

After the analysis of the conformational heterogeneity of the HHDHs studied, we used this data to evaluate the impact of the conformational dynamics on the entrance tunnels, as well as evaluate if these changes might be correlated with the different activity and substrate scope that these enzymes display. To do so, we computed the tunnels of the 10 most representative clusters of each *t-SNE* frame using the CAVER software. We then extracted how predominant the tunnels are for the different HHDHs and the Bottleneck Radius for each tunnel (BR, i.e., narrower region of the tunnel).

We obtained up to three tunnels that we called T1, T2, and T3 [Figure 5.5]. T1 is the central tunnel, which takes the shortest path from the active site to the solvent, going straight upwards, and has a similar shape for all HHDHs studied. This T1 is hypothesized to be affected in HheC by the C-terminal part of the neighboring chain, but the angle shown makes it not disturbed by the other chain. Also, the "breathing" motion observed in the dynamics of the enzyme does not affect the presence or shape of the T1 tunnel. The bottleneck radius of tunnel T1 ranges between 1.8 and 2.2 Å [Table 5.1], with the latter being the biggest in HheG subclass.

T2 and T3 are side tunnels that may appear in the different conformations if the side chains of the surrounding residues display a certain conformation that does not block the entrance: H11, F12, I84, Y185, F186, and the N-terminal residues (using HheC nomenclature). T2 is a tunnel that exits the active site from the front (see Figure 5.5) and is regulated by the residues H11 and F12. If the conformation allows it, a slightly different T2 tunnel (named T2') can be formed, which goes under the halide-binding site residues. T2 is present in HheC but is narrower by the effect of the N-terminal loop and helix residues. These data correlate well with the high activity of the D and G HHDHs toward bulkier and di-substituted epoxide substrates (Calderini et al., 2019).

T3 is only present in HheC and HheG, with a very small BR for HheC. T3 is defined by a T2 from the other side of the loop formed by the residues 80-90, being I84, the residue that permits the presence of T3 in HheC.

To compute the frequency of frames that show a specific tunnel, we assume that all frames in the same cluster share the same tunnels as the representative frame. We can make this assumption thanks to the clustering scheme used that groups together similar frames. It is also known how close they are in the *tIC* spaces after transforming them. We computed the weighted mean in the following way:

$$f = \frac{\sum_{i=1}^{n} \delta_i p_i}{M} * 100 \qquad \delta_i \begin{cases} \Rightarrow \text{ 0 if tunnel is not present} \\ \Rightarrow \text{ 1 if tunnel is present} \end{cases} \tag{5.1}$$

where M is the total number of clusterized frames, p_i is the number of frames in the cluster, and n is the number of clusters in each system.

From this data, we can extract that the tunnel T1 is wildly shown in all HHDHs, making it the main tunnel [Table 5.2]. T2 tunnel is hardly found in A2 and B HHDHs but is much more predominant in the others and has a frequency of more than 90% in the case of HheG. On the other side of the spectra, tunnel T3 is not present in most HHDHs: in only 36.2% of the frames in HheC, and it is much

TABLE 5.2: Computed tunnel frequency for each HHDH subclass. n.d. = not defined

HHDH	Tunnel T1	Tunnel T2	Tunnel T3
HheA2	92.4%	12.3%	n.d.
HheB	97.6%	25.7%	n.d.
HheC	96.9%	77.5%	36.2%
HheD2	88.0%	71.1%	n.d.
HheG	97.6%	91.8%	65.8%

more present in HheG (65.8%).

We hypothesized that tunnel T2 prevalence might be related to the conformational changes in the studied HHDHs. To understand better the motions and regions of the enzymes that regulate the presence and BR of T2, we applied random forest classifiers to elucidate the heavy atom distances that modulate tunnel T2 formation. The input data for the decision trees were the distances between all combinations possible of contacts between heavy atoms that shape the tunnels at any point. The table of residues containing residues that at any point shape the tunnel was obtained using CaverAnalyst. On the other hand, the feature that the decision tree has to predict is the presence or absence of the tunnel T2 for the selected frame through the random forest classifier algorithm. With this, we obtained the contacts between heavy atoms that define the presence or absence of tunnel T2 for each HHDH.

For HheA, T2 formation is greatly influenced by the side chains of the residues Tyr184 (from the halide binding loop), Arg84 (from the loop close to the catalytic residues that delimit the T3 shape in HheC), and other residues from the same regions. These two regions are the most flexible and are involved in the “breathing” movement.

For HheB, the presence or absence of tunnel T2 mainly depends on Tyr166. Only when Tyr166 is displaced out of the active site pocket, the tunnel T2 is available. Because HheB and the region of the Tyr166 are rigid, this explains the low frequency of T2 for HheB. The tunnel T2 in HheC is mainly defined by residues in the protein's most flexible regions: the N-terminal loop (5-14) and the halide binding site, specifically, the residue Pro183. For the HheD2, the tunnel T2 is usually short and only limited by the residues in the N-terminal loop, specifically Phe17, and residues close to the active site (65-80) also play a role in this system.

Finally, in HheG, the tunnel T2 is defined between the halide-binding site residues and the loop close to the catalytic residues. However in this enzyme, the catalytic pocket is wider than in the other HHDHs, and the distance between the catalytic residues and the halide-binding site of the protein is considerably bigger, thus explaining the high prevalence of this tunnel for HheG (92%). Also, the flexible N-terminal part in HheG does not play a role in tunnel T2.

Altogether, studying the flexibility between the HHDHs families has revealed essential regions on the enzymes that may play an important role in defining the different displayed properties. Some, like HheB, are much more rigid than others, but all of them show a "breathing motion" of the halide binding site. This motion brings closer the halide-binding site residues with the catalytic-site residues and thus allows the enzyme to adopt closed conformations of the loop. This open-closed transition was originally thought to be affecting the main tunnel (T1). However, after inspecting the tunnel dynamics in all HHDHs employing this new computational protocol using *t-SNE* coupled with Caver, we observed that tunnel T1 remains unaffected by this motion. Indeed it is tunnel T2 the most affected by this breathing, thus suggesting a key role of T2 for the observed differences in substrate scope and activities displayed in the HHDHs' subclasses.

Chapter 6

Sneaking into the molecular conformations that define thermal stability in halohydrin dehalogenase HheD2

6.1 Halohydrin dehalogenase type D: State-of-the-Art

From the new families obtained in the previous studies using a new query for BLAST search in the databases, the most extensive family of HHDHs that was found was the D(Schallmey et al., 2014). HHDHs type-D show a high melting temperature and can catalyze the reactions at higher temperatures than most of the HHDHs families previously known. They display low catalytic proficiency for converting ethyl 4-chloro-3-hydroxybutyrate into ethyl 4-cyano-3-hydroxybutyrate. However, there is one exception, which is HheD2. This HHDH shows lower thermal stability than the rest of the members of the same class and displays ten times greater catalytic activity than most HHDHs in the D family. The melting temperature of HheD2 is 38ºC, 17-24ºC inferior to that of the other enzymes in the same family(Wessel et al., 2021).

These differences are remarkable, considering the high sequence identity among all enzymes in the family (>67%). To rationalize the sequence variations that make the differences in behavior, the HheD2 was crystalized and studied in Prof. Anett Schalley's lab. As described in the previous chapter, there are some structural differences between HheD2 and other HHDHs from different families, and the sequence identity is low. Still, the tetrameric form, the halide binding site, and catalytic residues are conserved. In this project, we were intrigued to evaluate if conformational dynamics could explain the different thermal stability and activity of HheD2.

6.2 Computational details

6.2.1 System preparation

The selected HHDHs were prepared and simulated without any ligand and in tetrameric conformation. Amino acid protonation states were predicted using the PropKa software(Olsson et al., 2011). Then, the enzyme was solvated in a pre-equilibrated cubic box with a 10 Å buffer of *TIP3P* water(Jorgensen et al., 1983) molecules using the

AMBER16 leap module. The systems were neutralized by addition of explicit counterions (Na^+).

A two-stage geometry optimization approach was performed. The first stage minimizes the positions of solvent molecules and ions imposing positional restraints on solute, and the second stage is an unrestrained minimization of all the atoms in the simulation cell. The systems are gently heated using six 50 *ps* steps, incrementing the temperature 50 *K* each step (0–300 *K*) under constant volume and periodic boundary conditions. Extra heating step of 30 *K* was performed for the 330 *K* MD simulations.

In order to control the temperature, Langevin thermostat was used. All systems were equilibrated without restrains for 2 *ns* at a constant pressure and temperature.

6.2.2 Molecular dynamics

After system equilibration, all MD simulations were performed under *NVT* ensemble, performing at 60 *ns* per day in our in-house GPU cluster GALATEA (Nvidia GTX1080). In particular, five replicas of 500 *ns* were carried out for each system and temperature, adding up to 15 μs (7.5 μs at each temperature) of accumulated MD simulation time. MD simulations were done using the David E. Shaw modification of the Amber 99SB force field (ff99SB-ILDN)(Lindorff-Larsen et al., 2010).

All analysis done was carried out in Jupyter-notebook environment (python), using mdtraj(McGibbon et al., 2015), pytraj (Ambertools 16)(Roe and Cheatham III, 2013; Nguyen et al., 2016) and pyemma(Scherer et al., 2015) libraries. tICA features included in this manuscript are alpha-carbon coordinates.

6.3 Results and discusion

In Prof. Anett Shallmey's lab, they constructed several modifications in HheD2 to enhance thermal stability using two approaches. The first was using the FoldX online server(Schymkowitz et al., 2005) in the AlanineScan mode and comparing ΔGs for each residue if mutated to an alanine. In the end, the residue showing a lower $\Delta\Delta$G was D198. Looking at the 3DM databases, for HHDHs in this position, Asp is one of the less frequent amino acid, being the most common residues: isoleucine and valine. The most uncommon one is leucine. The mutants HheD2 D198I, D198V, and D198L were then selected for experimental validation(Wessel et al., 2021).

The second approach was to replace the region that shows the highest β-factors of the protein (highest mobility) with the ones found in the thermally more stable HHDHs, in this case, HheD3 and HheD5. Therefore the most solvent-exposed helices (αE and αF) in the halide-binding site of HheD2 were replaced with the helix of HheD3 and HheD5, thus creating the variants HheD2 helixD3 and HheD2 helixD5. Both strategies generated more thermally stable enzymes but at the expense of lowering their catalytic rates.

To understand the molecular mechanism that makes these enzymes more thermally stable and less active, we computationally evaluated the system HheD2, HheD2 D198V, and HheD2 helixD3. To better understand the protein's thermal stability, MD simulations were performed at 27 and 57 ºC by including an extra heating step in the heating process. At 57 ºC, we expected the mutants to have activity, as opposed to the WT (based on the experimental data). Five replicas of 500 ns were done for each system at different temperatures, leading to 15 μs of accumulated MD simulated time (2.5 μs (5*500 ns) x 3 systems x 2 temperatures). The first analysis done was to check the RMSF for the molecules at 27 and 57 ºC. The most flexible regions are the helices αE and αF in the halide-binding site and the loop on top of the catalytic site, similar to the same behavior explored in the previous chapter (i.e., the so-called "breathing motion"). However, looking at the data at 57 ºC, the helices in the halide-binding site present much higher flexibility[Figure 6.1].

Further conformational analysis was done by computing the *tICA* on the alpha-carbon coordinates of the enzymes. This revealed three different conformations that the enzymes can explore: X-ray-like, A, and B conformations [Figure 6.2]. These two new A and B conformations disrupt the halide-binding site differently. In A, the halide-binding site is completely unfolded and disorganized, making the enzyme not preorganized for catalysis. In the B conformation, the halide-binding helices site completely collapses on top of the catalytic residues, thus undergoing a drastic “breathing” motion leading to a fully occluded state not optimal for catalysis.

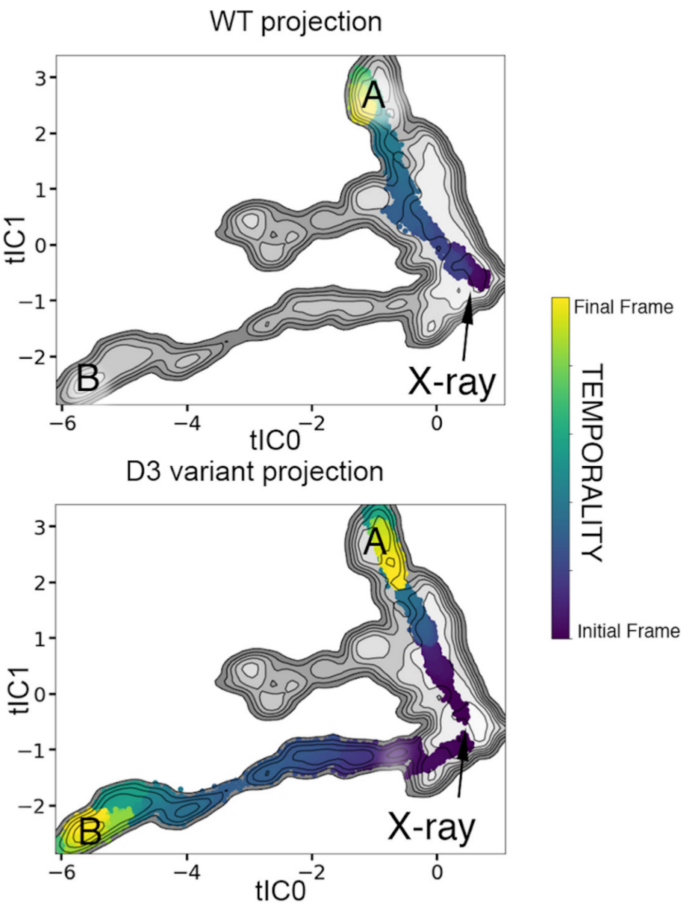

FIGURE 6.1: Common conformational space (shown in grey) for WT and helixD3 variant reconstructed by applying the dimensionality-reduction technique *tICA* to the combined (WT and helixD3) MD trajectories at 57 °C (five replicas of 500 ns, i.e., 2500 ns for each system). The x- and y-axis correspond to the first identified *tIC* components (*tIC0* and *tIC1*), describing the slowest kinetically relevant conformational changes observed along the MD runs. The different conformations were visited at high temperature (57 ºC) (color range: initial MD frames are shown in purple, whereas final ones are shown in yellow). All simulations start at the X-ray structure (marked with a black arrow). The simulation evolves towards minima A in the WT trajectory. For the helixD3 variant, one of the MD simulations also progresses towards A, whereas in the other MD trajectory, minima B is explored.

All these new conformations [Figure 6.2] are sampled in the MD simulations performed at higher temperatures, whereas the ones at standard conditions remain in the X-ray-like conformations [Figure 6.2]. The variant HheD2 D198V showed conformations closer to A. As mentioned before, in the A conformation, the enzyme has no functional halide-binding site. This D198V mutation disrupts the hydrohen-bonding between the residue D198 and Q160. This hydrogen-bond stabilized the loop region 170-186, which contains the helices αE and αF. Such hydrogen-bond disruption makes the halide-binding site substantially more flexible, thus explaining the lower activity shown, especially at higher temperatures.

However, HheD2 helixD3 experimentally showed excellent thermal stability(Wessel et al., 2021) and no drastic activity reduction. This engineered variant can also explore the A and B conformations in the MD simulations. This B conformation is characterized by the slow motion of the αE and αF helices towards residues 68–74, thus leading to the highly occluded state due to the drastic "breathing" motion. This is motivated by the hyper flexibility of the αE and αF helices and the prior breakage of the D198-Q160 hydrogen-bond. However, multiple new polar and hydrophobic interactions were found between the new colliding regions thanks to the modifications introduced in the halide-binding site helixes. This new conformation explains the enhanced thermal stability thanks to the increased buried surface area(Schymkowitz et al., 2005; Schallmey et al., 2013).

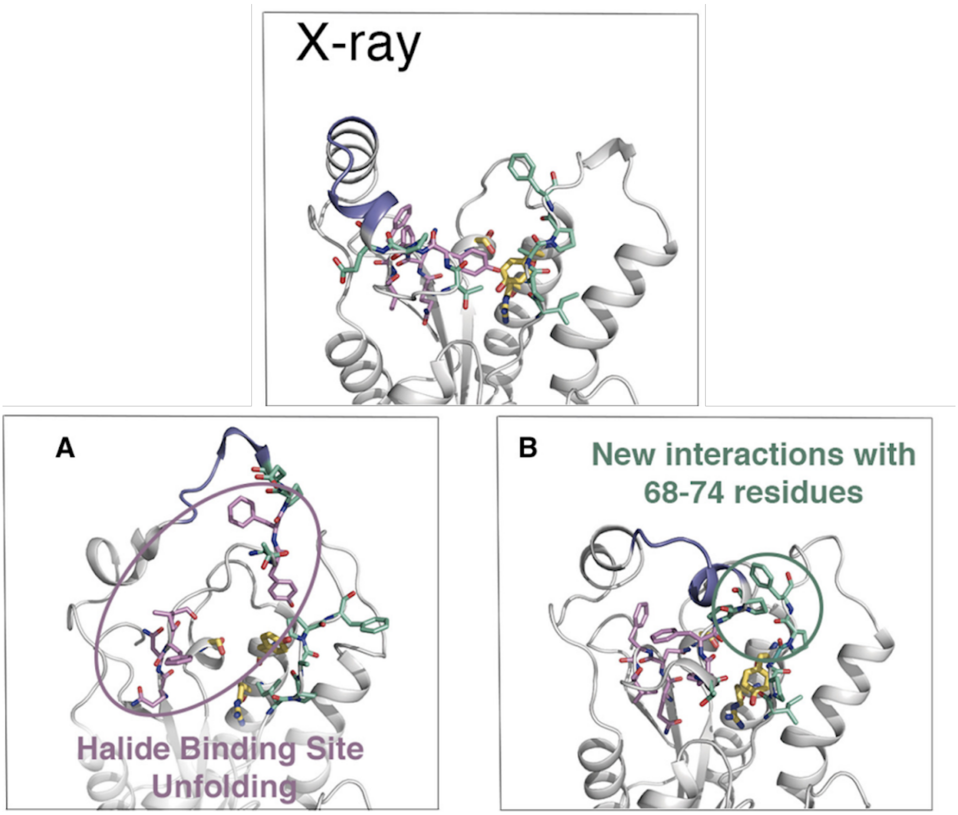
X-ray
A
Halide Binding Site Unfolding
B
New interactions with 68-74 residues

FIGURE 6.2: Representative conformations extracted from energy minima A and B compared with the starting X-ray structure. Catalytic triad residues are displayed in yellow, the halide-binding site residues in purple, and the teal residues establish new interactions in the thermostable conformation. In particular, *tIC0* describes the unfolding of the halide binding pocket, whereas *tIC1* describes the slow motion of the αE and αF helices towards residues 68–74.

In summary, the D198V variant disrupts the hydrogen-bond interaction between D198-Q160. This increases the flexibility of the residues 170-186, but due to the lack of mutations introduced in the HheD2 helixD3 variant, the B conformation cannot be stabilized. This observation explains the reduced thermal stability of the single mutant variant compared to the helix3 variant and results in very low catalytic activity. However, the rearrangement of the helices αE and αF in variant helixD3 at higher temperatures requires the breaking of the hydrogen-bond between D198 and Q160. Consequently, we proposed to Prof. Anett Schallmey's group to study the combination of the two mutants because the mutations may have a synergistic effect. Indeed, the new HheD2 D192V helixD3 exhibited 90% relative activity at 60°C but with a very low base activity. This can be explained by this new enzyme's ability to explore all A and B conformations freely.

Chapter 7

Mixing Biology, Chemistry, Computations and Experiments: Exploration of Halohydrin Dehalogenase C evolved variants.

7.1 Halohydrin dehalogenase type C: State-of-the-Art

Of all the attempts done to engineer HheC, the most successful is the one from Fox et al.(Fox et al., 2007) In this work, they performed DE on HheC from *Agrobacterium tumefaciens* to obtain the best catalyst for their use in the pharmaceutical industry, particularly for the synthesis of statin drugs. This study from Codexis used an algorithm called ProSAR to select neutral, beneficial, or harmful mutations during the different rounds of DE. With this approach, a new library including the beneficial and neutral plus novel random mutations are included in the DE process, thus generating improved variants (round) along the DE evolutionary pathway [See Figure 7.1]. This process finishes until an enzyme fulfilling the desired activity towards ethyl (S)-4-chloro-3-hydroxybutyrate [Overall reaction in Figure: 1.3], superior selectivity and stability for the tested conditions. In this study, they performed 18 rounds of DE, thus yielding the HheC mutated protein (named HheC R18) that includes 37 mutations. HheC R18 and the previous evolved variants have mutations all around the protein, and because of how DE works, what are most of the mutations doing in the enzyme is not known, but all are beneficial in some way.

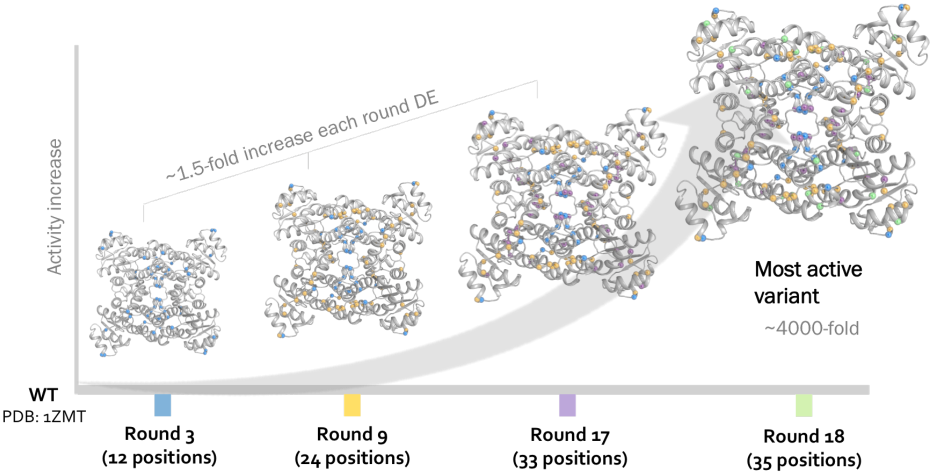

FIGURE 7.1: Scheme of the Directed Evolution (DE) from (Fox et al., 2007). In this scheme, the structures of the different variants obtained are shown. The number of mutations is displayed in parenthesis, and the position of the mutations is shown in spheres. In blue the mutations introduced on the round 3 or prior, in orange, the mutations introduced between round 3 and 9, in purple the mutations introduced between round 9 and 17, and finally the mutations introduced in the round 18 of DE in green.

Some of the mutations included in the last HheC R18 variant were previously studied as single points variants and are known for the effect they have on the protein(Schallmey et al., 2015; Tang et al., 2002; Guo et al., 2015). However, the role of many other mutations is hypothesized. One example is the mutation T143A, which was studied in a previous study and is known to enhance catalytic activity 11-fold(Schallmey et al., 2015). The mutations M245V and C153S enhance the stability of the enzyme by impeding unwanted disulfide-bond formation(Tang et al., 2002). The mutation P84V enhances the enantioselectivity towards the R epoxide(Guo et al., 2015). In a paper from Prof. Janssen, they studied a variant from the ProSAR experiment that was not selected, but showed better activity and much higher thermal stability(Schallmey et al., 2013). In that paper, they described the effect of many of the mutations present. Hereafter the mutations for the thermostable mutant that are found in HheC R18 are discussed. F86W, H201W, V205Y, Q87R, and P135S enhance the stability of the tetrameric conformation by generating new interactions between monomers, thus stabilizing the tetrameric conformation. Also, mutations Q37H, K38Q, G99D, and K121R increase electrostatic effects on the surface, making the protein more water-soluble and stable. F86W affects the binding of bulkier substrates by limiting W139 mobility. F82A and A83P allow a productive binding of the epoxide, thus enhancing the activity and changing selectivity. Mutation P185Y stabilizes the cis-peptide bond in the halide-binding site. This cis-peptide bond is vital for interacting with the halide, water, or another nucleophile in the active site. The mutations A83P, P84V, and P135S use the fact that prolines have less flexible dihedrals to either increase flexibility in desired areas by modifying the prolines for some other amino acids or adding rigidity by introducing the conformationaly restricred proline.

Even with all that information extracted from the crystal structure and using FoldX calculations, the effect of most of the 37 mutations found in R18 is still unknown. The complex interactions and synergistic effects they may have, has also not been explored.

7.2 Computational details

7.2.1 System preparation

HheC WT and variants were prepared with ligand and in dimeric and tetrameric conformation. Amino acid protonation states were predicted using the PropKa software(Olsson et al., 2011). Then, the enzyme was solvated in a pre-equilibrated cubic box with a 10 Å buffer of *TIP3P* water(Jorgensen et al., 1983) molecules using the AMBER16 leap module. The systems were neutralized by addition of explicit counterions (Na^+). The ligand was parameterized using the *antechamber* and *parmchk* programs from *ambertools16*(Case et al., 2005). The charge values were obtained by single point *Hartree-Fock*(Seaton, 1977) calculations using *Gaussian 09* software(Frisch et al., n.d.) using the *6-31G** basis set(Rassolov et al., 2001). Other parameters were extracted from the General Amber Force Field (*GAFF*)(Wang et al., 2004)

As in previous chapters, a two-stage geometry optimization approach was performed. The first stage minimizes the positions of solvent molecules and ions imposing positional restraints on solute, and the second stage is a unrestrained minimization of all the atoms in the simulation cell. The systems are gently heated using six 50 *ps* steps, increasing the temperature 50 *K* each step (0–300 *K*) under constant volume and periodic boundary conditions. Extra heating step of 30 *K* was performed for the 330 *K* MD simulations.

In order to control the temperature, Langevin thermostat was used. All systems were equilibrated without restrains for 2 *ns* at a constant pressure and temperature.

7.2.2 Molecular dynamics

All simulations were done using the Amber 99SB force field (*ff99SB-ildn*)(Lindorff-Larsen et al., 2010) After equilibration in the isothermal-isobaric ensemble (*NPT*), 3 replicas of 2000 *ns* were run for each system in the canonical ensemble (*NVT*). After that, the MD data is obtained is analysed and the free energy is expanded by selecting new starting points of shorter MD simulations from the less explored areas in order to sample all the conformational space as possible. This is known as adaptive sampling (Bowman, Ensign, and Pande, 2010). After several rounds of sampling, 126 μs were obtained from the dimeric systems and 280 μs between HheC and HheC R18 in the tetrameric conformation.

The graphics processing unit (GPU) version of pmemd in Amber16 was used for the MD simulations, which were executed on the in-house GPU cluster GALATEA.

7.2.3 MD analysis

Dimeric systems

To understand backbone movements that might be affected by the dynamics of the enzyme, distances between alpha-carbons were selected as features for *tICA* analysis of the dimeric systems. Because the number of features is equal for all systems, *tICA* space was computed together with all the variants and later split for its evaluation. We used Python packages Numpy (Harris et al., 2020), Pandas (pandas_dev_team, 2020), pyemma (Scherer et al., 2015), and Matplotlib (Hunter, 2007) for data manipulation, statistics, and visualization.

Tetrameric systems

Initially, the analysis intended in this chapter was to build a Markov State Model (MSM)(Chodera et al., 2007; Bowman, Huang, and Pande, 2009; Husic and Pande, 2018) of all the conformations obtained during the MD simulations to elucidate the changes in activity allong the different DE variants: WT, R9 and R18. Trying multiple collective variables (CV) were not successful. We observed that the most evolved variants are much more rigid and stable in only one conformation (especially for the tetramerical systems) compared to HheC WT or less evolved variants. The MSM was therefore not build because at least 2 conformations are needed. Although there is no MSM built, the extensive MD simulated data can be analyzed to rationalize the change in proficiency.

Data was extracted from the MD simulations and sequencial *tICA* dimensionality reduction was used by using the pyemma software(Scherer et al., 2015). The input data for this analysis i.e., CVs was focused on the set of binding distances along the MDs that have a major influence in the proper binding of the cyanide and epoxide substrates in the active site. In order to rank this set of parameters, all distances between carbon-alpha were computed and using random forest regressor (RFR)(Breiman, 2001), were ranked. The best 30 are selected as CVs. The initial data was randomly split into a training set (80%) and test set (20%). We used Python packages Numpy (Harris et al., 2020), Pandas (pandas_dev_team, 2020), Scikit-Learn (Pedregosa et al., 2011), and Matplotlib (Hunter, 2007) for data manipulation, machine-learning, and visualization.

The *tICA* space is build using the 30 CVs selected distances and the surface is clusterized using the kmeans(Jin and Han, 2010) algorithm and 400 random cluster-centers. After this, the Markov macroclusters are build with 2 conformations for the WT system. 2000 frames from each most stable conformation of each macrocluster were extracted and all contacts between all residues were computed for these frames. Smirnov-Kolmogorov non-parametric test("Kolmogorov–Smirnov Test" 2008) was used to explore the most divergent contacts on each macrocluster sampled, The 20 most divergent features were then used as input for generating the new *tICA* space.

7.3 Experimental details

Kinetic experiments were carried out by Ms. Sophie Staar (Günther) under supervision of Prof. Dr. Anett Schallmey, head of the Biochemistry group of the Technische Universität Braunschweig.

All other experiments were carried out by Mr. Miquel Estévez-Gay under supervision of Prof. Dr. Anett Schallmey in the Technische Universität Braunschweig with the help of Mr. Marcel Staar, or under supervision of Dr. Marc Ribó in the Universitat de Girona with the help of Mr. Alejandro Romero in the Enginyeria de proteines (Protein engineering) group.

7.3.1 Protein production

Escherichia coli strains DH5α and BL21 (DE3) Gold (Thermo Fisher Scientific, Darmstadt, Germany) were used as hosts for cloning and heterologous protein production, respectively. HheC WT and mutant genes were expressed from *pET28a(+)* and *pBAD*-based vectors, utilizing a T7 promoter. For the kinetics experiments, an N-terminal hexahistidine tag (His-tag) fusion was included. For stability and oligomerization experiments, the proteins did not include a His-tag.

For heterologous production of HheC, 500x2 *mL* TB media (4 $mL * L^{-1}$ glycerol, 12 $g * L^{-1}$ peptone, 24 $g * L^{-1}$ yeast extract) supplemented with kanamycin or ampicilin (based on the vector) was inoculated using 10% (v/v) of the respective overnight culture. Protein expression was directly induced by adding IPTG or L-arabinose (based on the expression system). After 3 *h* incubation (37 °C, 200 *r.p.m.*), cells were harvested by centrifugation (4400 *g*, 20 *min* at 4 °C) and cell pellets were stored at -20 °C until further use.

7.3.2 Protein purification

With His-tag

Protein pellets were resuspended in 50 *mM* Tris·SO4, 500 *mM* Na_2SO_4, pH 7.5 buffer and cell membranes were broken by sonication. After centrifugation (11000 *g*, 30 *min* at 4 °C) and filtration, purification of resulting enzymes was performed by immobilized metal affinity chromatography using a 5 *mL* HisTrap HP column (Cytiva, Marlborough, United States of America) and an Äkta Pure FPLC system (Cytiva) according to a published protocol(Koopmeiners et al., 2016). HheC-containing fractions with the highest protein purity, as determined by SDS-PAGE, were pooled and afterwards desalted using PD-10 columns (Cytiva) and Phosphate buffer (pH 7.5, 4 *mM* EDTA, 7.1 *mM* β-mercaptoethanol).

Without His-tag

Protein pellets were resuspended in 50 *mM* Tris·SO4, 500 *mM* Na_2SO_4, pH 7.5 buffer and cell membranes were broken by french press. After centrifugation (11000 *g*, 30 *min* at 4 °C) and filtration, purification of resulting enzymes was done by collecting the soluble phase and anion-exchange chromatography using a HiTrap column (GE Healthcare, Freiburg, Germany) on an Äkta Pure FPLC system (GE Healthcare) using NaCl up to 1 *M*. The corresponding target peak eluded at 0.37 *M* NaCl and was tested by SDS-page. After ion-exchange, size-exclusion chromatography was performed in orger to get a extra pure sample, swap the Cl^- ions with Tris-SO_4

50mM pH 7.5, 4 mM EDTA, 7.1 mM β-mercaptoethanol buffer and test the oligomerization state (more on the results section). To so so, Superdex 200 Increase 10/300 GL column (Cytiva) and an Äkta Pure FPLC system (Cytiva).

7.3.3 Thermal Stability assays

Thermofluor assay was used using the purified protein, including the *SYPRO orange* dye. Using the QuantStudio™3 real-time cycler (Thermo Fisher Scientific), the light emited by the dye was monitored while increasing the temperature gradualy up to 95 °C.

A JASCO J810 instrument equipped with a Peltier temperature control module attached to a thermal bath was used in order to monitor the proteins UV-absorbance (260 and 280 nm) while heating the sample.

7.3.4 Halide-release kinetics

Halide-release kinetics was monitored using pure ethyl (S)-4-chloro-3-hydroxybutyrate (*(S)-4-C-3-HB*) in the halide-release assay described in literature(Schallmey et al., 2015). The reaction was monitored by checking the absorbance at 480 nm. With this strategy, the amount of chlorine is monitored, thus knowing how much of the corresponding (S)-ethyl-2-(oxiran-2-yl)acetate (*2-OAA*) epoxide is produced.

7.3.5 Epoxide-opening kinetics

The epoxide ring opening reaction was monitored by monitoring the formation of the hydroxynitrile product by GC (GC-2010 Plus, Shimadzu). Product formation was quantified on the basis of a standard curve for product. Chemical background activities in control reactions without enzyme were subtracted before fitting.

In order to get the get the kinetic parameters of both epoxide and cyanide, the other substrate was kept at fixed concentration at least 5 times more concentrated than the K_{50} value.

7.4 Results and discussion

To rationalize the effect of the introduced mutations in HheC R18, our goal in this project was to study the effect of the mutations introduced in HheC and their effect on the enzyme conformational dynamics using multiple replicas long timescale MD simulations and also experimental evaluation.

7.4.1 Experimental evaluation of HheC WT and R18

It was not known experimentally whether the introduced mutations had an impact on the enzyme oligomerization state, thermal stability, and kinetic constants. In this chapter of the thesis, these parameters were explored and compared between WT and variants in order to know the effect of the mutations introduced.

Changes in the oligomerization state of HheC WT and R18

To experimentally evaluate the oligomerization state of HheC and HheC R18, we produced the two proteins using recombinant genes without including His-tag. The proteins were purified using ion exchange chromatography [Figure 7.2]. The amount of produced protein was significant, and with only this purification step (tested by SDS-PAGE), a significant amount of protein with enough purity was obtained. However, a calibrated size-exclusion chromatography was used to explore the oligomerization states of the proteins (WT and R18) and to further purify them [Figure 7.3]. Interestingly, HheC WT shows the peaks for the monomeric and tetrameric conformation. These peaks were then isolated and analyzed further to understand if this is a stable equilibrium between conformations or if these are stable. After a day, the oligomerization states were measured again, obtaining similar results. Significantly, HheC WT recovered the dimeric conformation after purifying, lyophilizing, and solubilizing it again. This event did not happen in HheC R18, where we observed tetrameric conformation after the ion exchange purification and after one hour after solubilizing. HHDHs are reported as part of the SDS family and are usually reported active as dimer or tetramer(Hylckama Vlieg et al., 2001), but HHDH's primary active conformation is the tetrameric state(Jong et al., 2003). The results suggest that HheC R18 has a much more stable tetrameric conformation.

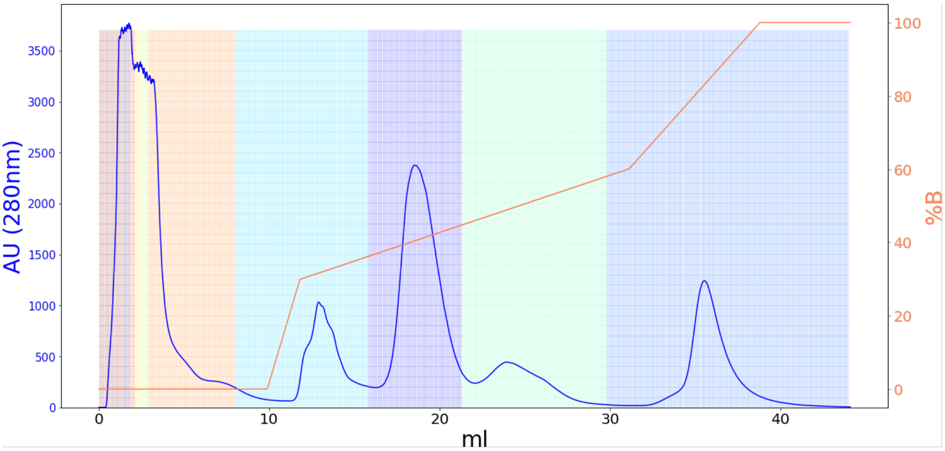

FIGURE 7.2: Output of an ion-exchange chromatography for the purification of HheC. On the X axis it is displayed the volume of liquid phase eluded, in blue, the absorbance at 280 nm and in orange the gradient of buffer A (without salts) and B (with salts).

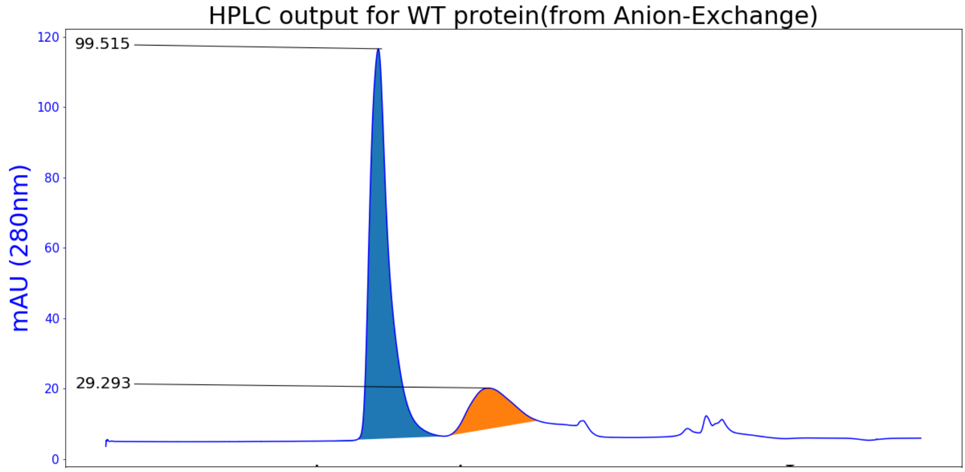

FIGURE 7.3: Output of a Size-Exclusion chromatography (gel filtration) for the purification of HheC and study of the oligomerization state. On the X axis it is displayed the volume of liquid phase eluded. The line in blue indicates the absorbance at 280 nm at every point of the filtration.

Thermal stability assays of HheC WT and R18

To obtain the thermal stability data, we purified HheC and HheC R18, followed by monitoring the absorbance at 260 and 270 nm while heating the samples. With this technique, one can monitor how the absorbance increases thanks to the unfolding events occurring in the protein that expose the usually buried aromatic residues, thus being solvent-exposed [Figure 7.4]. With this experiment, we observed how both HheC WT and HheC R18 showed high thermal stability, but measuring the exact melting temperature value was not straightforward with this procedure.

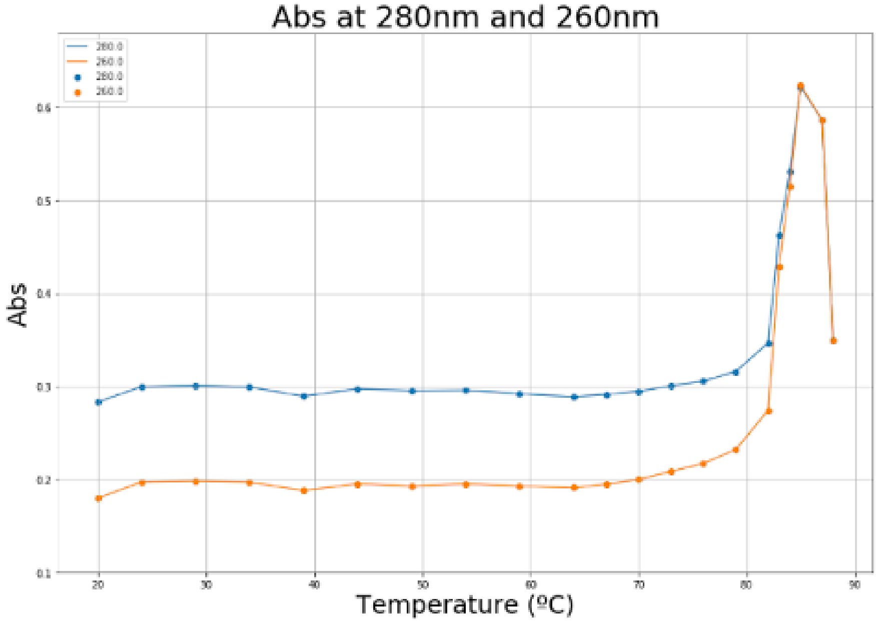

FIGURE 7.4: Spectra at 260 and 280 nm of HheC while increasing gradually the temperature. We can observe that the thermal resistance is high, but it is difficult to decipher $T_{1/2}$.

Additionally, we used the thermofluor technique to evaluate the thermal stability more accurately and obtain the melting temperatures for both WT and R18. To do so, we used the *SYPRO orange* die and a *qPCR* machine. We computed the melting temperatures ($T_{1/2}$) in triplicate. This $T_{1/2}$ is the temperature where half the protein is unfolded. This point is where the melting curve is at half the maximum value, and the slope is at its maximum value. Therefore to correctly compute it, we obtained the position where the first derivative is at the maxima. The median value of all three replicates indicated that all two present a melting temperature of approximately 73ºC.

Thanks to the experimental assays, we know that HheC WT shows high thermal stability, which was not modified along the rounds of DE. Finally and most importantly, what remained to be elucidated was how activity towards the main and promiscuous reaction was modified. It is known and reported by Fox et al.(Fox et al., 2007) that HheC R18 shows a much higher conversion in the overall reaction [Figure 7.5] (ethyl (S)-4-chloro-3-hydroxybutyrate to ethyl (R)-4-cyano-3-hydroxybutyrate using cyanide), but how kinetic parameters are affected for any of the two reactions was not previously studied.

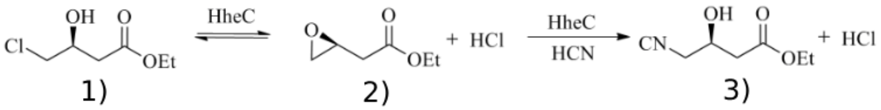

FIGURE 7.5: ChemDraw scheme of the overall reaction catalyzed by HheC. The reaction shows the conversion of ethyl (S)-4-chloro-3-hydroxybutyrate (1) to ethyl (R)-4-cyano-3-hydroxybutyrate (3) using cyanide. This is done in 2 reactions, the first one produces the corresponding epoxide (2) and releases clorine (Cl^-)

Kinetic characterization of HheC WT and R18

For the first dehalogenation reaction, the data shows a Michaelis-Menten distribution for both HheC WT and HheC R18 [Figure 7.6]. HheC WT shows better k_{cat} and K_M values, thus making HheC WT an overall 15 times more effective. Knowing that HheC R18 was evolved to perform both reactions consecutively but was specially engineered for enhancing the second promiscuous reaction, this worst dehalogenation activity found in HheC R18 is not unexpected. Interestingly, despite the tetrameric state of the protein, the shape of the kinetics did not change into a Hill equation, so no cooperativity between monomers is needed at this stage.

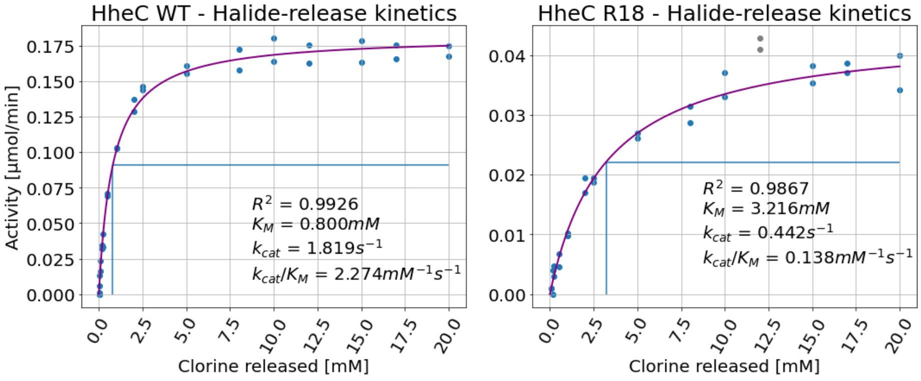

FIGURE 7.6: Kinetic data obtained from HheC WT and HheC R18. The data is represented in a Michaelis-Menten distribution. With this data, k_{cat} and K_M are obtained for the dehalogenation reaction.

Regarding the promiscuous epoxide-ring opening reaction using cyanide, Nickel chloride was used to track the cyanide consumption. This is done by measuring absorbance at 276 nm. For this reaction, pure 2-OAA is used as a substrate, and $NaCN$ as CN^- source. To get the kinetic parameters for the epoxide and cyanide, we kept one of the substrates at a high concentration to make it not affect the kinetics. The other was added at different concentrations to obtain the kinetic values. After measuring the kinetic values of the cyanide, it was clear that to measure the kinetics of the epoxide correctly, it was necessary to use a high concentration of cyanide. This was not done at five or ten times more than the K_M of the cyanide for safety reasons, so only two times the K_M values for the cyanide were used. With that, it is important to note that while kinetic values may not be highly accurate, they are still sufficiently precise for the purpose of comparison.

The results showed that HheC R18 has more than one hundred (100) times higher catalytic activity towards cyanide and one hundred and thirty (130) times more thowards 2-OAA [Figure 7.7]. In both cases, the significant increase is in the k_{cat} value. Another interesting observation is that for both cyanide and *2-OAA*, the kinetics follow a Hill–Langmuir formula and not Michaelis-Menten with Hill constant (n_H) of 3.36 and 1.43 values, respectively. Otherwise, regarding the kinetics values of HheC WT, only cyanide follows a Hill equation (n_H=3.03), and epoxide follows the typical Michaelis-Menten. We used Hill to fit the data in all the cases to make the data comparable.

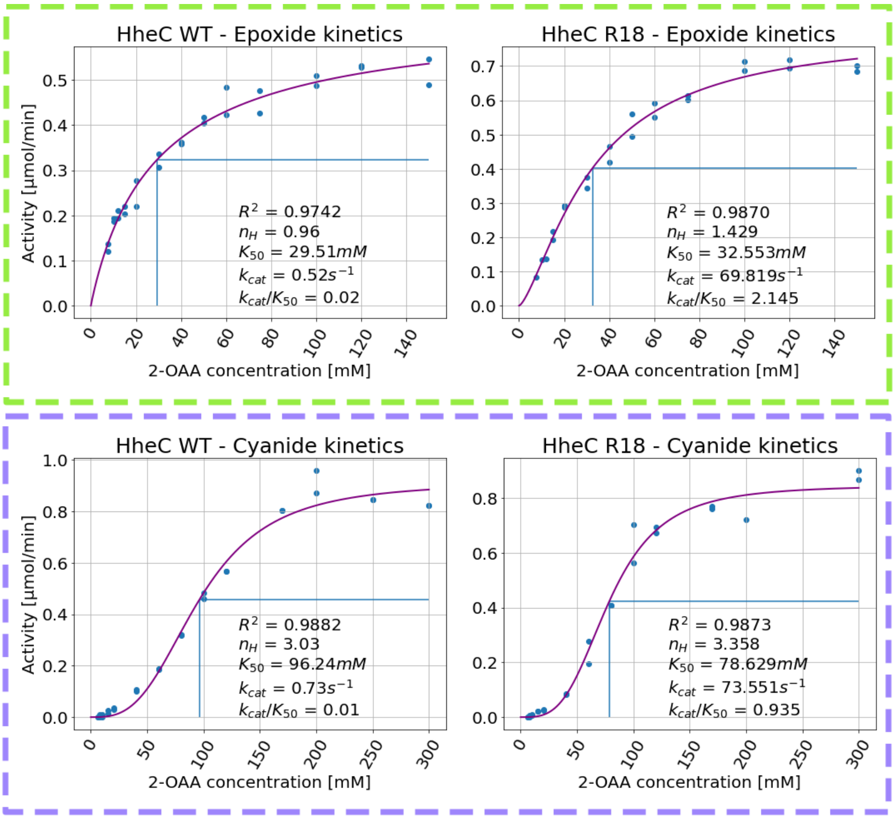

FIGURE 7.7: Kinetic data obtained from HheC WT and HheC R18. The data is represented in a Hill distribution. With this data, k_{cat} and K_{50} are obtained for both 2-OAA and cyanide on the epoxide-ring opening reaction.

Therefore the mutations introduced in HheC to obtain HheC R18 yielded a much higher activity towards the promiscuous reaction. The rates for the epoxide-ring opening reaction are equal or higher as in the HheC WT, making it not the promiscuous reaction anymore. The mutations increased the catalytic efficiency by increasing k_{cat} and not K_M or K_{50}. The other parameter that increased is the Hill constant (n_H). This parameter is what defines cooperativity between subunits(Weiss, 1997). For this second reaction, n_H for the binding of the epoxide is 0.96 for HheC WT and 1.43 for HheC R18. This means that there is a higher cooperativity between subunits on the R18 variant. One way to understand positive cooperativity ($1 < n_H <$ number of active sites) in the Hill equation is that it is easier to have an effective binding event in one monomer if there is effective binding in another active site (understanding effective binding as the event of having the substrate well positioned for catalysis). If there is negative cooperativity ($0 < n_H < 1$), it is more challenging to have effective binding if there is a substrate in another active site.

Monitoring the cyanide consumption, we measured that the Hill parameter is much higher in HheC WT but even higher for HheC R18, being 3.03 and 3.36, respectively. This means that for this reaction when effective binding happens for the cyanide, it is highly probable to have another cyanide in another active site. In the case of HheC R18, the value is almost at its maximum (4); which means that the binding event of cyanide is significantly increased when cyanide molecules are bound in the other active site, making it almost simultaneous binding.

7.4.2 Computational exploration of the conformational landscape of HheC WT and R18

More than one-hundred μs of simulated MD time in the tetrameric system (HheC WT and R18) for each variant was accumulated to evaluate the conformational dynamics of HheC WT and R18 and understand their different catalytic activities (see below). However, we also included the analysis of the dimeric (AB) system because we experimentally observed that HheC WT is also found as a dimer (as shown by the size-exclusion chromatography) and has been described previously in SDR-family enzymes and HHHDs.

Capturing global conformational dynamics of HheC WT, R9, and R18 in the dimeric state

All the systems simulated in this chapter have cyanide, and the *2-OAA* epoxide initially bound in the active sites. For the long 2000 *ns* MD simulations, both epoxide and cyanide unbind the active site early in the MD simulations for all systems.

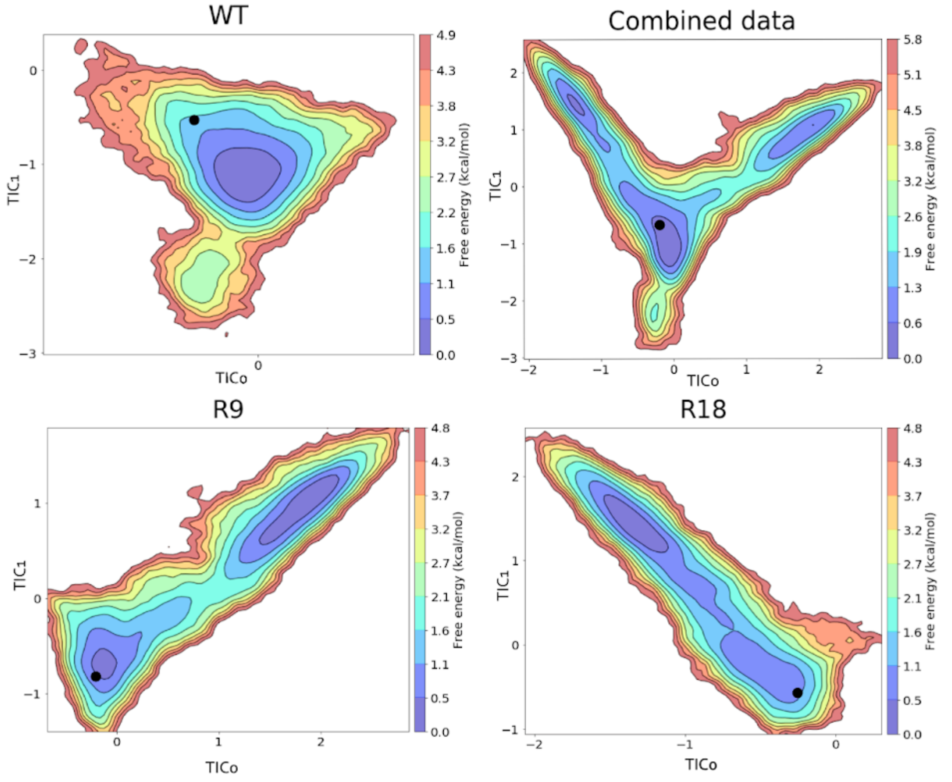

FIGURE 7.8: Free Energy Landscapes obtained from the MD simulations of the systems in the dimeric oligomerization State. With a black dot is marked the position of the X-ray of HheC WT in the *tICA* space. Individual plots for each system are created, and on the top-right there is the FEL generated using all data of all variants together.

HheC WT has a main conformation not far from where the crystal structure lies[Figure 7.8]. All conformations in HheC dimer are not drastically different from each other, and none show any considerable deformation or movement of the enzyme. HheC R9 shows two main conformations; it maintains the X-ray-like conformation, and another one where the big helix that holds the catalytic residues Tyr145 and Arg149 are displaced, which are not well positioned for catalysis. Regarding HheC R18, the X-ray-like structure is also explored, but now another predominant conformation has been sampled. In this new conformation, the halide-binding site is completely disrupted and reorganized due to a massive opening of the catalytic pocket. In this new conformation, catalysis is unlikely to happen because the halide binding site and catalytic residues are entirely disorganized. Compared to HheC R9, the barrier of going from the X-ray-like conformation to the disorganized one in HheC R18 is almost barrier-less7.8.

The results from the MD simulations in the dimeric conformation seem to be paired with the experimental results in the oligomerization state. HheC is stable in X-ray-like conformations. However, HheC R9 and R18 are not. Also, we know that the mutations included in the first half of the DE process already destabilized the dimeric conformation in HheC. Also, the dimeric conformations fail to stabilize a conformation with the substrates in the active site, also pairing with the reported data suggesting that the dimeric conformation is not active.

Capturing global conformational dynamics of HheC WT, R9, and R18 in the tetrameric state

We also explored HheC WT and R18 in the tetrameric oligomerization state. Our analysis also started with long MD simulations and progressively sampling other conformations in rounds of ten replicas of 200 ns simulations. In this case, only HheC WT and HheC R18 were computed (due to the higher computational cost of the tetrameric calculations). Interestingly, as opposed to the dimeric state backbone analysis using alpha-carbon distances shows a much better stabilization of HheC R18 than HheC WT. HheC R18 displays a single conformation resembling the X-ray structure with the halide binding and active site pockets properly preorganized for catalysis[Figure 7.9].

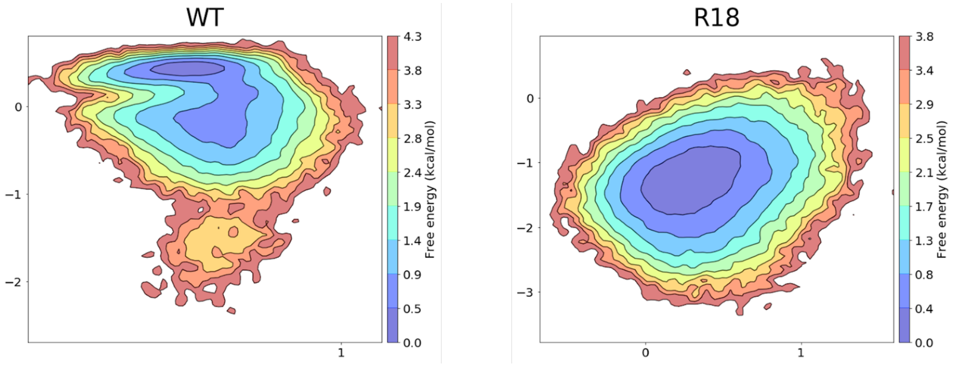

FIGURE 7.9: FEL obtained from the MD simulations of the systems in the tetrameric oligomerization State. Individual plots for each system are created. On the left, the FEL for the HheC WT and on the right the FEL of HheC R18.

It should also be mentioned that binding distances for cyanide and 2-OAA are much shorter for both systems than those found in the dimeric simulations, and there are no unbinding events at the start of the simulations. By computing the binding distances for each chain, one can find similar trends for monomers A and C in HheC WT. Still, HheC R18 improves binding distances in monomers B and D, specifically in cyanide binding. Because of this phenomenon and the higher Hill coefficient found for R18 in the cyanide epoxide-ring opening reaction, we explored the allosteric effect that this event might have using the Shortest Path Map (SPM) tool(Romero-Rivera, Garcia-Borràs, and Osuna, 2017b).

The SPM highlighted in HheC WT essential residues (Tyr178 and Leu179) in the halide-binding site and connected them to the catalytic and closeby residues of the opposite monomer (Tyr145, Ala104, and Leu105) [Figure 7.10]. During the path, residues Lys122 and Val113, mutated in HheC R18, are selected, and Asp97, Asn114, Val116, Met120, Arg123, Ser144, Ala159, Lys162, and Gly164 are residues that are mutated in HheC R18 and interact with residues in the SPM path. This path seems essential for describing the allosteric communication existing between monomers and mutating residues in this pathway seems to be incrementing the allosteric effect in HheC R18.

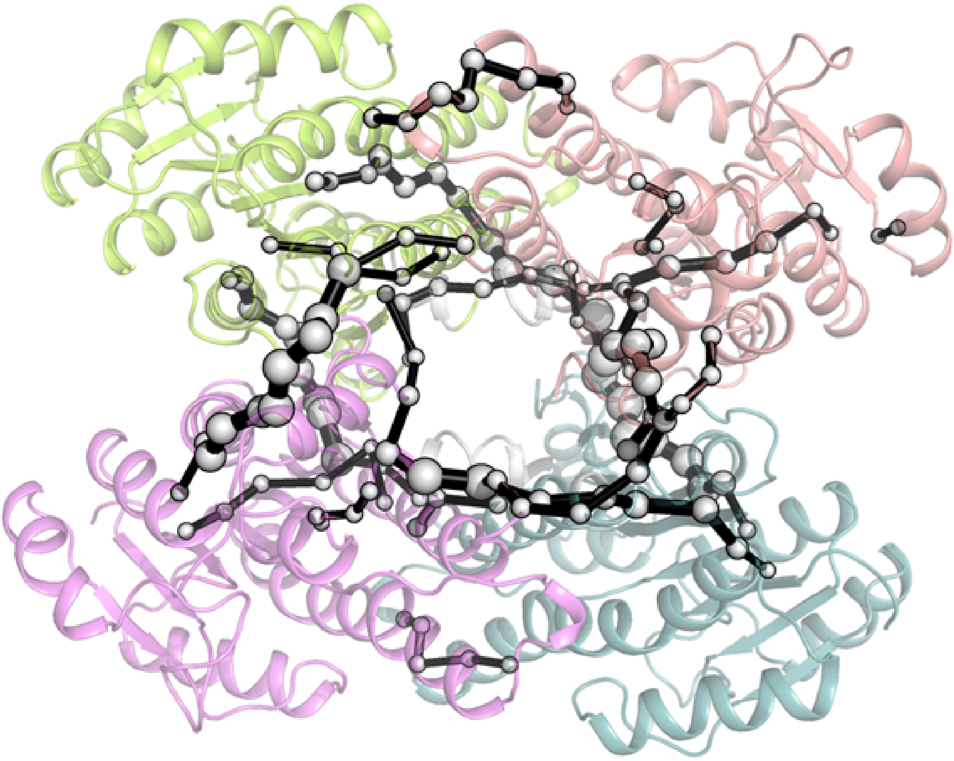

FIGURE 7.10: 3D representation of the SPM results obtained from the MD simulations of HheC WT.

Capturing active site conformational dynamics of HheC WT and R18

To rationalize the increase in k$_{cat}$ between HheC WT and HheC R18, we created a new set of features to focus more on catalytically important events (rather than global conformational dynamics). Due to having four active sites, the simulations were analyzed as monomers and aligned with each other. To have a better set of features, Random Forest Regressors (RFR) were used to extract those alpha-carbon distances that explain the better binding of cyanide and *2-OAA*.

For cyanide binding, the distance between CN^- and the N of Tyr177 in the halide-binding site is computed and used as a target in the RFR.

The output shows that three regions are of great importance: the loop and the end of the N-terminal alpha-helix (11-14, more specifically, Gly13), residues in the alpha-helix close to the catalytic Arginine (99-105), and the colliding residues in the loop close to the catalytic residues (specifically Ile81).

The 30 distances with higher prediction values from the RFR are then used as features for the new *tICA* space. This new *tICA* space describes the different conformations represented by the previously described areas that are sampled [Figure 7.11]. HheC R18 stabilizes one conformation, but HheC WT samples two other barrierless conformations. Because all residues have fundamental interactions between side chains, and by using alpha-carbon interactions, we are not explicitly considering the conformation-defining CVs. In order to achieve that, 2000 frames of each conformation in HheC WT were extracted, and all contacts between residues were computed. Then, to decipher the ones that vary the most between the different conformations, the distribution of the contact distances between conformations were compared using the Smirnov-Komogorov nonparametric test.

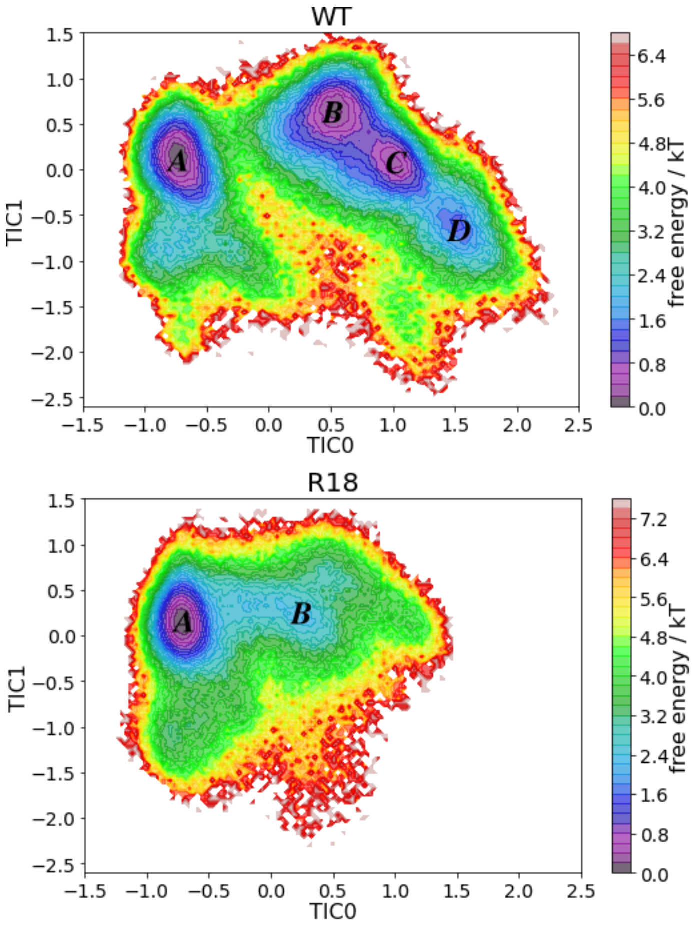

FIGURE 7.11: Free Energy Landscape (FEL) obtained from the MD simulations of the systems in the tetrameric oligomerization State using as features/collective variables the residue contacts selected by the *Smirnov-Kolmogorov* analysis. The new *tICA* space has been created with all data together and after splitted in order to obtain the individual plots for each system. On the top, the FEL for the HheC WT and on the bottom the FEL of HheC R18. HheC WT samples A, B, C and D conformations, and HheC R18 stabilizes A conformation mainly and part of the B conformation (but not stable). In red are the conformations in higher in energy, and in blue to purple the most sampled and stable conformations.

TABLE 7.1: Distances explored in the A-D conformations in HheC WT and in all simulated time. CN^- binding distance is the distance between the nucleophyle and the Tyr177, *2-OAA* binding distance is the distance between the oxigen on the epoxide and the catalytic Tyr, and finally, the preorganization distances are the mean distances between side-chains of the catalytic residues.

HheC WT	CN^- **binding**	**2-*OAA* binding**	**Preorganization**
A	8.9 ± 7.2Å	19.2 ± 13.1Å	5.1 ± 2.2Å
B	16.7 ± 14.5Å	24.4 ± 13.4Å	4.7 ± 1.6Å
C	13.1 ± 12.9Å	21.3 ± 13.3Å	4.8 ± 1.9Å
D	19.6 ± 13.2Å	24.3 ± 14.3Å	4.7 ± 1.7Å
Total	14.7 ± 12.3Å	20.9 ± 13.3Å	4.8 ± 1.8Å

By ranking the results, the contacts between residues that describe the most differences between conformations are unveiled. Unsurprisingly, most of the contacts are between residues in the previously mentioned regions, but some take more importance, such as Asn79, and Asp80, the catalytic residues, and nearby residues like Thr134 and Ile130.

To get a new *tICA* space with greater detail, the data obtained was used to describe a new space. This new space has more defined conformations, but HheC R18 stabilizes only one conformation compared to HheC, which explores multiple [Figure 7.11].

In this new space, the binding of the epoxide, the cyanide, and the preorganization of the active site are explored. Distances on each minimum were computed and compared.

In the A conformation (shared by HheC WT and HheC R18), HheC R18 shows smaller distances between catalytic residues, similar to the ones in the HheC WT X-ray structure (4.2Å). For HheC WT, the conformations B, C, and D show better preorganization distances for catalysis (arround 4.6-4.9Å) [See table 7.1]. On the other hand, looking at the cyanide-binding distances, HheC WT shows much shorter distances between the cyanide and the catalytic

TABLE 7.2: Distances explored in the A-B conformations in HheC R18 and in all simulated time. CN^- binding distance is the distance between the nucleophile and the Ala177, *2-OAA* binding distance is the distance between the oxigen on the epoxyde and the catalytic Tyr, and finally, the preorganization distances are the mean distances between side-chains of the catalytic residues.

HheC R18	CN^- **binding**	**2-*OAA* binding**	**Preorganization**
A	12.30 ± 11.26Å	12.27 ± 9.45Å	4.66 ± 1.72Å
B	14.64 ± 10.61Å	15.21 ± 9.79Å	4.92 ± 1.94Å
Total	13.51 ± 11.63Å	12.11 ± 9.11Å	4.70 ± 1.77Å

residues in the A conformation. This means that there is no conformation in HheC WT where we can observe properly preorganized catalytic distances and also good CN^- binding distances[Table 7.1]. For HheC R18, CN^- binding distance in minimum A is still smaller than the one in the other conformation B, but the mean value is higher than HheC WT. Finally, the distance between *2-OAA* and the catalytic residues is also smaller in conformation A [Table 7.2]. Although the mean distances are quite high (and also standard deviation) in HheC R18 there is a higher ammount of frames with both CN^- and 2-OAA below 4Å (3.25% of frames in conformation A for HheC R18, as oposed to 1.65% frames in conformation A for HheC WT). This data is in line with the superior activity of HheC R18 with respect to the WT .

Thanks to all the data obtained during the MD simulations and the thorough analysis done, the regions of the enzyme that are enhancing the catalytic activity, stability, and oligomerization state are now known. *tICA* analysis played a crucial role to decipher the slow movements that define the catalytically active conformations in HheC and the R18 variant. The SPM unveiled the allosteric pathways that explain the great cooperativity observed in the wet lab by studying the enzyme kinetics. Additionally, the positions that play a higher role in the communication pathway are now identified thanks to the correlation-based SPM analysis.

We expect that altogether this data and the further analysis of the mechanisms that are involved in the better performance of HheC R18 will be relevant to pinpoint and design new positions to rationally design improvedHheC variants displaying higher stability and broader substrate and reaction scope.

Chapter 8

Conclusions

In this thesis, we described the conformational changes among natural and laboratory-evolved HHDHs. We evaluated the effect of the differences in sequence due to natural and laboratory-introduced mutations and how these impact properties like activity or thermal stability. This analysis was done based on the synergistic combination of computations and experiments.

Computational pipelines based on Molecular Dynamics (MD) simulations, novel dimensionality reduction techniques in the field, and feature selection techniques were developed and applied along the thesis, which showed to be successful in unveiling and characterizing the key conformational changes and multiple conformations sampled during the MD simulations. These computations were complemented by multiple experimental assays to better characterize the physical properties of the laboratory-evolved HheC variants of importance for the synthesis of statin drugs.

Hereafter, a summary of the main conclusions extracted from this thesis is presented:

- In **Chapter 4**, a **computational pipeline** based on different dimensionality reduction techniques for identifying the key conformational changes explored through multiple MD simulations was first developed and tested in non-HHDH enzymes. This protocol, coupled with tunnel analysis calculations, was found to reliably match the experimental data provided by our collaborators.

- In **Chapter 5**, the previously developed protocol, coupled with **tunnels-analysis**, was applied in **wild-type HHDHs enzymes** representative from different families for the first time. We described three main tunnels in HHDHs and rationalized the main structural differences among families in the halide-binding site, N-terminal helices, and loops and how these changes affected the available tunnels. A "breathing" motion of the halide binding and active sites was found, which regulates the formation of one of the tunnels named T2. Based on that, we identified the most crucial residue contacts for defining these T1-T3 tunnels and proposed through machine learning techniques some possible positions for enhancing the substrate scope of HHDHs.

- In **Chapter 6**, the **thermal stability** of HheD2 and some related mutants HheD2 D198V and HheD2 helixD3 were studied. We computationally explored the different conformations that these enzymes can explore at 27 and 57 °C temperatures. By **rationalizing** the newly explored conformations, we unveiled the mechanism by which the HheD2 enzyme enhances its thermal stability. The effect of the mutations and helixD3 transfer was also investigated. D198V mutation breaks a crucial hydrogen bond that stabilizes the helices αE and αF and leads to the collapse of the halide-binding site into the opposite side of the enzyme. The exchange of the helices αE and αF stabilized this collapsed conformation. As both modifications showed a similar effect, we hypothesized that both mutations could have a synergistic effect. This was validated experimentally by Prof. Anett Schallmey's group.

- In **Chapter 7**, the laboratory-evolved HheC R18 was **characterized** both **experimentally and computationally**. The experimental characterization involved the measurement of the melting temperature, oligomerization state, and kinetic parameters for the natural and promiscuous reaction. These assays indicate that the introduced mutations along the Directed Evolution campaign did not alter the thermal stability (as both WT and R18 share the similar melting temperature), but instead affected the oligomerization state, cooperativity between monomers, and the kinetic constants (especially the k_{cat} for the promiscuous reaction). To understand the molecular mechanisms that make the protein more **stable as a tetramer**, have **higher k_{cat}** towards cyanide and (S)-ethyl-2-(oxiran-2-yl)acetate without losing **thermal stability**, we performed long timescale MD simulations with the dimeric and tetrameric systems. From the MD simulations, we observed that the evolved variants adopt non-competent conformations as dimers compared to the WT enzyme. The exploration of the allosteric pathways through the Shortest Path Map (SPM) that provided the evolved variant with higher cooperativity between active sites was used to identify the key amino acids and regions involved in modifying the catalytic efficiency of HheC and HheC R18. The analysis of key distances of active site preorganization, and cyanide and epoxide binding between both WT and R18 systems show a higher percentage of catalytically productive events for HheC R18 in line with its superior activity for the promiscuous reaction.

These conclusions fully accomplished the objective of exploiting the flexibility of the different HHDHs families, as well as the tunnels they present. Insights on the key residues and regions that might have a stabilizing effect, affect the substrate scope, or even activity of the catalyst. This was achieved by fulfilling the other objectives of creating a new computational protocol in order to describe the variance in the tunnels and conformations in a statistically accurate manner. With this, the most complex system in the family of HHDHs (the most studied and with the most reported mutations, HheC) was studied with the aim of understanding the role of the mutations introduced by the DE protocol.

HheC and variants were studied and new data was reported like thermal stability, activity, or oligomerization state as well as differences in the conformational dynamics, which was not reported in the literature. The mid-term goal is to design new HHDH variants with improved activity towards non-natural epoxides and nucleophiles. This is work will continue in order to prove the usefulness of the computational protocols reported here. All of this taking into account the limitations of the methods, the simplification required in forcefield simulations, limitation in the protonation state of residues, and still the difficulty to obtain a large amount of simulated time in order to get all possible conformations and make accurate and realistic predictions of experimental data.

All this work done and the data published in this thesis will hopefully help future researchers working in the HHDHs or even any other protein to design improved variants.

Bibliography

Abdelmohsen, Loai K E A et al. (2013). “Micro- and nano-motors for biomedical applications”. en. In: *J Mater Chem B* 2.17, pp. 2395–2408.

Alberts, B. et al. (2014). *Molecular Biology of the Cell*. Online access with subscription: Kortext. W. W. Norton. ISBN: 9780393536966. URL: https://books.google.com.ec/books?id=NEY6zgEACAAJ.

Anderson, Vernon E (2001). “Ground State Destabilization”. In: *Encyclopedia of Life Sciences*. John Wiley Sons, Ltd. ISBN: 9780470015902. DOI: https://doi.org/10.1038/npg.els.0000625. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1038/npg.els.0000625. URL: https://onlinelibrary.wiley.com/doi/abs/10.1038/npg.els.0000625.

Arqué, Xavier et al. (2019). “Intrinsic enzymatic properties modulate the self-propulsion of micromotors”. In: *Nature Communications* 10.1, p. 2826. ISSN: 2041-1723. DOI: 10.1038/s41467-019-10726-8. URL: https://doi.org/10.1038/s41467-019-10726-8.

Boehr, David D., Ruth Nussinov, and Peter E. Wright (2009). “The role of dynamic conformational ensembles in biomolecular recognition”. In: *Nature Chemical Biology* 5.11, pp. 789–796. ISSN: 1552-4469. DOI: 10.1038/nchembio.232. URL: https://doi.org/10.1038/nchembio.232.

Bowman, G. R., X. Huang, and V. S. Pande (2009). “Using generalized ensemble simulations and Markov state models to identify conformational states”. In: *Methods* 49.2, pp. 197–201.

Bowman, Gregory R, Daniel L Ensign, and Vijay S Pande (2010). “Enhanced modeling via network theory: Adaptive sampling of Markov state models”. en. In: *J Chem Theory Comput* 6.3, pp. 787–794.

Brändén, C.I. and J. Tooze (1999). *Introduction to Protein Structure*. Garland Pub. ISBN: 9780815323051. URL: https://books.google.es/books?id=miwWBAAAQBAJ.

Breiman, Leo (2001). "Random Forests". In: *Machine Learning* 45.1, pp. 5–32. ISSN: 1573-0565. DOI: 10.1023/A:1010933404324. URL: https://doi.org/10.1023/A:1010933404324.

Calderini, Elia et al. (2019). "Selective Ring-Opening of Di-Substituted Epoxides Catalysed by Halohydrin Dehalogenases". In: *ChemCatChem* 11.8, pp. 2099–2106. DOI: https://doi.org/10.1002/cctc.201900103. eprint: https://chemistry-europe.onlinelibrary.wiley.com/doi/pdf/10.1002/cctc.201900103. URL: https://chemistry-europe.onlinelibrary.wiley.com/doi/abs/10.1002/cctc.201900103.

Campello, Ricardo J. G. B., Davoud Moulavi, and Joerg Sander (2013). "Density-Based Clustering Based on Hierarchical Density Estimates". In: *Advances in Knowledge Discovery and Data Mining*. Ed. by Jian Pei et al. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 160–172. ISBN: 978-3-642-37456-2.

Case, David A et al. (2005). "The Amber biomolecular simulation programs". en. In: *J Comput Chem* 26.16, pp. 1668–1688.

Castro, C. E. and E. W. Bartnicki (1968). "Biodehalogenation. Epoxidation of halohydrins, epoxide opening, and transhalogenation by a Flavobacterium species". In: *Biochemistry* 7.9, pp. 3213–3218. ISSN: 0006-2960. DOI: 10.1021/bi00849a025. URL: https://doi.org/10.1021/bi00849a025.

Chodera, J. D. et al. (2007). "Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics". In: *J Chem Phys* 126.15, p. 155101.

Cleland, W.W. (1990). "3 Steady-State Kinetics". In: ed. by David S. Sigman and Paul D. Boyer. Vol. 19. The Enzymes. Academic Press, pp. 99–158. DOI: https://doi.org/10.1016/S1874-6047(08)60196-1. URL: https://www.sciencedirect.com/science/article/pii/S1874604708601961.

Cooper, Geoffrey M. (2000). *The Cell: A Molecular Approach*. Sunderland (MA): Sinauer Associates. ISBN: 9780878931194. URL: https://www.ncbi.nlm.nih.gov/books/NBK9839/.

Crick, Francis (1970). "Central Dogma of Molecular Biology". In: *Nature* 227.5258, pp. 561–563. ISSN: 1476-4687. DOI: 10.1038/227561a0. URL: https://doi.org/10.1038/227561a0.

Dey, Krishna K. et al. (2015). "Micromotors Powered by Enzyme Catalysis". English (US). In: *Nano Letters* 15.12, pp. 8311–8315. ISSN: 1530-6984. DOI: 10.1021/acs.nanolett.5b03935.

Dror, Ron O. et al. (2012). "Biomolecular Simulation: A Computational Microscope for Molecular Biology". In: *Annual Review of Biophysics* 41.1. PMID: 22577825, pp. 429–452. DOI: 10.1146/annurev-biophys-042910-155245. eprint: https://doi.org/10.1146/annurev-biophys-042910-155245. URL: https://doi.org/10.1146/annurev-biophys-042910-155245.

Elenkov, Maja Majerić, Bernhard Hauer, and Dick B. Janssen (2006). "Enantioselective Ring Opening of Epoxides with Cyanide Catalysed by Halohydrin Dehalogenases: A New Approach to Non-Racemic β-Hydroxy Nitriles". In: *Advanced Synthesis & Catalysis* 348.4-5, pp. 579–585. DOI: https://doi.org/10.1002/adsc.200505333. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/adsc.200505333. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/adsc.200505333.

Fox, Richard J. et al. (2007). "Improving catalytic function by ProSAR-driven enzyme evolution". In: *Nature Biotechnology* 25.3, pp. 338–344. ISSN: 1546-1696. DOI: 10.1038/nbt1286. URL: https://doi.org/10.1038/nbt1286.

Frenkel, Daan and Berend Smit (2002). "Chapter 6 - Molecular Dynamics in Various Ensembles". In: *Understanding Molecular Simulation (Second Edition)*. Ed. by Daan Frenkel and Berend Smit. Second Edition. San Diego: Academic Press, pp. 139–163. ISBN: 978-0-12-267351-1. DOI: https://doi.org/10.1016/B978-012267351-1/50008-0. URL: https://www.sciencedirect.com/science/article/pii/B9780122673511500080.

Frisch, M. J. et al. (n.d.). *Gaussian~09 Revision D.01*. Gaussian Inc. Wallingford CT 2009.

Gao, Changyong et al. (2019). "Surface Wettability-Directed Propulsion of Glucose-Powered Nanoflask Motors". In: *ACS Nano* 13.11, pp. 12758–12766. ISSN: 1936-0851. DOI: 10.1021/acsnano.9b04708. URL: https://doi.org/10.1021/acsnano.9b04708.

Gao, Wei and Joseph Wang (2014). "The Environmental Impact of Micro/Nanomachines: A Review". In: *ACS Nano* 8.4, pp. 3170–3180. ISSN: 1936-0851. DOI: 10.1021/nn500077a. URL: https://doi.org/10.1021/nn500077a.

Gordon, John C. et al. (2005). " H++: a server for estimating p Ka s and adding missing hydrogens to macromolecules ". In: *Nucleic Acids Research* 33.suppl$_2$, W368–W371. ISSN: 0305-1048. DOI: 10.1093/nar/gki464. eprint: https://academic.oup.com/nar/article-pdf/33/suppl_2/W368/7623463/gki464.pdf. URL: https://doi.org/10.1093/nar/gki464.

Gunasekaran, K., Buyong Ma, and Ruth Nussinov (2004). "Is allostery an intrinsic property of all dynamic proteins?" In: *Proteins: Structure, Function, and Bioinformatics* 57.3, pp. 433–443. DOI: https://doi.org/10.1002/prot.20232. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.20232. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.20232.

Guo, Chao et al. (2015). "Exploring the enantioselective mechanism of halohydrin dehalogenase from Agrobacterium radiobacter AD1 by iterative saturation mutagenesis". en. In: *Appl Environ Microbiol* 81.8, pp. 2919–2926.

Harris, Charles R. et al. (2020). "Array programming with NumPy". In: *Nature* 585.7825, pp. 357–362. ISSN: 1476-4687. DOI: 10.1038/s41586-020-2649-2. URL: https://doi.org/10.1038/s41586-020-2649-2.

Hasnaoui-Dijoux, Ghannia et al. (2008). "Catalytic promiscuity of halohydrin dehalogenase and its application in enantioselective epoxide ring opening". en. In: *Chembiochem* 9.7, pp. 1048–1051.

Ho, Tin Kam (1995). “Random decision forests”. In: *Proceedings of 3rd International Conference on Document Analysis and Recognition*. Vol. 1, 278–282 vol.1. DOI: 10.1109/ICDAR.1995.598994.

— (1998). “The random subspace method for constructing decision forests”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20.8, pp. 832–844. DOI: 10.1109/34.709601.

Hu, LiHong et al. (2009). “Do Quantum Mechanical Energies Calculated for Small Models of Protein-Active Sites Converge?” In: *The Journal of Physical Chemistry A* 113.43, pp. 11793–11800. ISSN: 1089-5639. DOI: 10.1021/jp9029024. URL: https://doi.org/10.1021/jp9029024.

Hunter, J. D. (2007). “Matplotlib: A 2D graphics environment”. In: *Computing in Science & Engineering* 9.3, pp. 90–95. DOI: 10.1109/MCSE.2007.55.

Husic, Brooke E. and Vijay S. Pande (2018). “Markov State Models: From an Art to a Science”. In: *Journal of the American Chemical Society* 140.7, pp. 2386–2396. ISSN: 0002-7863. DOI: 10.1021/jacs.7b12191. URL: https://doi.org/10.1021/jacs.7b12191.

Huynh, Kathy and Carrie L Partch (2015). “Analysis of protein stability and ligand interactions by thermal shift assay”. en. In: *Curr Protoc Protein Sci* 79, pp. 28.9.1–28.9.14.

Hylckama Vlieg, Johan E. T. van et al. (2001). “Halohydrin Dehalogenases Are Structurally and Mechanistically Related to Short-Chain Dehydrogenases/Reductases”. In: *Journal of Bacteriology* 183.17, pp. 5058–5066. DOI: 10.1128/JB.183.17.5058-5066.2001. eprint: https://journals.asm.org/doi/pdf/10.1128/JB.183.17.5058-5066.2001. URL: https://journals.asm.org/doi/abs/10.1128/JB.183.17.5058-5066.2001.

Ji, Yuxing et al. (2019). “Macroscale Chemotaxis from a Swarm of Bacteria-Mimicking Nanoswimmers”. In: *Angewandte Chemie International Edition* 58.35, pp. 12200–12205. DOI: https://doi.org/10.1002/anie.201907733. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/anie.201907733. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/anie.201907733.

Jin, Xin and Jiawei Han (2010). “K-Means Clustering”. In: *Encyclopedia of Machine Learning*. Ed. by Claude Sammut and Geoffrey I. Webb. Boston, MA: Springer US, pp. 563–564. ISBN: 978-0-387-30164-8. DOI: 10.1007/978-0-387-30164-8_425. URL: https://doi.org/10.1007/978-0-387-30164-8_425.

Jochens, Helge and Uwe T. Bornscheuer (2010). “Natural Diversity to Guide Focused Directed Evolution”. In: *ChemBioChem* 11.13, pp. 1861–1866. DOI: https://doi.org/10.1002/cbic.201000284. eprint: https://chemistry-europe.onlinelibrary.wiley.com/doi/pdf/10.1002/cbic.201000284. URL: https://chemistry-europe.onlinelibrary.wiley.com/doi/abs/10.1002/cbic.201000284.

Johnson, Kenneth A. and Roger S. Goody (2011). “The Original Michaelis Constant: Translation of the 1913 Michaelis–Menten Paper”. In: *Biochemistry* 50.39, pp. 8264–8269. ISSN: 0006-2960. DOI: 10.1021/bi201284u. URL: https://doi.org/10.1021/bi201284u.

Jolliffe, Ian T. and Jorge Cadima (2016). “Principal component analysis: a review and recent developments”. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374.2065, p. 20150202. DOI: 10.1098/rsta.2015.0202. eprint: https://royalsocietypublishing.org/doi/pdf/10.1098/rsta.2015.0202. URL: https://royalsocietypublishing.org/doi/abs/10.1098/rsta.2015.0202.

Jong, R. M. de et al. (2003). “Structure and mechanism of a bacterial haloalcohol dehalogenase: a new variation of the short-chain dehydrogenase/reductase fold without an NAD(P)H binding site”. In: *The EMBO Journal* 22.19, pp. 4933–4944. DOI: https://doi.org/10.1093/emboj/cdg479. eprint: https://www.embopress.org/doi/pdf/10.1093/emboj/cdg479. URL: https://www.embopress.org/doi/abs/10.1093/emboj/cdg479.

Jong, René M. de et al. (2005). “Structural Basis for the Enantioselectivity of an Epoxide Ring Opening Reaction Catalyzed by Halo Alcohol Dehalogenase HheC”. In: *Journal of the American Chemical Society* 127.38, pp. 13338–13343. ISSN: 0002-7863. DOI: 10.1021/ja0531733. URL: https://doi.org/10.1021/ja0531733.

Jong, René M. de et al. (2006). “The X-Ray Structure of the Haloalcohol Dehalogenase HheA from <i>Arthrobacter</i> sp. Strain AD2: Insight into Enantioselectivity and Halide Binding in the Haloalcohol Dehalogenase Family". In: *Journal of Bacteriology* 188.11, pp. 4051–4056. DOI: 10.1128/JB.01866-05. eprint: https://journals.asm.org/doi/pdf/10.1128/JB.01866-05. URL: https://journals.asm.org/doi/abs/10.1128/JB.01866-05.

Jorgensen, William L. et al. (1983). "Comparison of simple potential functions for simulating liquid water". In: *The Journal of Chemical Physics* 79.2, pp. 926–935. ISSN: 0021-9606. DOI: 10.1063/1.445869. eprint: https://pubs.aip.org/aip/jcp/article-pdf/79/2/926/6427020/926_1_online.pdf. URL: https://doi.org/10.1063/1.445869.

Jumper, John et al. (2021). "Highly accurate protein structure prediction with AlphaFold". In: *Nature* 596.7873, pp. 583–589. ISSN: 1476-4687. DOI: 10.1038/s41586-021-03819-2. URL: https://doi.org/10.1038/s41586-021-03819-2.

Jurcik, Adam et al. (2018). "CAVER Analyst 2.0: analysis and visualization of channels and tunnels in protein structures and molecular dynamics trajectories". en. In: *Bioinformatics* 34.20, pp. 3586–3588.

Khan, Faez Iqbal et al. (2017). "The Lid Domain in Lipases: Structural and Functional Determinant of Enzymatic Properties". In: *Frontiers in Bioengineering and Biotechnology* 5. ISSN: 2296-4185. DOI: 10.3389/fbioe.2017.00016. URL: https://www.frontiersin.org/articles/10.3389/fbioe.2017.00016.

Kirby, Anthony J. (1996). "Enzyme Mechanisms, Models, and Mimics". In: *Angewandte Chemie International Edition in English* 35.7, pp. 706–724. DOI: https://doi.org/10.1002/anie.199607061. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/anie.199607061. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/anie.199607061.

"Kolmogorov–Smirnov Test" (2008). In: *The Concise Encyclopedia of Statistics*. New York, NY: Springer New York, pp. 283–287. ISBN: 978-0-387-32833-1. DOI: 10.1007/978-0-387-32833-1_214. URL: https://doi.org/10.1007/978-0-387-32833-1_214.

Koopmeiners, Julia et al. (2016). "Biochemical and biocatalytic characterization of 17 novel halohydrin dehalogenases". en. In: *Appl Microbiol Biotechnol* 100.17, pp. 7517–7527.

Koopmeiners, Julia et al. (2017). "HheG, a Halohydrin Dehalogenase with Activity on Cyclic Epoxides". In: *ACS Catalysis* 7.10, pp. 6877–6886. DOI: 10.1021/acscatal.7b01854. URL: https://doi.org/10.1021/acscatal.7b01854.

Krissinel, Evgeny and Kim Henrick (2007). "Inference of macromolecular assemblies from crystalline state". en. In: *J Mol Biol* 372.3, pp. 774–797.

Lassila, Jonathan Kyle et al. (2006). "Combinatorial methods for small-molecule placement in computational enzyme design". In: *Proceedings of the National Academy of Sciences* 103.45, pp. 16710–16715. DOI: 10.1073/pnas.0607691103. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.0607691103. URL: https://www.pnas.org/doi/abs/10.1073/pnas.0607691103.

Leveson-Gower, Reuben B., Clemens Mayer, and Gerard Roelfes (2019). "The importance of catalytic promiscuity for enzyme design and evolution". In: *Nature Reviews Chemistry* 3.12, pp. 687–705. ISSN: 2397-3358. DOI: 10.1038/s41570-019-0143-x. URL: https://doi.org/10.1038/s41570-019-0143-x.

Lindorff-Larsen, Kresten et al. (June 2010). "Improved side-chain torsion potentials for the Amber ff99SB protein force field". en. In: *Proteins* 78.8, pp. 1950–1958.

Luo, Ming et al. (2018). "Micro-/Nanorobots at Work in Active Drug Delivery". In: *Advanced Functional Materials* 28.25, p. 1706100. DOI: https://doi.org/10.1002/adfm.201706100. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/adfm.201706100. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/adfm.201706100.

Ma, Steven K. et al. (2010). "A green-by-design biocatalytic process for atorvastatin intermediate". In: *Green Chem.* 12 (1), pp. 81–86. DOI: 10.1039/B919115C. URL: http://dx.doi.org/10.1039/B919115C.

Ma, Xing et al. (2015). "Enzyme-Powered Hollow Mesoporous Janus Nanomotors". In: *Nano Letters* 15.10, pp. 7043–7050. ISSN: 1530-6984.

DOI: 10.1021/acs.nanolett.5b03100. URL: https://doi.org/10.1021/acs.nanolett.5b03100.

Maaten, Laurens van der and Geoffrey Hinton (2008). "Visualizing Data using t-SNE". In: *Journal of Machine Learning Research* 9.86, pp. 2579–2605. URL: http://jmlr.org/papers/v9/vandermaaten08a.html.

MacKerell Jr., A. D. et al. (1998). "All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins". In: *The Journal of Physical Chemistry B* 102.18, pp. 3586–3616. ISSN: 1520-6106. DOI: 10.1021/jp973084f. URL: https://doi.org/10.1021/jp973084f.

Maier, James A. et al. (2015). "ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB". In: *Journal of Chemical Theory and Computation* 11.8, pp. 3696–3713. ISSN: 1549-9618. DOI: 10.1021/acs.jctc.5b00255. URL: https://doi.org/10.1021/acs.jctc.5b00255.

María, Pablo Domínguez de et al. (2005). "Understanding Candida rugosa lipases: an overview". en. In: *Biotechnol Adv* 24.2, pp. 180–196.

Martínez, L. et al. (2009). "PACKMOL: A package for building initial configurations for molecular dynamics simulations". In: *Journal of Computational Chemistry* 30.13, pp. 2157–2164. DOI: https://doi.org/10.1002/jcc.21224. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.21224. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.21224.

McGibbon, Robert T. et al. (2015). "MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories". In: *Biophysical Journal* 109.8, pp. 1528 –1532. DOI: 10.1016/j.bpj.2015.08.015.

McInnes, Leland, John Healy, and Steve Astels (2017). "hdbscan: Hierarchical density based clustering". In: *Journal of Open Source Software* 2.11, p. 205. DOI: 10.21105/joss.00205. URL: https://doi.org/10.21105/joss.00205.

Molgedey, L. and H. G. Schuster (1994). "Separation of a mixture of independent signals using time delayed correlations". In: *Phys. Rev. Lett.* 72 (23), pp. 3634–3637. DOI: 10.1103/PhysRevLett.72.3634. URL: https://link.aps.org/doi/10.1103/PhysRevLett.72.3634.

Motlagh, Hesam N et al. (2014). “The ensemble nature of allostery”. en. In: *Nature* 508.7496, pp. 331–339.

Nagasawa, Toru et al. (1992). “Purification and characterization of halohydrin hydrogen-halide lyase from a recombinant Escherichia coli containing the gene from a Corynebacterium sp.” In: *Applied Microbiology and Biotechnology* 36.4, pp. 478–482. ISSN: 1432-0614. DOI: 10.1007/BF00170187. URL: https://doi.org/10.1007/BF00170187.

Najib, Fadhil M and Omed I Hayder (2011). “Study of Stoichiometry of Ferric Thiocyanate Complex for Analytical Purposes Including F Determination”. In: *Iraqi National Journal of Chemistry* 42, pp. 135–155.

Nguyen, Hai et al. (Jan. 2016). *PYTRAJ v1.0.0.dev1: Interactive data analysis for molecular dynamics simulations*. Version v1.0.0.dev1. DOI: 10.5281/zenodo.44612. URL: https://doi.org/10.5281/zenodo.44612.

Olsson, Mats H. M. et al. (2011). “PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions”. In: *Journal of Chemical Theory and Computation* 7.2, pp. 525–537. ISSN: 1549-9618. DOI: 10.1021/ct100578z. URL: https://doi.org/10.1021/ct100578z.

Orosz, Ferenc and Beáta G. Vértessy (2021). “What’s in a name? From “fluctuation fit” to “conformational selection”: rediscovery of a concept”. In: *History and Philosophy of the Life Sciences* 43.3, p. 88. ISSN: 1742-6316. DOI: 10.1007/s40656-021-00442-2. URL: https://doi.org/10.1007/s40656-021-00442-2.

Osuna, Sílvia (2021). “The challenge of predicting distal active site mutations in computational enzyme design”. In: *WIREs Computational Molecular Science* 11.3, e1502. DOI: https://doi.org/10.1002/wcms.1502. eprint: https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/wcms.1502. URL: https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wcms.1502.

pandas_dev_team (Feb. 2020). *pandas-dev/pandas: Pandas*. Version latest. DOI: 10.5281/zenodo.3509134. URL: https://doi.org/10.5281/zenodo.3509134.

Pavelka, A. et al. (2015). "CAVER: Algorithms for Analyzing Dynamics of Tunnels in Macromolecules". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 13.3, pp. 505–517. ISSN: 1545-5963. DOI: 10.1109/TCBB.2015.2459680.

Pedregosa, Fabian et al. (2011). "Scikit-learn: Machine learning in Python". In: *the Journal of machine Learning research* 12, pp. 2825–2830.

Pettersen, Eric F. et al. (2021). "UCSF ChimeraX: Structure visualization for researchers, educators, and developers". In: *Protein Science* 30.1, pp. 70–82. DOI: https://doi.org/10.1002/pro.3943. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pro.3943. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/pro.3943.

Ponder, Jay W. and David A. Case (2003a). "Force Fields for Protein Simulations". In: *Protein Simulations*. Vol. 66. Advances in Protein Chemistry. Academic Press, pp. 27–85. DOI: https://doi.org/10.1016/S0065-3233(03)66002-X. URL: https://www.sciencedirect.com/science/article/pii/S006532330366002X.

— (2003b). "Force Fields for Protein Simulations". In: *Protein Simulations*. Vol. 66. Advances in Protein Chemistry. Academic Press, pp. 27–85. DOI: https://doi.org/10.1016/S0065-3233(03)66002-X. URL: https://www.sciencedirect.com/science/article/pii/S006532330366002X.

Rassolov, Vitaly A. et al. (2001). "6-31G* basis set for third-row atoms". In: *Journal of Computational Chemistry* 22.9, pp. 976–984. DOI: https://doi.org/10.1002/jcc.1058. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.1058. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.1058.

Rindfleisch, Sören et al. (2022). "Ground-state destabilization by electrostatic repulsion is not a driving force in orotidine-5-monophosphate decarboxylase catalysis". In: *Nature Catalysis* 5.4, pp. 332–341. ISSN: 2520-1158. DOI: 10.1038/s41929-022-00771-w. URL: https://doi.org/10.1038/s41929-022-00771-w.

Roe, Daniel R. and Thomas E. Cheatham III (2013). "PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data". In: *Journal of Chemical Theory and Computation* 9.7, pp. 3084–

3095. ISSN: 1549-9618. DOI: 10.1021/ct400341p. URL: https://doi.org/10.1021/ct400341p.

Rohl, Carol A. et al. (2004). "Protein Structure Prediction Using Rosetta". In: *Numerical Computer Methods, Part D*. Vol. 383. Methods in Enzymology. Academic Press, pp. 66–93. DOI: https://doi.org/10.1016/S0076-6879(04)83004-0. URL: https://www.sciencedirect.com/science/article/pii/S0076687904830040.

Romero-Rivera, Adrian, Marc Garcia-Borràs, and Sílvia Osuna (2017a). "Role of Conformational Dynamics in the Evolution of Retro-Aldolase Activity". en. In: *ACS Catal* 7.12, pp. 8524–8532.

— (2017b). "Role of Conformational Dynamics in the Evolution of Retro-Aldolase Activity". en. In: *ACS Catal* 7.12, pp. 8524–8532.

Romero-Rivera, Adrian, Marc Garcia-Borràs, and Sílvia Osuna (2017). "Computational tools for the evaluation of laboratory-engineered biocatalysts". In: *Chem. Commun.* 53 (2), pp. 284–297. DOI: 10.1039/C6CC06055B. URL: http://dx.doi.org/10.1039/C6CC06055B.

Rosano, Germán L. and Eduardo A. Ceccarelli (2014). "Recombinant protein expression in Escherichia coli: advances and challenges". In: *Frontiers in Microbiology* 5. ISSN: 1664-302X. DOI: 10.3389/fmicb.2014.00172. URL: https://www.frontiersin.org/articles/10.3389/fmicb.2014.00172.

Salomon-Ferrer, Romelia et al. (2013). "Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald". In: *Journal of Chemical Theory and Computation* 9.9, pp. 3878–3888. ISSN: 1549-9618. DOI: 10.1021/ct400314y. URL: https://doi.org/10.1021/ct400314y.

Schallmey, Anett and Marcus Schallmey (2016). "Recent advances on halohydrin dehalogenases-from enzyme identification to novel biocatalytic applications". en. In: *Appl Microbiol Biotechnol* 100.18, pp. 7827–7839.

Schallmey, Marcus et al. (2013). "Biocatalytic and Structural Properties of a Highly Engineered Halohydrin Dehalogenase". In: *ChemBioChem* 14.7, pp. 870–881. DOI: https://doi.org/10.1002/cbic.201300005. eprint: https://chemistry-europe.onlinelibrary.wiley.com/doi/

pdf/10.1002/cbic.201300005. URL: https://chemistry-europe.onlinelibrary.wiley.com/doi/abs/10.1002/cbic.201300005.

Schallmey, Marcus et al. (2014). "Expanding the Halohydrin Dehalogenase Enzyme Family: Identification of Novel Enzymes by Database Mining". In: *Applied and Environmental Microbiology* 80.23, pp. 7303–7315. DOI: 10.1128/AEM.01985-14. eprint: https://journals.asm.org/doi/pdf/10.1128/AEM.01985-14. URL: https://journals.asm.org/doi/abs/10.1128/AEM.01985-14.

Schallmey, Marcus et al. (2015). "A single point mutation enhances hydroxynitrile synthesis by halohydrin dehalogenase". In: *Enzyme and Microbial Technology* 70, pp. 50–57. ISSN: 0141-0229. DOI: https://doi.org/10.1016/j.enzmictec.2014.12.009. URL: https://www.sciencedirect.com/science/article/pii/S0141022914002191.

Schattling, Philipp S et al. (2017). "Double-Fueled Janus Swimmers with Magnetotactic Behavior". en. In: *ACS Nano* 11.4, pp. 3973–3983.

Scherer, Martin K. et al. (2015). "PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models". In: *Journal of Chemical Theory and Computation* 11.11, pp. 5525–5542. ISSN: 1549-9618. DOI: 10.1021/acs.jctc.5b00743. URL: https://doi.org/10.1021/acs.jctc.5b00743.

Schrödinger, LLC (2015a). "The AxPyMOL Molecular Graphics Plugin for Microsoft PowerPoint, Version 1.8".

— (2015b). "The JyMOL Molecular Graphics Development Component, Version 1.8".

— (2015c). "The PyMOL Molecular Graphics System, Version 1.8".

Schymkowitz, Joost et al. (2005). "The FoldX web server: an online force field". In: *Nucleic Acids Research* 33.suppl$_2$, W382–W388. ISSN: 0305-1048. DOI: 10.1093/nar/gki387. eprint: https://academic.oup.com/nar/article-pdf/33/suppl_2/W382/7622711/gki387.pdf. URL: https://doi.org/10.1093/nar/gki387.

Seaton, M. J. (1977). "Hartree–Fock method". In: *Nature* 269.5629, pp. 631–631. ISSN: 1476-4687. DOI: 10.1038/269631a0. URL: https://doi.org/10.1038/269631a0.

Segel, I.H. (2013). “Enzyme Kinetics”. In: *Encyclopedia of Biological Chemistry (Second Edition)*. Ed. by William J. Lennarz and M. Daniel Lane. Second Edition. Waltham: Academic Press, pp. 216–220. ISBN: 978-0-12-378631-9. DOI: https://doi.org/10.1016/B978-0-12-378630-2.00012-8. URL: https://www.sciencedirect.com/science/article/pii/B9780123786302000128.

Semisotnov, G. V. et al. (1991). “Study of the “molten globule” intermediate state in protein folding by a hydrophobic fluorescent probe”. In: *Biopolymers* 31.1, pp. 119–128. DOI: https://doi.org/10.1002/bip.360310111. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bip.360310111. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/bip.360310111.

Shareefdeen, Z. (2022). *Hazardous Waste Management: Advances in Chemical and Industrial Waste Treatment and Technologies*. Springer International Publishing. ISBN: 9783030952624. URL: https://books.google.es/books?id=uXtsEAAAQBAJ.

Slater, C. Stewart et al. (2010). “Solvent Use and Waste Issues”. In: *Green Chemistry in the Pharmaceutical Industry*. John Wiley & Sons, Ltd. Chap. 3, pp. 49–82. ISBN: 9783527629688. DOI: https://doi.org/10.1002/9783527629688.ch3. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9783527629688.ch3. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9783527629688.ch3.

Srinivasan, Bharath (2021). “Explicit Treatment of Non-Michaelis-Menten and Atypical Kinetics in Early Drug Discovery**”. In: *ChemMedChem* 16.6, pp. 899–918. DOI: https://doi.org/10.1002/cmdc.202000791. eprint: https://chemistry-europe.onlinelibrary.wiley.com/doi/pdf/10.1002/cmdc.202000791. URL: https://chemistry-europe.onlinelibrary.wiley.com/doi/abs/10.1002/cmdc.202000791.

Susskind, L. and G. Hrabovsky (2014). *Classical Mechanics: The Theoretical Minimum*. Theoretical minimum. Penguin Books. ISBN: 9780141976228. URL: https://books.google.es/books?id=-WOCngEACAAJ.

Szefczyk, Borys et al. (2004). “Differential Transition-State Stabilization in Enzyme Catalysis: Quantum Chemical Analysis of Interactions in the Chorismate Mutase Reaction and Prediction of the Optimal Catalytic

Field". In: *Journal of the American Chemical Society* 126.49, pp. 16148–16159. ISSN: 0002-7863. DOI: 10.1021/ja049376t. URL: https://doi.org/10.1021/ja049376t.

Tang, Lixia et al. (2002). "Improved stability of halohydrin dehalogenase from Agrobacterium radiobacter AD1 by replacement of cysteine residues". In: *Enzyme and Microbial Technology* 30.2, pp. 251–258. ISSN: 0141-0229. DOI: https://doi.org/10.1016/S0141-0229(01)00488-4. URL: https://www.sciencedirect.com/science/article/pii/S0141022901004884.

Tang, Lixia et al. (2003). "Steady-State Kinetics and Tryptophan Fluorescence Properties of Halohydrin Dehalogenase from Agrobacterium radiobacter. Roles of W139 and W249 in the Active Site and Halide-Induced Conformational Change". In: *Biochemistry* 42.47, pp. 14057–14065. ISSN: 0006-2960. DOI: 10.1021/bi034941a. URL: https://doi.org/10.1021/bi034941a.

Vaissier Welborn, Valerie and Teresa Head-Gordon (2019). "Computational Design of Synthetic Enzymes". In: *Chemical Reviews* 119.11, pp. 6613–6630. ISSN: 0009-2665. DOI: 10.1021/acs.chemrev.8b00399. URL: https://doi.org/10.1021/acs.chemrev.8b00399.

"Chapter 2 Catalytic processes in industry" (1999). In: *Catalysis: An Integrated Approach*. Ed. by R.A. van Santen et al. Vol. 123. Studies in Surface Science and Catalysis. Elsevier, pp. 29–80. DOI: https://doi.org/10.1016/S0167-2991(99)80005-6. URL: https://www.sciencedirect.com/science/article/pii/S0167299199800056.

Vavra, Ondrej et al. (2019). "CaverDock: a molecular docking-based tool to analyse ligand transport through protein tunnels and channels". en. In: *Bioinformatics* 35.23, pp. 4986–4993.

Wang, Jiajia et al. (2019a). "A Review on Artificial Micro/Nanomotors for Cancer-Targeted Delivery, Diagnosis, and Therapy". In: *Nano-Micro Letters* 12.1, p. 11. ISSN: 2150-5551. DOI: 10.1007/s40820-019-0350-5. URL: https://doi.org/10.1007/s40820-019-0350-5.

Wang, Junmei et al. (2004). "Development and testing of a general amber force field". In: *Journal of Computational Chemistry* 25.9, pp. 1157–1174. DOI: https://doi.org/10.1002/jcc.20035. eprint: https:

//onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.20035. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.20035.

Wang, Lei et al. (2019b). "Lipase-Powered Mesoporous Silica Nanomotors for Triglyceride Degradation". In: *Angewandte Chemie International Edition* 58.24, pp. 7992–7996. DOI: https://doi.org/10.1002/anie.201900697. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/anie.201900697. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/anie.201900697.

Wang, Yajie et al. (2021). "Directed Evolution: Methodologies and Applications". In: *Chemical Reviews* 121.20, pp. 12384–12444. ISSN: 0009-2665. DOI: 10.1021/acs.chemrev.1c00260. URL: https://doi.org/10.1021/acs.chemrev.1c00260.

Watanabe, Fumiaki et al. (2015). "Crystal structures of halohydrin hydrogen-halide-lyases from Corynebacterium sp. N-1074". In: *Proteins: Structure, Function, and Bioinformatics* 83.12, pp. 2230–2239. DOI: https://doi.org/10.1002/prot.24938. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.24938. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.24938.

Weiss, James N. (1997). "The Hill equation revisited: uses and misuses". In: *The FASEB Journal* 11.11, pp. 835–841. DOI: https://doi.org/10.1096/fasebj.11.11.9285481. eprint: https://faseb.onlinelibrary.wiley.com/doi/pdf/10.1096/fasebj.11.11.9285481. URL: https://faseb.onlinelibrary.wiley.com/doi/abs/10.1096/fasebj.11.11.9285481.

Wessel, Julia et al. (2021). "Insights into the molecular determinants of thermal stability in halohydrin dehalogenase HheD2". In: *The FEBS Journal* 288.15, pp. 4683–4701. DOI: https://doi.org/10.1111/febs.15777.

"What is Independent Component Analysis?" (2001). In: *Independent Component Analysis*. John Wiley & Sons, Ltd. Chap. 7, pp. 145–164. ISBN: 9780471221319. DOI: https://doi.org/10.1002/0471221317.ch7. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/0471221317.ch7. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/0471221317.ch7.

Wijngaard, A J van den, P T Reuvekamp, and D B Janssen (1991). "Purification and characterization of haloalcohol dehalogenase from Arthrobacter sp. strain AD2". In: *Journal of Bacteriology* 173.1, pp. 124–129. DOI: 10.1128/jb.173.1.124-129.1991. eprint: https://journals.asm.org/doi/pdf/10.1128/jb.173.1.124-129.1991. URL: https://journals.asm.org/doi/abs/10.1128/jb.173.1.124-129.1991.

Wingfield, Paul T. (2015). "Overview of the Purification of Recombinant Proteins". In: *Current Protocols in Protein Science* 80.1, pp. 6.1.1–6.1.35. DOI: https://doi.org/10.1002/0471140864.ps0601s80. eprint: https://currentprotocols.onlinelibrary.wiley.com/doi/pdf/10.1002/0471140864.ps0601s80. URL: https://currentprotocols.onlinelibrary.wiley.com/doi/abs/10.1002/0471140864.ps0601s80.

Zanghellini, Alexandre et al. (2006). "New algorithms and an in silico benchmark for computational enzyme design". In: *Protein Science* 15.12, pp. 2785–2794. DOI: https://doi.org/10.1110/ps.062353106. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1110/ps.062353106. URL: https://onlinelibrary.wiley.com/doi/abs/10.1110/ps.062353106.