Metallopeptides and metalloenzymes in biocatalysis: computational insights for rational design

Laura Martínez Castro

DOCTORAL THESIS

Metallopeptides and metalloenzymes in biocatalysis: computational insights for rational design

Laura Martínez Castro

2025

Doctoral program in Chemistry

Department of Chemistry

Universitat Autònoma de Barcelona

Directors:

Dr. Jean-Didier Pierre Maréchal

Dr. M. Eugenio Vázquez Sentís

A mis abuelas y abuelos

"The more you know, the more you realize you know nothing."

Socrates

"Journey before destination."

The Stormlight Archive

Acknowledgements

Quiero agradecer en primer lugar a mis directores Eugenio y JeanDi, por su confianza, por su ayuda, por su guía y por haberme dado la oportunidad de crecer tanto profesional como personalmente. Gracias por facilitar la conjunción teórico-experimental y dejarme disfrutar de los dos mundos. También a los demás profesores que han compartido su tiempo y conocimiento conmigo, especialmente a Gregori Ujaque, Laura Masgrau, Agustí Lledós y José Luis Mascareñas, ha sido un auténtico privilegio trabajar con vosotros. Quiero mencionar especialmente a Giuseppe Sciortino, que empezó siendo mi doctorando de referencia y ahora ya es profesor, guía y, si me lo permite, amigo. No puedo olvidarme de Elena, sin ti este doctorado no sería viable y tu amabilidad y trato personal han sido un salvavidas.

El día a día no lo podría haber sobrellevado de no ser por mis compañeros. Mercè, las aventuras griegas no habrían sido lo mismo sin ti, qué suerte aprender sobre tantas cosas juntas (no científicas sobre todo). Xavi, aunque me dieses miedito al principio no puedo alegrarme más de la evolución de nuestra relación hasta el punto de que conozcas mi casa y que mi madre me pregunte por ti. Laura Tiessler, me sigue sorprendiendo tu amabilidad y generosidad, eres una fuente de inspiración, aunque a ti no te lo parezca. Lorena, no puedo estar más agradecida de haber empezado esto juntas, de contar con tu inestimable ayuda siempre, de apoyarnos mutuamente en nuestra intensidad y de saber que, aunque no hubiéramos llorado juntas en el baño, habríamos acabado siendo amigas igualmente. Jose, te echamos de menos por el despacho, gracias por estar siempre disponible para echar una mano. Felipe, gracias por hacerme un huequito en ese corazón tan grande que tienes y compartir tu perspectiva de la vida. Dídac, qué suerte que nos cambiáramos de despacho y hayamos podido compartir tantos desayunos y charlas mañaneras más o menos filosóficas. Sonia, gracias por llegar y hacer que el despacho se sintiese un poco más hogar. Fernanda, trabajar contigo ha sido un soplo de aire fresco, ¡qué bien que te quedes! También quiero agradecer al comando comida, especialmente Álex, Pedro y Bernat, por tener conversaciones tan distendidas y siempre sacarnos una sonrisa. Iker, Ming, Kessen, Fede, Gabriella, Reuben, Lea, Matteo, Lluís y todos los que habéis pasado aunque sea unos días en el despacho habéis dejado huella y ha sido un placer cruzar caminos.

En Santiago he tenido la suerte de ponerme la bata y aprender de los mejores. David (padre), mi experiencia investigadora empezó contigo y ha sido un privilegio continuarla tantos años. Soraya, qué divertida, inteligente e inspiradora eres, me siento muy afortunada de haber aprendido contigo. Carmen, gracias por tu paciencia y amabilidad, me ha encantado trabajar a tu lado. Alberto, gracias también por tu paciencia y complicidad que me han hecho sentir un poco menos patosa. Ana, Axel, Patri, Diego, David, Jacobo, Andrés y todo el P3L5, gracias por haberme hecho sentir en casa cada vez que paso por vuestro laboratorio. Y también agradecer a la planta 3 del CiQUS que me hayáis acogido siempre como una más, sois una familia preciosa y extremadamente divertida.

Quiero agradecer también a María Macías y a todo su laboratorio por haberme tratado tan bien los meses que estuve aprendiendo con ellos y el trato que recibo siempre que voy, es un placer a nivel profesional y personal.

One of the best parts of this journey has been the stay at the ITE-FORTH institute in Crete. Chara, you are more than inspirational and cannot wait to meet again and follow your splendid career. Giorgos, I deeply thank you for your enthusiastic discussions on not only scientific matters. Savvas, you are truly kind, hardworking and fun, thank you for making us feel at home. To Sofia, Pozidis, Mary, Heleni and everyone in the lab, thank you for being so kind and welcoming, it was a pleasure learning from you all.

I have been extremely lucky to work in close collaboration with Prof. Stephen Bell, whom I would like to thank for the trust he has placed in me for our projects together.

Además, me gustaría agradecer a Antonio Martínez por valorar tanto el trabajo que hemos hecho para la parte computacional de nuestra colaboración.

Todo este camino lo he podido iniciar gracias a haber crecido muy bien rodeada. Alba, esta tesis, literalmente, no habría sido posible sin ti. Gracias por estar y dejarme estar siempre. A mis chuliñas, sois casa sin importar la distancia y el tiempo. Che, Jorge, gracias por darle perspectiva a todo, que nada es tan serio en esta vida. Laura, qué bien habernos encontrado y que seamos las mejores, noto tu apoyo incondicional incluso a kilómetros de distancia. A mis químicos de Santiago, que es un orgullo verlos avanzar a todos, cada uno a su manera: Sergio, Noa, Loli, Dani, Gonzalo, Álvaro, Marcos, Gemma. Santi, ya lo celebraremos. A mis compis de piso a lo largo del tiempo, Nacho, Sara, Soraya, qué suerte que acabásemos coincidiendo. Mis chicos y chicas de Monto, habéis sido un pilar fundamental y sin duda vuestras tardes como policías o distrayéndome han sido imprescindibles en esta tesis, gracias por ser familia. Jan, Rosa, Ferran, Angela, sois modelos a seguir y me siento muy afortunada de que estemos compartiendo este camino. Patxi, nin 10000 km fan que che sinta distante, gracias. Cattani and, especially, Semadeni family, you most certainly inspired me when I was only a teenager, but your unconditional love continues to overwhelm me even after so many years and miles of distance.

Me reservo lo más importante para el final, la familia. Xavi, gracias por tu paciencia, por escoger ser mi compañero en las buenas y en las malas y por ser siempre el refugio donde puedo sentirme yo. Gracias también a la familia Herrero Barbero, especialmente a mis *suegris* y *cuñis*, vuestro cariño y apoyo incondicionales han sido un motor imprescindible para esta etapa. Quero mencionar á familia Vázquez Martínez, que están na niña vida coidándome dende que teño memoria, e en especial a Lucas e Julia, a miña fillola, por compartir ratos tan divertidos comigo. Tamén á miña cuñada Alba, que trae unha tranquilidade especial á nosa familia. Y por supuesto a mi familia Martínez Castro, con orgullo puedo decir que soy como soy gracias a ellos, especialmente en cuanto a emociones intensas provenientes de la rama Castro. Gracias a Marisa e meu padriño, quen sementou o xermen da ciencia e a informática fai moito tempo, e que entre os dous o foron regando co seu apoio. Abuela Fina, abuela Puri, abuelo Pepe, abuelo Herminio, este esforzo vai por vós. Mamá, papá, Pablo, gracias infinitas polo amor incondicional, por ensinarme qué é o verdadeiramente importante e por facerme como son.

Creo realmente que somos pedacitos de la gente que nos vamos encontrando, y soy muy afortunada de poder contar con piezas tan valiosas. Muchas gracias de corazón.

Abstract

This thesis explores the relationship between biomolecular structure and catalytic function in the context of metal-mediated biocatalysis, combining multi-scale computational modelling with experimental validation. The overarching goal is to develop predictive tools and mechanistic insights that support the rational design of both artificial metallopeptides and natural metalloenzymes. A synergistic integration of theoretical modelling and experimental testing enables efficient iteration in the development and refinement of design methodologies.

Two complementary research directions are pursued: the *de novo* design of bioorthogonal metallopeptides capable of catalyzing abiotic reactions in biological environments, and the in-depth mechanistic investigation of heme-containing enzymes that mediate natural oxidative transformations. Specifically, two β-sheet-based peptide scaffolds were successfully designed to coordinate metal complexes and catalyze bioorthogonal depropargylation reactions both *in vitro* and *in cellulo*. The importance of conformational dynamics in enzymatic function is highlighted by detailed studies addressing substrate specificity, product release, and catalytic cycle regeneration.

These rational design approaches are further extended to complex biomolecular systems, including DNA–metallopeptide assemblies, contributing to the development of computational frameworks for effective metal interaction modelling. Overall, this work advances the field of computational biocatalyst design and demonstrates the potential of multiscale strategies to address fundamental challenges in metallobiochemistry.

Resum

Aquesta tesi explora la relació entre l'estructura biomolecular i la funció catalítica en el context de la biocatàlisi mediada per metalls, combinant modelatge computacional multiescala amb validació experimental. L’objectiu general és desenvolupar eines predictives i coneixements mecanístics que donin suport al disseny racional tant de metal·lopèptids artificials com de metalloenzims naturals. La integració sinèrgica del modelatge teòric amb l’experimentació permet una iteració eficient en el desenvolupament i perfeccionament de metodologies de disseny.

S’aborden dues línies de recerca complementàries: el disseny *de novo* de metal·lopèptids bioortogonals capaços de catalitzar reaccions abiòtiques en entorns biològics, i l’estudi mecanístic en profunditat d’enzims que contenen hemo i que duen a terme transformacions oxidatives naturals. En concret, es van dissenyar amb èxit dos esquelets peptídics basats en estructures de làmines β per coordinar complexos metàl·lics i catalitzar reaccions bioortogonals de despropargilació tan *in vitro* com *in cellulo*. La importància de la dinàmica conformacional en la funció enzimàtica es destaca mitjançant estudis detallats que aborden l’especificitat del substrat, l’alliberament del producte i la regeneració del cicle catalític.

Aquestes estratègies de disseny racional també s’estenen a sistemes biomoleculars complexos, incloent-hi conjunts ADN–metal·lopèptid, contribuint al desenvolupament de marcs computacionals efectius per al modelatge d’interaccions metàl·liques. En conjunt, aquest treball fa avançar el camp del disseny computacional de biocatalitzadors i demostra el potencial dels enfocaments multiescala per afrontar reptes fonamentals en la metal·lobioquímica.

Resumen

Esta tesis explora la relación entre la estructura biomolecular y la función catalítica en el contexto de la biocatálisis mediada por metales, combinando el modelado computacional multiescala con la validación experimental. El objetivo general es desarrollar herramientas predictivas y profundizar en conocimientos mecanísticos que respalden el diseño racional tanto de metalopéptidos artificiales como de metaloenzimas naturales. La integración de los modelos teórico con la experimentación permite una iteración eficiente en el desarrollo y perfeccionamiento de metodologías de diseño.

Se abordan dos líneas de investigación principales: el diseño *de novo* de metalopéptidos bioortogonales capaces de catalizar reacciones en entornos biológicos, y el estudio mecanístico de enzimas dependientes del grupo hemo y que llevan a cabo transformaciones oxidativas naturales. En concreto, se diseñaron con éxito dos estructuras peptídicas basadas en láminas β para coordinar complejos metálicos y catalizar reacciones bioortogonales de depropargilación tanto *in vitro* como *in cellulo*. Se profundiza en la importancia de la dinámica conformacional en la función enzimática mediante estudios detallados sobre la especificidad del sustrato, la liberación del producto y la regeneración del ciclo catalítico.

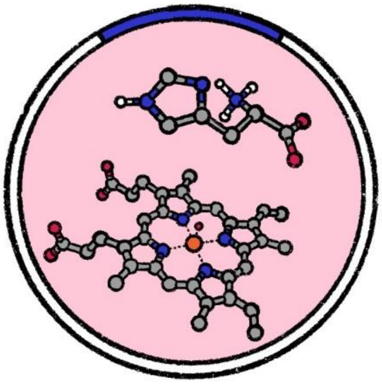
Estas estrategias de diseño racional se extienden además a sistemas biomoleculares complejos, incluidos los ensamblajes ADN–metalopéptido, contribuyendo al desarrollo de marcos computacionales eficaces para el modelado de interacciones metálicas. En conjunto, este trabajo impulsa el campo del diseño computacional de biocatalizadores y demuestra el potencial de los enfoques multiescala para abordar desafíos fundamentales en la metalobioquímica.

List of Abbreviations

meta-substrate	3-methoxybenzoic acid
para-product	4-hydroxybenzoic acid
para-substrate	4-methoxybenzoic acid
ATCUN	Amino-terminal Cu(II) and Ni(II) binding motif
AI	Artificial Intelligence
AM1-BCC	Austin Model 1-Bond Charge Correction
CPU	Central processing unit
CD	Circular Dichroism
CpdI	Compound I
cMD	Conventional Molecular Dynamics
cryo-EM	Cryogenic electron microscopy
COD	Cyclooctadiene
CYP, P450, CYP450	Cytochrome P450
DFT	Density Functional Theory
3WJ	DNA three-way junction
ECP	Effective core potentials
EPR	Electron paramagnetic resonance
FAV	Flavin Adenine Dinucleotide
FMN	Flavin Mononucleotide
Fmoc	Fluorenylmethyloxycarbonyl
FF	Force field
aMD or GaMD	Gaussian accelerated Molecular Dynamics
GAFF	Generalized Amber force field
GAFF	Genetic algorithm
GPCRs	G-protein coupled receptors
GPU	Graphics processing unit
HSAB	Hard and Soft Acids and Bases
HF	Hartree-Fock
HPLC-MS(ESI)	High performance liquid chromatography mass by electrospray ionization

HDX-MS	Hydrogen-deuterium exchange mass spectrometry
LiGaMd	Ligand Gausssian accelerated MD
ML	Machine Learning
MD	Molecular Dynamics
MM	Molecular Mechanics
CPR	NADPH-cytochrome P450 reductase
NMR	Nuclear Magnetic Resonance
PME	Particle Mesh Ewald
Pin1	Peptidyl-prolyl cis-trans isomerase NIMA-interacting
PCM	Polarizable continuum model
PES	Potential Energy Surface
PCA	Principal Component Analysis
QM	Quantum Mechanics
ROS	Reactive oxygen species
REMD	Replica Exchange Molecular Dynamics
RS	Resting state
RP-HPLC	Reverse phase HPLC
CYP199A4	Rhodopseudomonas palustris derived cytochrome P450
RMSD	Root mean square deviation
RMSF	Root mean square fluctuation
SCF	Self-Consisent Field
smFRET	Single molecule Förster resonance energy transfer
SPPS	Solid phase peptide synthesis
SMD	Solvation model based on density
tBu	Tert-butyl
TFA	Trifluoroacetic acid
TrpZip	Tryptophan zipper
Yap65	Yes-associated protein

Table of contents



CHAPTER 1
Introduction

1.1. Metals in nature

Metals have been intrinsic to human lives throughout history, starting with the Bronze Age over 5000 years ago. These elements are found at every level of complexity, from large architectural structures like the Glòries Tower in Barcelona, made of aluminum, to biomedical prosthetic devices and even molecular complexes used in pharmacological therapies for cancer treatment.

Metallic compounds represent approximately 25% of the Earth's crust in diverse forms, such as silicates or oxides, and account for up to 2% of the human body mass. Regardless of how small their quantities may be, these elements are essential for all life forms. For instance, the alkaline earth metal calcium represents 1.5% of the body mass as a crucial component of bone tissue, and it also plays vital roles related to muscle contraction or nervous system regulation. Many distinct functions have been associated with metals, including acting as structural elements, transport, catalysis, electron transfer, pH and electric potential regulation or facilitators of protein folding. The most abundant species belong to the alkaline and alkaline earth metal groups, but transition metals are also necessary as trace elements for specific purposes, like oxygen transport carried out by iron, structural and catalytic roles associated with zinc or connective tissue formation mediated by copper. However, a precise equilibrium of these elements — known as homeostasis — is vital, as any deviation from the optimal levels can lead to numerous diseases, ranging from cardiomyopathies to neurological disorders.[1]

From a chemical point of view, properties of transition metals are dictated by their partially filled *d* orbitals, which contain easily polarizable electrons that can participate in a wide variety of bonding and redox interactions. For this reason, transition metals can access multiple oxidation states, adopt different coordination geometries, or participate in electron transferring events in a way organic molecules cannot achieve. Moreover, unpaired electrons can lead to multiple spin states and spin-crossover events that can trigger structural or environmental changes involved in complex catalytic cycles.[2–4]

Pearson's Hard and Soft Acids and Bases (HSAB) theory contemplates charge density and polarizability to classify metal centers as hard or soft acids, which will be prompted to interact with bases of similar character. Transition metals fall into the soft or borderline-soft category, so they will tend to interact with soft donor atoms like sulfur or nitrogen, which conditions their coordination preferences; however, their coordination preferences can shift depending on their oxidation states.[5] These characteristics confer transition metals remarkable versatility and reactivity which have been harnessed by nature to perform complex transformations under mild conditions. In this thesis, two of these elements with radically different implications in nature are under the spotlight: iron and palladium.

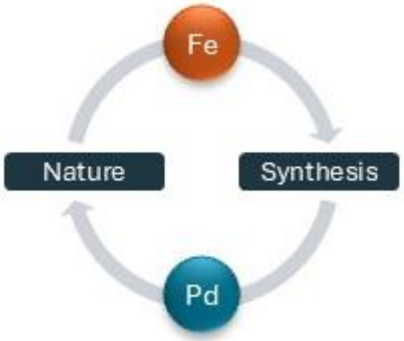


Figure 1.1. Scheme of two perspectives on biocatalysis: nature-inspired catalytic systems harnessed for synthetic purposes and incorporation of conventional synthetic catalysis into biological contexts.

Iron is one of the essential metallic elements that enables the reactivity of multiple enzymes in the form of cofactors or metallic clusters. Its dysregulation is directly involved in many diseases, either due to its deficiency — like anemia — or excess. The latter has to do with the redox capability of iron to generate reactive oxygen species (ROS), which are compounds containing unpaired electrons that make them extremely reactive and can disrupt key cellular elements like amino acids, DNA nitrogenated bases or membrane-forming lipids, eventually leading to cellular death and progressive organ damage. However, when properly regulated, the redox activity of Fe is a great tool that can be finely tuned by its ligands (first coordination sphere) and its environment (i.e., active site amino acids forming the second coordination sphere). This versatility, in addition to the elevated levels of Fe(II) available in early life, previous to the generation of the oxygenated atmosphere, explains the significant number of Fe-containing biomolecules present in all organisms. Scientific research has been focused not only on understanding the mechanistic entanglements of Fe-containing enzymes but also on adapting their reactivity and repurposing them for convenient use in a wide range of applications like diagnosis, therapy or biocatalysis.[6,7]

Palladium is widely used as a catalyst in organic synthesis, especially in C-C bond forming reactions such as Suzuki, Negishi, Heck, Sonogashira or Stille couplings but also in other types of processes like hydrogenation or C-C cleavage. Unlike with essential first-row transition metals, there are no palladium-dependent processes in nature, so in recent years scientists have focused on taking advantage of its varied reactivity and translating it into biological contexts. Its new-to-nature chemical behavior can increase the range of transformations achieved in living systems, creating exceptional tools for medicinal chemistry.[8] For instance, promising antimicrobial activity against *Mycobacterium tuberculosis* and other bacteria has been reported for methyl and octyl gallate Pd(II) complexes,[9] and anticancer alternatives to cisplatin based on oxime-functionalized ligands are outperforming the marketed drug during *in vitro* assays in various human cancer lines.[10] Remarkably, padeliporfin (Tookad®) is the first Pd-based agent approved for photochemotherapy against low-risk prostate cancer. Nevertheless, the scarce number of successful therapeutic examples highlights the complexity of the task. Some of the challenges encountered include solubilization of palladium complexes in aqueous cellular media, internalization into living organisms, ensuring the bioorthogonality of the catalytic activity, maintaining stability within the complex biological milieu and mitigating potential cytotoxicity. Numerous creative workarounds have been developed, like nanoparticle encapsulation of the complexes, modification of organic ligands or conjugation with biomolecular scaffolds ranging from small peptides to large enzymes or antibodies.[11–15]

This thesis focuses on the study of these metals as biocatalysts through their cooperation with biomolecular scaffolds.[16–18] Generally, the adaptation of biomolecules to generate efficient biocatalysts with applications in industry or biomedicine can be approached from four different perspectives: 1) start from a nature-provided entity and optimize it to improve its efficiency or soften its working conditions; 2) adapt the existing scaffold to modify its reactivity, by broadening or switching its substrate scope for example; 3) modify an existing non-catalytic biomolecule to add the new functionalities; 4) develop non-existing protein frameworks *de novo* specifically tailored for a unique catalytic objective.

The first two approaches are the basis of enzyme engineering, which is the specific area dedicated to modifying protein scaffolds for enhanced efficiency or altered reactivities. The most popular strategies for this purpose are based on directed evolution, and in recent years rational design has become an accessible alternative based on the detailed knowledge of the enzyme of interest and all the events related to its biocatalytic activity. However, strategies based on rational designs are mainly focused on the first approach, finding ways to improve the efficiency of engineered enzymes and less reports are found on substrate scope alteration.

The last two tactics have been explored in the context of bioorthogonal chemistry, which is the integration of synthetic chemical reactions into complex biological environments like cells or living organisms without interfering with native biochemical processes. These strategies face additional challenges such as the need for rapid and efficient reactions that can proceed in aqueous media under physiological conditions (e.g., temperature or pH) or using reagents that are non-toxic for the host system. Incorporating metal-based reactivity into biomolecular scaffolds further complicates the picture, as further research is required to efficiently solve key questions such as predicting structural stability of the complex, designing the proper metal center or ensuring cellular uptake. Despite these handicaps, progress in this field is leading to biomedical advances such as *in situ* drug generation or release, paving the way towards more personalized and effective therapeutic treatments. [19,20]

This work presents novel examples of approaches 2 and 3, attempting to gain insights for the rational modification of iron-containing enzymes to alter their selectivity and implementing palladium active sites in peptide scaffolds for bioorthogonal reactivity.

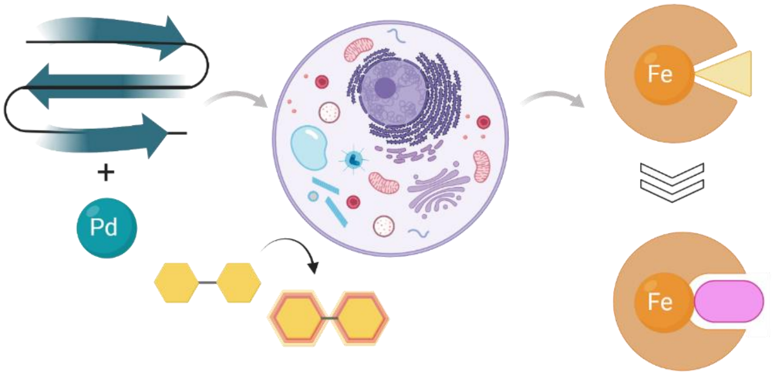


Figure 1.2. Schematic representation of the approaches on biocatalysis implemented in this thesis

1.2. Metalloproteins: reactive centers of the cell

1.2.1. Proteins, nature’s multipurpose machines

In living organisms, metals are rarely found as independent atoms but rather in complex with proteins. These macromolecules are composed of combinations of 20 essential amino acids, and they are present in all forms of life, from viruses – arguably an alive entity– to humans. The amino acid sequences are the primary structure of proteins, which are folded into secondary structural arrangements like α-helices, β-sheets, and other non-repetitive motives like loops. They further organize in space into a tertiary structure that renders functional entities, and when more than one chain interacts to complete the scaffold, it is known as the quaternary structure (Figure 1.3). The protein three-dimensional organization and its dynamics determine what activities can be performed, so massive scientific efforts have been dedicated to establishing clear relationships between structure and function.

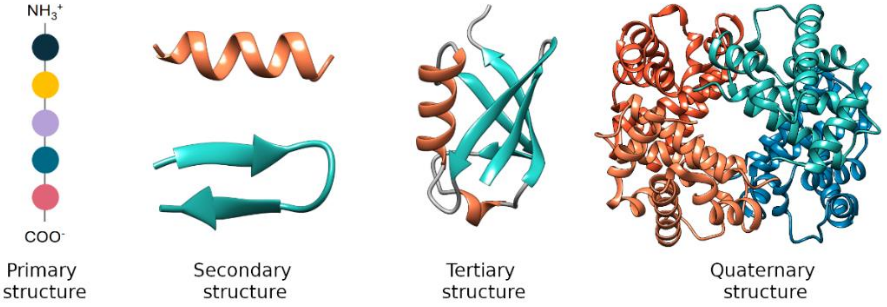


Figure 1.3. Levels of structural organization in proteins.

Many experimental are available for this purpose like X-ray crystallography, cryogenic electron microscopy (cryo-EM), circular dichroism (CD), nuclear magnetic resonance (NMR), electron paramagnetic resonance (EPR), single molecule Förster resonance energy transfer (smFRET) or hydrogen-deuterium exchange mass spectrometry (HDX-MS) that can be complemented with theoretical approaches like classical molecular dynamics (MD).[21–27] Moreover, when trying to study key components of proteins, one cannot focus only on the active site, as allosteric sites and networks between distal points are known to not only influence but also act as crucial elements in biomolecule activity. Identification of these complex relationships is elusive, and significant research has been devoted to gaining clearer insights.[28,29]

The range of functions carried out by these biomolecules is extremely wide, so a reduced selection will be mentioned to illustrate their relevance in biochemical processes.[30] For instance, proteins are basic structural components of the cytoskeleton, formed by tubulin, actin, and other fibrous polypeptides, and are responsible for the elasticity and structural soundness of the connective tissue through elastin and collagen. Many hormones and most receptors involved in signaling pathways are peptides like insulin or complex polymers like G-protein coupled receptors (GPCRs). Moreover, the immune response to external pathogens is triggered by antibodies and regulated through cytokines. Proteins are also closely related to the management of genetic information, as they are involved in many DNA-related processes like assisting chromatin folding (histones), triggering and engaging in transcription (transcription factors like p53, RNA polymerase), hosting translation (ribosomes), guiding the folding of the resulting proteins (chaperones) or facilitating replication (helicases or DNA polymerases).

Multiple vital functions of the organism are related to the capability of proteins to specifically bind other molecules. Molecular recognition is necessary for the transport of vital compounds like oxygen carried by hemoglobin, the storage of nutrients like nitrogen in cereal seed's prolamins or to perform catalytic transformations on several substrates ranging from small compounds to other polypeptides.[31,32] Proteins that belong to the last category, involved in carrying out any type of chemical reaction without being consumed in the process, are known as enzymes, which carry out multiple specific tasks like cutting other proteins (proteases), condensing multiple molecules into one product (synthases) or performing redox reactions for electron transfer, homeostasis or photosynthesis among many others. [33]

The mechanisms of substrate binding in proteins are crucial for understanding enzyme functioning and are described through different molecular recognition models. Early on, the lock-and-key model was proposed based on the assumption that enzymes maintained a rigid conformation that had to perfectly complement the ligand, just like a key only opens one lock. Although intuitive, this model did not account for protein dynamic flexibility, and other alternatives were considered. The induced fit model implies that the biomolecule undergoes a conformational change upon substrate binding to accommodate the ligand. Finally, the conformational selection model states that proteins exist in a dynamic state representing a pool of pre-existing organizations, so ligands bind to those that are most complementary to them. These models are not mutually exclusive, and small changes in proteins or their conditions can induce shifts from one mechanism to another.[34,35]

1.2.2. Adding metals to the mixture

Metalloproteins are calculated to account for 30-50% of the human genome, highlighting the essential role of metal ions in biological systems.[36] For instance, many of the aforementioned examples of functionalities are carried out by metal-containing biomolecules, especially in the case of enzymes, in which metal ions enable catalysis by stabilizing reaction intermediates, providing Lewis acidity to polarize substrates or facilitating atom or electron transfer.[37] In the latter, transition metal-dependent enzymes are especially relevant due to the numerous oxidation states accessible for these elements. Extensive research has been dedicated to understanding the mechanisms of natural metalloenzymes, trying to mimic their highly efficient and selective properties in engineered or repurposed alternatives tailored for synthetic approaches. Indeed, fields like biotechnology or biomedicine use metalloenzymes for green biocatalysis or therapeutic and diagnostic applications.

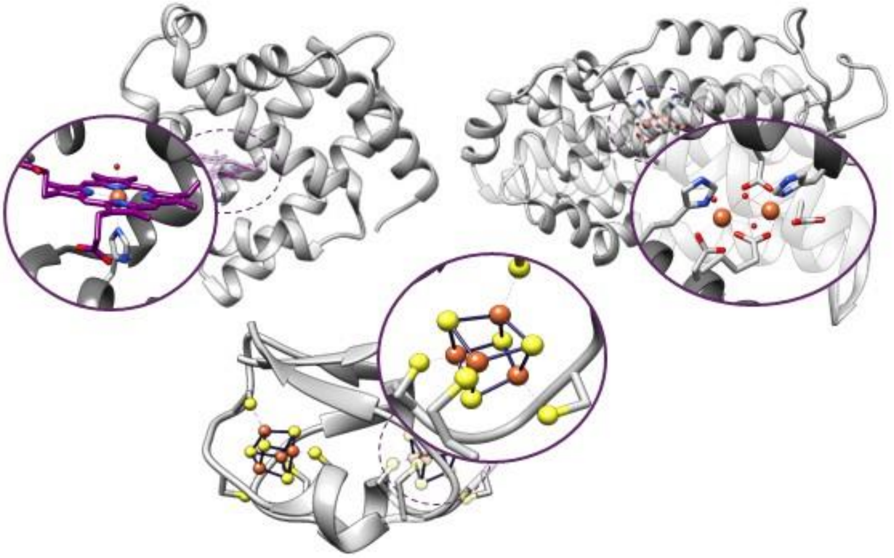


Figure 1.4. Structures of Fe-containing proteins. Myoglobin as an example of heme-binding protein, ferredoxin as an example of [4Fe-4S] cluster cofactor and ribonucleotide reductase protein R2 as an example of non-heme diiron center.

Precisely, iron-containing metalloproteins stand out for their versatility and biological importance. This metal can be incorporated into the protein scaffold in different ways, leading to specific chemical properties and reactivity (Figure 1.4). One of the most common are heme-containing enzymes, where Fe is embedded in a porphyrin macrocycle and axially coordinated to the protein.[38] Some examples are hemoglobin and myoglobin in charge of oxygen transport and storage, peroxidases and catalases that decompose ROS for detoxification or cytochromes, which participate in electron transfer for photosynthesis or cell respiration and will be examined in more detail later in this chapter. On the contrary, non-heme iron enzymes directly coordinate the free metal through amino acid side chains, mainly histidine, cysteine or carboxylates. These enzymes typically present dioxygenase reactivity, transforming their substrates through the incorporation of molecular oxygen like, for example, lipoxygenases oxidize polyunsaturated fatty acids. Iron-sulfur clusters (Fe-S) are inserted into protein scaffolds through cysteine coordination and can present different stoichiometries, typically [2Fe-2S], [3Fe-4S] or [4Fe-4S]. These enzymes are mainly involved in electron transfer chains or radical mediated catalysis. Other examples include multiple Fe centers or H-cluster active site proteins, which combine a binuclear [2Fe] cluster and a [4Fe-4S] cluster to catalyze reversible hydrogen production or uptake, which makes them very convenient for renewable energy technologies.

1.2.3. Cytochrome P450 Enzymes

On the topic of heme-containing enzymes and object of the present work, cytochromes P450 (or CYP) are a superfamily of metabolic enzymes present across all kingdoms of life that were first discovered in the 1960s. Their physiological roles are related to their ability to oxidize organic compounds and can be comprised into two main fields: metabolism of xenobiotic molecules and biosynthesis of signaling molecules involved in homeostasis and development. For instance, the importance of these biomolecules can be observed in plants, which contain numerous genes dedicated to CYPs to eliminate herbicides or synthesize protecting metabolites that are regulated according to their environmental needs; also, insects have developed resistance to synthetic or plant-derived insecticides through P450 metabolic detoxification. In the case of humans, these cytochromes are essential in drug metabolism, being present in over 75% of the reactivity linked to drug management in the body. Moreover, they are also involved in synthesis of hormones and vitamins or transformation of fatty acids.[39–41]

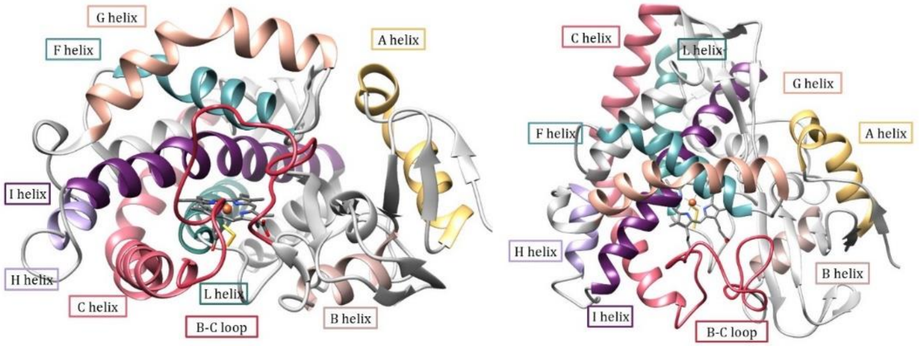


Figure 1.5. General structure of cytochromes P450 taking a prokaryotic species as an example. The different regions are labeled according to the depicted colors.

Bacterial CYPs undertake similar tasks although distinctive structural features are observed when compared to other CYPs. Generally, eukaryotic cytochromes are membrane-bound enzymes that connect through an N-terminal helix that is absent in the prokaryotic enzymes, which are soluble proteins. However, except for this N-terminal anchor, most CYPs share a very similar overall structure regardless of their species of origin (Figure 1.5). In fact, the closer the residues are to the binding site the higher their conservation, with two specific regions that exhibit contact with the heme moiety maintained throughout all variants. One of these regions is the β segment containing the Cys residue that binds the Fe atom, which protects the crucial amino acid from side reactivity and fixes it in the proper position for heme anchoring. The other group of conserved amino acids is involved in O_2 activation, located on helix I right above the metal center.[42]

Focusing on their reactivity, one of the main topics in this thesis, these enzymes function as monooxygenases by taking molecular oxygen from the atmosphere to transform their substrates, commonly aromatic derivatives or lipophilic compounds. In a general scheme, P450s insert one oxygen atom into a stable bond from the target molecule while reducing the other O, yielding a water molecule alongside the product. The catalytic cycle of the Fe-bound heme coenzyme is well known as it has been subject to detailed research. At different points of the cycle, the enzyme requires the stepwise reduction of some of the species involved in catalysis, so CYPs work with associated redox partners to provide the necessary reductive electrons. There are many partner proteins available depending on the specific cytochrome P450, although they can be classified into two main types: class I involves a ferredoxin reductase-ferredoxin tandem of proteins to transfer the electrons from NAD(P)H to the heme and is more common in prokaryotes; class II is frequently found in eukaryotic cells and depends on the NADPH-cytochrome P450 reductase (CPR), a flavoprotein that binds both FAD and FMN to create the electron chain from the electron carrier to the final enzyme.[43]

Another key aspect of cytochrome P450 enzymes is how they incorporate the molecules involved in their reactivity into the active site. The heme cofactor is buried deep within the protein core, so access of substrates, dioxygen, or solvent molecules is not trivial in the functioning of these enzymes. Several pathways have been identified, which often involve substantial conformational changes that can be hypothesized in crystallographic structures or explicitly observed through molecular dynamics simulations.[44] In most cases, key regions related to these recruiting movements are the F and G helices, the F-G loop, or the B-C loop. Manipulation of these channels has been examined as a way to control and modify biomolecule activity not only in CYPs but in a wide range of proteins such as other oxidoreductases, hydrolases or transferases.[41]

Given the potential of cytochromes P450 to activate inert C-C bonds to perform otherwise costly oxidative transformations, these enzymes have been subject to extensive research in many fields such as bioinorganic and bioorganic chemistry, regarded as an apparently ideal model for enzyme engineering with the aim of developing efficient biocatalysts. Indeed, Frances H. Arnold was awarded the 2018 Nobel Prize in Chemistry for pioneering work in the directed evolution of enzymes, including CYPs, to unlock new-to-nature transformations.[45–47] Despite this progress, the transition to rational design of P450 enzymes remains a formidable challenge.[48] Most current rational design approaches are limited to improving catalytic efficiency or stability in hope of enhancing industrial applicability. Some semi-rational strategies combine experimental mutagenesis with molecular docking rationalization to modulate substrate specificity, but still little is known about the complex mechanistic requirements related to substrate scope expansion or specificity modification (Figure 1.6).[49]

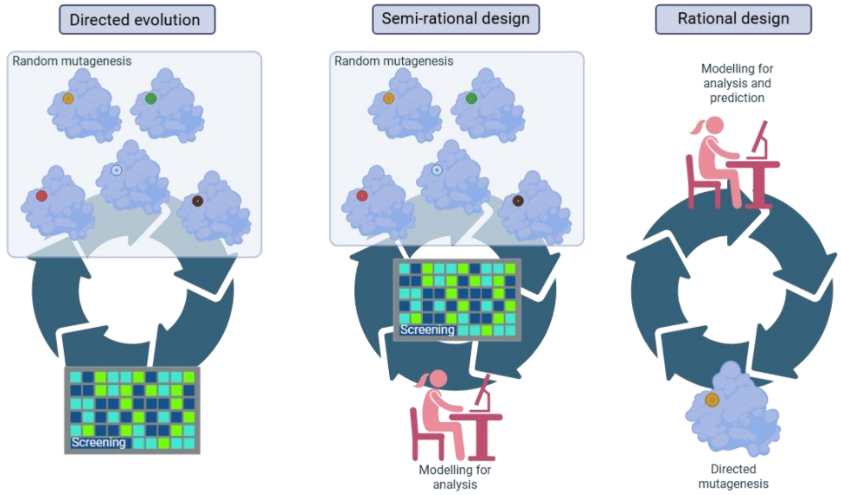


Figure 1.6. Schematic representation of basic steps in the different approaches of biocatalyst development.

In this line, one of the underexplored areas in rational engineering of CYPs is the role of conformational dynamics in their activity. Although the intrinsic flexibility of these enzymes is well-known, its impact on substrate binding, orientation and enzyme specificity has only recently become central in cytochromes P450 research. Computational approaches have traditionally focused on electronic mechanistic approaches centered in the active, but a few recent studies incorporating dynamic influence on catalytic processes have started to shed light on the importance of structural dynamics.[50,51] This thesis addresses this issue by applying multiscale molecular simulations to study substrate binding, product release and how dynamic events are relevant to the catalytic mechanism of cytochromes P450, offering insights that could support broader application of dynamics-based design strategies.

Some of the future perspectives focus on converting P450s into self-sufficient catalysts, expressing redox partner proteins combined with the heme domain or using artificial fusion systems. Moreover, advances from the computational perspective on addressing their structural complexity and predicting the impact of potential modifications will be tremendously beneficial for the development of these enzymes as versatile biocatalysts.[52]

1.3. Metallopeptides: the best of both worlds

1.3.1. Peptides: reducing size to increase opportunity

As described above, enzymes are a complex machinery refined by evolution (or design) that, despite numerous groundbreaking advances in decoding their structure–function relationships, can still be challenging to fine-tune for very specific needs. However, peptides are alternative, smaller biomolecules which are also built by amino acids, and they offer some advantages such as easier synthesis and customization while maintaining biocompatibility. There are many examples of naturally occurring peptides including insulin and glucagon, involved in glucose regulation; endorphins, which modulate the nervous system; and amyloid-β peptides, implicated in the pathology of Alzheimer's disease. Moreover, some protein domains can be isolated as shorter oligomers that retain their local folding patterns, providing useful miniaturized models to study key structural features that can be extrapolated to larger enzymes. These oligomers can be readily synthesized by organic means like solid phase peptide synthesis, a technique that earned its inventor the Nobel Prize in Chemistry in 1984 and has since been conveniently automatized. With proper design, *de novo* peptides can be crafted for multiple tailored purposes, although accurate prediction of their structural behavior remains a significant challenge. Nevertheless, the characteristics of peptides — including their modular architecture and capacity for self-assembly — make them powerful tools with applications in diverse fields such as therapeutic treatments, molecular imaging, innovative electro- or photoactive materials, drug delivery and catalysis, to name a few.

1.3.2. Mini metalloenzymes

Regarding catalysis, peptides present additional interest as chiral molecules capable of inducing asymmetric transformations. They can function by mimicking enzyme active sites, acting directly as organocatalysts themselves, modulating the reaction environment through their 3D structure or self-assembly or coordinated with reactive metallic centers.[53–56] Metallopeptides offer several advantages compared to their isolated oligomeric and metal components (Figure 1.7). On one hand, peptides are biocompatible molecules with defined structural features that are easy to synthesize and modify, although they have access to a limited range of transformations. On the other hand, transition metals can catalyze a wider spectrum of reactions but typically with little selectivity and under restrictive reaction conditions. This combination of characteristics makes these systems ideal candidates for bioorthogonal catalysis, as they can be introduced in biological contexts without interfering in ongoing processes as their transformations are unprecedented in cells.

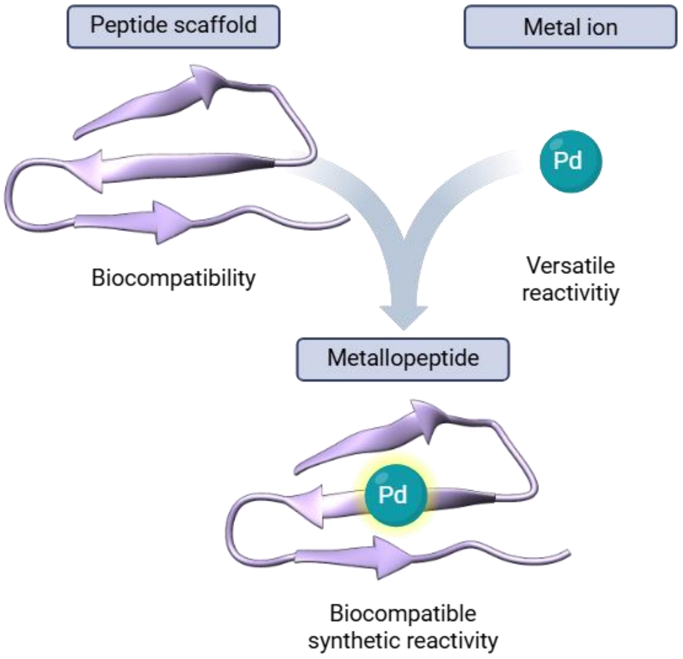


Figure 1.7. Schematic representation of metallopeptide advantages.

Tailored peptide synthesis enables the construction of first and second coordination spheres specifically for the selected metal, considering the ion's geometry along its catalytic mechanism. Different approaches can be employed to couple both species, from the use of prosthetic groups like heme to incorporate the metallic ion, to the insertion of naturally metal binding amino acids like histidine or noncanonical alternatives like bipyridine. Moreover, long-range interactions can be modulated through structural arrangement of the residues, which can also provide active site assistance, for example, through proton delivery. This structural control further enables the design of specific environments created by the peptide scaffold, for instance changing the redox potential of the substrates themselves or creating hydrophobic pockets to enhance reactivity.[57–62]

Unfortunately, this perfectly tunable scenario is still far from reality and cannot be readily accomplished exclusively through proper design. Confident predictions of short peptide conformation are challenging due to their flexible nature that usually lacks long-range stabilizing interactions and can be highly dependent on external factors like pH or ligand binding. Precisely, metal complexing effects are an added difficulty, as they can drastically change the structural arrangement, making it difficult to create proper binding sites within the peptide scaffold.[13] Moreover, most structure and metal binding prediction tools are optimized for larger entities and struggle modelling short, flexible, intrinsically disordered peptides, a problem partially derived from the smaller amount of experimentally resolved peptide structures.[57,63,64]

Some strategies to overcome these difficulties are based on using well-characterized stably folding peptides as templates to implement the metallic active center (Figure 1.8).[65] Different 3D structural arrangements have been proven successful in metallopeptide design: α-helix-based structures are most commonly used for their simplicity in terms of synthesis and stability, like the case of α-helical bundles or coiled coils.[66,67] Among β sheet examples, the most widely studied are amyloids β for their implication in neurological disorders like Alzheimer's or Parkinson's diseases linked to metal binding and aggregation propensity.[68] Moreover, some specific short sequences have shown remarkable metal-binding properties, like the ATCUN motif (amino-terminal Cu(II) and Ni(II)-binding) composed of three amino acids placed at the N-terminus of peptides that selectively bind Ni and Cu.[69,70]

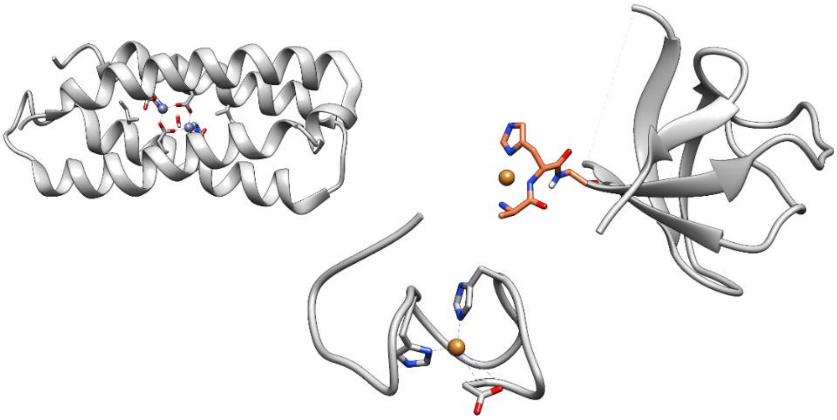


Figure 1.8. Examples of metal-binding scaffolds presenting different secondary structures.

Regarding the creation of functional metal binding sites, numerous computational tools based on naturally occurring motifs are available; however, these are often limited to predefined sequence templates or characteristic structural arrangements like those found in Zn-fingers.[71] Other options are based on structural fitting for a given coordination geometry, assessing the binding affinity through docking scoring of purely geometrical constraint goals. However, specific key points connecting metal binding with catalytic activity are not well characterized or properly implemented in prediction protocols, even less so when talking about *in cellulo* or *in vivo* biocatalysis.[66]

This thesis presents an alternative *de novo* metal binding site design strategy, which focuses on evaluating the structural suitability of a given scaffold for the insertion of metal-coordinating residues. In addition, fold stability is assessed through molecular dynamics based on specifically parametrized force fields for the electronically characterized Pd complex. Combination of this multiscale protocol with experimental testing allows the study of critical points essential for successful metal binding and dynamic fold stability prediction and their relationship with biocatalytic activity, offering a step forward towards the pursuit of effectively tailored metallopeptide design.

1.3.3. Molecular Modeling in biological systems

The 2024 Nobel Prize in Chemistry was awarded to D. Baker, D. Hassabis and J. M. Jumper for their work in protein design and structure prediction software that have granted access to some of the most complex processes in nature like creating newly functional molecules or the assembly of operational protein structures from lineal sequence information. This is indicative of the undeniable role that computational advances play nowadays in all aspects of life, including science, aiding in the development of innovative biotechnologies to address current problems like environmental sustainability, food security or health diagnostics and therapies.

The application of computation to scientific research has been under constant and exponential growth since the first appearance of computers in the 1950s. In the chemistry field, it has evolved from allowing simulations of just a few atoms due to hardware restraints to reproducing molecular dynamics of complex biological systems within days or even hours thanks to GPU-accelerated platforms. Likewise, hundreds of gigabytes of structural and activity data can be quickly processed in computational screenings for drug discovery, requiring large memory access and powerful connections. The most recent advances in Artificial Intelligence (AI) and Machine Learning (ML) have also had an impact on the scientific community, especially in the field of proteins as has been mentioned at the beginning, with promising outlooks for the future of research at many levels.

Computational chemistry techniques are usually classified according to their level of theory, where higher accuracy in the calculation of the energy of the system implies higher computational cost. Quantum mechanical (QM) methods are based on the Schrödinger equation to describe electron behavior and due to their calculation cost, they can only be applied to small systems. For larger molecules, molecular mechanics (MM) are preferred, as they represent atoms as spheres connected through classical springs with a given charge, decreasing considerably the computational cost and allowing the simulation of large-scale molecules and events. When needing to represent electrons in larger systems, semi-empirical QM or hybrid QM/MM methodologies are available, arriving at approximations that introduce empirical parameters to reduce computational cost in the former or representing the system at two levels, a small area or interest modeled through QM and leaving the interactions comprising the rest of the molecule under an MM model in the latter.

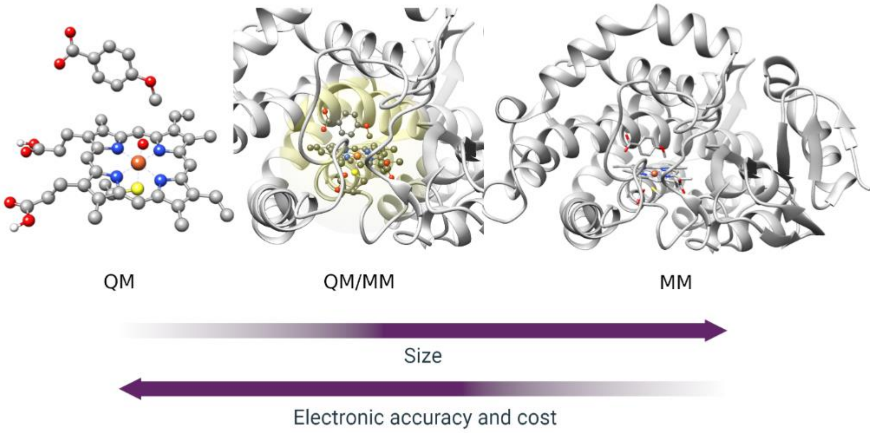


Figure 1.9. Schematic representation of the different levels of theory in Computational Modelling

On a general basis, molecular modeling comprises any (computational) technique that recreates reality for the study of a desired property relevant to a chemical or biological system. Perfect recreation of the real world is inherently impossible through modelling, so keeping this in mind when selecting the computational approach for the task at hand is crucial. For instance, a detailed map of the floor plan of Santiago de Compostela's cathedral could be the perfect preparation for a visit to its roofs, but it would not be the tool of choice when planning a three-day trip to the Galician capital, where a general map including restaurants and points of interest would be more useful. Similarly, the precision represented in a molecular model needs to adapt to the property under study to obtain relevant results. Very accurate quantic descriptions of electronic structures can be obtained for relatively small compounds, perfect to understand key factors influencing reactivity mechanisms. By contrast, reducing a complex system like an enzyme to the specific center where the reactivity takes place may not be representative of its full function and mechanism, as a global vision of the conformational changes affecting its interactions is needed. When tackling complex, multidimensional problems like structure-function relationships or biocatalytic activity, multi-scale protocols that combine these different approaches are indispensable. In fact, the importance of multiscaling was specifically recognized in 2013, when the Nobel Prize in Chemistry was awarded to A. Warshel, M. Karplus and M. Levitt for the development of integrative approaches that combined different levels of theory, allowing a paradigmatic change in the study of not only biologically relevant events but any scientific field.

On metalloprotein and metallopeptide studies, computational modelling has become an indispensable tool, particularly for rational design approaches that must integrate multiple levels of theory to provide meaningful insights into fold stability, metal coordination and catalytic reactivity. Among the available methodologies, Molecular Dynamics (MD) stands out as the most suitable technique for exploring time-dependent conformational changes that are intrinsic to biomolecular function. These motions are directly linked to fold stability and their simulation provides information on how binding other small molecules might influence structural dynamics and, hence, function.

To overcome the limitations of conventional MD, enhanced sampling techniques like accelerated MD and Replica Exchange MD (REMD) are employed to ensure exploration of wider conformational landscapes, avoiding trapping in local minima of the potential energy surface.[72,73] Moreover, free-energy methods such as free-energy perturbation approaches, thermodynamic integration or biased techniques like umbrella sampling or metadynamics allow the quantification of the thermodynamic cost of binding or other conformational changes.[74–77] All these methodologies are very reliant on the quality of the force field employed, which defines the classical interactions between atoms. Despite numerous robust force fields that have been developed for common biomolecules such as proteins, DNA or carbohydrates, accurate modeling of transition metals remains a significant challenge due to their diverse oxidation state and coordination geometries. In this thesis, the bonded model approach has been used to study how metal coordination affects both peptide flexibility and metalloprotein configuration (Figure 1.10).

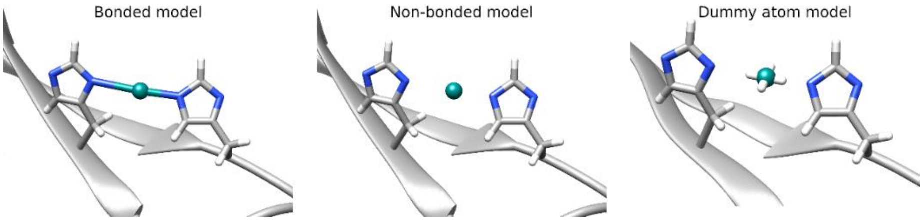


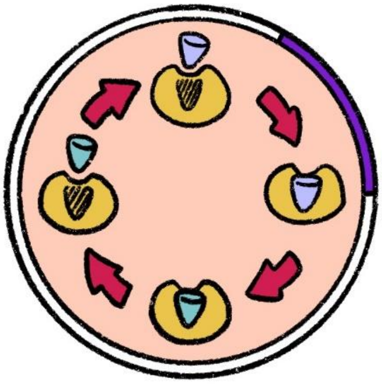
Figure 1.10. Most common models for representation of metal binding in metalloprotein design.

Another essential component of biomolecular modelling is the representation of intermolecular interactions. Molecular Docking methods aim to identify the most favorable binding poses between two interacting species — such as an enzyme and substrate, protein and cofactor, or peptide and metal complex — by evaluating the different conformations through a scoring function. One of the primary challenges in docking is the need to integrate large conformational flexibility with efficient sampling of plausible binding modes. To address this, docking platforms implement various strategies, including rigid docking, partial flexibility for either ligand or receptor, and constraint-driven reductions in the conformational search space. Tools such as Rosetta, GaudiMM, and HADDOCK incorporate these approximations.[78–81] Metal representation requires, again, specific strategies like the dummy atom model implemented in GOLD to guide coordination geometry (Figure 1.10), or Rosetta's metal-aware protocols, which allow metal–ligand interactions to be evaluated more accurately.[82–84] Furthermore, other programs are aimed at predicting metal binding based on protein sequence or structure to aid in the identification of metal binding sites or in the design of novel metalloproteins. In this thesis, BioMetAll was employed to predict point mutations that enable proper metal coordination, and GOLD was used to identify optimal binding poses of a Pd(II) complex with the designed peptides.

When investigating biocatalytic transformations, it becomes essential to model electronic behavior explicitly. This has led to the development of QM approaches and hybrid QM/MM methods, which account not only for the immediate coordination environment (first and second spheres) but also for the broader structural context of the biomolecular scaffold. Examples include the use of QM clusters to represent the active site of the biocatalyst, ONIOM-based QM/MM calculations or theozyme modelling, which represents the optimized TS structure into the binding site of the protein, allowing the investigation of how enzyme conformational changes influence the stabilization of transition states and modulate reaction outcomes.[85–87]

The more recent development of Artificial Intelligence (AI) or Machine Learning (ML) brings extraordinary properties for identifying complex patterns across multiple datasets, which have been harnessed not only for sequence-structure predictions but also to extract accurate force field parameters, structure-activity relationships or intermolecular interaction evaluations. These methods heavily rely on data quality, so some of the challenges that are currently faced are related to providing sufficiently extensive, varied and curated information to be productively analyzed.[88–92]

These examples illustrate how the integration of multiscale computational methods is crucial to gaining a comprehensive understanding of complex biochemical events. In this thesis, a new multiscale protocol for metallopeptide design was developed, combining structure-based metal binding site prediction, metal-peptide interaction modelling and metal-peptide complex structural dynamic simulations. Molecular modelling of the catalytic cytochrome P450 CYP199A4 revealed the relevance of multiscaling approaches in the comprehensive study of enzyme activity. Finally, these methods were further applied in other DNA-based and supramolecular systems to extend the expertise on their applicability.



CHAPTER 2
Objectives

This thesis aims at contributing to the comprehension of metal-mediated biocatalysis by combining molecular modelling and experimental techniques to study how the structural features of biomolecules can impact catalytic processes. More specifically, two different approaches to biologically relevant events are engaged. On the one side, translation of new reactivities into biological contexts through metallopeptide design is attempted by combining a reactive Pd(II) complex with different peptide scaffolds. On the other side, study of the transformations performed by nature-optimized heme-binding enzymes aims at decoding the molecular details of its reactivity to enable rational enzyme engineering.

Regarding the first chapter on metallopeptide design, the overall goal is the development of functional bioorthogonal palladopeptides based on β-sheet structures capable of catalyzing depropargylation reactions within cellular environments. For this purpose, experimental and computational approaches are combined in a multi-scale design pipeline applied to two different peptide scaffolds: WW domains and tryptophan zipper hairpins. The main objectives are:

- Development of a multi-scale computational protocol. Devise a pipeline to identify the best positions to insert metal-coordinating residues and predict their structural impact on the peptide scaffolds. This includes the use of the in-house developed *BioMetAll* program for sequence generation and screening along with Molecular Docking and conventional and accelerated Molecular Dynamics simulations to gain deeper understanding on the structural features of the designed peptides.
- Employ the spot-synthesis technique to assess the depropargylation reactivity of a library of 264 sequences generated through combinatorial arrangement of two His residues in a short, stable β-hairpin structure.

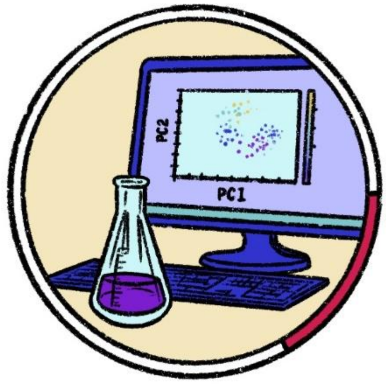
- Experimental evaluation of the designed metallopeptides. Synthesize and characterize the selected minienzymes to assess their structure dynamics through methods like CD or NMR and test their catalytic efficiency *in vitro*, Furthermore, this can help identify promising mutants to further test their *in cellulo* applications.

The second chapter focuses on establishing structural and dynamical bases for substrate selectivity and product management of the bacterial cytochrome P450 CYP199A4 through multi-scaling molecular modelling methodologies. Some of the methods include DFT catalytic profile exploration, conventional and accelerated Molecular Dynamics and extensive bioinformatic analysis of the simulations. The general primary objectives are:

- Identify key structural features in substrate/product binding and orientation.
- Determine molecular factors driving regioselectivity.
- Characterize entrance/release pathways in the enzyme scaffold.
- Investigate solvent involvement in the mechanism, specifically related to product stabilization as suggested by crystallographic data.

This approach aims to provide mechanistic insights into how CYP199A4 processes closely related molecules such as methoxybenzoic acid substrates or the equivalent hydroxybenzoic acid products. The end goal is to apply this information in rational design of engineered enzymes for alternative transformations.

On an additional note, the expertise acquired throughout this work is applied on alternative systems, such as non-canonical DNA scaffolds, where the interactions of peptides and metallopeptide are further investigated using molecular modelling approaches.



CHAPTER 3
Methodological background

This chapter provides an overview of the basic principles underlying the methods used in this thesis. As the work comprises both computational and experimental approaches, the fundamental ideas of the main techniques applied will be briefly described for ease in the understanding of the subsequent chapters, although not in deep detail.

3.1. Computational Chemistry

Modelling biochemical systems can be approached from two main levels of theory, depending on the questions being addressed and, therefore, the degree of detail represented in the selected model of reality. **Quantum Mechanics (QM)** provides the highest level of accuracy considering the electrons explicitly, but it comes at a high computational cost. Meanwhile, **Molecular Mechanics (MM)** uses a simpler, classical representation of atoms and bonds but at a lower expense, allowing the study of larger systems for longer timescales. **Hybrid approaches like QM/MM** attempt to arrive at a compromise by limiting the QM computation to a small region where the electronic exchanges take place while maintaining an MM description of the rest of the system.

Nowadays, the implementation of multi-scaling protocols that integrate diverse theoretical techniques is the most powerful and informative strategy to approach complex problems. Layered protocols allow access to solutions that integrate complementary perspectives, from electronic structure details to large-scale conformational dynamics. More specifically, multiscale modeling becomes indispensable when dealing with metal-containing systems, as their intricate electronic structure and coordination chemistry require specialized treatment. Accurate representation of metal behavior depends on quantum mechanical characterization that can be extrapolated to classical approaches through parametrization, which enables metal representation in Molecular Docking or Molecular Dynamics methodologies.

3.1.1. Quantum Mechanics (QM)

Quantum Theory arises from the concept that small enough particles —like electrons— do not behave like macroscopic objects but instead show wave-particle duality. To describe wave-like behavior, the wavefunction (Ψ) is used; this is a mathematical function that contains all possible information about the system it describes. Physically, it is related to the probability of finding the particle at a given point in space.[93]

To further interpret this function, quantum operators can be applied to obtain values describing observable physical properties. The central operator in quantum theory is the Hamiltonian ($\widehat{H}$), which gives the total energy of the system through the Schrödinger equation

$$\widehat{H}\Psi = E\Psi \tag{1}$$

where Ψ is an eigenfunction of the Hamiltonian and the allowed energies for a system correspond to the eigenvalues obtained (E). This equation can be represented in terms of time, allowing the prediction of the evolution of any system.[94]

In Theoretical Chemistry, solving the Schrödinger equation is desirable to understand and describe the system of interest with the highest level of accuracy. While this can be achieved for a single particle like in the case of hydrogenic atoms (He^{+}, Li^{2+}, etc.), the equation cannot be analytically solved for any other elements with more than 1 electron, let alone molecules. This challenge is known as the many-body problem, and various approximations have been developed to overcome it and obtain sufficiently accurate solutions.

To specifically address the many-body problem, **Self Consistent Field (SCF) methods** are employed in computational chemistry. The basic premise is that every electron of the system moves within an **average electrostatic field** generated by the other electrons. However, since the position of the electron is determined by the field and, in turn, the field depends on the position of the electrons, an iterative procedure –the *SCF loop*– is started to approximate a self-consistent solution. The typical steps include: 1) guessing of an initial distribution, 2) computation of the resulting field, 3) solving one-electron equations to find updated distributions, 4) comparison with step 1 and repetition until convergence is reached, i.e., the difference between iterations is negligible.

Depending on whether the method relies on the wavefunction or on the electron density to calculate the self-consistency, two main groups of SCF approaches can be found: **Hartree-Fock (HF)** or **Density Functional Theory (DFT)**.

Hartree-Fock

The Hartree-Fock methods are based on the HF wavefunction and the application of the Fock operator to evaluate the energy. HF wavefunctions are the asymmetric product of the best spin-orbitals possible to describe the electrons in the system. The Fock operator contains several terms to account for the different types of energies affecting a system, including kinetic energy of each electron, potential attractive energy between negative electrons and the positive nucleus, repulsive energy among each electron and the charged cloud created by the other particles and the exchange operator that accounts for interchangeable electrons that derive from the antisymmetry of the wavefunction.

By applying the operator to an ensemble of HF wavefunctions, the energy of the system can be calculated. The set of functions that are selected to represent the HF wavefunction are known as the basis set of the calculation. Several types are available, usually composed of one radial and two angular components to describe the shape of the orbitals. Additionally, quantum effects like **polarization** due to unoccupied orbitals or **diffusion** suffered by the most external electrons are included to accurately describe electron behavior. In heavy metal ions, characterized mainly by the properties of the valence electrons, less costly methods are considered for the representation of internal core electrons that do not play key roles in their reactivity such **as effective core potentials (ECP)** or **pseudo potentials**. Moreover, scalar relativistic effects are included in many of these ECPs, as in the Stuttgart-Dresden-Bonn pseudo potentials.

One of the main disadvantages of HF methods is their inability to contemplate *electronic correlation*, as they simply account for the average electric field experienced by each electron, assigning an effective potential. As a result, they ignore instantaneous electron-electron interactions, leading to a slight difference between the real energy of the system and the calculated one. This discrepancy is known as **correlation energy**. To address this limitation, several post-HF methods have been developed. Some examples are configuration interaction, perturbative approaches such as Møller-Plesset and others, but these methods are often considerably more computationally demanding.[95]

Density Functional Theory

Alternatively to solving the Schrödinger equation for the wavefunction — which depends on 3N spatial coordinates, where N is the number of electrons—, **Density Functional Theory (DFT)** translates the problem in terms of the electron probability density $\rho(x, y, z)$, which is a function of only three spatial coordinates. The theoretical foundations of DFT are established by the two **Hohenberg-Khon theorems**:[96]

1) The ground state energy of a molecule can be computed as a functional of the electron probability density $\rho(x, y, z)$ and therefore all properties can be obtained from ρ.

$$E_{\mathrm{gs}} = E_{\mathrm{gs}}[\rho(x, y, z)] \tag{2}$$

2nd) The ground state density provides the minimum energy of the system with the correct functional. Following variational principle, any energy calculated from a trial density will be equal or higher than the true ground state energy.

The first issue is that the true functional $E_{gs}[\rho(r)]$ is unknown, so the **Khon-Sham theorem (KS)** addresses this lack of a mathematical relationship between the energy and the electron probability density by contemplating a reference system (s) with some particularities.[97] It contains as many electrons as the real system of interest, but they do not interact amongst themselves. Moreover, each electron in the reference system experiences a potential energy that makes the electron probability density of s equal to that of the real system. This allows some simplifications to solve the Hamiltonian to obtain the energy of the systems.

To describe the energy dependent on ϱ(r), the functional can be decomposed in several terms, similarly to those described earlier for the Fock operator in HF calculations:

$$E[\rho(r)] = T_{e,s}[\rho(r)] + V_{ne}[\rho(r)] + V_H[\rho(r)] + V_{xc}[\rho(r)] \\ = \widehat{H}^{KS} \tag{3}$$

The first component corresponds to the kinetic energy of the independent electrons in system s, which can be obtained as the sum of the energy of each electron (4); the second and third contributions follow the Coulomb law of electrostatic potential, accounting for the attraction energy between nucleus and electrons and the classical electronic repulsion (5); finally, the last term corresponds to the *exchange-correlation potential*, which encompasses energy corrections not accounted for in previous terms and will be explained further on.

$$T_{e,s}[\rho(r)] = \sum_{i=1}^{N} \int \theta_i(r) \left(-\frac{\nabla^2}{2}\right) \theta_i(r)\ dr \tag{4}$$

$$V_{ne}[\rho(r)] = \sum_{A=1}^{M} \int \frac{Z_A}{|r - R_A|} \rho(r) dr \qquad V_H[\rho(r)] \\ = \frac{1}{2} \iint \frac{\rho(r_1)\rho(r_2)}{|r_1 - r_2|} dr_1 dr_2 \tag{5}$$

This operator is applied to the KS orbital functions θ. These are a Slater determinant of spin-orbital functions that verify that the probability density of the system is the sum of probability densities of the individual orbitals (6). As in HF, these orbitals can be described as a linear combination of basis sets. Keeping this in mind, equation 18 can be solved in an iterative manner following SCF principles.

$$\rho = \rho_s = \sum_{i=1}^{n} \left|\theta_i^{\mathrm{KS}}\right|^2 \tag{6}$$

$$\hat{H}_i^{\mathrm{KS}} \theta_i^{\mathrm{KS}} = \varepsilon_i^{\mathrm{KS}} \theta_i^{\mathrm{KS}} \tag{7}$$

Exchange-correlation functional

One of the central components of DFT is the **exchange-correlation functional $E_{xc}[\rho(\boldsymbol{r})]$**. This functional cannot be determined exactly, and thus, the accuracy of a DFT calculation strongly depends on the quality of the chosen approximation. The exchange-correlation energy comprises two contributions: the exchange energy, which comes from Pauli's exclusion principle, similar to the exchange operator in HF, and the correlation energy, which accounts for the instantaneous repulsion forces between electrons.

As a result, $E_{xc}[\rho(\boldsymbol{r})]$ is often decomposed into these two components, approximating each through different mathematical approaches. Typically, increasing the accuracy of these approximations also increases the associated computational cost, following the "Jacob's ladder" of functionals coined by Perdew.[98] The Local (Spin) Density Approximation (LDA or LSDA) relies exclusively on electron density and is effective for systems with homogenous densities that vary slowly with position, though it is not accurate enough for the calculation of properties like atomization energies. The Generalized Gradient Approximation (GGA) improves LDA by including the gradient of the density, enhancing accuracy for more diverse systems; some examples are Becke's exchange and the PW91 or LYP correlation functionals. Meta-GGA functionals go further by incorporating second derivatives or kinetic energy density (τ), as seen in B95 and TPSS. Hybrid GGA functionals, such as B3LYP, combine exact Hartree-Fock exchange with GGA corrections, weighted by empirical parameters, and are widely used due to their balance between accuracy and computational cost.[99]

Dispersion corrections

Since DFT is based on the *local* electronic density –sometimes including its gradient, as seen previously –, it inherently struggles to model long-range interactions like dispersion forces. These weak interactions, also known as **London dispersion** or **van der Waals forces**, appear between instantaneous dipole-induced dipole moieties and are of special importance in large systems, such as biomolecules or supramolecular complexes.

To address this defect, many methods have been developed, including nonlocal van der Waals functionals (vdW-DF), localized atomic potentials (LAP) or meta-hybrid density functionals very specifically parametrized (DFs). However, one of the most relied on options is **DFT with Dispersion correction (DFT-D)**, due to its robustness and applicability across most elements of the periodic table.

DFT-D methods include an empirical dispersion component to correct the energy obtained through simple DFT calculations. The general form, as developed by Grimme, adds a correction term based on the London formula for long-range pairwise interactions:

$$E_{total} = E_{DFT} + E_{dispersion} \qquad E_{ij}^{London\ disp} = -\frac{C_{6,ij}}{R_{ij}^{6}} \tag{8}$$

For DFT-D2, the London dispersion energy is computed for all pairs of atoms, regulated by a damping function $f_{\mathrm{damp}}(R_{ij})$ that gradually reduces the correction as interatomic distances decrease. Additionally, the whole correction energy is scaled by an empirical scaling factor S_6, optimized for different exchange-correlation functionals:

$$E_{\text{disp}}^{D2} = -S_6 \sum_{i,j>i}^{N} \frac{C_{6,ij}}{R_{ij}^6} f_{\text{damp}}(R_{ij}) \tag{9}$$

This approach is further improved in the case of Grimme's DFT-D3, which includes higher order and more environment-sensitive coefficients, as well as more realistic damping functions that better describe the short- and medium-range interactions. The small increase in computational power required in comparison to DFT-D2 is justified for the improvement in the results obtained. This is the dispersion correction method used for the majority of the present work.

$$E_{disp}^{D3} = -\sum_{n=6,8} S_n \sum_{i,j>i}^{N} \frac{C_n^{ij}}{R_{ij}^n} f'_{\text{damp}}(R_{ij}) \tag{10}$$

Solvent effects

In quantum mechanical calculations, it is possible to represent solvent effects including solvent molecules explicitly. Nevertheless, this comports a high computational cost, and it remains a challenge to determine the optimal number of molecules to represent a proper solvating environment. For this reason, a more common approach is to use *implicit solvent* techniques like **polarizable continuum model (PCM)**, which achieve a compromise between accuracy and computational demand. In this case, the solvent is treated as a continuum with a specific dielectric constant ε in which a cavity is created for the solute, with a size determined by the van der Waals radii of its atoms. The charge distribution of the molecule induces a polarization in the solvent which, in turn, influences the solute. A Self Consistent

Reaction Field methodology is used to iteratively modify polarization in the solvent and cavity until self-consistency.

3.1.2. Molecular Mechanics (MM)

Molecular Mechanics approaches solve the size limitations present in QM methods. As they do not represent electrons explicitly but instead consider atoms as spheres with charge and mass, this reduces the computational cost greatly, allowing the simulation of larger systems like biomolecules. Certainly, accuracy is sacrificed in this trade off, but classical laws present a sufficiently precise description of atom type, bond nature and other physicochemical properties to characterize a system.

The Born-Oppenheimer approximation is assumed so the nucleus positions (R) are enough to determine the energetics of the system. MM is based on **force fields (FF)**, which are the collection of parameters following classical physics laws that describe the behavior of the particles. The energy of a force field is composed of several terms that can be divided into bonded and non-bonded contributions. Within the FF, each nucleus is assigned an atom type that already contains information on the hybridization and local environment of the atom and will condition which bonded and non-bonded parameters are applicable to the specific particle.

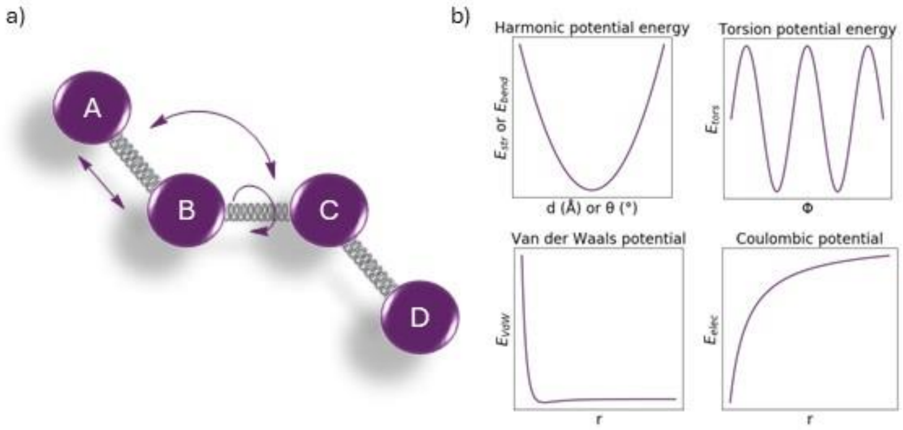


Figure 3.1. a) Representation of classical model of a four-atom molecule composed of charged spheres as atoms and springs as bonds. b) Schematic representation of potential energy functions for bonded (top) and non-bonded (bottom) components of a force field.

$$E_{tot}(R) = E_{str} + E_{bend} + E_{tors} + E_{VdW} + E_{elec} \tag{11}$$

Equation (11) contains the five basic terms of the FF, the first three comprising the bonded components and the last two the non-bonded. We will now focus on each contribution.

$$E_{str} = \sum_{bonds} K_d(d - d_0)^2 \tag{12}$$

$$E_{bend} = \sum_{angles} K_\theta(\theta - \theta_0)^2 \tag{13}$$

$$E_{tors} = \sum_{torsions} V_\omega[1 + cos(k\omega + \phi)] \tag{14}$$

E_{str} (12) refers to the energetic cost of stretching (or contracting) a bond between two atoms A and B (Figure 3.1); E_{bend} (13) is the energy associated to changing the angle between two atoms A and C interconnected through atom B. Both are described by Hooke's law, i.e. a harmonic potential. The values K_d and K_θ are force constants, and the corresponding distance or angle variable (d, θ) is compared to a reference value d_0 or θ_0.

E_{tors} (14) comprises the energy of rotating the angle described by the planes containing atoms A-B and atoms C-D. It follows a sinusoidal potential where ω is the torsion angle (as if the bond between atoms B and C were rotated along its axis). V_ω stands for the amplitude of torsion, k the periodicity and ϕ the phase shift.

There are cases where an additional bonded factor may be needed. For instance, with sp^2 atoms it is necessary to ensure that all their bonds stay in the same plane. This is known as improper torsion potential and is characterized through a function dependent on either the angle or the distance between the atom and the plane.

$$E_{VdW} = \sum_{\substack{non-bonded\\ ijpairs}} \epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \tag{15}$$

$$E_{elec} = \sum_{\substack{non-bonded\\ ij\ pairs}} \frac{1}{4\pi\epsilon_0} \frac{q_i q_j}{r_{ij}} \tag{16}$$

The non-bonded van der Waals potential E_{VdW} (15) describes weak interactions between atoms that are not connected through a direct bond. It follows a Lennard-Jones 12-6 potential to characterize the repulsion and attraction components and is dependent on the distance r_{ij} between the atoms.

Finally, the E_{elec} term (16) describes the electronic interactions following Coulomb's Law. It is a sum over all pairwise interactions and is dependent on the partial charge q_i assigned to each atom as well as on the distance r_{ij} separating them, regulated by the dielectric constant ϵ_0. The corresponding partial charges can be obtained from empirical fittings or can be computed. One of the most common computational procedures is Restrained Electrostatic Potential (RESP), in which the charges are distributed so they will reproduce a QM calculated electrostatic potential for a given system; hyperbolic restraints are applied to avoid excessive charges on buried atoms and equivalence restraints ensure same charges in analogous centers.

Similarly to the point charges, all parameters in force fields can be derived from empirical data or QM calculations. Recently, even new approaches through machine learning techniques have appeared. In all cases, these variables can be tailored to specific biological systems like proteins, nucleic acids or lipids or they can be a general set of rules to describe any type of molecule. Numerous FF are available, e.g. AMBER, GROMOS or CHARMM, providing multiple case-scenarios to select the most appropriate for the phenomenon wished to study.

Within the term Molecular Mechanics there are many types of techniques involved, with the common grounds of treating electrons implicitly. Now we will explain those used extensively in this thesis.

Molecular Dynamics

Overview

Molecular Dynamics (MD) is a computational simulation method to predict the evolution of a system through time where classical physics laws of motion govern the behavior of the atoms as they are considered spheres with a given mass, charge, position and velocity. By modeling its interactions over time, a biomolecule -such as a protein- can explore its **potential energy surface (PES)**, depicting the dynamic ensemble of conformations accessible at a given temperature.

To generate the trajectory of the system, Newton's second law of motion (17) is numerically integrated in a step-by-step fashion, with short time steps typically one or two orders of magnitude shorter than the fastest vibrational frequency that needs to be accurately resolved. At each time step, the forces applied are computed from the potential defined by the force field.

$$-\frac{dV}{dr} = m \cdot \frac{d^2r}{dt^2} \tag{17}$$

It is worth mentioning that to avoid the restricting frequency of H bonds, which vibrate at fastest speed, the SHAKE algorithm is usually implemented to keep these connections fixed. That way, the time step can be increased from 1 to 2fs for long simulations.

Starting from a given position extracted from either empirical data (e.g. X-Ray, NMR) or models (e.g. homology modelling, AlphaFold) a set of velocities are randomly assigned according to a Maxwell-Botzmann distribution, ensuring thermal equilibrium. Knowing these initial conditions, the simulation updates atomic positions, velocities and forces for each time step until the desired simulation time is reached.

Integration schemes

Different integrational schemes have been developed to compute the trajectory of the molecular system. Initially, **Verlet algorithm** would approximate coordinates through a Taylor series expansion (18) and velocities would be approximated retrospectively (19).

$$r(t+\Delta t) = 2r(t) - r(t-\Delta t) + \frac{F(t)}{m}\Delta t^2 \tag{18}$$

$$v(t+\Delta t) = \frac{dr_i(t)}{dt} = \frac{\mathrm{r(t+\Delta t) - r(t-\Delta t)}}{2\Delta \mathrm{t}} \tag{19}$$

However, this model presents some drawbacks. To calculate the upcoming position, the previous coordinates ($r - \Delta t$) are needed; this represents a challenge at the starting point of the simulation, which can be solved by using fictional previous coordinates estimated from the initial velocities. Moreover, the Taylor series is truncated at the third order which can lead to errors in calculation, and velocities are not accounted for explicitly.

To obtain both position and velocities at each time step, the **leapfrog algorithm** was developed, except that both values are obtained with a difference in phase as positions are calculated at $t + \Delta t$ and velocities at $v + \frac{1}{2}\Delta t$.

Nowadays, the most common algorithm is the **Velocity Verlet**, which not only stores both position and velocity for each time step but also eludes the need for initial velocity estimation as it is *self-starting*. The basic steps are as follows:

1. Calculate half-step velocity

$$\boldsymbol{v}\left(\frac{\Delta t}{2}\right) = \boldsymbol{v}_0 + \frac{\boldsymbol{F}_0}{2m}\Delta t \tag{20}$$

2. Update coordinates

$$\boldsymbol{r}(\Delta t) = \boldsymbol{r}_0 + \boldsymbol{v}\left(\frac{\Delta t}{2}\right)\Delta t \tag{21}$$

3. Calculate force

$$\boldsymbol{F}(\Delta t)$$

4. Update velocity

$$\boldsymbol{v}(\Delta t) = \boldsymbol{v}\left(\frac{\Delta t}{2}\right) + \frac{\boldsymbol{F}(\Delta t)}{2m}\Delta t \tag{22}$$

Obtaining both accurate positions and velocities is not only convenient but essential when calculating velocity dependent properties of the system like energy or coupling thermostats to maintain constant conditions of temperature, pressure, etc.

On the first note, **energy** along the trajectory of a MD simulation can be approximated with the total Hamiltonian (23) derived from position and momentum of each particle of the N atoms system.

$$\hat{H} = \hat{T} + \hat{V} = \sum_{i=1}^{N} \frac{p_i^2}{2m_i} + V(r) = \frac{1}{2} p \cdot m^{-1} \cdot p + V(r) \tag{23}$$

Thermodynamic ensembles represent the microscopic state of the system including its thermodynamical properties. Different conditions can be simulated: *canonical* ensemble (NVT), where a fixed number of atoms (N) is kept under constant volume (V) and temperature (T); *microcanonical* (NVE), with fixed energy instead of temperature; *exothermic-isobaric* ensemble (NPT) in which pressure (P) is static; *grancanonical* ensembles (μTV) with a fixed chemistry potential (μ).

Thermostats (and barostats) ensure that the mentioned properties are kept constant by altering a variable accordingly. For instance, a heat bath adapts velocities to maintain the average temperature, or a pressure bath alters the volume of the system. Several methods are available such as Berendsen or Nosé-Hoover, which have formulas for both thermal and pressure stability, while others are specific like Langevin or Andersen.

Solvent implications

To mimic reality, simulations include solvent either in an explicit or implicit manner. For explicit representations, the molecular system is embedded in a solvent box with a certain volume. In this case, forces cannot be distributed equally along the whole of the solvent box, since the bulk and the surface would not experience the same interactions. To solve this issue, **periodic boundary conditions (PBC)** are implemented by creating an infinite lattice of system-solvent boxes that behave equally. If a molecule leaves its starting cage, instead of flying off into vacuum it will enter the neighboring box, with an exact image appearing in the initial quadrant.

Nevertheless, this elegant solution presents a challenge to replicate **long-range interactions** that may encompass longer distances than the boundary. Therefore, a commonly applied solution is the use of cut-off distances to limit the separation between atoms accounted for. For Lennard-Jones potentials, to soften the transition in the near cut-off distance, methods like **shifted potentials** are introduced, in which the interaction is gradually reduced. For Coulomb electrostatics, Ewald sum is used in refined algorithms like **Particle Mesh Ewald (PME)**, which calculates the long-range interactions in a reciprocal space with the computationally approachable Fast Fourier Transform and interpolates the charges back to the particles.

Enhanced sampling

As mentioned before, the PES represents the interconnections of all conformations available for a specific biomolecule, including the information on energetic cost for transitions from one structure to another. These transitions are key in biomolecular function, especially for proteins. Some large conformational changes may have a high energy barrier which, in nature, implies a slow transformation. However, time scales in computation are limited to ranges of μs, which forbids the access to certain wells of the PES. To overcome this problem**, enhanced sampling techniques** have been developed to allow the full exploration of the PES.

Some methods are based on biasing the trajectory by guiding the exploration along certain reaction coordinates or collective variables (CVs), e.g. restricting the distance between two atoms or fixing the root-mean-square deviation (RMSD) of a region of the molecule. Some examples are metadynamics, umbrella sampling or steered molecular dynamics. However, selection of appropriate CVs for the desired event can be somewhat tricky and can limit the sampling.

On the other hand, methodologies that avoid biasing are available, based on altering some condition of the simulation (temperature, pressure, potential…) to reduce the energy barrier between wells in the PES. *Replica*-exchange molecular dynamics (REMD) or Gaussian accelerated Molecular Dynamics (GaMD) belong to this category. We will further discuss the latter as it has been employed in several projects of this thesis.

Accelerated MDs apply a potential boost to the real potential of the system, which can introduce statistical noise for energetic reweighting (recalculating the real potential energy of the system). Gaussian accelerated MDs handle this problem by specifically introducing a harmonic potential boost following a Gaussian distribution, allowing the reweighting through accessible transformations like cumulant expansion to the second order. The potential boost can be applied to specific terms, like the total potential or the dihedral; also, a dual-boost option where both terms are enhanced is available.

GaMDs act by adding a harmonic boost potential when the real potential $V(r)$ of the system is under a given energy threshold *E*. The resulting altered potential $V^{*}(r)$ has the form

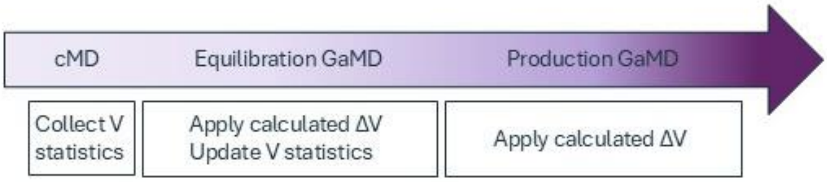
$$V^{*}(r) = V(r) + \Delta V(r)$$

$$\Delta V(r) = \begin{cases} \frac{1}{2} k_0 \frac{1}{V_{max} - V_{min}} (E - V(r))^2, & V(r) < E \\ 0, & V(r) \geq E \end{cases} \tag{24}$$

Both *E* and *k₀* are parameters that can be adjusted given that they meet certain criteria: 1) two potentials v_1 and v_2 should maintain their relative relation (i.e. $v_1 > v_2$ then $v_1* > v_2*$); 2) potential difference between two certain potential values should be reduced; 3) k_0 factor must be between 0 and 1, so that increasing k_0 results in higher potential boost applied.

The energy threshold value *E* can be set to a value between V_{max} and $V_{min} + \frac{1}{k}$, with $k = k_0 \frac{1}{V_{max} - V_{min}}$, and k_0 will be calculated correspondingly. It is noteworthy that k_0 must be small enough for proper reweighting, which is regulated through a user-defined limit $\sigma_0\ applied$ to the potential boost standard deviation ($\sigma_{\Delta V}$):

$$\sigma_{\Delta V} = k(E - V_{av})\sigma_V \leq \sigma_0 \tag{25}$$



Scheme 3.1. Key steps in GaMD simulations.

Methodology

The program employed to carry out GaMD simulations in this thesis is AMBER. Before launching the enhanced sampling exploration, a short, conventional MD simulation is recommended to obtain a well-arranged, stable conformation to start with. Within the GaMD protocol there are three main parts, with a schematic representation shown in Scheme 3.1. Firstly, a few conventional MD steps are run, where the initial potential statistics are gathered. Then, the equilibration stage is entered, with some steps of preparation where the potential boost is already applied; then, 50ns of equilibration dedicated to updating the statistical parameters as well as calculating the potential boost parameters E and k take place. Note that during this second equilibration the potential boost is still applied. Finally, the GaMD production phase is entered, with the previously calculated parameters fixed to apply the desired potential boost.

Variations of GaMD to particularly model ligand binding events have also been introduced in the AMBER suite. These LiGaMDs include another boost on the ligand's dihedral potential.

Molecular Docking

Overview

Molecular docking predicts the interactions and preferred orientation of a molecule bound to another, such as an organic ligand bound to a protein or a metal complex to a peptide. It estimates binding modes and affinities through a scoring function at a rather low computational cost that allows fast results.

It has widespread applications, from elucidating structural information not available empirically to structure-based drug design. In recent years, the latter has gained interest, especially from pharmaceutical companies, as techniques like **Virtual Screening** allow the contrasting of large libraries of ligands against a receptor, giving molecular scaffolds with higher probability to work as future marketable drugs.

Docking programs require two main types of algorithms: a **search algorithm** to explore the available conformational space and to find the possible orientations and a **scoring function** to evaluate and rank these poses based on estimated binding affinity.

Sampling

In the case of the receptor, exploration can be reduced to a specific area containing the known binding site or it may need to be extended to the whole structure to identify which are suitable spots for molecular interaction. The latter has a higher computational cost and is known as **blind docking**.

Initially, molecular docking would simulate binding processes in a lock-and-key mechanism, considering both ligand and receptor as static molecules that would fit together in a **rigid docking**. Nevertheless, other mechanisms are present in nature like induced fit, in which the presence of the ligand can trigger a conformational adaptation on the receptor to better complement the interactions. From a computational perspective, the latter is far more challenging to reproduce but several methods have been developed to approximate it, introducing mobility of the interacting molecules in the calculation.

Ligand flexibility can easily be achieved through rotatable bonds or pre-generating a pool of conformers to test. **Receptor flexibility**, on the other hand, can be included through *side chain flexibility* (for proteins), *softened potentials* that allow some atomic overlap by reducing steric clashes penalties or *ensemble docking* with a set of structures extracted from MD or Normal

Modes calculations. Additionally, the **ligand-receptor complex** can be minimized to achieve more stable interactions before evaluating the resulting pose.

Three main types of algorithms have been developed to drive the search through the conformational space. **Shape matching** looks for structural complementarity between ligand and receptor in a rigid manner, e.g. DOCK software. In a **systematic search** all possible conformers of the ligand are docked in the binding site; this is very exhaustive and can be more expensive computational-wise. Some solutions to reduce costs include fragmentation, applied in the Ludi software. Finally, **stochastic methods** randomly generate the geometries tested and then evaluate them, which greatly reduces computational cost. More than one algorithm can be found in this category, such as *Monte Carlo* (MC), *tabu search*, *swarm optimization* (SO) or *evolutionary algorithms* (EAs). The latter is employed by GOLD or GaudiMM, the programs used in this thesis.

EAs are based on nature's concept of evolution. More specifically, **genetic algorithms (GAs)** use "genes", just like those transmitted from parents to children, to store conformational characteristics of ligand and receptor. Through an iterative process, the initial generation of these genes is recombined and mutated (through crossover and mutation operators respectively) to obtain a new generation of children. Each new set is evaluated through the designated scoring function and those with highest scores are then included in the subsequent iteration.

Scoring

Scoring functions use physicochemical parameters to estimate binding energy. These functions can be very complex to provide realistic results, but this comes at a high computational cost. Therefore, in molecular docking accuracy is sacrificed to find a good balance with speed, which is one of the technique's fortes.

Binding energy approximations can be derived from different approaches:

Force field based: non-bonded potential terms are used to estimate ligand-receptor interactions, namely electrostatic and van der Waals. Solvent effects can be accounted for through a dielectric constant dependent on point charge distance, but entropic effects cannot be contemplated. In GOLD,[82] the *Goldscore* function follows this method:

$$Goldscore = S_{hb_{ext}} + S_{vdw_{ext}} + S_{hb_{int}} + S_{vdw_{int}} \tag{26}$$

Empirically based: energetic terms are weighed based on empirical values, which should come from a high-quality data set. Another scoring option available in GOLD belonging to this category is *Chemscore*:

$$\begin{aligned}\Delta G_{binding} = \Delta G_0 + \Delta G_{hb} S_{hb} + \Delta G_{metal} S_{metal} \\ + \Delta G_{lipo} S_{lipo} + \Delta G_{rot} H_{rot}\end{aligned} \tag{27}$$

Knowledge based: parameters are derived from statistical analysis of available empirical data, assuming that higher frequency of a certain interaction means more favorable energy. *Potential of Mean Force* (PMF)[100] is a common approach within this group.

Machine learning (ML) algorithms like random forest or deep neural networks use different data sets to train the model and extract the parameters for the functions. These have shown quite good performance, accounting for intricate interactions that are difficult to represent with other methods. Nevertheless, the results are very dependent on the training set used.[101]

3.1.3. Metals in Molecular Mechanics

Overview

Regardless of the tuneability of FF and docking schemes, one of the species that remains a challenge to model in MM approaches is metals. As mentioned before, electrons are not considered explicitly in these methods, so accounting for the various oxidation states and coordination geometries when generating the parameters for metallic compounds is not as straight-forward as for some other standard elements. To overcome this issue, three main models have been developed: bonded, non-bonded and dummy atoms.

Bonded models represent the metal interactions as a fixed covalent bond, with the corresponding covalent and non-covalent parameters derived from QM calculations. This approach allows the exploration of metal effects from a structural point of view, but it is limited in showing dynamic variations in the coordination mode.

Non-bonded models simply account for the metal connectivity through Lennard-Jones long range and electrostatic interactions. This allows flexibility but is unable to reproduce a coordination bond with its nuances regarding ligand or protein residue affinity.

Dummy atom approaches depict the coordination geometry around the metal with spheres representing fictitious atoms that also interact with the biomolecule of interest. This technique provides a convenient compromise, fixing geometry but allowing ligand exchanges. However, it requires large empirical parametrization in a rather complicated procedure that handicaps its applicability.

Metals in Molecular Dynamics

For the work contained in this manuscript regarding Molecular Dynamics, the bonded model is the most suitable, as the interest lies in exploring the effects of metal coordination on the structure of proteins or peptides and not the dynamics of the metal coordination itself. For this reason, the tool of choice when parametrizing metals is MCPB.py provided by Amber, which is based in the Seminario method for constant extraction and RESP fitting to obtain partial charges. The program consists of a pipeline of several steps: 1) identifying the coordinating sphere of the metallic ion; 2) providing the QM input files to achieve geometry optimization, vibrational frequencies and partial charges; 3) extracting the force field parameters; 4) fitting the charges and creating the corresponding files; 5) combining all the intricate information regarding connectivity and parameters for the full system.

For step 2, two different models are used. A smaller version with the essential atoms to elaborate the bonded parameters and a larger representation to account for the charges. In step 3, for actual values calculation the Seminario method is applied. It computes the force constants for the harmonic bond and angles from the Cartesian Hessian matrix.

Metals in Molecular Docking

Most docking programs struggle to contemplate metal ions or metallic complexes as possible ligands, usually due to the lack of readily representation of metal coordination in scoring functions. Some approaches applied by programs like *AutoDock* include force field-based representation of metals through van der Waals and Coulombic interactions, lacking proper description of coordination geometry or complex electronic behavior. Other options are the manual set up of a predefined geometry or the implementation of geometric restraints to satisfy given coordination rules.

For this thesis, docking calculations are carried out with *GOLD*. The selected solution for metal interactions was introduced by considering coordination bonds between a fictious hydrogen atom that acts as acceptor and the corresponding residue side chain that acts as donor.[84] The fictious hydrogens are organized around the metal in the corresponding coordination geometry, and the parameters for each pseudo-covalent bond are included in the parameter file of the program. These interactions will be accounted for in the hydrogen bond intermolecular term ($S_{hb_{int}}$in (26)) of the Golscore scoring function.

Predicting metallic interactions

Numerous efforts have been dedicated to predicting metallic binding sites in proteins. Several programs are based on sequence analysis of the target protein to identify typical metal-binding motifs, like the case of *MetalDetector*[102] or *ZincFinder*.[103] Other cases exploit characteristic conformational arrangements associated with metal binding sites like the structural Zn. Oher tools that are not based on one specific motif but on structural factors suitable for metals are available like *BioMetAll*.[104]

BioMetAll is based on metrics collected from the MetalPDB[105] database regarding any metal-binding biomolecule. It has analyzed two values specifically: the coordination distance between the metal and the α-carbon of the donor residue and the angle between the metal, the α- and β-carbons. In this way, it can predict backbone arrangements that could favor metal binding, not based on previously observed patterns or side chain disposition. This allows *de novo* metal-center creation in multiple scaffolds, not limited to naturally occurring tendencies. The program allows further customization of the search, like setting a minimal number of coordinating residues or the inclusion of backbone atoms as potential donors.

One of the most interesting features for this thesis is the "mutations" tool, where BioMetAll is given a binding motif of interest —for instance the motif formed by two His residues— and suggests pertinent positions to mutate to the desired amino acids to create the metal binding site. This application will not perform the mutations or introduce the metal atom in the model, but it comes in handy for necessary previous steps of locating potentially coordinating spots from a 3D structural scaffold in a very fast manner. Most recent versions allow further customization including other metrics of the side chain or filtering for specific metallic elements. Moreover, it can rank the identified coordinating sites based on a ML-based scoring function.

3.2. Experimental Chemistry

The empirical methodology followed throughout this thesis will be described in detail in the experimental sections of the corresponding chapters. However, it is worth dedicating a special mention to the basic principles of **Solid Phase Peptide Synthesis (SPPS)**, as it is the synthetical foundation for all the experiments carried out.[106,107]

SPPS has become one of the most widely used techniques for peptide assembly since its first appearance in the 1960s due to its versatility, efficiency and rapid execution. The fundamental principle involves anchoring of the first amino acid to a solid mount -i.e. a resin-, followed by iterative addition of protected amino acids until the full sequence is obtained (Figure 3.2).

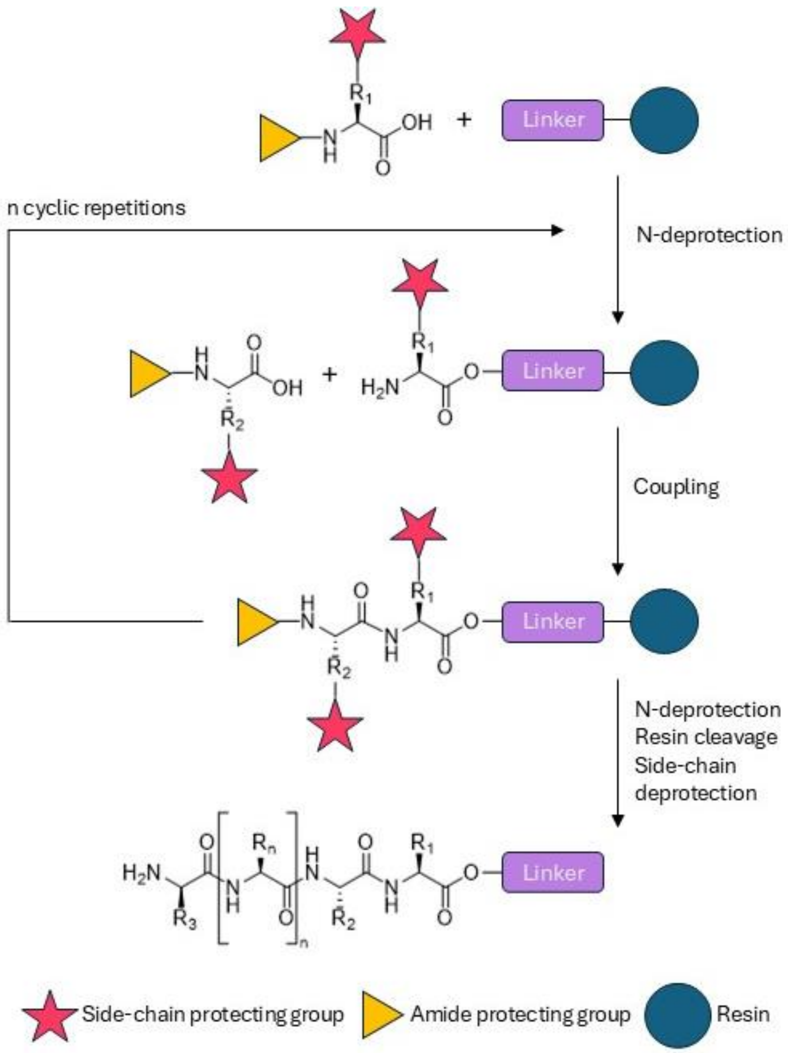


Figure 3.2. Schematic representation of the process of solid phase peptide synthesis.

Coupling of each new residue takes place through the carboxylic acid of the incoming monomer, so the resin must have a free amino group to initiate the process. To prevent cross reactivity, each incoming residue has the amino terminus temporarily protected. Moreover, any potentially reactive side chain groups must be orthogonally capped, meaning that they will only be deprotected under specific conditions that will not interfere with the amide bond formation between residues.

Fmoc/tBu is the most common strategy employed, named after the protecting groups used for N-terminal and side chain capping respectively. Fluorenylmethyloxycarbonyl (Fmoc) is deprotected in basic media after treatment with piperidine, leaving the amine group accessible for the next amino acid coupling. *Tert*-butyl (tBu) is removed from all side chains with trifluoroacetic acid (TFA) after the whole sequence is synthesized.

The key step in the process is the amide bond formation between amino acids, which requires activation of the C-terminus. Most common procedure is the use of **coupling agents** that generate active esters *in situ*. Depending on the desired reactivity, several compounds are available ranging from carbodiimides to phosphonium salts (Figure 3.3).

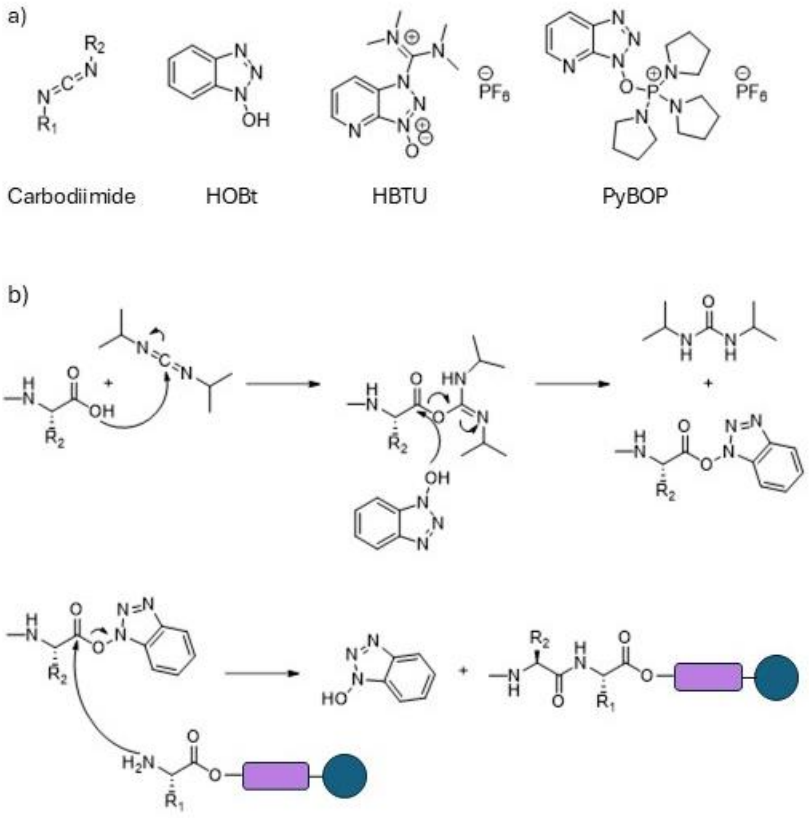
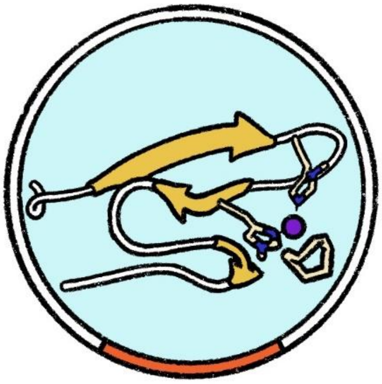


Figure 3.3. a) Examples of most common coupling agents. b) Amino acid activation for SPPS with diisopropylcarbodiimide (DIC) and 1-hydroxybenzotriazole (HOBt)

SPPS is often referred to as *excess chemistry*, as it relies on using reactants in large amounts to ensure coupling completion and quantitative yields in each step of the synthesis. However, despite its efficiency, this technique requires purification follow up, to eliminate possible side products derived from truncated sequences or other undesired reactions. It can become a complex process as the subproducts may present similar chromatographic properties to the target peptides. Nevertheless, this synthetic technique can be readily automated, allowing the synthesis of up to 60 amino acid long sequences in a very convenient, streamlined process, which has made peptide chemistry broadly accessible.



CHAPTER 4

Catalytic metallopeptides: *de novo* design of peptide scaffolds for Pd intracellular reactivity

4.1. Overview

In the context of expanding the arsenal of bioorthogonal tools available, in this chapter two different ways of approaching the exploration of new peptide scaffolds to achieve efficient catalytic reactivity within the cellular environment are exposed. Firstly, starting from a well-characterized, structurally stable protein domain, computational rational design is applied to devise potential mutants to coordinate a Pd(II) metal ion. The second project unifies a combinatorial synthetic protocol with computational rationalization to understand key aspects of successfully active palladopeptides.

For both studies, the catalyzed reaction is the depropargylation of the fluorogenic probe HBTPQ'. The compound is non-fluorescent until the reaction takes place, providing a convenient way to follow the catalysis not only during *in vitro* assays but also when observing the inside of cells through fluorescence microscopy. Propargyl protecting groups have been used in peptide uncaging, pro-drug activation or antibody-drug conjugates reversible formation strategies to take advantage of *in cellulo* transformations carried out by Pd or Cu derivatives, hence the ability to catalyze this deprotection without interfering with any other cellular events is a powerful tool in medicinal chemistry.[108–110]

The specific biomolecules attached to the metal complex determine the stability of the system in living media and its capability of internalization, among other crucial factors discussed previously (see Introduction chapter). Numerous examples of coiled coil structures or α-helices that have been attached to metal centers to create successful catalytic metallopeptides can be found in the literature. Recently, Mascareñas et al. reported a helical peptide able to enter the cell and catalyze the desired depropargylation reaction, although it is unable to function if the cells are first incubated with the catalyst and then exposed to the substrate, which indicates rapid deactivation of the mini-metalloenzyme in the cellular media.[13]

For this thesis, β-sheet-based structures are studied as alternative scaffolds that provide larger and more tunable surfaces to create a protected and yet accessible metal binding center. To achieve a rational design approach, a computational protocol has been developed to analyze these structures, identify potential regions to create suitable metal binding sites and assess their structural stability.

4.2. Protocol for rational design of a mini-enzyme based on the WW domain

4.2.1. The biological scaffold

The peptide of interest in this work is known as WW domain, a 33-40 amino acid long sequence characterized by two very conserved tryptophan residues (W) that give its name to this fold. The WW domain is present in many natural proteins such as Yap65 (Yes-associated protein) or Pin1 (peptidyl-prolyl cis-trans isomerase NIMA-interacting) involved in protein-protein interactions, specifically binding proline-rich sequences. WW is among the smallest naturally occurring protein domains that can fold into a stable conformation without additional structural aid such as disulfide bridging.[111] It self-organizes into a triple strand anti-parallel β-sheet, all three threads connected by three or four amino acid-long turns. The characteristic Trp residues are in the first and third strands, and a strictly conserved Pro can be found in the third strand as well. Interactions between β1 and β3 are known to be detrimental for the stability of the fold, while the second strand contains highly conserved hydrophobic residues usually involved in binding to its target proteins.[112] Moreover, it presents a slight bending inducing a "bowl-like" conformation, with the N and C terminus interacting in the convex part of the structure and the β2 strand hydrophobic core facing the concave side.

In sight of a stable fold that would provide the opportunity to create a more protected environment in the concave part of the structure, WW domains were selected as the peptide scaffold to modify by introducing Pd-binding residues —i.e. histidines– to create a proper catalytic center for the depropargylation reaction.

4.2.2. Objectives

The end-term goal of this project is to apply a rational design protocol to generate a bioorthogonal palladopeptide capable of entering cells and perform its depropargylation reactivity efficiently without rapidly losing activity. However, this thesis focuses on two essential steps to achieve this goal:

a) Development of a multi-scale computational protocol for the design of suitable peptide scaffolds for metal binding site incorporation resulting in a functional mini-metalloenzime.
b) Use the designed palladopeptide to perform depropargylation catalysis in relevant yields.

From a computational perspective, the protocol should be fast and user-friendly, balancing the time cost with accurate results. The rapid approach will be provided by the in-house developed *BioMetAll* program, which will be tested as the base for the design protocol generating putative metallopeptide sequences through its "*mutations*" tool and combining it with Molecular Docking to filtrate the results. Further refinement and accuracy will be attempted by extending to protocol into a second analytical phase through the application of conventional and accelerated Molecular Dynamics (cMD and GaMD respectively). These techniques are more time consuming, especially when modelling the metal-bound systems due to parametrization, but they provide broad sampling to assess the behavior of the selected peptides with and without the metal clip restriction. Additionally, extensive analysis will aid in the rationalization and identification of key components of structure stabilization.

From the experimental side, synthesis, coordination and catalytic activity of the designed peptides will be carried out to select the best option for *in* *cellulo* tests. Moreover, it will provide structural insights to deepen the understanding of the WW-based metallopeptides as well as granting the opportunity to validate the multi-scale computational protocol.

4.2.3. Methodology

General steps

The general outline of the steps taken in this work include 1) an analysis of potential mutation sites with BioMetAll; 2) propose double mutants and assess them through fast MM approaches like BioMetAll and docking; 3) synthesis of the sequences and testing catalytic activity *in vitro* and *in cellula*; 4) obtain structural data to fully characterize the best mutant (Figure 4.1).

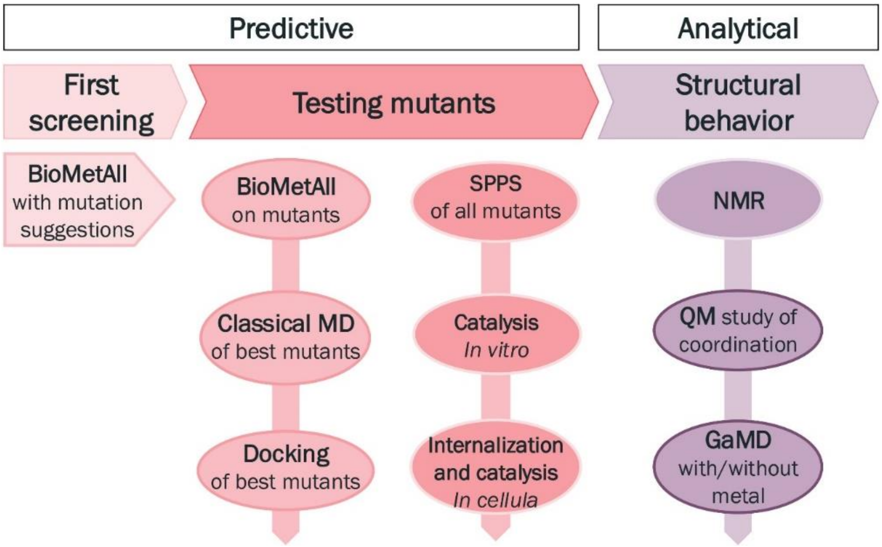


Figure 4.1. Scheme of the methodological steps followed in this work.

Computational predictive protocol and analysis

To generate the pool of structures for the *BioMetAll* analysis, short stabilizing MD simulations were carried out. The initial structures were the X-ray coordinates in PDB entry 1E0M, 1E0N and 1ZR7. For WW0 simulation, the initial structure was set up from 1E0M, mutating and eliminating the corresponding residues in UCSF Chimera. The setup of the system was organized with the program *leap* from Amber suite, using force field FF14SB for the peptides and GAFF for any remaining atoms. The sequences were immersed in a cubic box of TIP3P water molecules extending 10 Å from the surface of the peptides. Chloride ions were added to neutralize the charges. MD simulations were run in the OpenMM engine following a standard protocol: 2000 steps of energy minimization, heat up of water molecules and side chains from 100 to 300 K and MD production under periodic boundary conditions. For the crystallographic structures, 50 ns of production were simulated, while 100 ns were collected for reference peptide WW0. Clustering of these trajectories was performed using the k-means algorithm implemented in *cpptraj* from Amber.

BioMetAll was set up to search for mutations at any position to achieve a His-His coordinating motif while avoiding distances shorter than 2 Å of backbone or side chains. Results were analyzed with an in-house developed script, counting all mutations suggestions. This version of the program evaluates the relevance of a coordination position based on the number of probes that represent viable metal interaction with a given residue. Mutation suggestions were considered valid if their probes represented at least 60% of the highest values.

For the designed mutants with the best results in the second *BioMetAll* screening, new MD simulations were sent using the Amber engine. Nevertheless, the protocol from previous MD trajectories was maintained, except that the production was extended to 200 ns. Again, the clustering of these trajectories was carried out with *cpptraj*, and most representative frames were selected for docking studies.

Docking calculations of the metal ion into the selected mutants were done with *GOLD*, using Goldscore values as evaluators of each pose. Due to Pd and Pt resemblance in reactivity, the metal interactions were modelled through Pt, which has been parametrized using benchmarked Xray data and can be readily implemented in the params file.[113] The square planar coordination geometry was arranged through the dummy H atom protocol described in the methodology chapter.

To study the energetic differences between coordination combinations of the Pd(COD) complex with the two His residues, Gaussian 16 was employed. DFT calculations were performed using B3LYP hybrid functional adding Grimme's D3 dispersion correction. For all non-metallic atoms the 6-31+G(d,p) basis set was used and for the Pd, the Stuttgart Dresden pseudopotential was selected with the corresponding set of *f* polarization functions. The SMD continuum model represented the water solvent and the convergence criterium of the SCF calculations was set at a tight level (1×10^5) to ensure proper frequencies for the relevant vibrational modes. Optimization was carried out free of any restrictions. Zero-point, thermal and entropy corrections were considered for the free energy values considered for the comparison between coordination modes.

The MCPB.py protocol was followed to extract the bonded parameters of the Pd to include in the MD simulations of the metallopeptides. These parameters were calculated with the Seminario method, while the charges were obtained from a RESP model including the COD ligand and the complete His residues. Initially, classical MD simulations were run in Amber, performing three minimization phases of 1000 steps each, minimizing only the solvent in the first one, adding the side chains and metal complex in the second and freeing the whole system for the last one. After heating from 100 to 300 K allowing mobility of only the solvent and H atoms, two equilibration steps were simulated. Firstly, 100 ps of NVT ensembles restraining the peptide backbone and then 500 ps under NPT conditions with the same atomic restrictions. Finally, simulations were run for 200 ns of production.

For the GaMD simulations, the initial coordinates and velocities are taken from the conventional MD simulations that reach convergence. Afterward, 26×10^6 steps of equilibrations (see theoretical background chapter) were followed by up to 500 ns of accelerated molecular dynamics production. These simulations were run in triplicates.

Convergence of all simulations was analyzed through several measurements with MDTraj[114] such as RMSD compared to the first frame and comparing all frames, RMSF, PCA and cluster counting.

To assess conformational evolution of the structures, Ramachandran plots were generated extracting the dihedral angle data with *cpptraj* and depicting the values within density estimates extracted from the RamachanDraw program. MDTraj, NumPy and Matplotlib were used to process the data and print the plots.[115–117]

Experimental synthesis and *in vitro* catalysis

For the synthesis of the peptides, the microwave-assisted peptide synthesizer *Liberty Lite* was employed. A standard Fmoc/tBu procedure was followed, with H-Rink amide ChemMatrix as resin and using five-fold excess of amino acid in each coupling for 4 min at 90 ºC. DIC was utilized as activator, Oxime as base and DMF as solvent. Fmoc protecting groups were truncated using 20% piperidine in DMF at 75 ºC for 1 min. Cleavage from the resin and elimination of side chain protecting groups were carried out with a deprotection cocktail containing 900 µL TFA, 50 µL CH_2Cl_2, 25 µL H_2O and 25 µL TIS, using 1mL for every 40 mg of resin and exposing it for 2 h. The resulting mixture was added to ice-cold diethyl ether for 30 min and the precipitate was centrifuged and washed again with ice-cold diethyl ether. The final solid was dried under argon and redissolved in water.

Peptide purification was carried out through reverse phase HPLC on a semipreparative RP-HPLC with Agilent 1100 series LC with UV-vis detector using a Phenomenex Luna-C_{18} (250 × 10 mm) column. Conditions for the separation were a linear gradient starting at 5 going up to 75% of solvent B during 40 min at a 4mL/min flow rate. Solvent A was water with 0.1% TFA and B acetonitrile with 0.1% TFA. Fractions containing pure products were freeze-dried.

The WW mutants were analyzed under analytical UHPLC-MS on an Agilent 1200 series LC/MS with a Phenomenex SB C_{18} (1.8 μm, 2.1 × 50mm) analytical column to confirm correct sequence. Conditions consisted of a linear gradient 5 to 95% B for 20 min at a 0.35 mL/min flow rate. Again, solvent A was water with 0.1% TFA and B acetonitrile with 0.1% TFA. UV absorption was set at 222, 270 and 330nm. Electrospray ionization mass spectrometry (ESI/MS) was carried out with an Agilent 6120 Quadrupole LC/MS model in positive scan mode using direct injection of the pure peptides into the MS detector.

To generate the metallopeptides, incubation with [$PdCl_2$(COD)] in water at a 1:1 ratio for 1h took place in a 2mL HPLC vial. The metal solution was prepared as 1 μL of 20 mM in DMSO with 1.0 equiv. For coordination, 10 μL of a solution of HBTPQ’ 20mM in DMSO, 1.0 equiv was added to 990 μL of PBS and the resulting solution was mixed with the previously prepared metallopeptide. The reaction mixture was stirred at 1000 rpm for 24 h at 37 ºC. Coordination was assessed by taking 50 μL of the reacting solution and diluting it to 100 μL with MeOH and analyzing it through HPLC-MS(ESI). The calibration curve method was employed to measure reaction yields, using coumarin as internal standard. Every value was the average of two independent measurements.

4.2.4. Results

Initial computational screening

The work of Macias *et al.* 2000 was taken as a starting point for the metallopetide design, where they devised a prototype WW domain sequence to enhance stability through triple β strand-favoring interactions (PDB code 1E0M). This sequence was selected as a template scaffold, modifying it to prevent uncontrolled coordination. For instance, His14 and His23 were substituted by Pro and Thr respectively. Moreover, Asp9 was replaced by Thr for synthetic purposes,[118] and five terminal residues were truncated (two at the *N*-terminal and three at the C-terminal) to favor the WW domain fold. The new sequence generated was the reference peptide for the following experiments, designated **WW0**.

The proposed WW0 structure was examined through CD experiments, which displayed the characteristic spectrum of a properly folded WW domain, with a positive ellipticity band at 230 nm and a negative band at about 210 nm.[119] Additionally, MD simulations corroborated this result, showing a very stable peptide for the 200ns of trajectory (Figure 4.2).

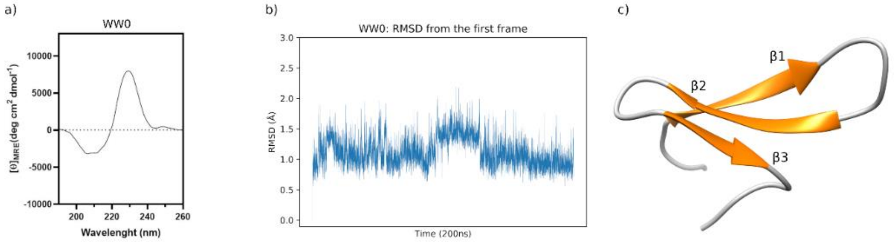


Figure 4.2. Structural analysis of the WW0 reference peptide. a) CD spectrum of WW0, b) RMSD along a 200ns trajectory, c) most representative cluster along the simulation of WW0's antiparallel triple β strand structure.

To search for proper positions to include coordinating residues, a study of mutation suggestions was carried out with BioMetAll. To generate a pool of structures for the analysis, findings presented in Macias et al. were taken into account. They identified two very conserved residues near each terminus of the peptide, a proline and a tryptophan respectively, which appear in different combinations in various natural WW domains, i.e. having both present in their sequence or either Trp or Pro. Four NMR structures from the Protein Data Bank (PDB) comprising all the possible combinations were selected for our analysis, including natural sequences and the previously described prototype (PDB codes 1ZR7, 1E0N and 1E0L for the naturally occurring and 1E0M for the designed sequence). Furthermore, short stabilizing MD simulations were carried out for systems 1E0M, 1E0N and 1ZR7, and representative poses of the two most populated clusters of each trajectory were added to the ensemble. Lastly, two snapshots of the predominant clusters in the WW0 simulation were also considered. Altogether, 12 different structures were analyzed for the BioMetAll study.

The program was instructed to search for spots in the structure that would create a suitable metal binding site if mutated to His residues. Several positions appeared suggested with high frequency near the turn between strands β1 and β2, as well as some residues located closer to the central part of the peptide along all three strands (Figure 4.3a). Moreover, several mutation suggestions were located at the convex part of the structure.

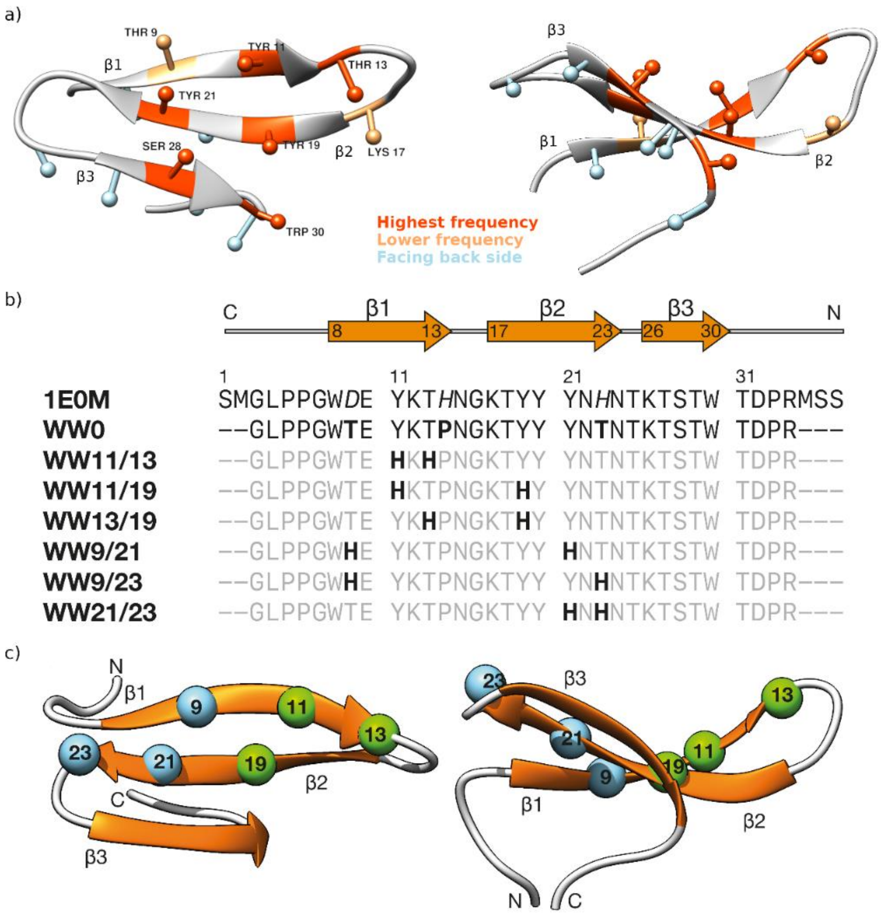
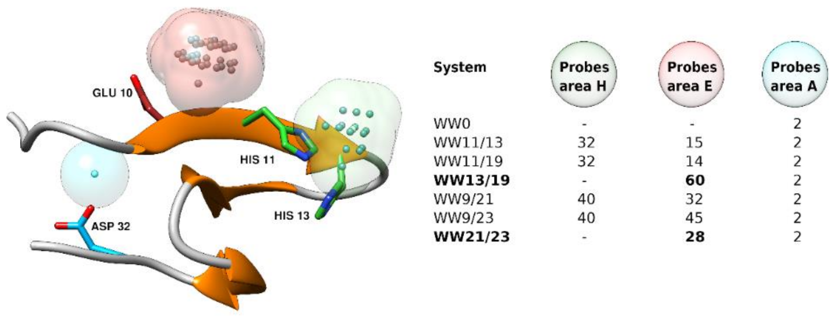

a)
THR 9
TYR 11
THR 13
β1
TYR 21
SER 28
TYR 19
β2
LYS 17
β3
TRP 30
β3
β1
β2
Highest frequency
Lower frequency
Facing back side
b)
C
β1
β2
β3
N
8
13
17
23
26
30
1
11
21
31
1E0M SMGLPPGWDE YKTHNGKTYY YNHNTKTSTW TDPRMSS
WW0 --GLPPGWTE YKTPNGKTYY YNTNTKTSTW TDPR---
WW11/13 --GLPPGWTE HKHPNGKTYY YNTNTKTSTW TDPR---
WW11/19 --GLPPGWTE HKTPNGKTHY YNTNTKTSTW TDPR---
WW13/19 --GLPPGWTE YKHPNGKTHY YNTNTKTSTW TDPR---
WW9/21 --GLPPGWHE YKTPNGKTYY HNTNTKTSTW TDPR---
WW9/23 --GLPPGWHE YKTPNGKTYY YNHNTKTSTW TDPR---
WW21/23 --GLPPGWTE YKTPNGKTYY HNHNTKTSTW TDPR---
c)
N
β1
9
11
13
23
21
19
β2
C
β3
23
β3
21
13
11
9
19
β1
β2
N
C

Figure 4.3. a) Mutation positions obtained from BioMetAll study colored according to suggestion frequency. b) Selected sequences for the experiments represented along the template peptide 1E0M and reference peptide WW0. The coordinating His residues are highlighted in bold for clarity. Names of the sequences correspond to the positions mutated to His residues. c) Selected positions for mutation. In green, those positions that represent the best suggestions from BioMetAll and in blue the less favored combinations.

These were discarded as the metal would be more exposed to the solvent and could interfere with some WW fold-stabilizing interactions (Figure 4.3a, in blue). With the added interest of evaluating *BioMetAll* performance, six final sequences were devised: three combining the highest frequency positions near the β1-β2 turn (WW11/13, WW11/19 and WW13/19, green in Figure 4.3c) and another three on the opposite side (WW9/21, WW9/23 and WW21/23, blue in Figure 4.3c) including high and low frequency positions as well as a mutation not proposed. Modifications in the third strand were avoided, as this is known to be the most flexible segment of the structure and fixing it in a coordination sphere might introduce too high of a restraint.

The new mutants together with the reference WW0 were modeled and evaluated again with *BioMetAll*, this time simply searching for suitable coordination spots. Three main areas were identified, but only one (denominated area H) would result in the desired 2His-Pd complex. One of the other regions seemed to be an artifact (area A), as very low probes appeared even in the WW0 peptide. The third zone (area E) points to coordination between Glu10 and His11, which would require rather constrained orientations in the side chains and is not suitable for Pd binding (Figure 4.4).



System	Probes area H	Probes area E	Probes area A
WW0	-	-	2
WW11/13	32	15	2
WW11/19	32	14	2
WW13/19	-	**60**	2
WW9/21	40	32	2
WW9/23	40	45	2
WW21/23	-	**28**	2

Figure 4.4. Second BioMetAll analysis of the selected mutants. Different metal coordination areas depicted as colored clouds with the probes provided by the program. Correct binding area H in green, undesirable coordination area E in red and possible artifact area A in blue.

The two mutants showing preference for area H coordination were **WW13/19** and **WW21/23**, both including the His residues located at the highest frequency suggested mutations from the first BioMetAll study. To further assess their fitness as metal-coordinating peptides, MD simulations were carried out for both sequences, showing stability of the triple β strand conformation. From the most populated clusters along the trajectories, the most representative frame was selected as receptor for a molecular docking calculation with the metal ion as ligand. For mutant WW13/19 promising poses were identified with Goldscore values of 49, showing possible bonding of the metal with both His residues. In the case of WW21/23 even better scores of 65 were obtained, indicating coordination of the Pd with not only His but also other surrounding residues like Thr9. This could result in poor accessibility of the metal center for the substrate of interest in the catalytic reaction, but further simulations would be necessary to address this question, eluding the "simple, fast prediction" approach aimed for this design protocol.

To summarize, from the preliminary computational design two mutants, **WW13/19** and **WW21/23**, were discriminated against hundreds of possible combinations and were identified as potential metallopeptides.

Experimental synthesis and application

All seven sequences designed previously were synthesized through microwave-assisted SPPS. Any impurities were removed through reverse-phase HPLC and correspondence with the sequence was checked through HPLC-MS(ESI). To coordinate the Pd ion and obtain the metalloprotein, each peptide was incubated in water with [$PdCl_2$(COD)] in a 1:1 ratio for 1h. Coordination was assessed through HPLC-MS(ESI), which was achieved for all mutants except for WW0. All sequences were further characterized through circular dichroism (CD), comparing metal-free and metal-incubated results. For the case of WW0, the peptide maintains the characteristic spectrum of a WW fold mentioned earlier even after incubation with the Pd complex, which agrees with the absence of coordination observed in HPLC-MS results. For the rest of the mutants, CD spectra suggest that the fold differs from the triple β strand independently of the presence of the metal. The only exception is sequence WW13/19, which displays a well-defined WW-fold CD spectrum upon addition of the Pd (Figure 4.5a). This is expected to favor the formation of the metal center and stability of the minienzyme and would point to the results predicted through computational means.

Catalysis was tested with all palladopeptides for the depropargylation reaction of the model fluorogenic probe **HBTPQ’** (Figure 4.5b). The probe was synthesized in a well-established four-step protocol with a 54% yield. To obtain the metallopeptide for the catalysis, a 20μM solution in Milli-Q water of every mutant was mixed with [$PdCl_2$(COD)] in a 1:1 ratio. After 1 h incubation, a 200 μM solution of HBTPQ’ in PBS was added to the mixture, and the reaction vessel was shaken at 37 °C for 24 h. To measure the yield, a calibration curve was elaborated with coumarin as the internal standard. All sequences presented catalytic activity with over 30% yield, even non-coordinating WW0. This was due to the presence of free [$PdCl_2$(COD] in the reaction environment, which displays about 65% yield on its own (Figure 4.5c).

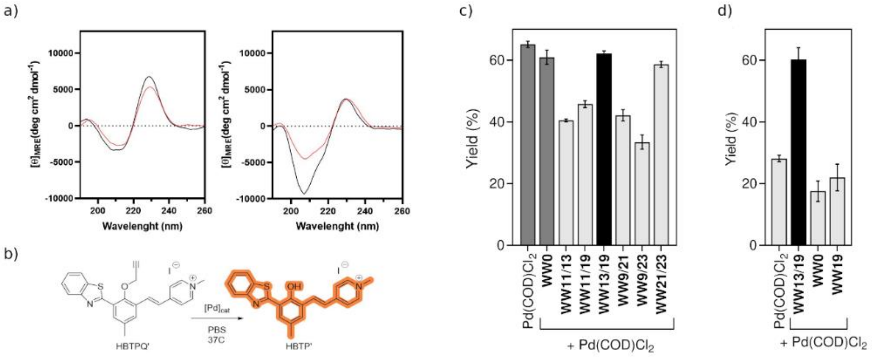


Figure 4.5. a) CD spectrum for reference peptide WW0 (left) and WW13/19 (right). Black trace line corresponds to measurements of the free peptide, red trace line to measurements after incubation with $PdCl_2$(COD) for 1 hour. b) Depropargylation reaction tested using the metallopeptides as $[Pd]_{cat}$. The product HBTP’ is fluorescent. c) Results of in vitro catalysis. Firstly, the metallopeptides are generated incubating the sequences with $PdCl_2$(COD) for 1 h to then add a 200 μM solution of HBTPQ’ and mix for 1h. Control

experiments with the Pd(II) salt and the non-coordinating WW0 sequences are depicted in dark grey. d) Same experiment but after ultradiafiltration to remove the Pd(II) salt excess. One extra control experiment was conducted with the 1His-containing sequence WW19.

To ensure no residual catalysis derived from the Pd complex, an extra step of ultra-diafiltration with Amicon centrifugal filters was added before setting the 24h reaction, to separate any molecules in the solution weighting less than 3kDa. In this case, WW0 yield decreased significantly, in line with previous observations discarding coordination of the His-free peptide (Figure 4.5d). Moreover, a variant with only one His residue in position 19 was tested to assure the need for double coordination (WW19), which showed yields in the range of WW0, confirming that the double anchor of both His residues is necessary. Regardless of the procedure, the mutant achieving the best performance was WW13/19, with a reasonably good yield of 62% close to that of the Pd complex alone, followed by WW21/23. Remarkably, these results are in excellent agreement with the computational predictions presented in the previous section.

The following experiments were not carried out by the author of this thesis but will be briefly commented on to give proper context to the chapter.

Given that WW13/19[Pd(II)] was the best catalyst *in vitro*, its potential for application inside mammalian cells was studied. Firstly, internalization experiments were performed with a tetramethylrhodamine-labeled version of WW13/19 to monitor cellular uptake through fluorescence microscopy. HeLa cells were incubated with either the palladated and free TMR-WW13/19 peptide. Notably, the metallopeptide showed considerably better results than the free structure. Quantification through cell cytometry confirmed that TMR-WW13/19[Pd(II)] was up to 6000 times more efficient in entering the cell. Additionally, further testing indicates that, once inside the living unit, the catalyst accumulates in endosomal vesicles suggesting that internalization most likely occurs by micropinocytosis (see Appendix A for further details).

Finally, intracellular catalysis was tested with the same HBTPQ' fluorescent probe. WW13/19[Pd(II)] was able to catalyze the reaction, either incubating the cell lines with the probe before or after exposure to the metallopeptide. This indicates that the minienzyme is robust enough to stay functional inside the cell environment. Moreover, a turnover of 9 was calculated, proving that the designed mutant can undergo catalytic cycles. As a control, incubation of cells with the palladium complex or WW0 mixed with [$PdCl_2$(COD)] did not show any fluorescence under the same conditions (see Appendix A for further details).

In order to gain more structural insight, NMR spectra of WW0 and WW13/19 with and without metal were obtained by the group of M. Macias (see Appendix A). These confirmed all the findings: no alteration appeared upon addition of the Pd complex to sequence WW0, which already presents the characteristic spectrum of a well-folded WW domain; on the other hand, the vague chemical shift dispersion of WW13/19 shows clear rearrangement after adding the metal, indicating that 1) the free peptide does not have a well-defined conformation and 2) coordination drives WW domain-like folding. Moreover, spectra show that the actual coordinating species corresponds to Pd(COD), with the metal center maintaining a square planar coordination with the two histidine residues and the organic ligand.

Dynamical analysis

Once WW13/19 was determined as the best mutant for coordination and catalysis, and the metal center was identified as Pd(COD)His_2 through NMR, a new set of calculations was set up to identify the specific coordination pattern through the imidazole moieties. All combinations of δ and ε nitrogen bonding to Pd were optimized at the DFT level of theory, obtaining vibrational frequencies to ensure simulation of minima in the PES (Figure 4.1). The combination His13ε and His19δ yielded the lowest energy, although the worst value was only about 2kcal/mol higher. This would suggest that tautomeric exchange is possible, which is backed by NMR results: the undefined shape of the Trp-Hε resonance peaks are indicative of coexistence of different Pd-N coordination. It is worth noting that the combination His13δ His19δ was impossible to obtain, as the optimization would reproduce proton transfer between residues, so this possibility was discarded.

a)

Model	ΔG (kcal/mol)
εδ	0
δε	0,764
εε	1,760
δδ	-

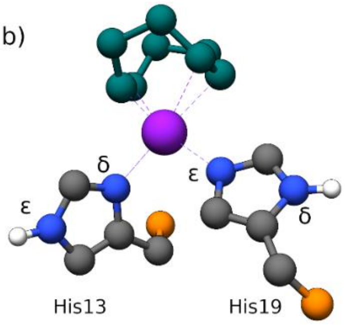


Figure 4.6. a) Coordination modes tested and their corresponding relative Gibbs free energy from DFT calculations. b) Sample structure of the binding center for the Pd(II) metal (purple) with the two His residues and the cyclooctadiene ligand (green).

Having the optimal coordination geometry, parameters were extracted to reproduce accurate metal-peptide interactions and MD simulations of the catalyst were run. In this case, enhanced sampling was performed to ensure proper exploration of the conformational landscape. GaMD simulations of WW0, WW13/19 and WW13/19[Pd(COD)] showed stable WW fold in all cases for the first 100 ns of trajectory. For both WW0 and the metallopeptide, after these initial steps the third strand increased its flexibility, alternating between double and triple β strand conformation (Figure 4.7a and b). As mentioned before, this behavior is still typical of

WW domains. Simulations of the free WW13/19 peptide show that after the initial 100ns the starting conformation is lost and the domain is unfolded, concordant with NMR results.

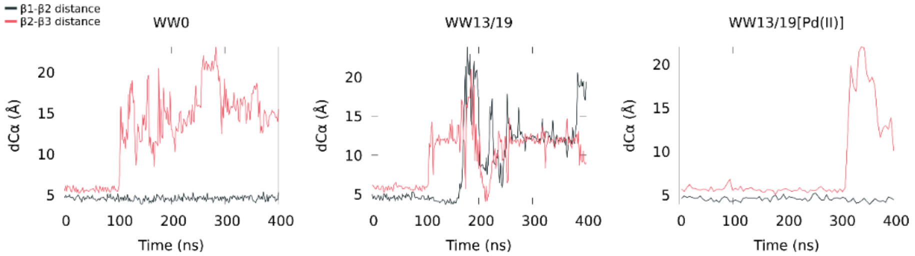


Figure 4.7. The evolution of distance between each β strand for WW0, WW13/19 and WW13/19[Pd(II)]. Measurements are taken between the Ca of the central residue in each strand. The black trace line represents Glu8-Tyr18 distance and the red line Tyr18-Ser26 distance along GaMD simulations.

These simulations provide valuable insights into the previously mentioned differences in cellular uptake between free WW13/19 and its metallated counterpart. Literature examples report that conformational restrictions such as cyclization or stapling can enhance cell penetration.[120,121] Therefore, the structural rigidity derived from the metal coordination to the core of the peptide may contribute to the improved internalization obtained for the palladopeptide.

4.2.5. Conclusions

The computational protocol for the design of a metal binding scaffold based on the WW domain was successfully applied. The remarkable results obtain with *BioMetAll* were crucial in the predictive steps, firstly suggesting suitable mutation spots and then filtering the best performing sequences. Extending the protocol to molecular dynamics simulations allowed a deeper understanding of key factors such as metal-enhanced stability of the peptide fold, probably facilitating cell penetration events.

Straightforward synthesis of the sequences allowed the validation of the computational protocol and confirmed structural tendencies predicted through modelling. *In vitro* assays demonstrated the efficient catalytic activity of the metallopeptide, rendering yields close to that of the Pd complex alone.

Finally, the end objective of the project was achieved by successfully performing a depropargylation reaction *in cellulo* using the designed metallopeptide built from the WW domain and the Pd(COD) complex. This biocatalytic activity within the complex biological media is not observable without the peptide, highlighting the effective cooperative behaviour of the metallic ion within the biomolecular scaffold.

4.3. Combinatorial libraries and molecular modeling: understanding key structural aspects in the design of stable metallopeptides

4.3.1. The biological scaffold

For this study, the selected peptide scaffold is known as tryptophan zipper (TrpZip) due to the four tryptophan residues that face one side of the peptide intercalating each other, responsible for the stability of the characteristic β-hairpin fold adopted by these sequences. The typical length of these peptides varies from 12 to 16 residues, depending on the turn sequence, and they fold cooperatively with a remarkably favorable folding free energy due to the mentioned cross-strand pairing of Trp residues. The amino acids forming the turn can vary, including turn-inducing residues like asparagine, glycine or D-proline; for this work, the peptides present a type II' β turn adopted by amino acids Gly and Asn.

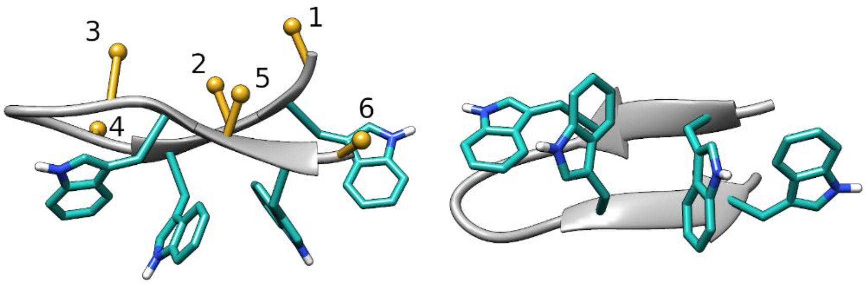


Figure 4.8. Representation of the TrpZip peptide structure. Mutation positions studied are numbered and highlighted in yellow.

Similarly to WW domains, these peptides represent the shortest sequence to adopt a stable tertiary structure without the need of additional structural restrains like disulfide bonds. Moreover, their two faces are very well distinguished, bearing one of them all the indole side chains responsible for stabilization interactions and leaving the other one open to modifications to insert metal-binding residues to create the reactivity site. Precisely, six alternating positions, three on each strand, are suitable to insert different arrangements of amino acids, setting the perfect template for combinatorial testing including two His residues and four other amino acids. To reduce the combinations to a manageable amount, only four different possibilities were selected based on their propensity to induce β-sheet secondary structures. These residues are alanine, arginine, threonine and valine, which additionally, they cannot compete with the His residues for Pd binding.

4.3.2. Spot synthesis technique

The technique employed to carry out the combinatorial approach of this study is based on SPOT libraries, which were devised in 1992 to perform several reactions in parallel.[122] For this procedure, a cellulose-based matrix is divided in small assigned "spots" where the reagents are specifically located to generate the desired chemicals (peptides for instance) in a way that each well can be positionally identified. In this way, large screening processes are simplified by comprising hundreds of separate "reaction vessels" in a piece of paper.

The spot-synthesis procedure has been widely used in protein-protein interaction studies such as antibody specificity or enzyme substrate analysis among others etc.[123] Particularly for the TrpZip peptides, the depropargylation reaction was carried out on the cellulose matrix and easily followed through fluorescence microscopy, allowing a fast and simple screening approach for catalytic purposes, which has only been attempted once before to compare dendritic and lineal peptides for esterolytic activity in a smaller sample.[124]

4.3.3. Objectives

Three main objectives are aimed at with this work:

a) Combinatorial assessment of best sequences for metallopeptide catalysis through spot technique.
b) Application of the devised computational protocol to predict the best His insertion positions for palladopeptide design.
c) Structural assessment of the designed mini-enzymes and their catalytic performance.

Synergic combination of the spot-library results with the computational prediction tool will be attempted by using the best performing sequences as a filter for the *BioMetAll* screening. Furthermore, metal-binding influence on the structures will be studied through MD to complete the protocol and try to reproduce experimental findings with the objective of confirming applicability of this multi-scaling approach in analysis of Pd-based mini-enzymes.

4.3.4. Methodology

Library generation and spot catalysis

To generate all the combinations for the spot synthesis, a python script was developed to create the sequences following the given criteria: all Trp residues would be maintained in their original positions, as well as the turn residues Gly and Asn; additionally, positions 1, 3, 5, 8, 10 and 12 would be scanned to insert a combination of two His residues with the four residues Ala, Arg, Thr, Val selected for their higher β-sheet inducing character.[125] Under these conditions, 264 different sequences were obtained and synthesized through the spot technique into a cellulose support.

The libraries were synthesized into CelluSPOT slides by the company *Intavis Peptide Services GambH*. The resulting cellulose-supported sequences were transferred to a microscope slide to perform the catalytic screening assay. The protocol started with incubation of the peptides with 10mL of a 1mM solution of the palladium source dichloro(1,5-cyclooctadiene)palladium(II) –PdCl2(COD)– agitating for 1h followed by three washes with PBS during 10min to eliminate any excess of Pd. Then, the fluorescent probe HBTPQ' was added as 50 µL of a 200 µM solution to the slide, which is then covered and sealed with *cytosyl* coating. The emission intensity was recorded under the microscope at times 0, 4 and 24 h. The system was kept under humid atmosphere to prevent the slide from drying out. The experiment was performed on three CelluSPOT slides.

Peptide synthesis and *in vitro* catalysis

The two study case sequences D14 and D4, were synthesized following Fmoc/tBu SPPS protocols, using all amino acid derivatives and reagents from *Sigma-Aldrich* and *Iris Biotec*. All amino acids had Fmoc as *N*-terminal protecting group, and each side chain had standard protection: Fmoc-Ala-

OH, Fmoc- Val-OH, Fmoc-Arg(Pbf)-OH, Fmoc-Trp(Boc)-OH, Fmoc-Thr(t-Bu)-OH, and Fmoc-His(Trt)-OH. All synthesis were carried out with a *Liberty Lite* automatic microwave assisted synthesizer from CEM corporation.The resin used was H-Rink-Amide ChemMatrix from *Biotage AB*, with a load of 0.57 mmol/g. Fmoc-peptide synthesis protocols were followed, on a 0.1 mmol scale coupling in 5-fold excess of each amino acid for 4 min at 90 ºC, using DIC as the activator, Oxime as base and DMF as solvent.

Deprotection of the Fmoc groups was carried out with 20% piperidine in DMF during 1 min at 75 ºC. To cleave the peptide from the resin, the solid was mixed with a deprotection cocktail for 2 h. The cocktail contains 900 μL of TFA, 50 μL of CH_2Cl_2, 25 μL of H_2O and 25 μL of TIS, in a proportion of 1mL of solution for every 40 mg of resin. The resin was filtered and the solution was added to ice-cold diethyl ether for 30 min. The precipitate is centrifuged, washed with ice-cold ether and dried under argon. Finally, it is redissolved in water.

Purification of the sequences was carried out on a semipreparative RP-HPLC *Agilent* 1100 series LC equipped with UV-vis detector. The column used was the *Phenomenex Luna*-C^{18} (250 × 10 mm) reverse-phase. Standard purification conditions started with 5% of B solvent, increasing in a linear gradient until 75% of B solvent over 40 min at a flow rate of 4 mL/min. Solvent A was water 0.1% TFA and solvent B acetonitrile 0.1% TFA. Pure peptide-containing fractions were ice dried.

Metallopeptides were generated by mixing each peptide with a solution of $PdCl_2$(COD) in equimolar amounts for 15 min. For the in vitro catalysis, 10 μL of HBTPQ’ 20 mM in DMSO (1.0 eq) were diluted with 990 μL in PBS in a 2 mL HPLC vial with screw cap. 1 μL of 20mM solution of the previously prepared metallopeptides was added (0.1 eq) and mixture was stirred at 1000 rpm for 24 h at 37 °C. After ther reaction takes place, an aliquote of 50 μL was diluted to 100 μL with MeOH and analyzed vy RP-HPLC-MS. Results were evaluated through a calibration curved using coumarin as internal standard. Two independent measures were averaged for each value.

Computational prediction protocol

From the computational perspective, the previously studied protocol was applied in this new system, starting with a preliminary screening for metal coordination (within the mentioned positions) with *BioMetAll*, then docking of the metal with the selected candidates, followed by MD with and without the metal complex and finally GaMD simulations to complete exploration.

For the *BioMetAll* screening, two sets of TrpZip structures were studied. Firstly, only three crystallographic sequences of other TrpZip peptides were analyzed, with PDB codes 1LE0, 1LE1 and 1LE3. Then, a second pool of randomly selected sequences from the spot assay were screened, namely D13, D14, E2, E3, E9, F2, F7, F9, F10, F11 and F13, which contain some of the best, worst and medium performing peptides. These structures were represented in UCSF Chimera and relaxed in the same platform through 110 steps of minimization (100 steepest descent and 10 conjugate gradient), 1000 steps of heating from 0 to 298 K and finally 1000 steps of production under NVT conditions. Finally, a third study case with the 6 best performing candidates, that is F14, D17, D6, H12, D15 and D14, was carried out, generating the mutants, embedding them in a 10 Å cubic box of TIP3P waters, and relaxing them with AMBER in CPU through a minimization, heating and equilibration for 1000 steps each and 5000 steps of production. In all cases, *BioMetAll* was instructed to look among the non-Trp residues to insert mutations to generate His-His coordinating centers.

Computational dynamic analysis

For the study case sequences D14 and D4, conventional MD simulations were run, setting up the system with *leap* from AMBER, using the Amber14SB force field for the protein and GAFF for any remaining atoms. The peptide was solvated in a cubic box of TIP3P water molecules of 25 Å distance to the biomolecule and chloride ions were included to neutralize the system. All variants of different sequences and terminals followed the same protocol, starting with 1000 minimization steps restraining the peptide atoms, followed by another 1000 steps with restrains only on the backbone, 1000 steps of free minimization, then 100 ps of heating from 0 to 300 K, equilibration under NVT conditions for 100 ps with restrains on the backbone and another 500 ps of NPT equilibration. Finally, 500 ns of NPT production were obtained.

Afterwards, *GOLD* docking calculations were performed with the naked metal ion into the most representative clusters of each simulation to setup the initial coordinates for the MD simulations. All automatic settings were selected for the genetic algorithm, and the dummy H atom method was followed to configure the Pd ion in a square planar coordination. Metal parameters for Pt were used as in the previous section.[113]

Pd-containing systems were parametrized with MCPB.py following the Seminario method to extract the force constants and RESP technique to obtain atomic charges. The corresponding metal geometry with the binding residues and COD ligand was optimized in Gaussian16 with the B3LYP hybrid functional applying Grimme's D3 dispersion. The basis set for the metal included the SDD pseudopotential and f-polarization functions, and for the remaining atoms the 6-31+G(d,p) basis set was selected. All calculations were carried out in water as implicit solvent with the SMD model. Finally, the conventional MD simulations followed the same protocol as for the free-peptide systems.

For the accelerated MDs, the simulations start from stable MD coordinates and undergo 26×10^6 steps of equilibration to gather the different parameters for the potential boosts. Then, 500 ns of production were simulated both with and without metal. All MD simulations including conventional and accelerated were run in triplicates.

As in the previous section, trajectories' convergence was assessed through RMSD, RMSF, PCA and cluster counting obtained with MDTraj. Ramachandran plots were generated to represent the conformational changes along the simulations using cpptraj to extract the dihedral angles data and MDTraj, NumPy and Matplotlib to generate the plots, taking the density estimates from the RamachanDraw program. Furthermore, timeline plots of the evolution of secondary structure of each residue were firstly generated with VMD and then edited with gnuplot.[115–117,126,127]

4.3.5. Results

SPOT synthesis, catalysis and selection of study case

A library of 264 sequences containing 2 metal-coordinating His residues was devised from a combinatorial approach. The TrpZip characteristic type β-II' turn formed by Gly and Asn as well as the positions of the Trp residues were maintained in all the designs, and the remaining 6 positions were filled with two His residues and 4 different residues that are known to induce β-sheet stability, namely Ala, Arg, Thr and Val (Figure 4.9). The library was then synthesized into a cellulose-based support, with each sequence occupying a marked "spot" in the slide that made it easy to identify.

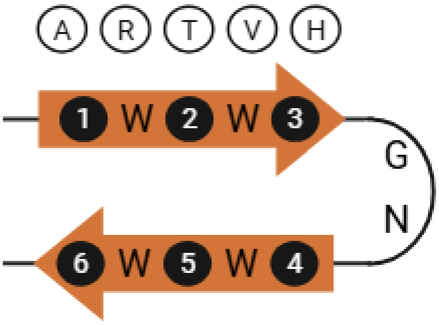


Figure 4.9. Schematic representation of hairpin structure with the screening positions

The CelluSPOT library was treated to transform the sequences into metallopeptides and test their catalytic activity with the HBTPQ' probe. The slides were incubated for 1h with dichloro(1,5-cyclooctadiene) palladium(II) salt in aqueous solution to obtain the minienzymes and then the fluorogenic probe was added. The reaction was followed under a microscope, irradiating the sample at 330nm and measuring the emission at 635 nm. The triplicate measurements resulted in very reproducible patterns, especially for identifying the best and worst performing sequences.

Analyzing the top 15 performing peptides in the spot assay, the most repeated combinations of His positions correspond to 1-3, 3-10, 5-8 and 8-10. The intention of this work is to stabilize the hairpin structure with the metallic clip, to ensure some degree of rigidity which has been shown to favor cell penetration. For that reason, sequences with the His residues in opposites strands of the structure are of greater interest for the global aim of the project. Therefore, sequence D14 with His residues in positions 3 and 10 was selected as a study case, together with the negative control peptide D4, which presents quite a lower emission in the spot analysis despite only differing in the amino acids in positions 1 and 5: Ala1/Arg5 for D14 and Arg1/Ala5 for D4.

Both peptides were synthesized through SPPS and incubated with the metal salt, studying their coordination through mass spectroscopy. A M/z peak corresponding to the metalated peptide was obtained only for D14, while D4 exclusively presented signals consistent with the free peptide even after 24 h of incubation. Fluorescence titration confirmed that the affinity of sequence D14 for Pd(II) is significantly higher than that of peptide D4, with dissociation constants of 1.0 μM and over 1 mM respectively (Figure 4.10a). Additionally, CD experiments show that only D14 changes its spectrum upon addition of $PdCl_2(COD)$ (Figure 4.10b).

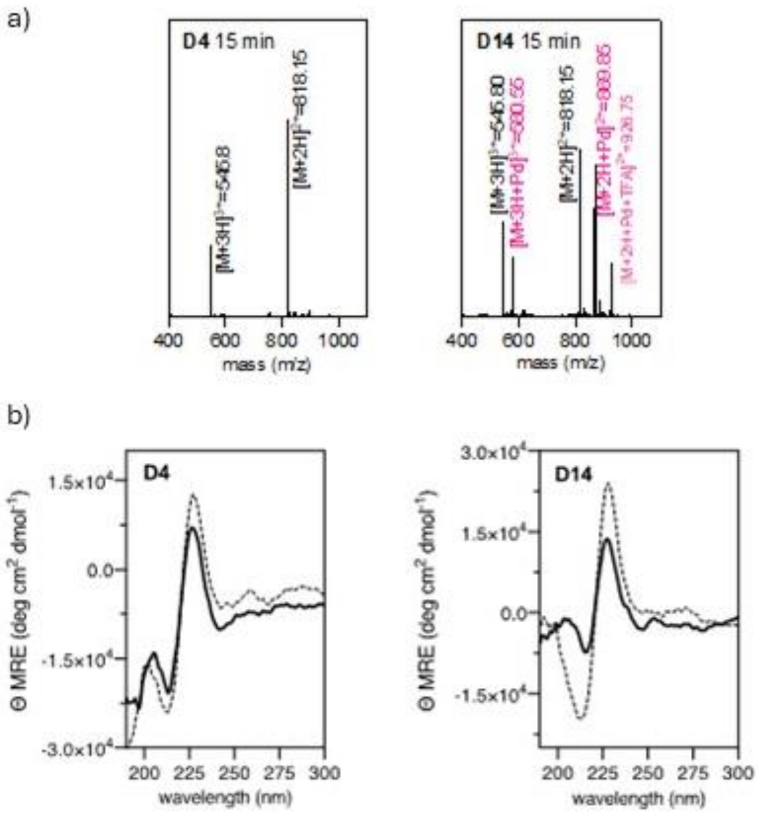


Figure 4.10. a) EM-ESI+ (m/z) spectrum. Peaks corresponding to Pd binding are indicated in red. b) CD spectrum of free peptide (solid line) and after metal incubation (dashed line).

NMR spectrum of free D14 shows poorly dispersed peaks indicating that the uncoordinated peptide mostly presents an extended conformation in solution. Nevertheless, increasing concentrations of the Pd(II) salt induce a better dispersion of the peaks, indicative of a defined secondary structure coherent with a β-hairpin fold. Moreover, no peptide aggregation was observed in the mM range for the D14 TrpZip (Figure 4.11).

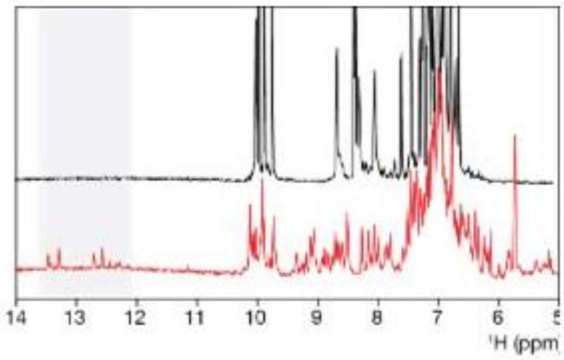


Figure 4.11. 1D NMR spectrum of free **D14** sequence (black line), and after incubation with Pd(II) salt. The highlighted region shows peak dispersion in the His region due to coordination of the metal.

In vitro assays of both metallopeptides revealed promising depropargylation yields of 56% for [Pd(II)]D14 while much lower values of 16% were obtained for the D4 sequence as it was expected from the poor metal binding affinity.

BioMetAll screening

With the aim to compare the combinatorial screening with the computational approach described in the previous chapter, an initial BioMetAll screening was performed on three TrpZip structures from the PDB (codes 1LE0, 1LE1 and 1LE3. Between the six positions acceptable for coordination 5 of them were found as putative sites, with preference for combinations 8-10, 1-3 and 3-12. A second screening was carried out giving as input a larger library of structures, extracted from the spot library sequences without evaluating their performance. In this case, the best combinations of positions to insert a His-His binding motif were 1-3, 3-10, 5-8 and 8-10. Remarkably, these positions coincide exactly with the results obtained from the top 15 performing sequences in the spot assay.

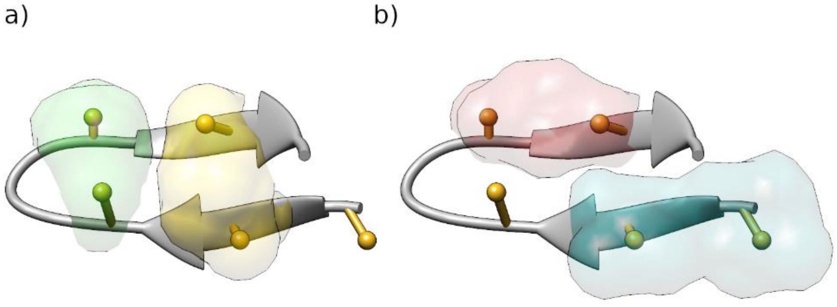


Figure 4.12. Best combinations of positions to insert His residues according to both spot-synthesis screening and BioMetAll predictions. a) Depicts combinations between the two strands, b) depicts combinations between the same strands.

Then, to test the synergy between experimental and theoretical approaches and try to narrow down the options, the 6 best performing peptides from the spot library were selected, adding an extra layer of filtering to the *BioMetAll* analysis from the experimental results.

It must be highlighted that the program was instructed to analyze all the residues except for the Trp and point out the best positions to create a double His coordination site. Also, despite using sequences that already contain His, the program evaluates equally all positions based on backbone preorganization and does not consider the side chain coordinates. It is worth noting that in the first run with only three structures the program is able to discard Asn, which is located in the turn and was already not contemplated from the experimental design point of view. Furthermore, position 3 is always found amongst the selected points for mutation regardless of the structures provided for the *BioMetAll* testing, and in the last two experiments based on the combinatorial sequences, the combination 3-10 appears as one of the top suggestions both in *BioMetAll* and in the spot assay.

BioMetAll was able to reproduce the top performing combinations of His placement in the sequence, even without prioritizing the known top catalysts from the experimental results.

Study case: D14 *vs* D4

B- hairpin stability in free peptides: terminal capping influence

To understand what dictates affinity for metal complexes and performance of the resulting metallopeptide, modeling of both study case sequences D14 and D4 was undertaken. Furthermore, since the experimentally studied peptides present the C-terminal amidated as a result of cleavage from the resin used in SPPS, the study was amplified to assess the influence of terminal capping on the interaction network and the general behavior of the structures, as previous studies have shown that capping might affect some aspects of β-sheet forming structures although their relevance with respect to binding of the metal is not well understood.[128] All possible terminal capping combinations (both N- and C-terminals protected with acetylation and amidation respectively, only N- or C- terminal protected and both deprotected) were reproduced in a computational model and submitted to conventional MD simulations of 500 ns in triplicates, so that their secondary structure and contact evolution could be analyzed (Figure 4.13).

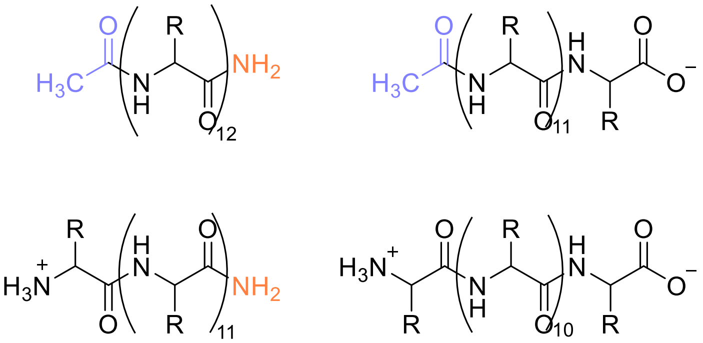


Figure 4.13. Capping combinations studied. From left to right and top to bottom: both N- and C- terminals protected, N-terminal acetylated, C-terminal amidated, both N- and C-terminals deprotected.

All terminal residues —Ala1/Thr12 in D14 and Arg1/Thr12 in D4— displayed an unorganized secondary structure independent of the presence or absence of a capping group. The second and second to last amino acids —both Trp residues in every peptide— maintained a generally stable β-sheet conformation along every simulation except for the Ac-peptide-COO$^-$ case, where it showed a higher tendency to become unorganized, a behavior that was more remarked in the D14 sequence. Another general trait independent of the terminals and sequences was observed regarding the residues right before and after the turn —Arg5/Val8 for D14 and Ala5/Val8 for D4—, which interchanged between β-sheet and turn conformations along the trajectory.

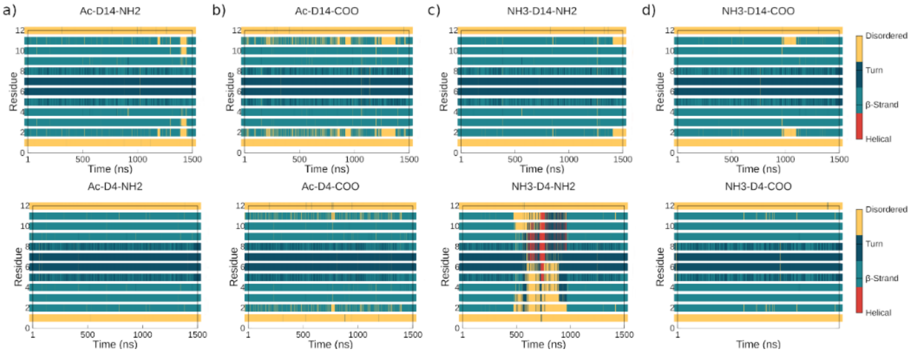


Figure 4.14. Timeline of the 3 replicates of 500ns for each terminal combination for D14 (top) and D4 (bottom) peptides. The colors represent simplified types of secondary structure for clarity.

Regarding the contacts, a polar interaction network was identified among the most stable simulations, connecting specifically residues 5 to 8, 3 to 10 and 2 to 11 through hydrogen bonds between the backbone N-H and O atoms (Figure 4.15). Depending on the terminal, an extra hydrogen bond appeared for the fully protected Ac-peptide-NH_2 and the NH_3-peptide-NH_2 sequences; moreover, for the unprotected NH_3^+-peptide-COO^-, a strong electrostatic interaction was observed between the charged terminals.

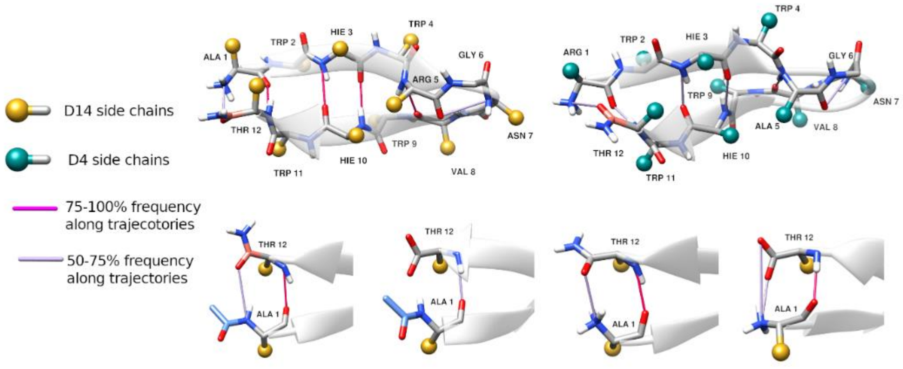


Figure 4.15. Contacts analysis along MD simulations. Top left, D14 peptide. Top right, D4 peptide. Bottom, different contacts between the respective terminal capping.

From these results, the most stable β-hairpin structure was obtained for the fully unprotected NH_3^+-peptide-COO^- systems (Figure 4.14d), which presented some disordered structure in residues 2 and 11 for sequence NH_3-D14-COO^- at the beginning of one of the replicates but otherwise stayed very compact for the remaining simulation time. This can be explained by overstabilization due to the ionic interactions between the oppositely charged ammonium and carboxylate groups. This trend also prevailed in an analysis of distance between terminals, which maintained an average distance of 4 Å or higher in all peptides except for the free-terminal systems, which present a distance close to 3.5 Å (Table 4-1). Moreover, the difference in average distance between the catalytic D14 sequence and non-catalytic D4 is the lowest for unprotected terminals, which highlights the strong driving force represented by the electrostatic interaction.

The fully protected sequences also present quite a stable structure except for one of the replicates of Ac-D14-NH_2 again, which shows disordered organization for residues 2, 3, 10 and 11 during about a fifth of the simulation time for that replicate (Figure 4.14a).The extra hydrogen bonding between the amide hydrogen of the capped *N*-terminal residue and either the capped C-terminal carboxylic oxygen or its hydroxyl sidechain elongate the polar interactions network along the peptide backbone, adding a β-hairpin stabilizing effect. On the opposite side, the systems that generally show less preservation of the hairpin arrangement are the Ac-peptide-COO^- (Figure 4.14b), showing recurrent disorganization (residues 1, 2, 11 and 12) and turn (5, 6, 7 and 8) conformation more frequently. During these simulations no apparent interactions with the charged C-terminal take place.

Even more extreme is the specific case of NH_3^+-D4-NH_2 simulations, that clearly show how during one of the replicates the hairpin conformation is disassembled completely after less than 100 ns, creating a dynamic interchange between helical and random coil conformations. Due to this replicate, the average terminal distance difference between both NH_3^+-peptide-NH_2 sequences is the highest and the stability of the hydrogen bond network along the trajectory is the lowest among all structures. However, if only the two trajectories that maintain a stable structure are considered, the mean distance descends to 3.92 Å (0.48 Å lower than D14). Since these peptides represent the experimentally tested capping arrangement, these results point out a possible relationship between preorganized structure stability and catalytic activity, but further studies were required to determine if this replicate could be discarded as an outlier.

Table 4-1. Average distance between the nitrogen of the N-terminal and the carboxlic carbon of the C-terminal along MD simulations.

	D14 (Å)	**D4** (Å)	**Δ(D14-D4)**
Ac-peptide-**NH_2**	4.76	4.23	0.53
Ac-peptide-**COO^-**	6.25	5.70	0.55
NH_3^+-peptide-**NH_2**	4.10	5.97	-1.87
NH_3^+-peptide-**COO^-**	3.54	3.61	-0.07

Enhanced dynamics and metal influence

In order to further assess the distinct behavior of NH_3^+-D14-NH_2 and NH_3^+-D4-NH_2 peptides initially spotted both in the uncoordinated systems simulations and in the experimental observations, two approaches were conducted: extension of the conformational exploration with enhanced GaMD and modeling of the metal-coordinated structures for conventional and accelerated dynamics studies.

As mentioned earlier, cMD simulations of both free peptide sequences show a general tendency to β-hairpin stability except in one of the replicates for D4. When extended with enhanced GaMD sampling, both sequences show inability to stabilize the hairpin conformation and alternate between helical or unordered structures; the singular exception appears in one of the replicates of peptide D14, which maintains a stable β-hairpin throughout the 500ns of simulation with some disordered structure in residues 1, 2, 11 and 12 (Figure 4.16a, first replicate).

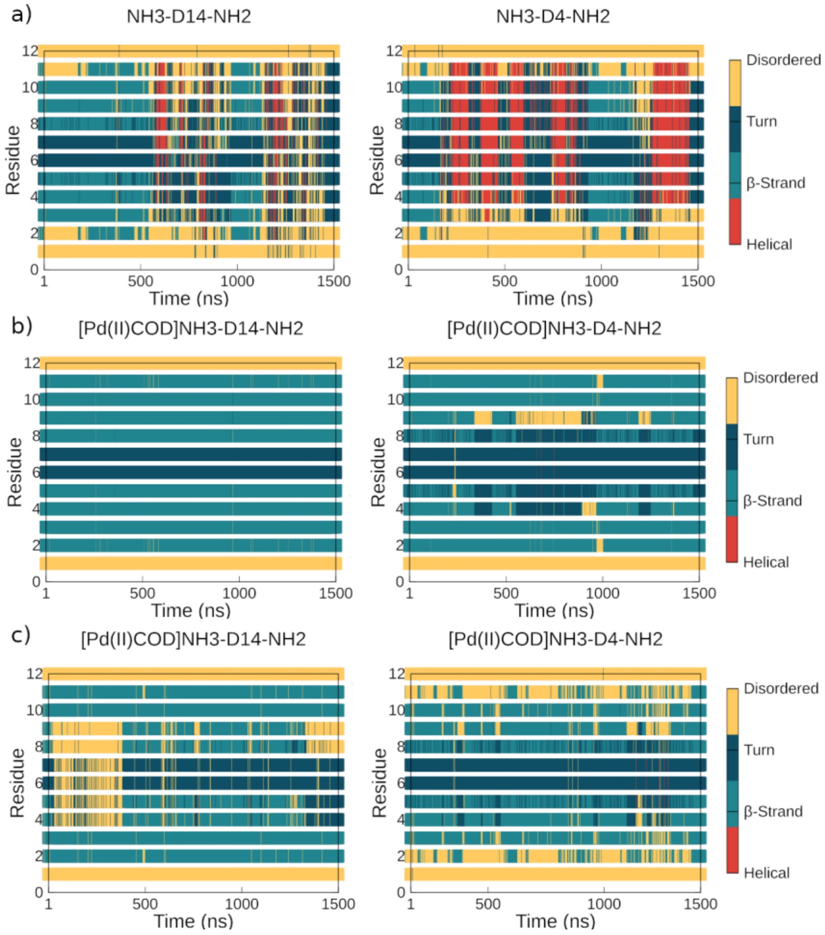


Figure 4.16. a) Timeline of the 3 replicates of GaMD with D14 (left) and D4 (right) peptides. Each replicate is 500 ns. b) Timeline of the 3 replicates of conventional MD with Pd(II) complex-bound D14 (left) and D4 (right) peptides. Each replicate is 500 ns. c) Timeline of the 3 replicates of GaMD with Pd(II) complex-bound D14 (left) and D4 (right). Each replicate is 500 ns.

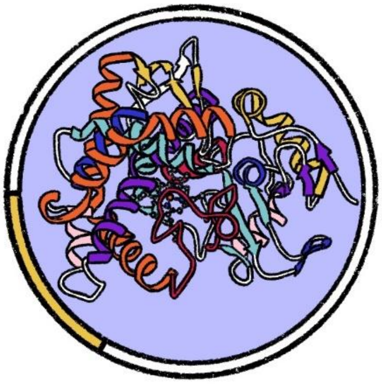
On the contrary, the conduct observed for the metallopeptides is radically different attending to the peptide sequence. During conventional simulations of [Pd(II)D14], the β-hairpin fold appears extremely stable throughout the whole trajectory, with the two strands in antiparallel sheet conformation and the Gly6-Asn7 turn well defined. For the enhanced sampling, the β-II' turn appears momentarily extended to neighboring residues like Trp4, Arg5, Val8 or Trp9 or deformed into coiled coil but the overall structure is rather stable. However, [Pd(II)D4] shows extension of the β-II' turn organization through most of the MD simulations and in GaMDs all residues experience conformational strain along the simulation (Figure 4.16b and c).

These results point out that some prearrangement specifically favoring coordination might take place, but most likely both sequences would show proper organization for Pd binding. However, the difference outstands once the coordination takes place, as D14 sequence readily adapts to the constraints imposed by the metallic complex but D4 is not able to sustain a stable conformation as a metal ligand. These would explain why no formation of the metallopeptide is obtained for NH_3^+-D4-NH_2 sequence, as it is unable to adopt a stable conformation with the metallic clip between the His residues.

4.3.6. Conclusions

The spot synthesis was successfully applied to scan the catalytic activity of a library of 264 TrpZip sequences. The best and worst performing peptides were consistently identified through fluorescence microscopy. Parallelly, the devised computational protocol predicted the best performing combinations for the insertion of the metal-binding His residues through *BioMetAll* analysis, even when no experimental information was used to bias the input sequences towards the best performing peptides. This confirmed the potential of this tool in the design of metalloenzymes based on structural information exclusively.

One of the best sequences, D14, catalyzed the depropargylation reaction *in vitro*, confirming the effectiveness of both experimental and theoretical screening processes. Furthermore, dynamic simulations provided a rationale to the lack of coordination and, hence, catalysis of the very similar D4 sequence. This was derived from the conformational constraints imposed by the metal complex, which were well tolerated by the D14 structure but prevented the D4 sequence from adopting a stable β-hairpin conformation.



CHAPTER 5

Metalloenzymes: structural insights into cytochrome P450 catalysis from multiscale study of CYP199A4

5.1. Overview

5.1.1. Cytochromes P450's catalytic cycle

As introduced earlier, cytochrome P450 enzymes (also referred to as P450s, CYP or CYP450) have been widely studied for their capacity to perform oxidative reactions that are otherwise hard to obtain through synthetic means like activation of inert C-C bonds. These enzymes are monooxygenases that utilize molecular oxygen as a reactant, allowing the performance of oxidative reactions in good yields and with high atom economy. For these reasons, they have been extensively studied as biocatalysts for many years, although, due to their complexity, many questions remain unsolved. In particular, two issues are of major interest:

1. Decoding the rules of substrate specificity. CYP450s present in nature can either be extremely specific or widely promiscuous. From a biocatalytic point of view, assessing which substrate(s) can bind to a given P450 is fundamental.
2. Characterizing catalytic selectivity. CYP450 can perform a wide range of reactions. Although the most common is the hydroxylation of C-C bonds, its oxidative power can lead to other transformations like epoxidation, *N*-demethylation or *O*-demethylation. Those reactions also tend to present regio- and stereospecific patterns which, in many cases, could result from a combination of factors like the electronic properties, shape or size of the substrate, the energetic profile of specific steps of the catalytic mechanism, the binding of the substrate or its pre-catalytic conformation.

Understanding the relative weight of each molecular variable that drives CYP reactivity is a real challenge and often requires individualized analysis for each P450. An extensive amount of work has been reported in the literature in this respect. Major efforts have been dedicated to describing the main steps of the catalytic cycle, which in its initial resting state (RS) presents the porphyrin ring characteristic of type-B heme molecules coordinated to ferric Fe, with the two axial coordinating positions occupied by the S atom of a cysteine residue in the proximal side and a water molecule in the distal side. During the cycle, to activate the dioxygen molecule two electrons need to be transferred from associated electron transfer partners that use NAD(P)H to sequentially deliver the reductive particles at determined steps of the cycle.[39,129] Although the main type of reactions catalyzed by cytochromes P450 are hydroxylations, in this work the focus is placed on *O*-demethylation reactivity. A representation of the currently accepted version for the catalytic cycle in this reaction is reproduced in Figure 5.1.

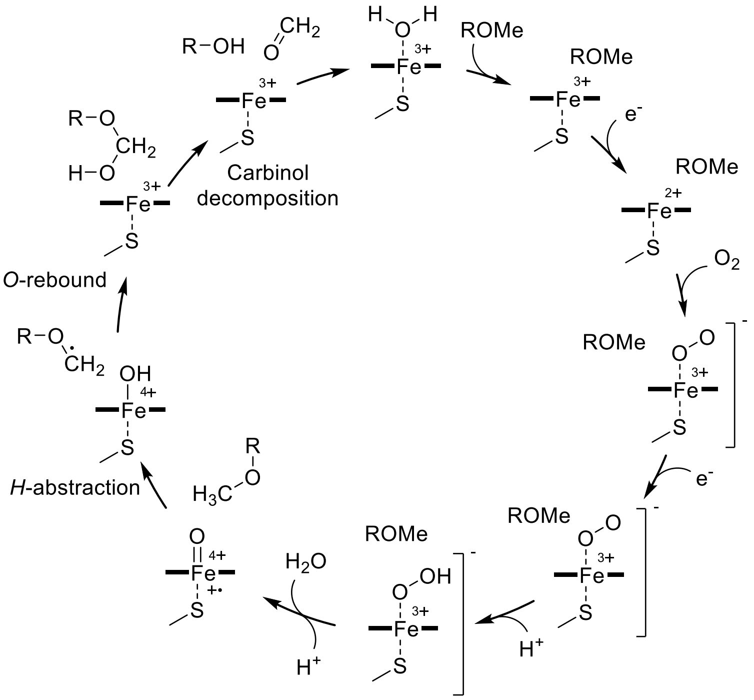


Figure 5.1.*O*-demethylation catalytic cycle

The resting state of the enzyme presents the water molecule bound at the sixth coordination site of the octahedral iron atom in oxidation state +3, occupying the distal part of the catalytic pocket. This water molecule is displaced upon substrate entrance, triggering a spin state shift from low to high spin. Then, the first reduction of Fe(III) to Fe(II) takes place, thanks to the transfer from the reductase partner of the P450. Then, coordination of the molecular oxygen occurs at the now vacant distal site. At this point, the second electron transfer occurs, reducing the dioxygen to produce a peroxo intermediate. Generation of the reactive species is obtained by double protonation of this peroxo complex, firstly obtaining a hydroperoxo species also known as *Compound 0* (Cpd0) and then breaking the O–O bond in a heterolytic manner to render *Compound I* (CpdI).[130] CpdI is responsible for most of the reactivity performed by this enzyme as a high-valent iron-oxo species with the Fe in a formal oxidation state IV and a porphyrin radical. CpdI reacts with the substrate to first abstract a H atom generating a substrate radical that then reacts with the *Compound II* (CpdII) in an oxygen rebound step. Finally, a carbinol decomposition step takes place, yielding the hydroxylated compound and a formaldehyde side product.[131] At the final stage, a water molecule is rebound at the axial coordinating position as the product leaves, regenerating the resting state and restarting the catalytic cycle.[132] However, this final step has not been well characterized yet. Slight modifications to the described final steps occur depending on the catalyzed reaction and the substrate.[133]

Many questions remain unclear with respect of CYPs reactivity, more specifically regarding dynamic structural influence, molecular origin of their specificity or resting state regeneration after catalysis, all of which can be critical points in the rational design of enzymes. In collaboration with the group of Prof. Stephen Bell from the University of Adelaide, Australia, we aim at decoding some of the key factors influencing these questions from the computational perspective.

5.1.2. Experimental background on CYP199A4

All the work developed in this thesis has been focused on the cytochrome CYP199A4, a soluble enzyme derived from *Rhodopseudomonas palustris.*[134] R. *palustris* is a Gram-negative bacteria found in nature that can adapt to widely diverse environments such as marine coastal sediments, earth-worm droppings, or pond water. To do so, it has developed the ability to switch among four types of metabolic routes attending to the energy and carbon source: *photoautotrophic* and *photoheterotrophic*, where light is the energy source and carbon is obtained from either CO_2 or organic compounds respectively, and the equivalent *chemoautotrophic* and *chemoheterotrophic* metabolism types in which organic molecules are transformed for energy. Further flexibility is displayed by the bacterium, as it can grow in aerobic or anaerobic conditions, it can fix nitrogen for ammonia generation or can use diverse organic and inorganic electron donors. This is possible thanks to an extraordinarily rich number of metabolism-dedicated genes that encode a wide variety of enzymes to adapt to fluctuating sources of carbon, nitrogen, light and oxygen. For this reason, R. *palustris* is an attractive candidate as a study model in many biotechnological fields such as biocatalysis or bioremediation. [135–138]

The group of Prof. Stephen G. Bell has studied some of the enzymes involved in R. *palustris* metabolic pathways, more specifically cytochromes P450 like CYP199A4, CYP199A2 or CYP195A2 among others. Over the years, they have created a rich database of experimental information including more than 70 crystallographic structures of different proteins as their wild type sequences or mutated versions. Most systems have been characterized either in a ligand-free state or bound to several substrates, inhibitors or products. The Australian group has identified benzoic acids as the primary substrate for the protein, which has high selectivity for the negatively charged, planar moiety but displays distinct affinities depending on the substituents of the benzene ring.[139] They also investigated the performance of CYP199A4 for a large variety of reactions, which include *O*- and *N*- dealkylation, sulfoxidation, amide and cyclic hemiacetal formation, hydroxylation, desaturation or epoxidation.[140,141]

Of particular interest, the best performing substrate is 4-methoxybenzoic acid, which undergoes *O*-demethylation to yield 4-hydroxybenzoic acid (Figure 5.2a). This *para*-substituted species presents a strong binding with a dissociation constant of 0.22 μM and an excellent catalytic activity with a product formation rate of 1220 min^{-1}. Strikingly, its close *meta*-substituted analogue 3-methoxybenzoic acid presents a far weaker dissociation constant of 69 μM and does not produce any metabolite (Table 5.1). This observation is particularly surprising because of the similarity between both species; a certain amount of metabolic deviation between both systems could be expected but not to the extent characterized experimentally. Moreover, other *meta*-substituted reactants like 3-methylamino or 3-methylthiobenzoic acids present better affinities and higher product formation rates, although this happens for *N*-demethylation and sulfoxidation reactivities respectively.[141] The *para*-substituent regioselectivity of CYP199A4 has also been observed in other organic compounds of interest like veratric acid, a lignin derivative precursor of other molecules of industrial interest like vanillin.[142] This molecule is substituted both on *meta*- and *para*- positions although the reaction takes place selectively on the latter.

Table 5.1. Substrate binding and catalytic activity data for CYP199A4 with different substrates. % HS corresponds to spin shift induced by water displacement from the distal site of the heme group. Kd is the dissociation constant. NADH and PFR stand for NADH

oxidation rate and product formation rate respectively, measured in nmol·(nmolCYP^{-1})·min^{-1}. Benzoic acid is abbreviated as BA. All values are extracted from Coleman *et al.* 2018 except for veratric acid, which was extracted from Bell *et al.* 2010.

Substrate	% HS	Kd (M)	NADH	PFR
4-methoxy BA	≥95	0.22 ± 0.02	1340 ± 28	1220 ± 120
3-methoxy BA	40	69 ± 2	498 ± 5	-
3-methylamino BA	10	31 ± 1	255 ± 2	175 ± 1
3-methylthio BA	30	33 ± 0.5	56 ± 0.2	37 ± 1
Veratric acid	70	29.5 ± 3.1	1101 ± 59	1061 ± 88

Further characterization of the substrates binding and reactivity are also available (Table 5.1). For instance, the required spin shift resulting from water displacement from the RS upon substrate binding can be experimentally measured, as it induces a blue shift (from ~418 to ~390 nm) in the Soret band in the UV-vis spectrum of the enzyme.[141,143,144] This value, referred to as %HS, is indicative of the binding affinity of the compounds, similarly to the dissociation constant K_d, and is also indicative of catalytic cycle initiation after the displacement of the Fe-coordinated water molecule. Moreover, catalytic cycle activation is measured through NADH oxidation and product formation rate, which are given as nmol·nmolCYP^{-1}·min^{-1}. A closer look at these values for both the *meta-* and *para-* substituted methoxybenzoic acids reveals a significant difference in the percentage of high spin switch induction %HS, with over 95% for 4-methoxybenzoic acid and only 40% in the case of 3-methoxybenzoic acid coherently with the K_d observations. Moreover, this is directly correlated with the NADH oxidation and product formation rate, which are in the ranges of 1340 and 1220 respectively for the *para*-substituted substrate meanwhile in the *meta*-substituted analogue experiments only the NADH oxidation rate could be measured, which is more than 3 times lower.

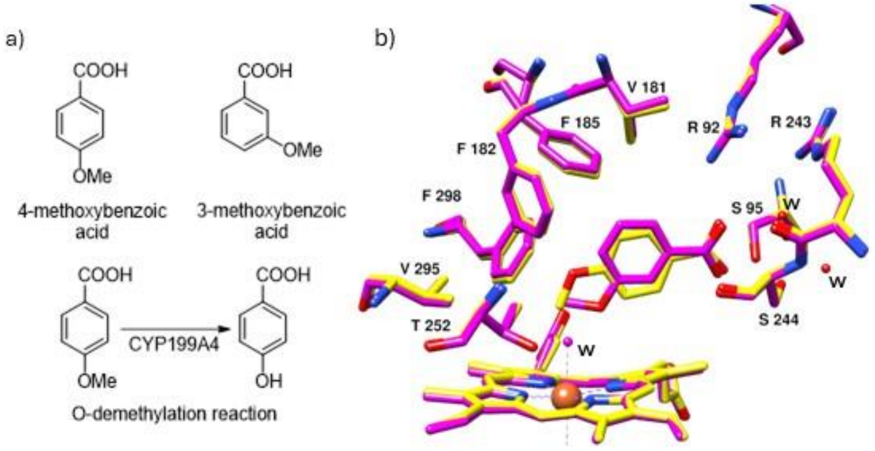


Figure 5.2. a) Substrates studied in this work for the *O*-demethylation reaction catalyzed by CYP199A4. b) Superimposed crystallographic structures of the *para-* (yellow) and *meta-* (magenta) substituted benzoic acid-bound systems.

These differences can also be appreciated in the crystallographic structures characterized for both substrates (Figure 5.2b). As expected from the reduced %HS induction, 3-methoxybenzoic acid crystallizes together with the Fe-binding water molecule in the binding site while the *para*-substituted equivalent presents no trace of the aqueous ligand. Regarding the orientation of the ligands, the carboxylic acid is almost in the same position and the arrangement of the residues in the binding site around this functional group is nearly the same for both structures. However, this leads to a small shift in the orientation of the benzene ring, which is slightly displaced from the heme group in the case of the *meta*-substituted substrate when compared with the *para-*, hence leaving the substituents with opposite orientations with respect to the Fe atom: in 4-methoxybenzoic acid the substituent's O atom is far from the metal (5.2 Å) leaving the C and reactive H atoms closer to the Fe (4.1 Å for the C and 3.1 Å for the closest H); meanwhile the O atom in the meta substituent is closer to the metallic center (4.3 Å) and although the carbon is at a similar distance from the metal, the closest hydrogen is almost one Armstrong further from the Fe ion (4.4 Å and 3.9 Å respectively).

Furthermore, some water molecules around the binding site other than the sixth coordinating ligand are crystallized in almost identical positions for both structures. The most remarkable are two molecules located near the carboxylic acid group of the substrates, one acting as a bridge between the ligand and Arg243 and the other possibly connecting the side chain of Ser244 in helix I with the backbone amide of residue Leu96 in the B-C loop. Interestingly, a chloride ion is also found in the exact same place in both crystals, presenting interactions with residues Tyr177 in the F helix, Gln203 in G helix, Arg243 in the I helix and two water molecules (Figure 5.3). Experiments altering the ionic strength and content of chloride demonstrated that its binding does not strengthen the substrate's binding to the enzyme, suggesting that its role could be related to isolating the active site from the solvent.[145]

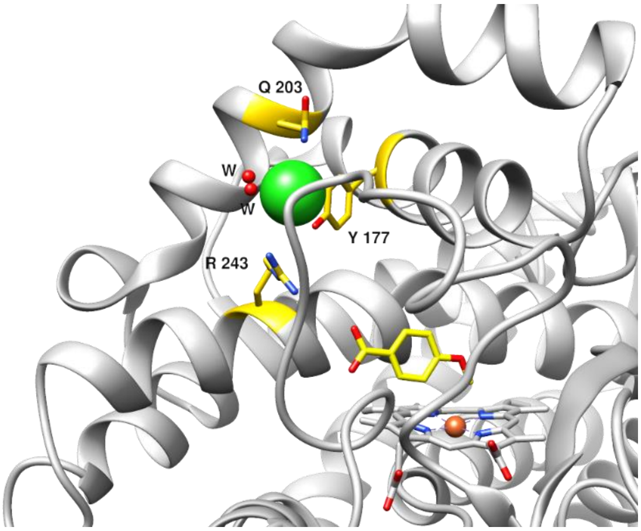


Figure 5.3. Crystallographic structure of 4-methoxybenzoic acid-bound CYP199A4 depicting chloride ion location (in green). The substrate in the binding site and the Cl^- interacting residues are highlighted in yellow.

In the recent years, CYP199A4 also appeared as a system particularly interesting to discuss in further details the solvation effect on binding and some catalytic steps – especially when it comes to the regeneration of the resting state. Indeed, among the structural data gathered by Bell and coworkers, a product-bound CYP199A4 X-ray structure has been recently characterized. As mentioned earlier, P450 studies have given major attention to substrate binding processes, so details regarding product release are not as well understood. In fact, very few other examples of product-bound cytochrome P450 structures are available in the literature, like for example $P450_{cam}$, $P450_{epoK}$, MycG, CYP101, CYP11A1 and a CYP17A1 mutant, which in some cases undergo several oxidations and have been characterized with more than one product (that is at the same time a substrate for the subsequent oxidative transformation).[146–151]

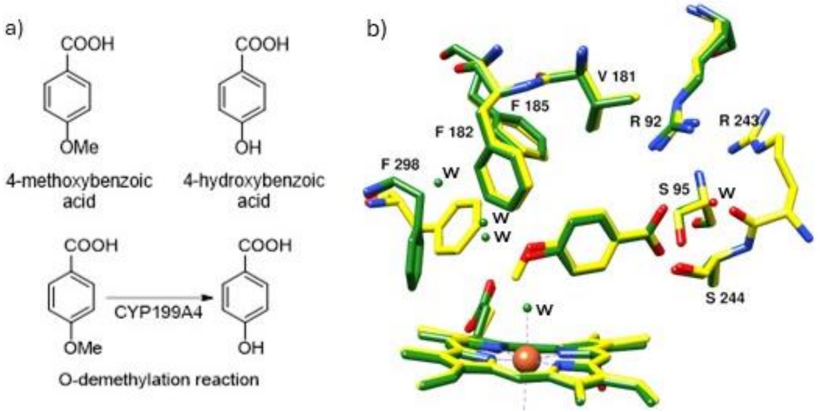


Figure 5.4. a) Substrate and product of the *O*-demethylation reaction catalyzed by CYP199A4. b) Superimposed crystallographic structures of the substrate (yellow) and product (green)-bound systems.

The same kind of experimental analyses regarding binding affinity and structural arrangement described in the previous sections for substrates have been carried for the corresponding 4-hydroxybenzoic product (Figure 5.4). Despite their similar structure, % HS induction goes from over 95% with the substrate to around 5% with the product, and the dissociation constant increases from 0.22 μM to 458 μM, signaling over 200-times worse affinity. As expected, no further product formation rate is obtained for the 4-hydroxybenzoic acid and the NADH consumption rate is reduced more than ten-fold. The orientation of both substrate and product in the active site is almost identical and the most noticeable differences are observed near the heme reactive center:

1. The sixth coordinating water ligand is crystallized above the Fe ion only in the product-bound system.
2. The residue Phe298 shows a folded conformation retracting from the binding site while extended in the substrate-bound complex
3. The freed space between the hydrophobic residue and the ligand's hydrophilic substituent is occupied by three water molecules that apparently establish a hydrogen bond network with the heme propionate, residue Thr395, the backbone of Phe298, the product hydroxyl moiety and the sixth aqua bound to the iron. Such network is absent in the substrate-bound species.

It is clear that CYP199A4 experimental studies offer a rich landscape for further understanding P450 mechanism, specificity and selectivity. Questions yet unanswered regarding the molecular origins of regiospecificity of this enzyme or what specifically drives efficient product release represent very interesting aspects to be decoded by computational approaches. This thesis aims to address these questions from a molecular modelling perspective, in hopes of gaining molecular insights on CYP199A4 species that can also set innovative dynamic-based approaches for the evolution of rational enzyme engineering.

5.2. Objectives

The study of CYP199A4 has been divided into two comparative analyses: ***meta*** against ***para***-substituted methoxybenzoic acids and *para*-substituted **substrate** *vs* the corresponding hydroxybenzoic acid **product**.

In the first topic, the main questions to address are:

a) What drives drastic regioselectivity in CYP199A4 towards *para*-substituted methoxybenzoic acid against its *meta*-substituted analogue?

b) What structural elements are key in substrate binding and transformation?

The focus of this work is placed especially on the structural elements, although an energetic assessment of the rate limiting step in the catalytic reaction is also carried out. Tools like enhanced sampling are applied to explore the dynamic behavior of substrate-bound systems and identify any relevant discrepancies between the two molecules of interest.

The second half will be the study of product differentiation in the enzyme interactions and behaviors with three main objectives:

a) Analyze the crystallographic structural differences in dynamic settings.

b) Characterize the product release mechanism for CYP199A4.

c) Assess the role played by the solvent hinted by the X-ray structures.

Accelerated dynamics simulations will reveal the conformations accessed by the product-bound enzyme. In this case, these will also be complemented with conventional MD analysis of more subtle movements of the residues inside the binding site in the presence of the product and solvent molecules. Careful design of analytical tools will be essential in the comprehension of the data generated.

5.3. Computational details

For this part of the thesis two main computational methodologies have been fundamentally employed: standard and accelerated classical molecular dynamics (cMD and GaMD) and DFT calculations on cluster models. The former are aimed at elucidating stability and dynamical behavior during substrate binding and channeling events. The latter are attempting to decode key steps in the catalytic mechanism. Details on each aspect are given in the following part of this section.

5.3.1. Molecular Dynamics set up

For the substrate-bound system, several models were devised to enable conformational exploration of different key steps of the catalytic cycle of the enzyme (Figure 5.5). These include the initial resting state (RS) with a sixth water ligand bound to Fe^{3+}, the first activation step (S1) with the sixth coordinating position vacant at ferric iron, a second activation step (S2) after redox partner reduction of Fe(III) to Fe(II) and finally the reactive species Compound I (CpdI). The two activation steps were considered separately to observe possible differences between the entrance of the substrates in the binding site and the arrangement right before molecular oxygen bonding. Moreover, penta-coordinated Fe^{3+} is obtained again when the product is leaving, right before the RS is regenerated to restart the catalytic cycle, so it represents different stages depending on the ligand found in the binding site.

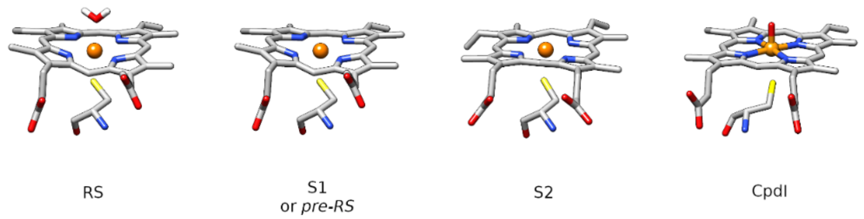


Figure 5.5. Models of the different catalytic steps parametrized for MD and GaMD simulations

Systems were set up from the corresponding X-ray structures, maintaining crystallographic waters and ions. Special attention was dedicated to the chloride ion present between the B-C loop, I, F and G helices, as initial calculations following standard protocols cleaning any solvent molecules from the crystallographic coordinates showed loss of active site configuration.

For the *meta*-substituted benzoic acid, original PDB 6PQ6 coordinates were used, for the *para*-substituted substrate the PDB code is 4DO1 (chain A) and for the *para*-substituted product two different options were simulated: the initial coordinates were taken either from 8VOC or 4DO1, manually altering the methoxy substituent for the corresponding hydroxyl group.

To obtain the parameters for the Fe containing heme group, MCPB.py protocol was followed. First, the corresponding models of Fe^{2+} or Fe^{3+} bound to porphyrin, the sulphur atom from the Cys residue and the oxygen in the water molecule when necessary were optimized and vibrational frequencies were obtained to ensure obtention of a potential minimum. These calculations were carried out with B3LYP hybrid functional including Grimme's D3 dispersion correction, applying the SDD pseudopotential for Fe and 6-31G(d,p) basis set for the remaining atoms. Charges for the models were obtained through the RESP method. For the

CpdI, the force field parameters were extracted from Cheatham et al.[152] The ligands were parametrized using antechamber, applying AM1-BCC method to obtain the charges and GAFF to describe each atom and its interactions. All the natural amino acids were parametrized under the Amber14SB force field.

5.3.2. Molecular Dynamics simulations

For the simulations, the systems were solvated by embedding the models (including crystallographic waters and Cl^- ion) in a cubic box of TIP3P waters leaving a 10 Å separation between the biomolecule and the limits of the box. Sodium ions were added to neutralize the system. All systems were submitted to conventional MD simulations, then to Gaussian accelerated MD and, in some cases, to Ligand Gaussian accelerated MD. The initial MD comprises the following protocol: 1000 steps of water minimization, 1000 steps of water and side chain minimization, 1000 steps of whole system minimization; heating from 0 to 300 K in 10000 steps with fixed hydrogens; 100 ps of NVT equilibration and 500 ps of NPT equilibration; finally productions were run in NPT under periodic boundary conditions (8 Å box standardize in Amber). The length of the simulations varied from 100 to 500 ns depending on the system.

The GaMD simulations protocol is the same described in the previous chapter, starting from a converged MD simulation for the 52 ns of equilibration before production. For the LiGaMD trajectories, same length equilibration is acquired starting from converged MD coodinates and extended into production. Length of GaMD simulations varies from 400 to 1200 ns, while LiGaMD times went frm 100 to 300 ns.

MD convergence was followed by RMSD to the first frame and all-to-all, PCA, RMSF and cluster counting. Water content data, distances and other analytical measurements were extracted with Cpptraj and the occupancy percentages were calculated through a Python script specifically designed for the system.

5.3.3. DFT study of the rate-determining step in the catalyzed *O*-demethylation reaction

DFT calculations were carried out on a simplified model of the active site including the substrate molecule and the CpdI, formed by the porphyrin ring with protonated propionates, the Fe ion bound to the oxo ion and the side chain of the Cys residue up to the α-carbon (Figure 5.6). The program employed was Gaussian 16 using the B3LYP hybrid functional adding Grimme's D3 dispersion for geometry optimization and frequency calculation. The basis set used for non-metallic atoms was 6-31G(d,p) and for the Fe atom the SDD pseudopotential with the corresponding *f* polarization function. The implicit solvent is chlorobenzene to approximate the dielectric constant of the binding site (ε=5.7).[131] The energies were refined using Def2QZVP for Fe and Def2TZVP for the remaining atoms.

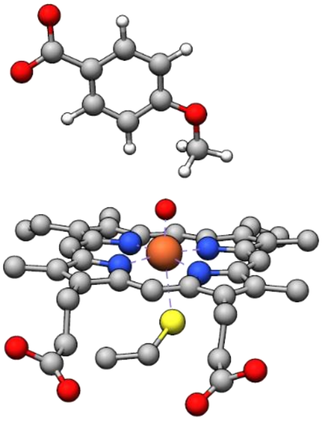


Figure 5.6. DFT model for the *O*-demethylation reaction study.

5.4. Study of *meta- vs para-* substituted benzoic acid substrates

In sight of the high specificity for the *para*-methoxy benzoate substrate displayed by CYP199A4, this work aims at decoding the key aspects involved in such specificity from a molecular modelling point of view. Three main sources of differentiation were regarded: stability of pre-catalytic conformations in the binding site, alternative ligand entrance channels for each substrate and intrinsic electronic properties linked to the RDS energies.

5.4.1. Substrate-CYP199A4 stability, key interactions and occupancy

GaMD simulations were performed in triplicates of 500ns for the systems representing the relevant catalytic steps with each substrate. Through the analysis of contacts displayed along the different dynamic simulations, three main interaction networks were identified for CYP199A4 (Figure 5.7):

1. **Hydrophilic patch**. Constituted by residues Arg92, Ser95, Arg243, Ser244 and Ser247, it is located mainly between the B-C loop and the N-terminal of helix I. In the X-ray structure a crystallographic water is included, establishing hydrogen bonding interactions mainly with the carboxylate moiety in the ligand and residues Ser95, Arg243 and Ser244. These amino acids stabilize the carboxylate moiety of the ligand, keeping the orientation observed in the crystallographic structures for *para*-substituted benzoic acids.
2. **Hydrophobic patch**. A second pocket comprises residues Leu98, Val181, Phe182, Phe185, Phe298, Phe299 — which interact with the

benzoic moiety of the ligand — as well as hydrophilic amino acids Ser76 and Arg300—that stabilize the carboxylate.

3. **π-stacking (orientation 3)**. The third region displays a strong tendency for π- π and π -CH3 interactions between the benzene and the porphyrin ring.

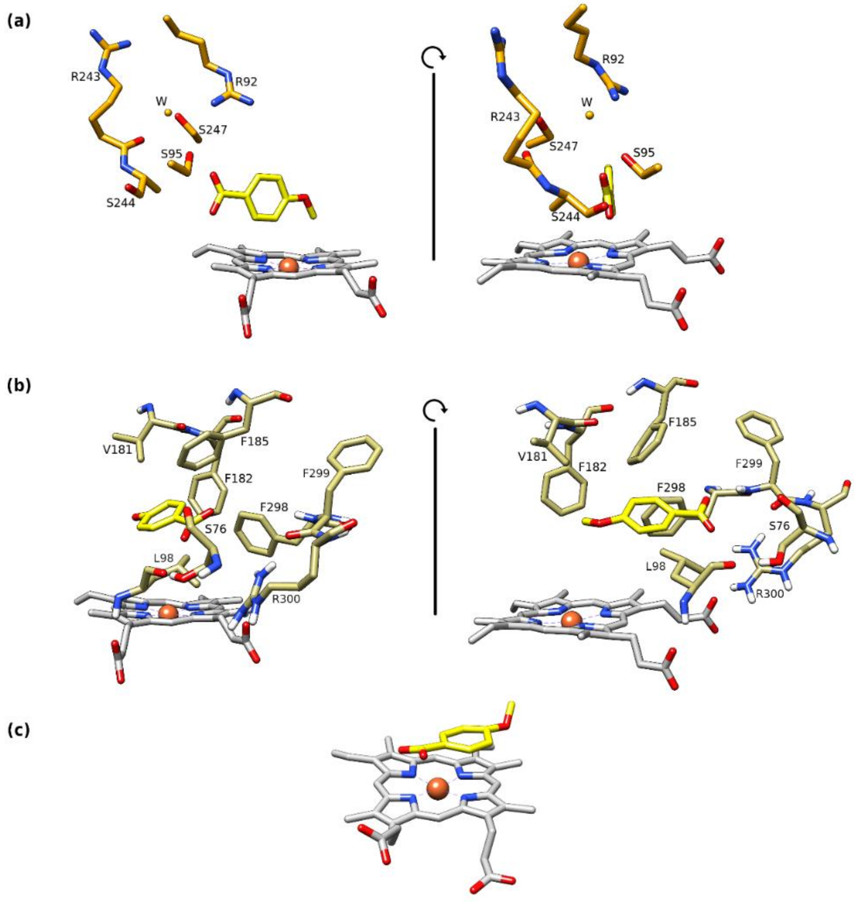


Figure 5.7. a) Hydrophilic patch residues in orange that interact with the carboxylic moiety of the ligand. In yellow, *para*-substituted ligand for reference. b) Residues in the hydrophobic patch depicted in ochre, with yellow *para*-substituted substrate for reference. c) Heme porphyrin π-stacking orientation of *para*-substituted ligand.

Occupancy of these areas differs according to ligand and catalytic step represented and allows evaluation of the difference in behavior between both substrates (Figure 5.8).

Along GaMD simulations, the *para-* substrate shows a higher preference for the hydrophilic patch (orientation 1), although there is variability depending on the catalytic step, as depicted in Figure 5.8. In RS, two main areas are occupied during the simulation, either the substrate stays in orientation 1 or it leaves the binding site, roaming free in the solvent with occasional interactions with the protein surface. In the S1 simulations, the ligand leaves the binding site in all replicas (67%), and before leaving, the orientation is either type 1 (hydrophilic stabilization) or 2 (hydrophobic stabilization), representing 19 and 14% of time, respectively. For ferrous S2, the benzoic acid has a prevalent occupation of the π-stacking orientation for 48% of simulations, while the second most occupied is orientation 1. However, the clearest trend is observed in the precatalytic step with the CpdI, where the substrate shows 99% occupation of the hydrophilic patch, maintaining the reactive H within a mean distance to the oxo ion of 2.78 Å and a mean Fe-O-H angle of 120º. These values are within the optimal reported measurements, which for H-abstraction comprehend a distance around 3 Å and an angle between 110 and 130º.[153]

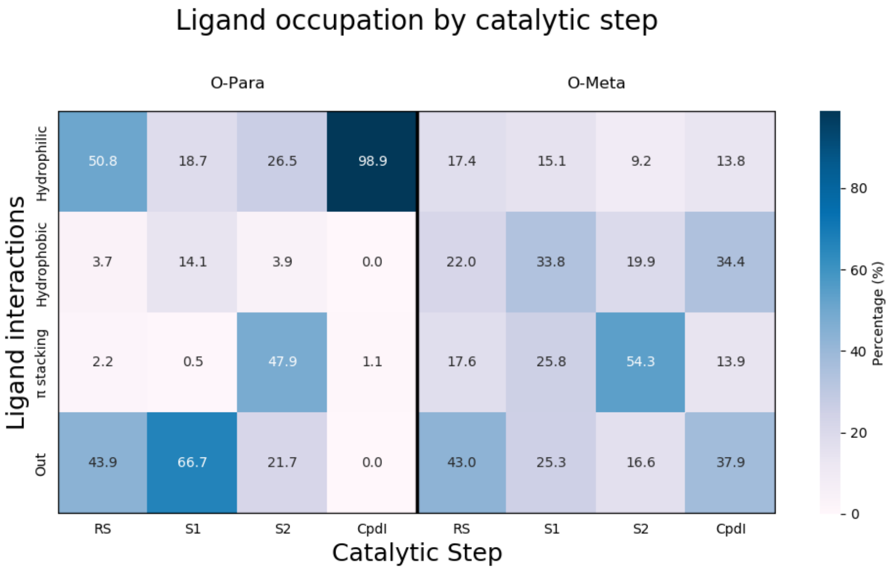


Figure 5.8. Heatmap of the different interaction networks occupancy along the GaMD simulations for each system according to the catalytic step. The catalytic step is indicated underneath each column. The first four columns correspond to simulations with 4-methoxy benzoic acid, and the last four columns to 3-methoxy benzoic acid-bound systems.

In the case of the *meta*-substituted benzoic acid, the general tendency shows a much more dispersed distribution of occupancies along the trajectories. Starting with the RS, up to 43% of simulation time is spent outside the binding site, and the rest of the time is equally divided between the other three areas. During S1 the preference changes for the hydrophobic patch, while for S2, the most occupied orientation is parallel to the heme group, although in any case all different orientations are visited. Finally, in the CpdI step the substrate leaves in half of the simulations, so the exploration outside the protein makes up 38% of simulation time, followed by 34% occupation of the orientation 2 and 14% for both orientation 1 and 3; the mean distance between the oxo ion and the reactive H atoms is 10.64 Å while the mean Fe-O-H angle is 118°. All these observations already point towards a far weaker capacity of the *meta*-substrate to reach catalysis-favoring orientations and a higher tendency to be egressed.

Both the crystallographic structure and the CpdI simulations indicate that the hydrophilic patch stabilization is key for CYP199A4 catalysis of benzoic acids, as it ensures the correct distance and orientation of the reactive H-C bond of the substrate with respect to the CpdI oxidative species. The tight stabilization of *para-* substrate in this region contributes to the preferential selectivity and catalytic efficiency of the enzyme towards this molecule compared to the *meta-* ligand.

One of the regions that displays a marked different behavior between CpdI simulations with *meta-* and *para-*substituted substrates is the B-C loop. The mean RMSF of this region with the *meta-*substrate is 1.11Å, while the same measure for *para-*substrate simulations is 0.65 Å. This is coherent with the fact that many of the residues involved in hydrophilic patch stabilization are located in the B-C loop. Moreover, higher exploration of the hydrophobic patch or alternative locations in the binding site correlate with higher RMSF values, like the case of *meta-*bound S2 CYP199A4 trajectories that display a mean RMSF value of 1.60 Å (Figure 5.9).

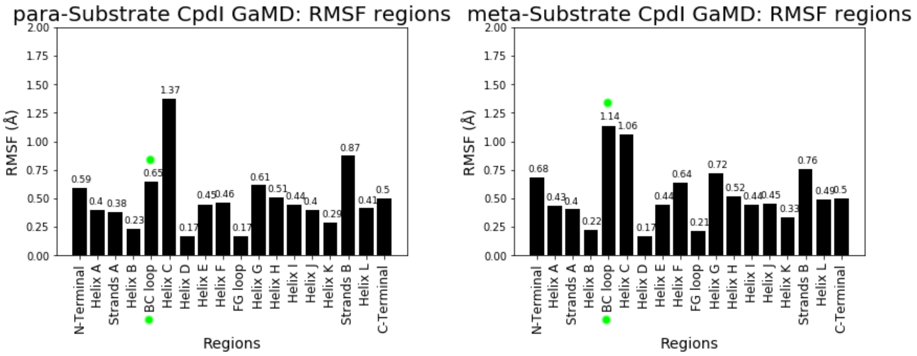


Figure 5.9. RMSF for *para-* and *meta-*substituted substrates (left and right respectively) along GaMD simulations of the CpdI CYP199A4 system. The B-C loop has been highlighted with a green dot.

Additionally, these simulations point out that both the RS and S1 have a higher tendency to expel the *para-* substrate. This highlights the necessity to readily eliminate the sixth aqua ligand from the Fe vicinity for the substrate to enter and the catalytic cycle to start, which is already hinted in the X-ray structures of the methoxy-substituted ligands. Though it may seem an unnecessary set of calculations, it is remarkable to see how simulations agree with the mechanism suggested for P450 in which substrate entrance is necessarily linked to the removal of the water from the distal position of the iron.

The lack of preference for hydrophilic stabilization and almost equal distribution among orientations found along *meta-* substrate simulations demonstrates less affinity of this ligand for the enzyme active site, which is especially relevant during the step comprising the reactive species CpdI. Altogether, these results indicate that the reason for *para-* selectivity over *meta*-substituted substrates in CYP199A4 has a strong structural component linked to stabilization of pre-catalytic orientations along the catalytic cycle and, more specifically, in presence of the oxidizing species CpdI.

5.4.2. The entrance pathway

Entrance pathways have been proven to play a crucial role in guiding substrate orientation in many enzymes, hence influencing biocatalytic activity. Given that the performed dynamic simulations naturally show egression events of both ligands, these were considered representative of the pathways in and out of CYP199A4. Therefore, a more detailed analysis was focused on identifying relevant differences between both substrates regarding their access to the active site, hypothesizing possible influence in the regioselectivity of this enzyme.

The general tendency observed in egression-depicting trajectories is for the ligands to start on the binding site either in hydrophilic patch stabilization or π-stacking and it shifts upwards, exploring largely the region between helices I, F and the B-C loop, alternating interactions between residues Pro93, Pro94, Ile97, Leu82, Glu83, Val181, Phe185, Leu240, Ser244, Leu245, Ser247. Once the I helix has been surpassed, the ligand explores one of three possibilities: 1) it continues along the I helix/B-C loop interface towards the surface, which will be further referred to as path A; 2) it moves towards the F and G helices interacting with residues such as Gln203, Leu180 or Ile196 to finally egress to the solvent, which will be referred to as path B and is the most frequent option; 3) the ligand shifts towards the F-G loop, staying in the area by alternating π-stacking interactions with Trp91 or polar contacts with Asn193, Ser83, Glu88 or Arg192. The latter only shows complete egression of *meta*-substituted substrates, while the *para*-benzoate fails to find an opening suitable to access the surface and goes back deeper within the protein core to end up leaving through path B (Figure 5.10).

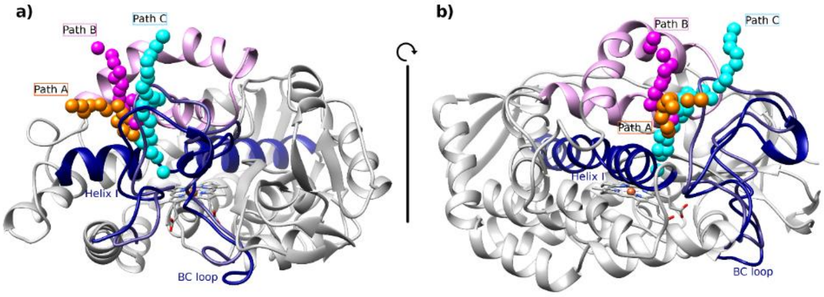


Figure 5.10. Pathways followed by *para*- and *meta*-methoxy benzoic acid along GaMD simulations. Representative trajectory for each path. Path C is only observed in *meta*-substituted substrate trajectories. Helix I and B-C loop are highlighted in dark blue, and helices F and G in light pink. Fron view (a) and side view (b) show the influence of B-C loop opening in the path accessibility.

All three substrate channel options are known entrances for CYP450 and coincide broadly with paths 2ac and 2c coined by R. C. Wade and coworkers,[44] representing a slightly lower or higher variant. Moreover, the crystallographic structure of substrate-free CYP199A2 — a homologous protein from a different strain of R. *palustris* with 86% sequence identity with CYP199A4 and very similar substrate specificity — presents a wider conformation with an open channel that connects directly the surface of the protein to the binding site, which perfectly matches the paths found in these trajectories (Figure 5.11 a and b). Additionally, the chloride ion characteristic of substrate-bound crystallographic structures of CYP199A4 (and CYP199A2) is precisely located in the intermediations of the described channels, interacting with residues Trp91, Arg92, Tyr177, Gln204, Asn207, Arg243 from the G and I helices and the B-C loop (Figure 5.11 c and d). This anion is believed to play a role in substrate isolation from the hydrophilic environment; however, during dynamic simulations of specially S2 and CpdI systems, it displays consistent unbinding from the protein surface, indicating potential Cl^- involvement in electrostatic equilibration of the binding site prior to catalytic cycle activation and ferric Fe reduction.

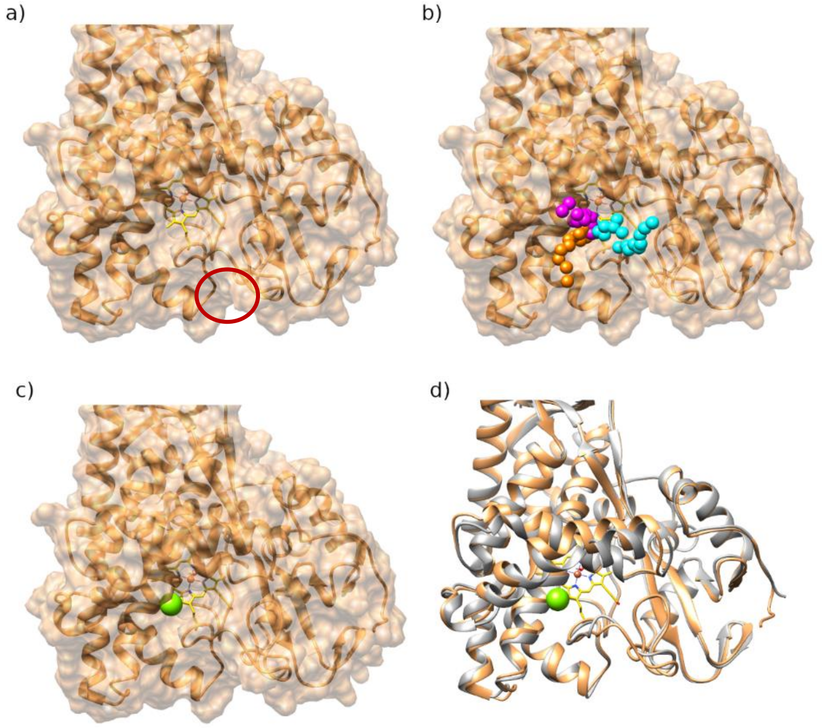


Figure 5.11. a) Crystallographic structure of CYP199A2 (PDB code 2FR7) with the surface depicted in orange; the open channel is highlighted in a red circle, showing direct access to the heme group depicted in yellow. b) Same CYP199A2 structure superimposed to the entrance pathways identified for CYP199A4. c) Same CYP199A2 structure showing the location of Cl⁻ ion in substrate-bound CYP199A4 structures. d) Superimposed structures of CYP199A2 and CYP199A4 (PDB code 4DO1) in orange and grey respectively, showing chloride location in the latter and channel opening difference.

Once again, the B-C loop plays a key role in the regulation of these pathways as it dictates which of the three alternatives are accessed by the substrate depending on its opening. For instance, some of the contacts that maintain a closed active site comprise van der Waals interactions between loop residues Gly81, Leu82 or Trp91 and F-G helix amino acids Ala184, Phe185 or Arg192 or even polar contacts between Arg92 and the substrate However, for substrate entrance the B-C loop breaks these interactions momentarily and opens up, clearing the pathway between the I, F and G helices. Furthermore, some of the B-C loop residues in the hydrophilic patch responsible for stabilization of the pre-catalytic orientation are also involved in guiding the ligands to the binding site, such as Arg92 or Ser95. In the case of Arg92, during pre-catalytic stabilization this residue is part of the hydrophilic interactions network, but when the ligands are leaving it often displays a rotameric shift to interact with Glu99, leaning the loop away from the rest of the protein and leaving the channel connecting the active site with the surface considerably more open. These results indicate that the highly flexible B-C loop plays a role in both substrate stabilization and entrance, signaling truly fine tuning between these two events (Figure 5.12).

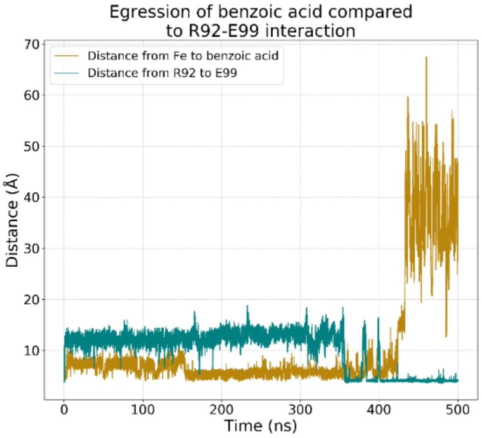


Figure 5.12. Evolution of distance between 4-methoxybenzoic acid and the binding site (yellow) and between B-C loop residues Arg92 and Glu99 along a GaMD simulation.

Path C, observed only in *meta*-bound systems, differs the most from the well-characterized CYP199A2 opening that all *para*-bound simulations display. Seeing the correlation between binding and entrance mediated by the B-C loop, the most plausible hypothesis would be that this pathway appeared due to destabilization of the active site resulting from *meta*-substrate inability to adopt a fixed conformation, triggering wider movements of the loop and therefore opening alternative pathways. Therefore, it is hypothesized that entrance pathways would be involved in recognition of the rigid, flat and negatively charged benzoic acid molecules but not so relevant in regioselective recognition of *para* and *meta*-substituted compounds.

One alternative pathway must be mentioned regarding the *para*-substituted benzoate egression. Two replicates, one modeling the RS and the other with S1 show a different trace that goes through the B-C loop. For the RS simulation the ligand starts stabilized in the hydrophilic patch but in the case of S1 it starts in the hydrophobic one, which is one of the least populated orientations for the *para*-substrate. Although this is a known pathway for other cytochromes P450 (pathway 2e),[44] it seems less likely for CYP199A4, as it appears only in two isolated replicates, it does not coincide with any crystallographic data and the starting points represent less occupied arrangements.

At this point in the study, both substrates appear to display very similar entrance paths while their stability in the binding site is far different, with the *meta*-substituted species displaying more flexible behavior and less propensity to generate catalytically competent geometries. Still, how the electronic properties of the system could influence the CYP199A4 selectivity for one system over the other cannot be studied with these methodologies.

5.4.3. Analysis of the rate-determining step of *O*-demethylation reactions

O-demethylation reactions in CYPs have been extensively characterized, among others, by S. Shaik and coworkers, who identified the initial hydrogen atom abstraction as the rate-determining step in the Compound

I-mediated mechanism. To assess the inherent energetic differences between the two substrates as a source of regiospecificity, a reactivity profile was constructed for each ligand in the presence of the oxidizing CpdI species.

For this purpose, a model of the CpdI species was created, representing the protonated porphyrin ring, Fe atom, O^{2-} ion and axial-coordinating Cys residue's side chain including the C_α. This species, together with the substrate presents a global charge of -1, with three unpaired electrons that can organize into two different spin states: high-spin (HS) quartet and (LS) doublet with broken symmetry. After evaluating the catalytic profile, doublet models provided lower energies in all cases, so these are the results that will be discussed.

Initial coordinates were taken from the crystallographic orientation of the ligands in the binding site, showing a perpendicular orientation to the porphyrin ring plane. However, in both cases, the optimized structures display the benzene ring moiety parallel to the heme group, favoring π-stacking interactions.

The natural orbitals representing the electronic distribution among the system locate the unpaired electrons similarly for both substrates (Figure 5.13). The orbitals comprise mainly the Fe and catalytic O atoms, extending to the cysteine S and slightly along the porphyrin ring, as well as to the reactive hydrogen-carbon bond and the O in the substrate substituent.

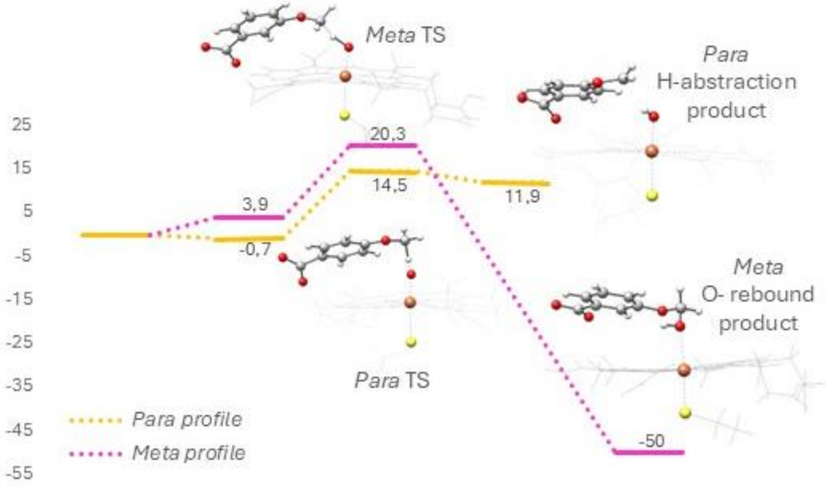


Figure 5.13. Reactivity profile for *H*-abstraction step in the *O*-demethylation reaction catalyzed by CYP199A4. Profile for the *para*-substituted substrate is depicted in yellow, while the profile for the *meta*-substituted substrate is in magenta.

The reactivity profile for both ligands presents crucial differences, the TS energy for each substrate being the most relevant: the *meta*-substituted substrate needs to surpass an activation barrier 6 kcal/mol higher than the *para*- substrate. Interestingly, the 3-methoxy benzoic acid readily goes from the H-abstraction transition state to the oxygen rebound product, which is the next step in the reaction, skipping the intermediate formation. At the point of writing this thesis, characterization of the *para*-substituted oxygen-rebound step has not been achieved, although it is a work in progress to evaluate possible further insights into substrate differentiation.

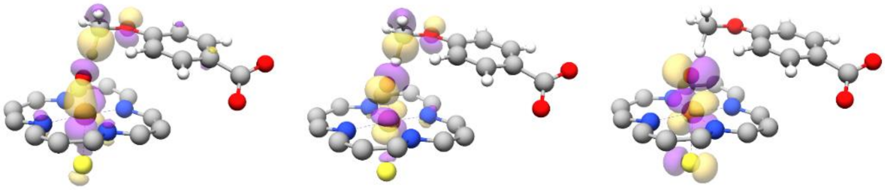


Figure 5.14. Natural orbitals identified for the distribution of the three unpaired electrons.

The difference in the energetic barrier for H-abstraction between substrates indicates intrinsic reactivity of the ligands as a clear driving factor in regioselectivity. Moreover, the porphyrin ring π-stacking interactions observed especially in S2 systems simulations is coherent with the lowest energy geometry obtained in DFT calculations.

Nevertheless, despite the energetic preference of a π-stacking stabilization between substrate and heme group, this orientation does not appear as the most populated in the simulations of the *para*-substituted substrate with the CpdI model nor in the crystallographic structure, highlighting the role of the active site amino acids in conditioning the allowed reactivity. In fact, Compound I is known as a "chameleonic species" with an electronic structure especially sensitive to its environment, so residues in the active site might play a role that is not accounted for in these calculations.

For this reason, the multiscale protocol applied in this thesis is necessary to provide a complete rationalization on the factors influencing substrate specificity in CYP199A4, from the early stages of substrate entrance to the binding site to the pre-catalytic and catalytic details.

5.4.4. Evaluation of mutation proposals to favor *meta*-substituted substrate binding

From the previous part of the study, it appears that the primary reason for the lack of *O*-demethylation for *meta*-substituted benzoic acid by CYP199A4 is likely due to the absence of stabilization of pre-catalytic geometries for this complex. With the aim of altering the regiospecificity of the enzyme in that direction, a series of *in silico* mutagenesis analysis were performed.

Examining the contacts along the previous set of simulations one of the residues that stand out is V181, as it sits on top of the *para*-substituted substrate and participates in stabilization of the catalytically competent geometry. A possibility is to increase the space available for the *meta*-substituted molecule to rearrange in an alternative position that still stabilizes the carboxylic moiety but orientating the substituent towards the Fe center, making room for the "bulkier" benzoic ring towards the upper part of the binding site. For this reason, amino acid Val181 was mutated into a smaller alanine, which still presents hydrophobic character but creates more space within the entrance of the cavity for the ligand to reorient.

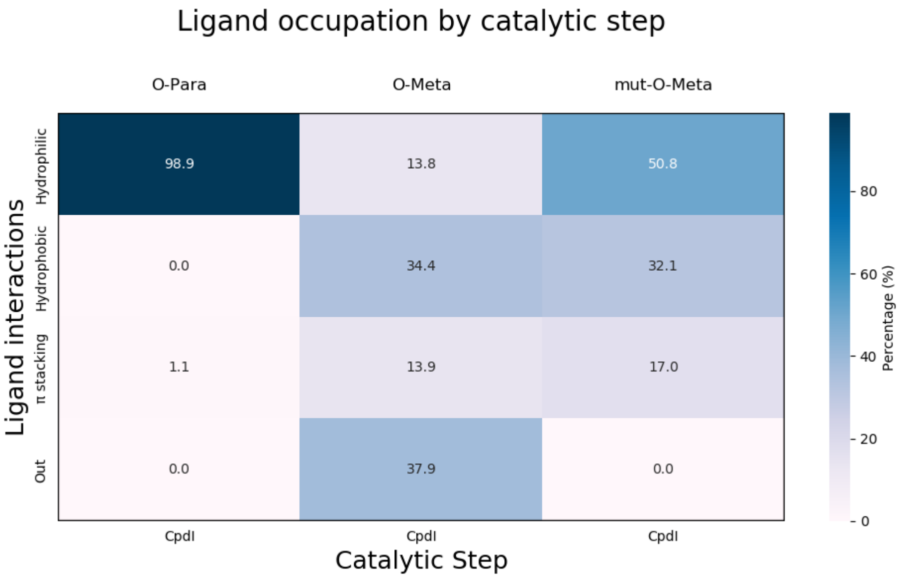


Figure 5.15. Heatmap of the different interaction networks occupancy along GaMD simulations for each system according to the catalytic step. The models all represent the CpdI catalytic step, indicated underneath each column. The first and second columns correspond to WT CYP199A4 with the 3-methoxy benzoic acid in the first one and 3-(methylthio) benzoic acid in the second one. The final column is the simulation with the V181A mutant with 3-methoxy benzoic acid as substrate.

MD were carried out with this V181A mutant and the *meta*-substituted substrate to test the hypothesis. The enhanced sampling trajectories show an increase of hydrophilic patch occupation compared to the WT, going from 14% to 51% occupation (Figure 5.15). Moreover, no tendency to abandon the binding site is observed, so all the simulation time is divided between hydrophilic, hydrophobic and π stacking arrangements. Regarding the pre-catalytic orientation, the mean distance from the reactive hydrogens to the oxo ion goes from 10.64 Å to 4.3 Å and the angle shows no relevant change going from 118° to 120°. The distance value is closer to the goal measurement of 3 Å although still not close enough for proper reactivity.

On a more detailed analysis looking at each replicate individually, radically different behaviors are obtained for each simulation. The first replicate displays a mean distance H_{ligand}-O_{CpdI} of 2.76 Å and mean Fe-O-H angle of 127°, which would favor the hydrogen abstraction for the demethylation reactivity. Most of the trajectory shows the ligand in a position very similar to the hydrophilic patch-stabilized *para* substrate but with the carboxylic moiety pointing at a lower angle, closer to Leu96 and Ser234. This leaves the substituent pointing downwards to the oxo ion and the bulkier part of the benzoic ring upwards to the binding site ceiling. Arg243 still creates an interaction network with Ser95 and one water molecule, although for this replicate the connections are extended to Tyr177 and up to two more solvent molecules that appear above the substrate in the space that used to occupy Val181 and is freed up by the mutation to alanine. Arg92 displays a strong interaction with Glu99 during the whole simulation, but the B-C loop does not show any opening movement as in the previous cases with the WT and instead maintains a closed conformation through interactions with the G and I helices (Figure 5.16).

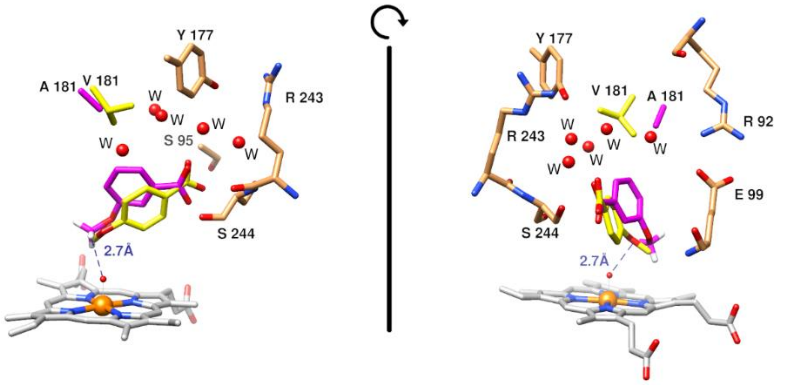


Figure 5.16. Frame from the GaMD of the mutant CYP199A4$_{V181A}$ with the *meta*-substituted substrate (magenta) compared to the crystallographic orientation of the *para*-substituted substrate (yellow) in WT CYP199A4.

For the other two replicates, the average H$_{ligand}$-O$_{CpdI}$ distance is 3.59 and 6.57 Å, which locate the reactive atoms too far from each other for the reaction to take place. In both cases, the substrate starts with the carboxylic moiety pointing to the upper part of the binding site, although after the first 100 ns the ligand explores alternative orientations, finding the hydrophilic-stabilized arrangement observed in the first replicate but rapidly switching to other options and finally achieving π-stacking with the porphyrin ring, with the substituent points away from the oxo species, unable to undergo any type of reactivity. With respect to the pathway residue interactions, in the second replicate they displayed a similar arrangement to replicate 1, except that the water molecules completing the interaction network were much less frequent; in the case of the last replicate, the organization of Arg and other pathway residues is similar to that of the *para*-bound simulations, which are unable to stabilize the *meta*-substituted substrate in a fixed position.

This mutation facilitates the exploration of alternative orientations for the stabilization of the *meta*-substituted substrate. One of the replicates shows that a desirable arrangement can be obtained with the V181A mutant, but the lack of reproducibility points out that stronger securing of the alternative “hydrophilic-stabilized” orientation would be advisable for proper catalysis.

At the time of writing of this manuscript, this mutation has been tested experimentally, showing no spin shift at all upon *meta*-substituted substrate binding experiments, indicating that the molecule cannot be accommodated in the active site in absence of Val181. Replicates need to be performed and mutations to other residues need to be investigated. An alternative mutation area could be the hydrophilic patch, to secure the orientation observed in the first replicate with the carboxylic acid pointing at a lower angle than the crystal.

5.5. *Para*-substituted benzoic acids: substrate *vs* product

Given that the demethylation of 4-methoxybenzoate is the best performing reaction catalyzed by CYP199A4 and the availability of a 4-hydroxy benzoate-bound CYP199A4 crystallographic structure, a second part of this enzymatic study was focused on elucidating key differences between substrate and product-bound cytochrome P450 to fully investigate the structural evolution along the function of this enzyme, starting with entrance and binding, catalysis, egression and recovery of the functional protein.

5.5.1. Product interactions and water content

Regarding the models selected for the catalytic cycle representation, it is important to highlight the differences caused by product presence. No product-CpdI simulations were carried out since the catalytic species does not exist in presence of the product of the reaction. Moreover, the S1 pentacoordinate ferric complex represents the final step after the reaction takes place and prior to RS recovery, instead of the initiation of the catalytic cycle after water displacement.

Accelerated simulations with 4-hydroxy benzoic acid show exploration of the same three areas described in the previous chapter for the *para*- and *meta*- methoxy-substituted variants, namely hydrophilic and hydrophobic patches and π-stacking with the porphyrin ring. The same analysis was performed in this study (Figure 5.17).

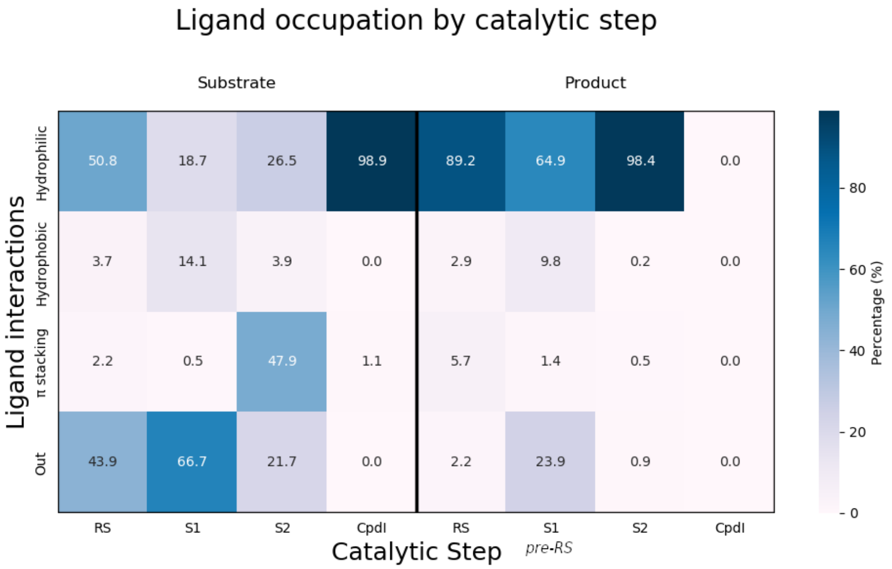


Figure 5.17. Heatmap of the different interaction networks occupancy along the GaMD simulations for each system according to the catalytic step. The catalytic step is indicated underneath each column. The first four columns correspond to simulations with 4-methoxy benzoic acid, the last four correspond to simulations with 4-hydroxy benzoic acid.

When compared to the *para*-substituted substrate, the product presents similar overall tendencies to either stay in the hydrophilic patch or leave the binding site, but differences in the exploration distribution can be appreciated. For RS simulations, a higher occupancy rate is observed for the hydrophilic patch compared to that of the substrate (89% *vs* 50%) while the remaining time is spread between hydrophobic, π stacking and trying to leave the binding site although it never reaches the surface of the protein. In the case of S2, hydrophilic stabilization is predominant during 98% of the simulation in contrast with the substrate-bound system, which showed almost a 50/50 distribution between π stacking and the hydrophilic patch. Nonetheless, for the S1 simulations the hydrophilic orientation represents 65% of the trajectories, while about 34% of the simulation corresponds to the product starting to leave the active site and accomplishing it in one of the replicates, which is similar to the substrate behavior that has a clear preference towards leaving the active center. It is important to highlight that, with the product-bound systems, S1 is representative of the last step in the catalytic cycle right before RS recovery. In fact, this is already hinted during the simulations, which show a water molecule from the solvent at a distance of 3.5 Å from the metal atom for 54% of the simulation time, resembling the conditions of the RS (Figure 5.18a). Additionally, this water becomes part of a hydrating network of two to three molecules around the ligand at different points of the trajectory. Even in the replicates where the surface of the protein is not reached, the product shows a tendency to abandon the active site accompanied by an increase in the water molecules around the hydroxyl group, filling the hydrophobic binding site with highly polar molecules (Figure 5.18b).

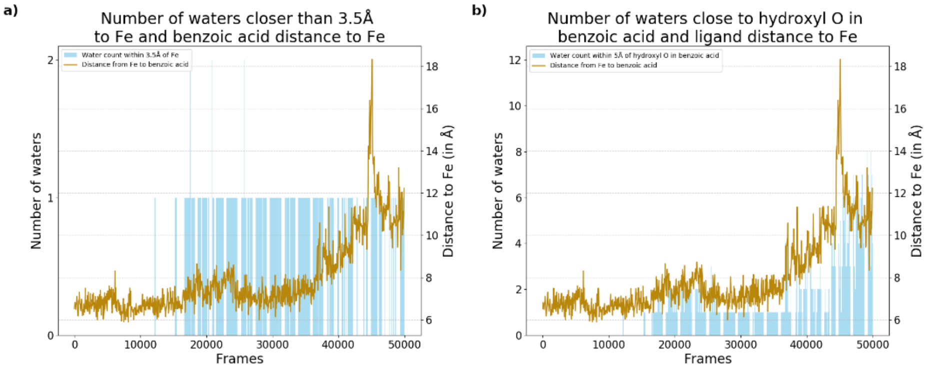


Figure 5.18.a) Number of water molecules closer than 3.5Å to the Fe atom along a ferric S1 GaMD simulation (blue bars) and distance between the product carboxylic carbon and the Fe atom (yellow line). b) Number of water molecules closer than 5Å to the hydroxyl oxygen (blue bars) along the same simulation and distance between the ligand and the binding site (yellow line, same atom references).

To further assess the difference in water content induced by both ligands, the water presence within 5 Å of the Fe atom was analyzed for S1 systems. About 30% of the substrate-bound trajectories present at least one water molecule in the binding site, while up to 60% of the product-bound simulations display solvation near the metal atom. Moreover, the average number of water molecules present in the active center in product-bound simulations is close to two with sporadic frames including up to five water molecules along the trajectories (Figure 5.19Figure 5.19).

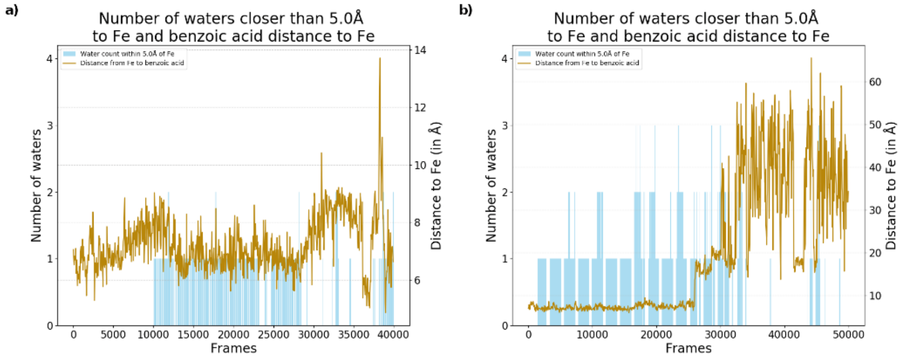


Figure 5.19. a) Number of waters within 5A of the Fe atom and distance from Fe atom to ligand carboxyl atom along ferric S1 GaMD simulations of a) substrate and b) product.

Two main points of interest were drawn from these results: 1) the RS seems to stabilize the product by stablishing hydrogen bonding between the sixth coordinate water and the hydroxyl group as observed in the crystallographic structure; 2) when released in a free dynamic approach, waters reproduce the RS during product-bound simulations, with a high occupancy of the axial coordinating position by a water molecule and even more solvation in the active site around the product substituent. This seems to indicate that, after the reaction, the whole binding site is ready to restart the catalytic cycle, with the new hydrophilic moiety of the product inducing higher solvation of the hydrophobic active center of the enzyme and possibly triggering product egression.

These findings are in good agreement with the product-bound X-ray structure, which shows four water molecules next to the 4-hydroxy benzoic acid and points out that water uptake is influenced by F298, which adopts a bent conformation freeing space in the binding site that can be taken up by the solvent molecules. For this reason, further analysis of active center evolution and its relation to water content was carried out.

5.5.2. Hydrophobic binding site arrangement

To study the relevance of amino acid F298, a set of MD and GaMD simulations starting from the product-bound crystal with the phenylalanine in bent conformation was carried out to compare it with simulations of the manually modified substrate-bound crystal with F298 in extended conformation. One more difference between both crystallographic structures to consider is the presence of the sixth coordinating water molecule only with the product. Given that the ferric S1 model reproduces RS tendencies with the product, the analyzed systems for this study include RS and ferric S1 of both product and substrate-containing simulations.

Conventional MD simulations of RS starting in a bent conformation show that the product is stable in all cases and F298 changes to extended configuration in one out of three replicates of 100 ns. When moving to GaMD simulations starting with bent F298, in two out of three replicates this residue extends into the binding site within the first 100ns. For the S1 systems, a similar behavior is observed: trajectories starting with a bent orientation switch to the elongated one in one in three cases for classical simulations and in all 3 for the accelerated dynamics. For the opposite initial conditions with the elongated configuration, the hydrophobic amino acid never shows the bending conformational change, neither in conventional nor accelerated simulations.

When checking the influence of F298 position on the water presence in the binding site, a small difference can be observed in the overall amount of water content, with the bent conformation allowing the appearance of up to one more water molecule (Figure 5.20a). Nevertheless, this seems to indicate that, regardless of the phenylalanine orientation, the binding site can accommodate the higher solvent content linked to product presence through different means other than F298 rearrangement (Figure 5.20b).

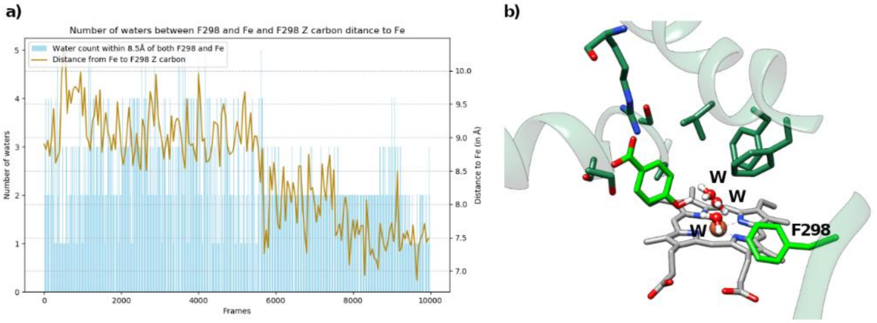


Figure 5.20. a) Comparison of water content (blue bars) and F298 distance to the Fe atom (yellow line) along a S1 MD simulation with the phenylalanine starting in bent conformation (distance > 8.5Å) and switching to extended conformation (distance < 8.5Å). b) Frame of the simulation with extended F298 allocating 3 water molecules in the binding site together with the product.

Overall, the comparative analysis points out that the bent rotamer obtained in the X-ray crystallographic structure is less occupied than the extended conformation for the dynamic simulations. A possible explanation is that molecular dynamics represent a more relaxed structure of any system compared to the static low-energy conformation observed in a crystal. For this reason, simulations with the elongated rotamer were considered valid to represent the product-bound CYP199A4 systems.

5.5.3. Entrance/egression pathway and water influence

Focusing on the pathway followed by each ligand to leave or enter the binding site, GaMD and LiGaMD simulations were carried out for substrate and product-bound structures. It must be noted that the type of calculations performed in this work only allow exploration of the egression pathway starting from the enzyme-bound position for any molecule, as the entrance would require vast exploration of the protein surface or biasing the trajectory through other means like Metadynamics. This method presents a compromise to avoid energetic biasing, allowing enough free exploration to find the actual preferred pathway for each molecule.

Only one main path is followed to abandon the active site of the protein irrespective of the ligand analyzed (Figure 5.21). As described in the previous chapter, it goes through an opening between the B-C loop and helices F, G and I. The initial steps require surpassing the I helix assisted by residues L98, V181, F182, F185, S244 and S247. At this point, two options appear to reach the surface of the protein depending on the opening of the B-C loop: path A reproduces interactions with amino acids on the left edge of the loop such as W91, P93, R243 and Q203; path B goes further to the right side of the loop, interacting more with the G helix and other residues like W91, R92, L180, A184, R192, Q203 or N207.

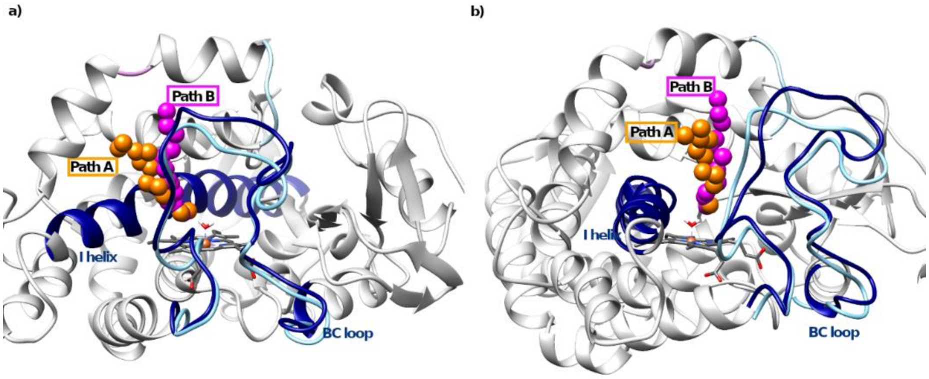


Figure 5.21. Pathways followed by para-substituted substrate and product when leaving CYP199A4 binding site. Depending on the B-C loop separation to the G-helices, one path or the other becomes more available. Two different perspectives shown: a) front view b) lateral view.

For product-bound systems, the latter is more common, showing stabilization with R243 and then pulling towards Q203 or even Q193, to leave through this B opening. Exiting of the 4-hydroxy benzoic acid is observed in all LiGaMD simulations independently of the catalytic step except for one replicate of the S2. Moreover, egression is attempted in all replicates of S1 systems during GaMD simulations and achieved in one of them. In the case of the substrate, only one replicate of the LiGaMD simulations in each catalytic step is able to reproduce egression, and reweighting experiments show a barrier 10 times higher for the substrate compared to the product.

Nevertheless, there is a methodological caveat to point out. When starting accelerated simulations from longer MD replicates (100 ns *vs* 500 ns of cMD), the egression can be observed for the substrate even during GaMD trajectories in RS and S1, as indicated in the previous chapter. It is known that accelerated dynamics are highly dependent on the starting coordinates, so these results indicate that, despite having achieved convergence of MD simulations and the system showing little to no conformational exploration, longer conventional trajectories are necessary to stabilize large systems like CYP199A4.

As it has been exposed in the previous chapter, this B-C loop opening that regulates the final steps of ligand egression is coordinated through several residues that are involved in both stabilization of the ligand in a catalysis-favoring orientation and in guidance through the entrance/leaving pathway, such as R92 or S95. For instance, the same rotameric shift towards E83 observed for Arg93 described earlier is reproduced in these systems as well (Figure 5.22a). Furthermore, B-C loop flexibility correlation with the ligand activity in the binding site is also replicated during the product-bound simulations: broad exploration of alternative orientations to the pre-catalytic one induces larger flexibility of the loop, showing higher RMSF values for its residues; meanwhile, static simulations like substrate-bound CpdI have much lower RMSF indices (Figure 5.22b). This confirms the tight correlation between entrance or egression events and stabilization of organic ligands inside CYP199A4.

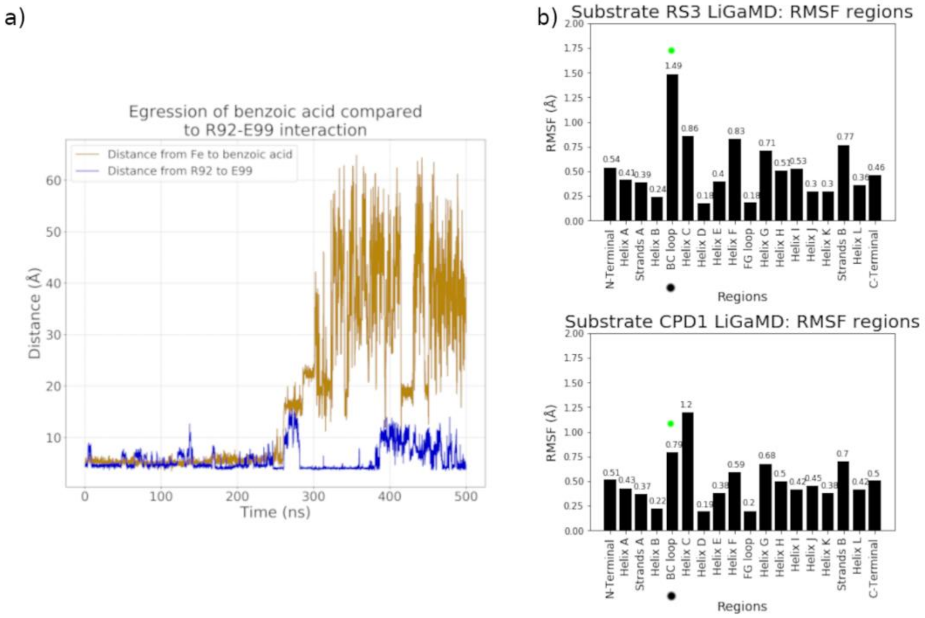


Figure 5.22. a) Comparison between product egression and B-C loop R92 interactions along a ferric S1 product-bound simulation. In dark blue, distance between R92 guanidinium carbon and E99 carboxylic atom and in yellow distance between the carboxylic atom in the benzoic acid product and the Fe atom. b) RMSF evolution of different regions along two distinct trajectories.

As described earlier, the exit of the ligand is accompanied by an increase in water content in the binding site and around the molecule itself. This has been previously demonstrated for product-containing simulations (Figure 5.18), but it is again emphasized in the cases where the substrate leaves. In particular, the S2 system under LiGaMD conditions displays substrate egression in one of the replicates, accompanied by an increase in water content within 5 Å of the Fe precisely when the ligand starts shifting towards the exit channel, reaching up to five water molecules when the substrate has completely left the surroundings of the protein. In the other two replicates where the benzoic acid stays in hydrophilic position for more than 80% of the trajectory, no waters are detected in the binding site under any circumstance (Figure 5.23a). Moreover, the substrate becomes surrounded by water molecules as it reaches the protein surface (Figure 5.23b).

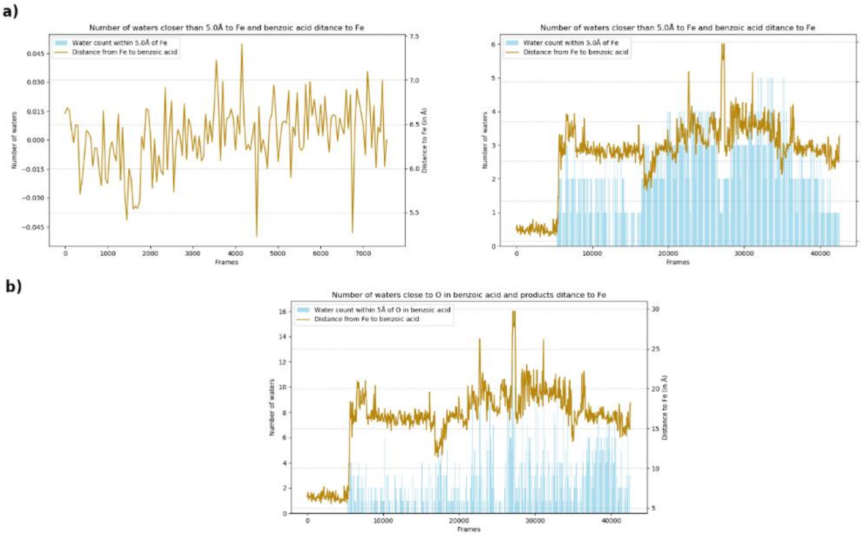


Figure 5.23. a) Two replicates of substrate-bound Fe^{3+} S1 systems, on the left the ligand does not abandon the binding site, on the right it does. b) Number of water molecules around the methoxy substituent of the substrate as it leaves from the binding site.

This further supports the hypothesis that water molecules in the binding site induce egression, being this solvation more accessible to the product-containing active center due to its hydrophilic substituent. In this line, two main water pathways to the binding site have been identified along GaMD simulations (Figure 5.24Figure 5.24). Both channels specifically lead to the axial coordinating position of the Fe ion. The first one starts between helices C and L, near residues R125, Y155, Y177 and E367, subsequently accessing the kink in helix I over A248 to finally reach the metal vicinity. The second water path goes through the B-C loop/heme propionate interface similarly to other previously reported paths.[50] The latter has only been identified in product simulations, highlighting the higher solvation tendency of the product-bound systems and once again the relevance of the loop for the correct functioning of the enzyme.

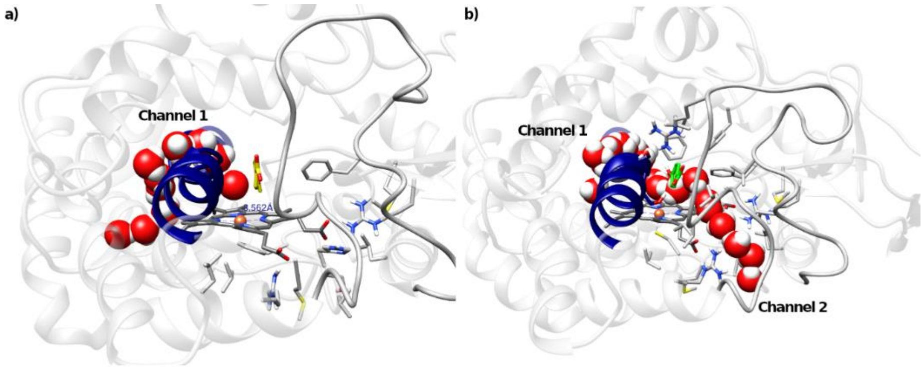


Figure 5.24. Water channels connecting the surface of the enzyme with the binding site found during substrate (a) and product-bound (b) simulations. Key regions have been highlighted: the B-C loop and the heme group are depicted in grey and the I helix in dark blue.

5.6. Conclusions

Several critical findings have been obtained from the studies reported in this chapter:

- Reactivity of CYP199A4 is dependent not only on the catalytic profile of each substrate but especially on pre-catalytic orientation, fixed by the carboxylate polar interactions and van der Waals hydrophobic stabilization of the benzene ring.
- The sixth coordinating water molecule plays a key role along the catalytic cycle: its displacement is essential for catalytic cycle activation, and its reincorporation is linked to product egression.
- The entrance pathway to CYP199A4 binding site is relevant in selectivity of the enzyme but not in regioselectivity, and it is shared between substrates and products.
- Rational prediction of suitable alterations to modify enzyme selectivity can be grounded on multiscaling protocols. However, effective substrate selectivity modification cannot be achieved through single point mutations. As this complex mechanism is governed by interactions on different areas of the structure such as the B-C loop, active site entrance or the hydrophobic binding site ceiling, cooperative transformations along key regions should be considered.

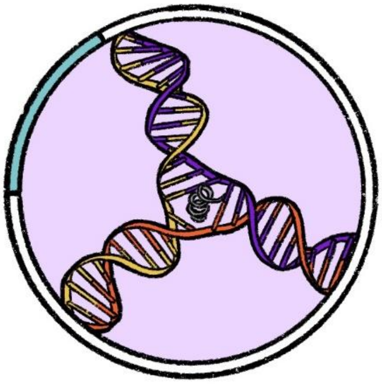
Regarding the work on the first part of this Chapter, the results from the comparison between *meta*- and *para*-substituted methoxy benzoic acids highlight the importance of integrational approaches. Combination of dynamical studies with energetic calculations emphasizes the double source of regiospecificity of CYP199A4: the *para*-substituted substrate achieves stabilization in pre-catalytic orientation and presents a lower energetic barrier for the rate determining step of the catalyzed reaction.

Moreover, RS and S1 catalytic steps models display lower stabilization of the catalysis-favoring orientation, which is coherent with the water-free Fe coordination obtained in the crystallographic structures and indicates that the binding of the preferred substrate rapidly triggers catalytic activation by displacing the water molecule and favoring reduction of the ferrous species. Meanwhile, the *meta*-bound simulations indicate more disperse distribution of the ligand in different parts of the active site or even outside of the protein environment, indicating that CYP199A4 is not tailored to fit the analogue molecule.

In the second part, molecular modeling of the 4-hydroxybenzoic acid-bound CYP199A4 shows an alternative way of accommodating the solvent inside the binding site compared to the X-ray structure. The increase of water content in the product-bound models compared to the substrate is still observed, although it results from a global rearrangement of the active site residues more than the sole Phe298 orientation.

Moreover, the hydroxylated product displays more egression events during the simulations, which occur through the same pathways observed for the substrates. These events are always accompanied by an increase of water molecules in the binding site and around the aromatic compound, directly linking the water content with the product release mechanism. Even S1 simulations replicate RS conditions in the presence of the product, highlighting its tendency to restart the catalytic cycle recovering the water-bound hexacoordinated ferric ion species. This is further emphasized by the identification of an extra access channel for solvent molecules into the active site exclusively in product-bound systems.

The application of classical and, especially, enhanced molecular dynamics to the study of this enzyme has provided complementary insights into the large amount of experimental data available. The ground base of knowledge about CYP199A4 reactivity has been enlarged, opening the door to further rational design for specific purposes like substrate regiospecificity modifications from a deep molecular understanding.



CHAPTER 6
Other projects

During the course of this thesis, the training on different methodologies and tools has been expanded through participation in complementary projects regarding additional systems. This chapter comprises a brief overview of three of them, including studies on DNA three-way junction recognition, characterization of a self-assembling supramolecular metallocage and exploration of large-scale conformational rearrangements of the bilobular protein SBD2.

6.1. Peptides for the selective recognition of three-way DNA junctions

Three-way junctions (3WJs) are DNA structures consisting of three interconnected double-helical arms that typically appear during defective replication or realignment of repetitive sequences in the genetic material known as microsatellites. Both microsatellite instability (MSI) and defects in the mechanisms that fix misaligned repeats have been linked to cancer propensity.[154–156] Moreover, 3WJs branching also plays a part in repeat expansion diseases, genetic conditions that affect especially the nervous system and appear due to the unfold and unproper re-annealing of repeat sequences. Therefore, recognizing these specific DNA assemblies for targeted action in anomalously replicating cancer cells or stabilizing canonical DNA conformation can lead to potential new therapies for detrimental diseases.

The branching point where the DNA strands interlock presents a hydrophobic cavity of about 12 Å of diameter. Different molecules that can selectively bind this active site in 3WJs have been reported, including metallopeptides, triptycenes and metal helicates, which are metallosupramolecular entities formed by multidentate ligand arrangement around metal ions resulting in single, double or triple-stranded helices.

Through collaboration with the group of Prof. E. Vázquez and M. Vázquez, two different molecules have been devised for molecular recognition and binding to the central cavity of 3WJs, and molecular modelling has been applied to aid the design and characterize the different interactions involved. On the one hand, an α-helical peptide with high hydrophobicity that presents adequate dimensions to fit the branching point was studied to determine structural evolution in presence and absence of the 3WJ. On the other hand, a Cu and Ni helicate was designed to incorporate an ATCUN motif for enhanced DNA targeted cleavage.

6.1.1. 3WJ recognition by α-helical peptide

To ensure fitting in the hydrophobic branching point of 3WJs while establishing further interactions with the negatively charged phosphate skeleton of the DNA, the α-helical peptide was specifically designed to combine a 10-Ala amino acids chain with 4 Lys residues at the C-terminus. To assess the structural stability of the sequence and any possible influence of the DNA on the fold, classical molecular dynamics were carried out for both the free peptide and the DNA-peptide complex. The amino acids were represented through the AMBER19SB force field, while BSC1 was selected for the DNA scaffold.

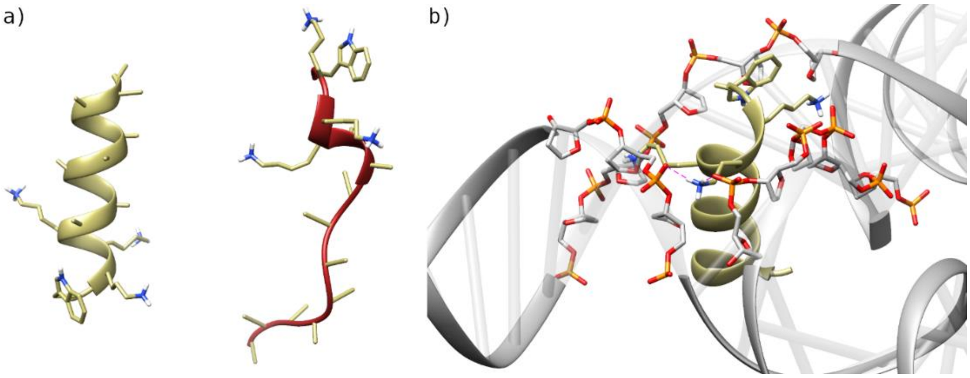


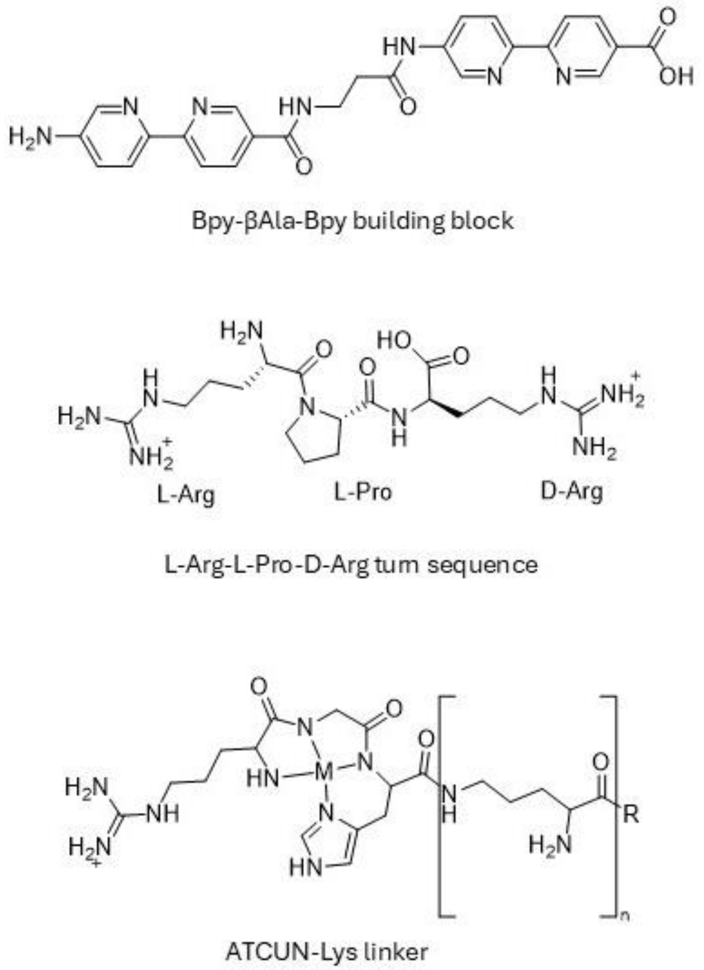
Figure 6.1. a) Evolution of helical peptide during MD simulations. In green the initial structure, in red the fold after 200 ns of simulation. b) Frame of the peptide-3WJ simulation. Inter-molecular interactions are highlighted in pink

Simulations of the peptide by itself present destabilization of the helical structure within the first 200ns of cMD simulations. Interestingly, when including the DNA environment not only the α-helix was stabilized but the peptide-3WJ interactions were stable throughout 1 μs of trajectory for two replicates (Figure 6.1). These stabilizing forces appeared mainly between the Lys side chains and the phosphate groups. Moreover, two different orientations were analyzed, starting with the Lys-containing C-terminus facing either the concave or convex side of the 3WJ. No distinction between the two arrangements was observed, as the DNA shows a tendency to flatten, eliminating the cavity and equilibrating the interactions for both orientations.

These results demonstrate not only that the peptide fits the cavity but also its suitability for stable recognition of the 3WJ branching point, which provides a hydrophobic pocket that greatly preserves the helicity of the amino acid sequence.

6.1.2. Helicate design for directed ATCUN-DNA interactions

This second project involves recognition of the 3WJ by a peptide helicate. More specifically, the selected structure is composed of three strands, each formed by two 2,2'-bipyridine (Bpy) amino acid derivatives connected through a β-Ala residue (Figure 6.2). These strands are connected through a type II or type II' turn formed by ^{L}Arg-^{L}Pro-^{D}Arg sequences specifically selected to induce the desired chirality while increasing the solubility of the supramolecule. This sequence folds into a triple strand helicate upon metal binding in two enantiomeric centers formed by the Bpy residues. Coordination with Cu(II) or Ni (II) induced ΔΔ and ΛΛ chirality respectively (Figure 6.3).



Bpy-βAla-Bpy building block

L-Arg-L-Pro-D-Arg tum sequence

ATCUN-Lys linker

Figure 6.2. Components of the helicate structure.

Moreover, the N-terminus of the non-natural peptide is elongated with a Lys linker that connects a RGH N-terminal sequence, which forms another metal binding center for an extra metallic ion known as ATCUN. These ATCUN motifs have been identified as ROS generating centers used in targeted DNA cleavage. Therefore, the objective of the project is to integrate the 3WJ recognition capacity of the helicate assembly with the DNA damaging function of the ATCUN motif to create a targeted drug. For this purpose, assessment of the optimal linker length to ensure ATCUN motif-double helix DNA interactions was carried out through computational means.

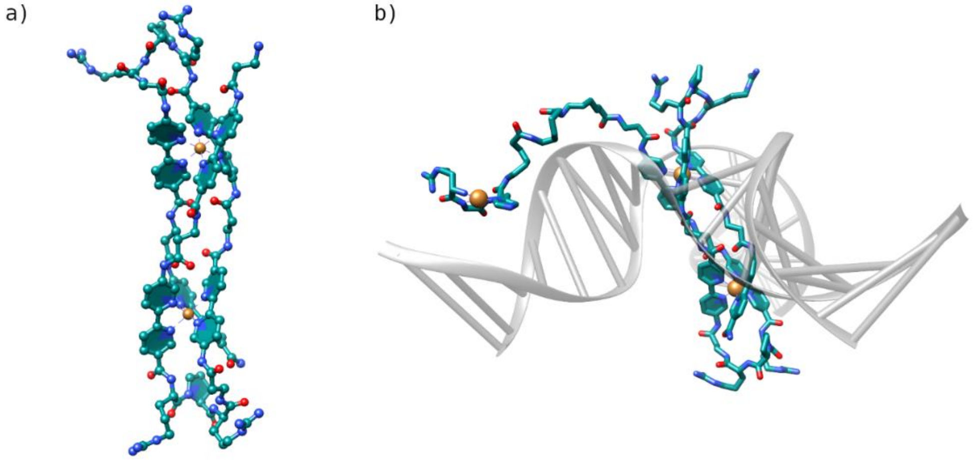


Figure 6.3. a) Helicate fold of the peptide, bipyridine residues are depicted with the aromatic ring filled. b) Best scoring pose of L3 with the ATCUN motif at one end and helicate peptide core docked inside the 3WJ.

Given the complex nature of the molecule, a multistep molecular modelling-based approach was engaged as follows:

1. Normal Mode analysis to search for structures of the 3WJ suitable for binding.
2. Docking of the M(II)helicate core to the most opened structure of 3WJ.
3. Molecular mechanics-based minimization to relax the obtained solution.
4. Covalent docking of 4 linkers with different lengths (from 1 lysine residue up to 4, designated as L1, L2, etc.) and the ATCUN motif to the best solution in 2.
5. Molecular mechanics-based minimization to relax the obtained solution.

For the Molecular Docking, GaudiM^2 was used for both helicate inclusion in the 3WJ and ATCUN interaction with the DNA branches. In step 2, unrestricted exploration of the site was allowed; moreover, given the rigid nature of the supramolecular assembly, only conformational exploration of the Arg side chains was considered. For the fourth step, a fixed interaction between the first Lys and the N-terminal of the first β-Ala was imposed; free conformational exploration for the Lys linkers was implemented through a specific rotation module programmed for this system. This was necessary since the Lys residues were non-conventionally connected between the side chain amino group and the acid of the following residue, so basic amino acid force fields did not recognize the molecular assembly. Steps 3 and 5 were carried out in the YASARA suite, using AMBER14 and the generalized YASARA force field to describe the system. In this case, a simplified bonded model was used for the metals, giving them fixed distances to the coordinating residues.

As a summary of the results, L1 and L2 linkers were too short to reach any grooves in the DNA arms of the 3WJ. L4 presented promising binding scores, although the long linker adopted extended conformations that placed the ATCUN motif too far from the DNA. In the cases that did show interactions within the nucleotide grooves the linker presented intricate, disfavoured arrangements. L3 was found the best size for the linker, providing reasonable scoring solutions that comfortably located the ATCUN motif within reach of the DNA grooves (Figure 6.3b).

This project allowed exploration of alternative MM approaches for complex systems with personalized Molecular Docking simulations and MM-based relaxation. The proposed linker was successfully tested experimentally, demonstrating suitability of the protocol for the optimization of metal complex-DNA interactions.

6.2. Self-assembling supramolecular metallocages

Self-assembling supramolecular entities are of great interest for their defined architectures that can redefine the characteristics of space within their limits.[157,158] Their properties are applied for catalysis in confined spaces, molecular recognition, materials science and many other fields. Particularly, metallocages are composed of one or more organic ligands that are connected through metal ions to generate the three-dimensional structure of interest.[158] Despite the focus on transition metal chemistry for their valuable coordinative and reactive properties useful in supramolecular assemblies, the incorporation of main-group metals can be beneficial as they present lower toxicity and are more abundant, reducing the costs of the innovative systems.[159]

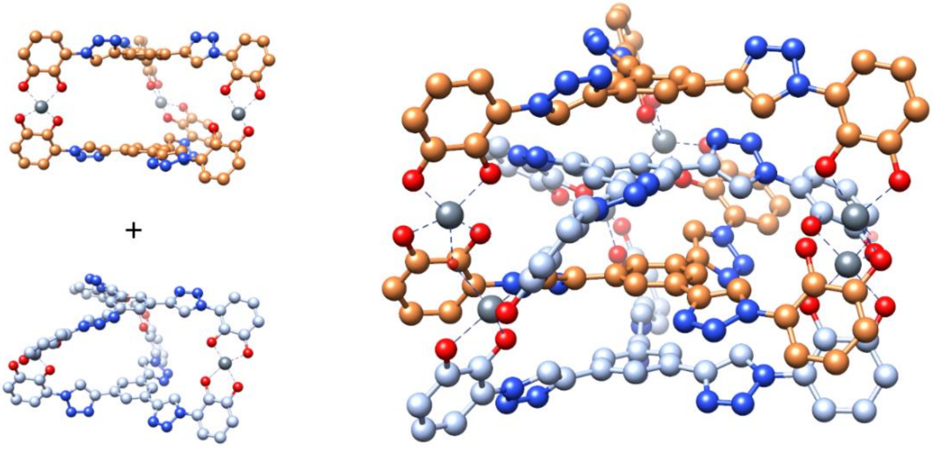


Figure 6.4. Separated metallocages (left) that assemble in a quadruple decker structure connected through catecholate nodes and hydroxyl bridges (right).

In this work, a new main-group based metallocage was characterized, revealing remarkable reversible assembly of two M3L2 halves into one quadruple-decker structure connected through metal-catecholate nodes and hydroxyl bridges (Figure 6.4). DFT optimization of the structure revealed non-covalent interactions (NCI) driving the assembly, mainly π-stacking between the ligand layers of the interlocking parts. Moreover, comparison between two metallic elements, namely Ga and Al, revealed absolute energy values more favorable for Ga in line with the experimental findings through NMR and ITC calculations.

6.3. SBD2 protein conformational change exploration

This work was performed during the international research stay at ITE-FORTH institute in Heraklion, Greece. The project was focused on the bilobular protein SBD2 (substrate-binding domain 2) involved in amino acid transport in bacteria. This enzyme is composed of two main domains that suffer a large closing conformational change to bind the ligand in the central active site, in this case the glutamine amino acid, although allosteric activation involving distal domains has been identified in similar scaffolds.[160] The objective was to detangle the elements triggering such motions, combining experimental structural determination techniques like smFRET or HDX-MS with molecular dynamics.

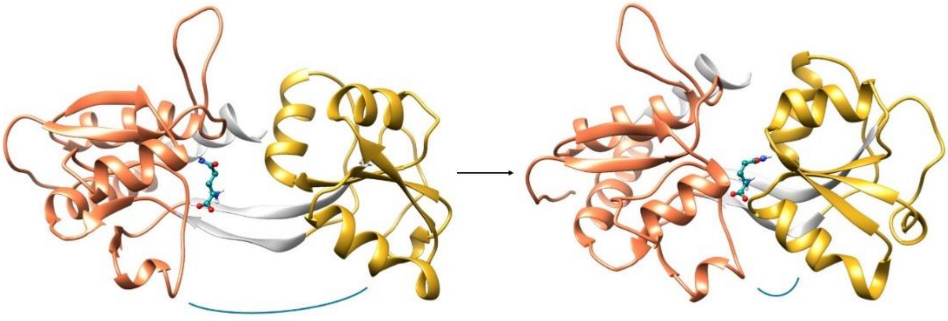
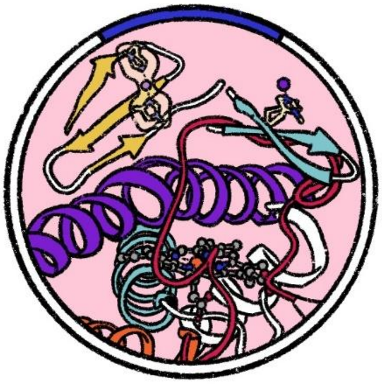


Figure 6.5. Left, open structure of SBD2 with glutamine bound. Right, closed structure of SBD2 with glutamine in the binding site. Each domain is depicted in orange and yellow, and the substrate is depicted in blue.

From a computational point of view, determination of the proper protonation state of the substrate for the simulations was a first crucial step in reproducing the behaviour of the protein. Closing motions were simulated through long-range cMDs of 1 μs, which allowed the analysis along the trajectories of determinant residues identified through mutational experiments (Figure 6.5). Additionally, the large conformational change was also reproduced in GaMD simulations within 100ns of the trajectory. The combination of both types of simulation techniques allowed a wider exploration of the conformational space in smaller amounts of computational efforts to provide broader analyses to complement and rationalize experimental observations.



CHAPTER 7
Conclusions

This thesis has explored the intricate relationship between structure and function in metal-mediated biocatalysis through the combination of experimental and computational approaches. Multi scale molecular modelling of the different systems has granted access to predictive protocols for metallopeptide design, as well as proved the relevance of dynamic perspectives in the analysis of enzymes.

The first project in Chapter 4 presents the groundwork for developing a computational protocol while iterating with experimental validation on the topic of *in vivo* efficient artificial metalloenzymes. Beginning with *BioMetAll*, the prediction of best positions for metal center formation was successfully achieved and the initial six candidates were narrowed down to two promising peptide sequences with catalytic potential. Experimental testing confirmed these predictions, and a subsequent round of molecular dynamics simulations reproduced the expected behavior of well-structured peptide scaffolds coordinated to a metal ion, also observed through CD and NMR techniques.

The second section in Chapter 4 applies the described protocol to a smaller β-hairpin peptide, identifying the best coordination spots once again through *BioMetAll*. In this case, the screening of a 264-sequence library through the spot synthesis technique was implemented, arriving at the same best His residue placements as in the computational prediction. The role of structural organization for proper metal binding to enable catalysis was confirmed using two similar structures as a case study, highlighting the importance of structural arrangement in metal binding processes.

Both β-sheet-based peptide structures have been effectively modified to incorporate a metal binding center capable of carrying out depropargylation reactivity *in vitro*. Moreover, further studies have demonstrated the capability of the designed sequences to effectively perform the desired reactivity *in cellulo*, enabling cell penetration specifically through the metallopeptide biomolecule.

Chapter 5 describes the application of modeling techniques to elucidate key molecular components of catalytic reactivity in complex molecules like cytochromes P450. The combination of static structural data with dynamic simulations has revealed key regions in ligand binding and proper orientation for catalytic success. The influence of entrance or egression pathways has also been addressed with both substrates and products. Moreover, some regions like the B-C loop have been attributed key roles in both entrance and stabilization of pre-catalytic orientations, signaling their potential for modifications in the pursuit of altered reactivity of the enzyme. Furthermore, the influence of solvent molecules in the overall mechanism, especially concerning product management and its impact on the regeneration of the biocatalyst resting state, has been assessed, identifying a significant role in the increase of local hydrophilicity upon product formation and restoration of the resting state in the catalytic cycle.

Furthermore, molecular modeling expertise has been extended to the simulation of non-canonical DNA structures to investigate peptide and metallopeptide interactions in diverse systems. These studies have enabled the analysis of how biomolecular interactions influence peptide dynamics and have provided insight into the optimal design of metal complex–DNA interactions.

Future perspectives for the projects presented include implementation of an additional step in the predictive protocol to account for catalytic reactivity of the metallopeptides to complement the structural analysis; in this way further insights into the complex relationship between metal binding and actual catalytic reactivity could be obtained and implemented to enhance the overall accuracy and applicability of the methodology. In the case of CYP199A4 engineering, further optimization through alternative mutations to increase *meta*-methoxybenzoic acid efficient transformation would be desired, contemplating changes in the B-C loop regions and extending the study to other potential regions.

More broadly, in the field of biocatalysis, emerging generative design strategies may be explored based on the foundations established in this thesis. Precisely, novel reactivities could be attained through tailored *de novo* design of protein scaffolds for the implementation of suitable metal binding centers for targeted catalytic functions.

APPENDIX A

Chapter 4: Supplementary information

A1. Protocol for rational design of a mini-enzyme based on the WW domain

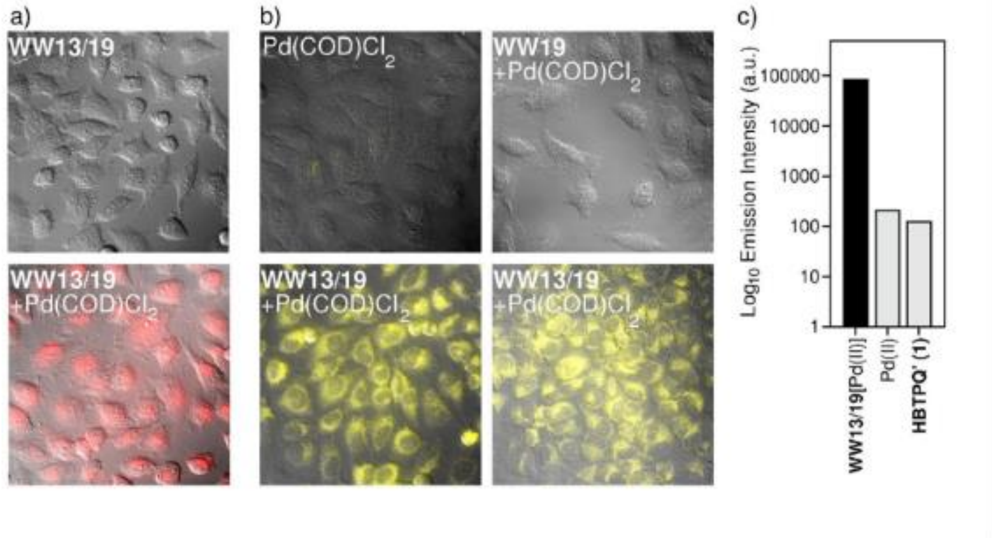


Figure A 1. Internalization of TMR-labeled peptide WW13/19 and its palladoprotein derivative WW13/19[Pd(II)]. (a) Fluorescence microscopy of HeLa cells with 5 μM solutions of the peptides WW13/19 (top) and WW13/19[Pd(II)] (bottom). (b) Cellular depropargylation of the HBTPQ' fluorogenic probe. Incubation with the Pd(II) salt (top left), WW19 incubated with the Pd(II) salt (top right), incubated first with [Pd(II)]WW13/19 and then the probe (bottom left) or the inverse order (bottom right). (c) Quantification of the intracellular emission.

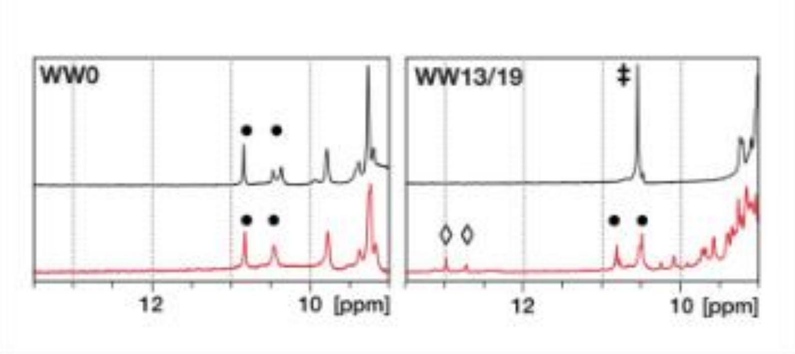


Figure A 2. 1H NMR spectra of the free peptide (black line) and after addition of the Pd(II) salt. Left, detail on WW0 peptide, which shows no change upon metal addition. Right, detail of WW13/19, which shows mostly unfolded conformation with only the resonance for the Trp residues (double dagger). Addition of the metal (red line) triggers folding of the peptide into typical WW domain structure, as demonstrated by the two Trp resonances (black dots). New imidazole resonances at ~13 ppm correspond to the H in the His residues that are now fixed due to Pd(II) coordination.

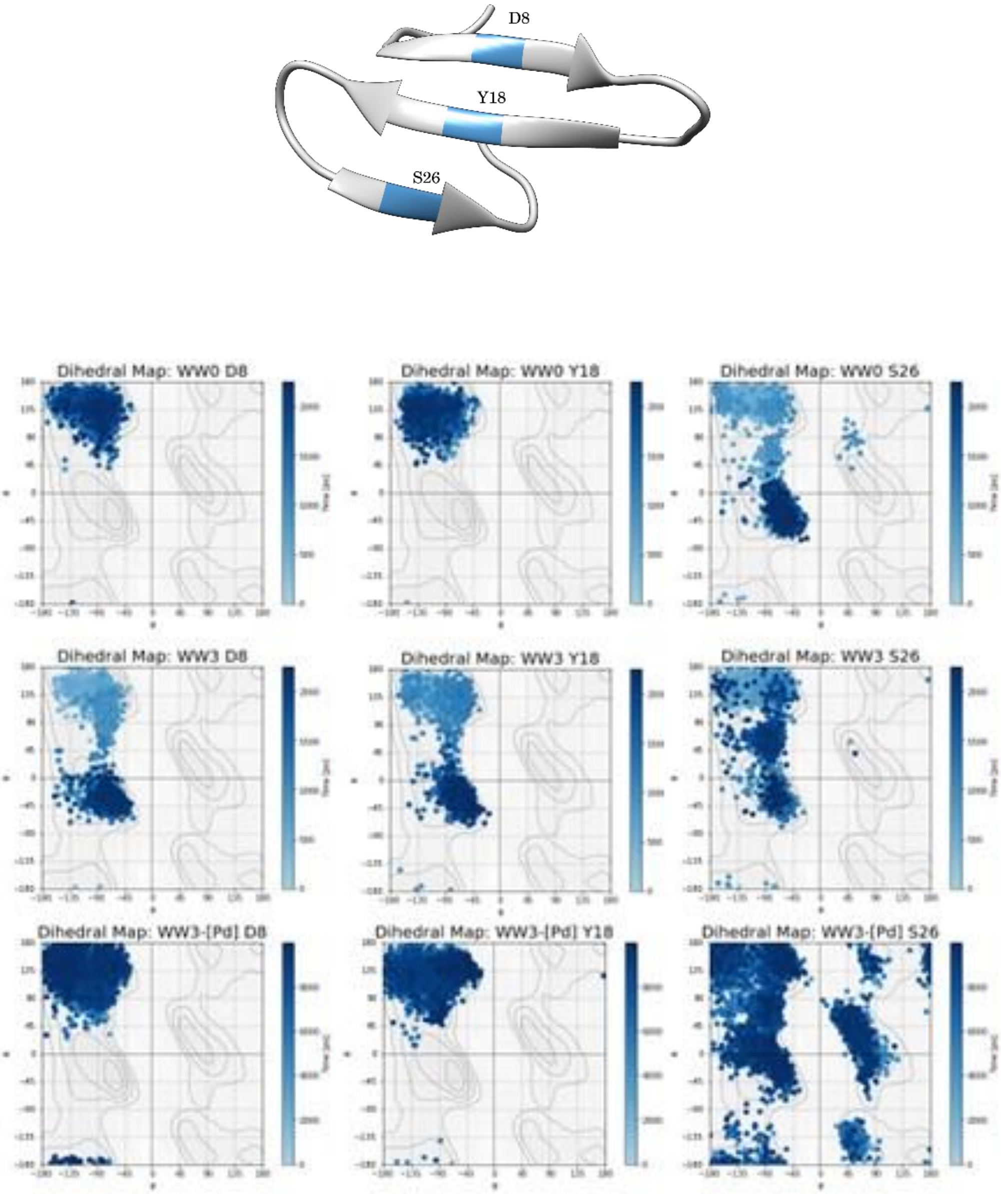


Figure A 3. Ramachandran plots of representative residues in the three β-strands (Asp8 in β1, Tyr18 in β2, and Ser26 in β3) along the GaMD simulations of WW0, WW13/19 and WW13/19[Pd(II)].

For convergence analysis, several tests are performed. Root Mean Squared Deviation (RMSD) between alpha carbons is calculated along the trajectory comparing each frame to the initial structure (a) and also comparing all frames (c), providing a vision of how much the protein changes and if it stabilizes a certain conformation. Principal Components Analysis (b) is carried out to see how much conformational space is explored and also if there are certain conformations that are visited more than once. A clustering analysis counts how many different clusters according to a cut off are in the trajectory, so the number should reach a plateau when the simulation is converged (d). Finally, Root Mean Square Fluctuation (e) for each region of the protein is calculated to see which areas are more flexible (bottom right). This applies equally to all the appendix.

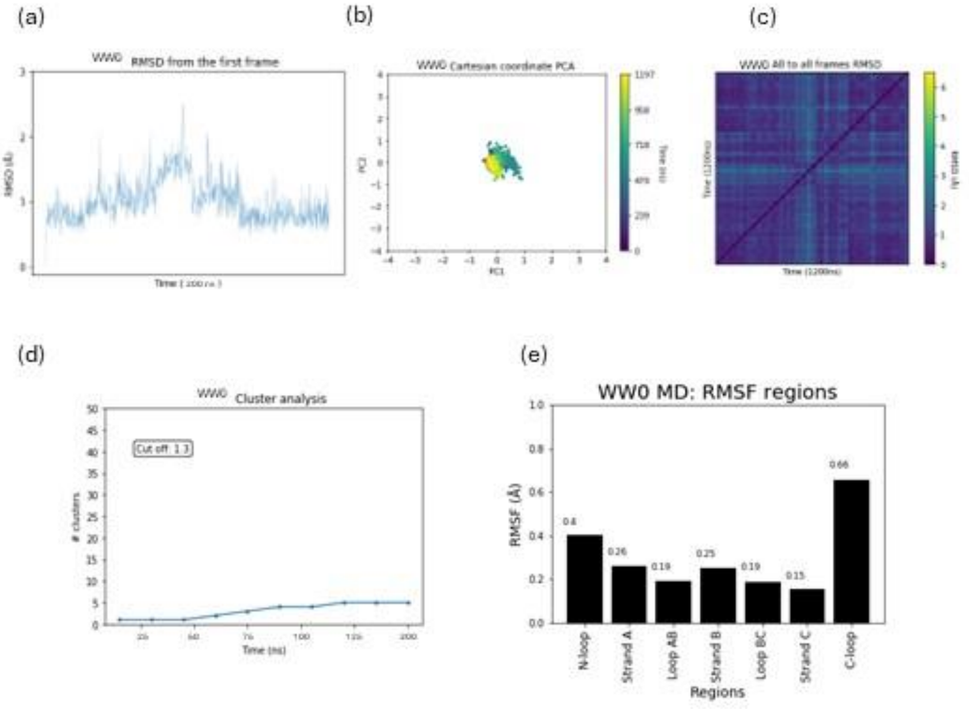


Figure A 4. Convergence analysis of MD simulations of WW0 peptide.

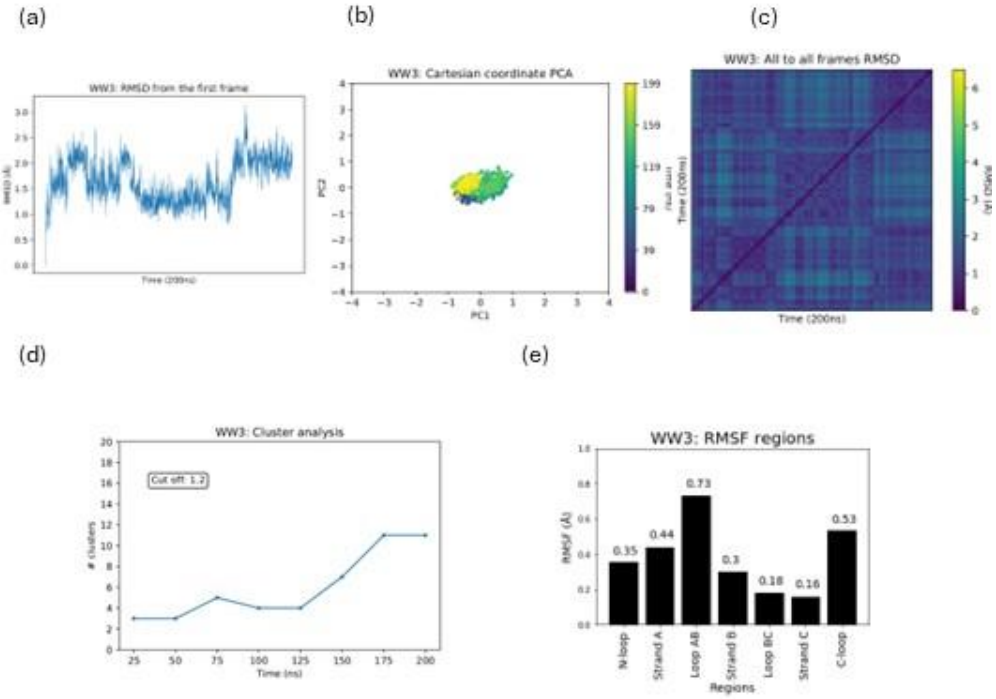


Figure A 5. Convergence analysis of WW13/19 peptide.

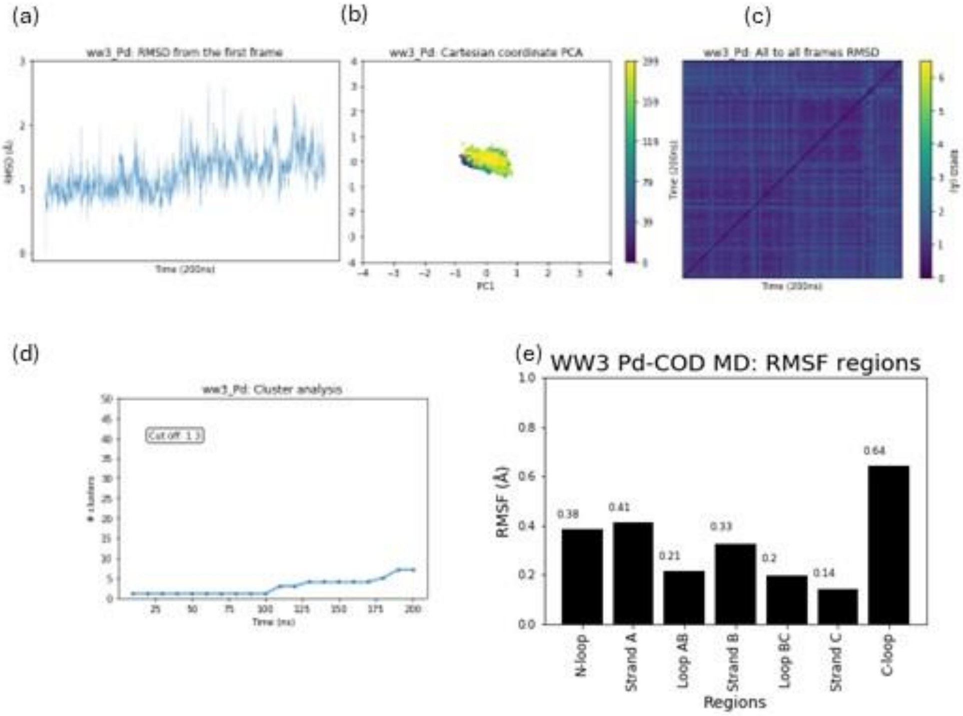
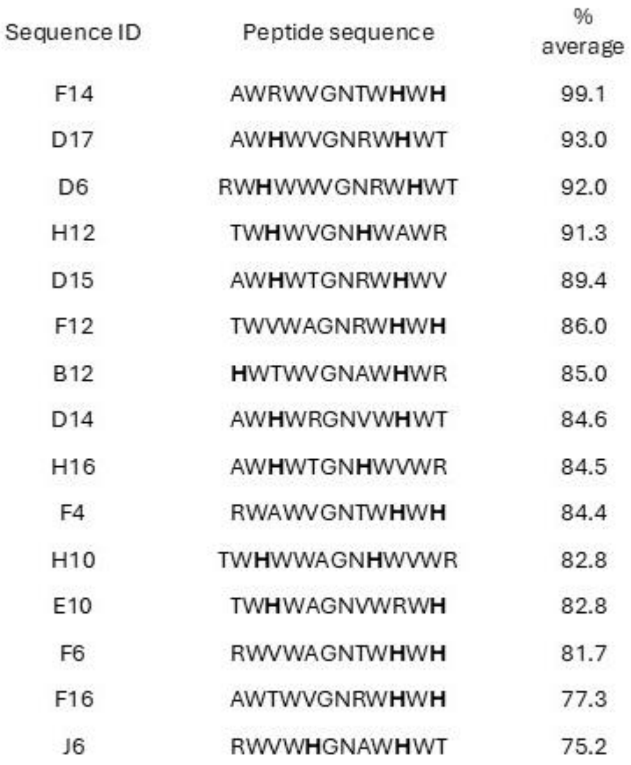


Figure A 6. Convergence analysis of [Pd(II)]WW13/19.

A2. Combinatorial libraries and molecular modeling: understanding key structural aspects in the design of stable metallopeptides



Sequence ID	Peptide sequence	% average
F14	AWRWVGNTW**H**W**H**	99.1
D17	AW**H**WVGNRW**H**WT	93.0
D6	RW**H**WWVGNRW**H**WT	92.0
H12	TW**H**WVGN**H**WAWR	91.3
D15	AW**H**WTGNRW**H**WV	89.4
F12	TWVWAGNRW**H**W**H**	86.0
B12	**H**WTWVGNAW**H**WR	85.0
D14	AW**H**WRGNVW**H**WT	84.6
H16	AW**H**WTGN**H**WVWR	84.5
F4	RWAWVGNTW**H**W**H**	84.4
H10	TW**H**WWAGN**H**WVWR	82.8
E10	TW**H**WAGNVWRW**H**	82.8
F6	RWVWAGNTW**H**W**H**	81.7
F16	AWTWVGNRW**H**W**H**	77.3
J6	RWVW**H**GNAW**H**WT	75.2

Figure A 7. Top 15 performing sequences in the spot synthesis analysis.

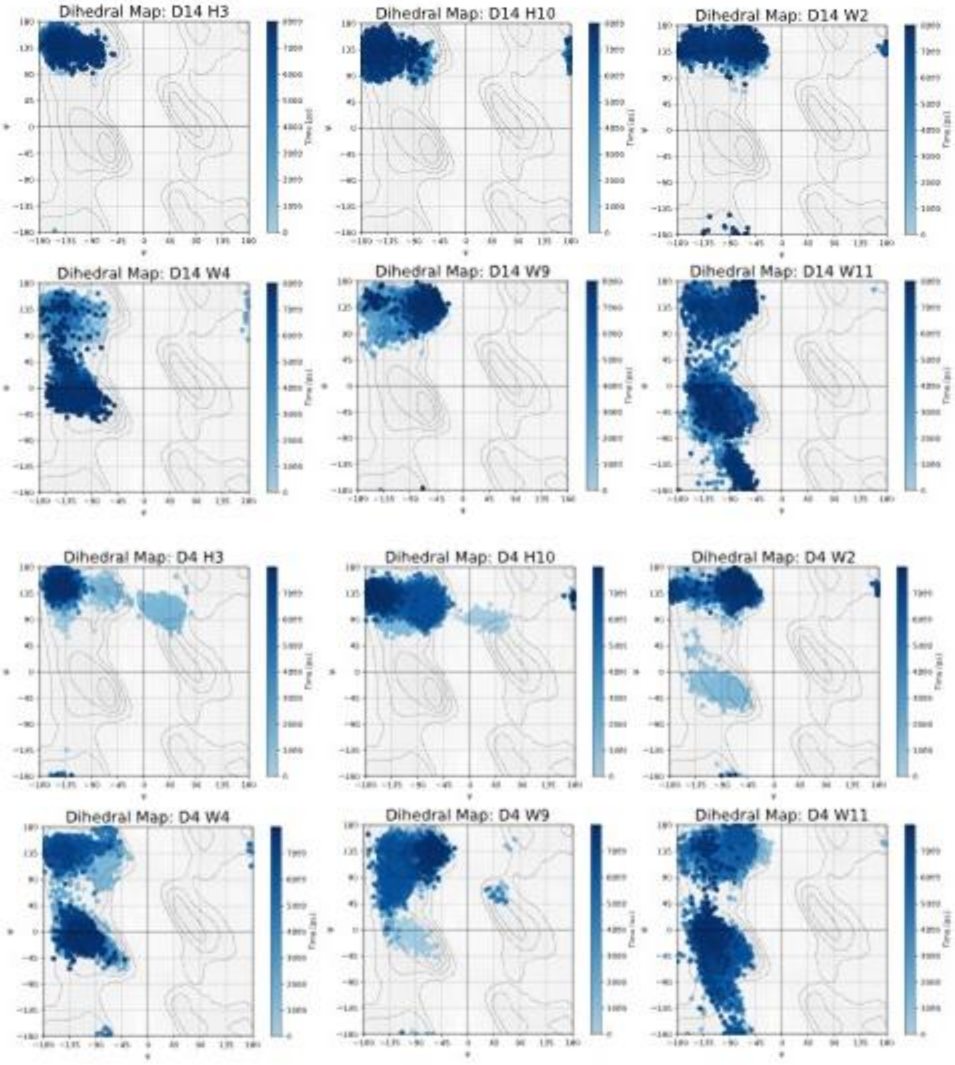


Figure A 8. Ramachandran plots of the two His residues and the four Trp for D14 (six at the top) and D4 (six at the bottom) along MD trajectories. These were calculated for the C-capped sequences.

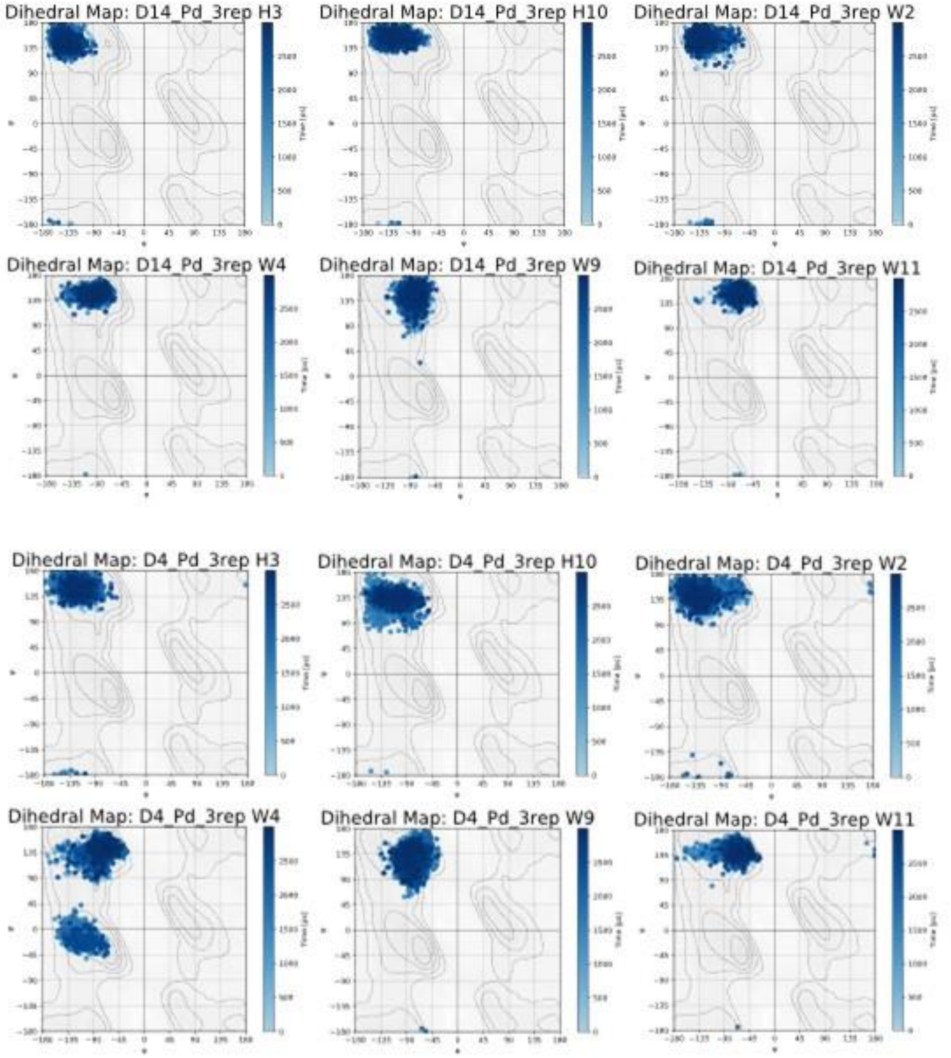


Figure A 9. Ramachandran plots of the two His residues and the four Trp for [Pd(II)]D14 (six at the top) and [Pd(II)]D4 (six at the bottom) along MD trajectories. These were calculated for the C-capped sequences.

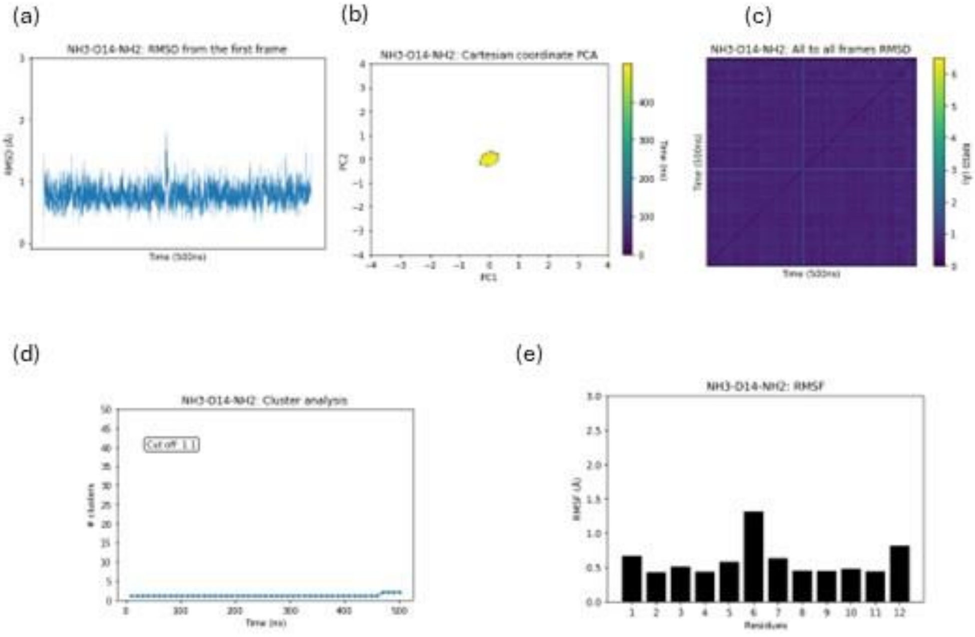


Figure A 10. Convergence analysis of NH3-D14-NH2 MD, 1st replicate.

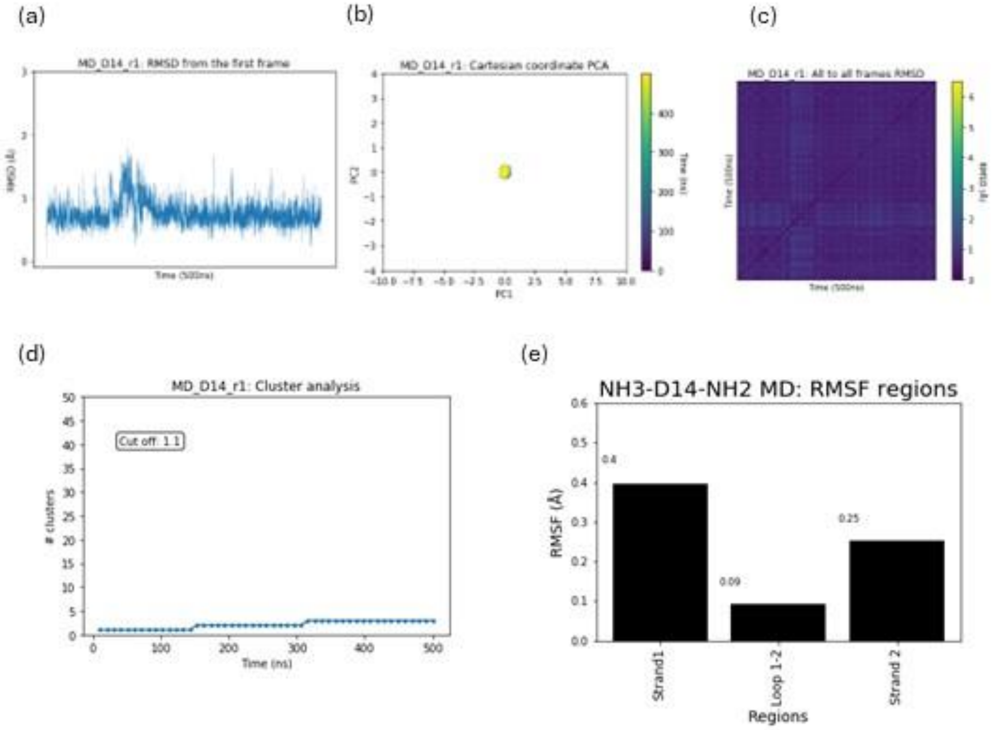


Figure A 11. Convergence analysis of NH3-D14-NH2 MD, 2nd replicate

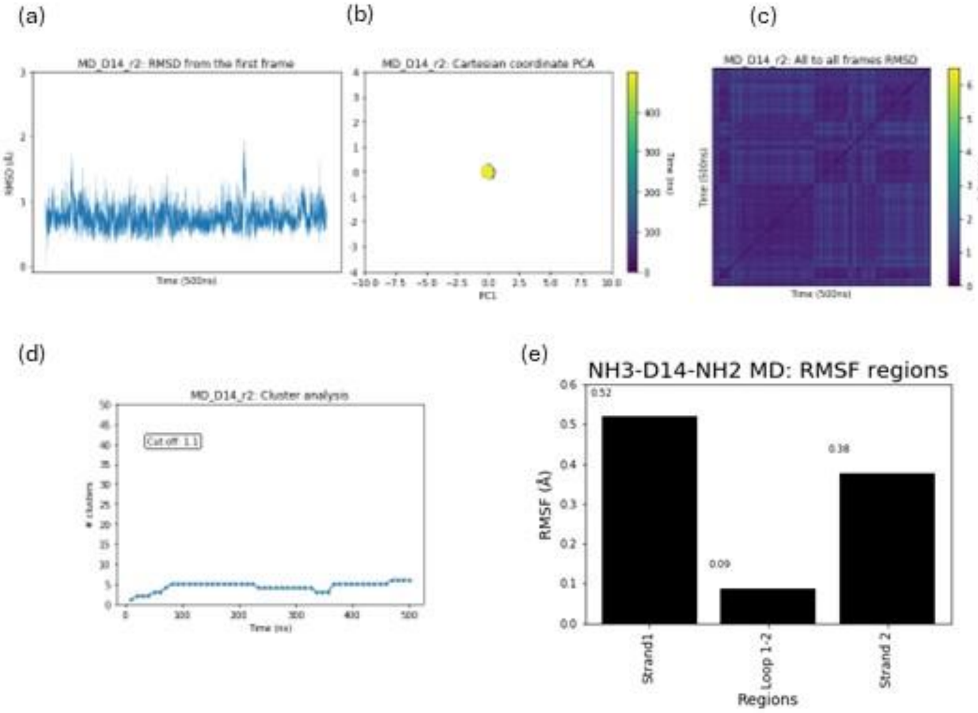


Figure A 12. Convergence analysis of NH3-D14-NH2 MD, 3rd replicate.

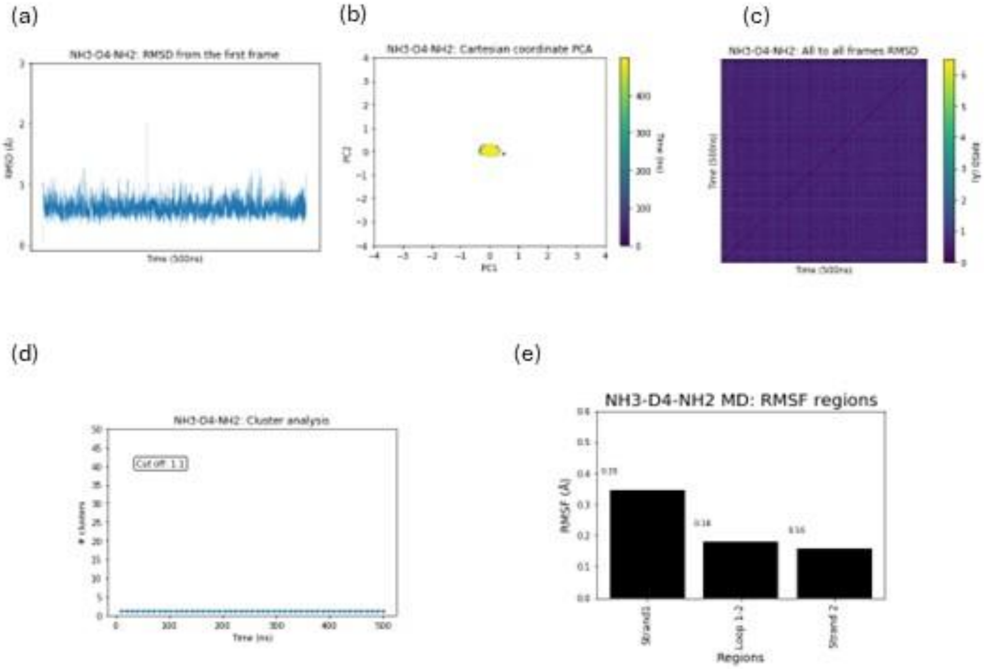


Figure A 13. Convergence analysis of NH3-D4-NH2 MD, 1st replicate.

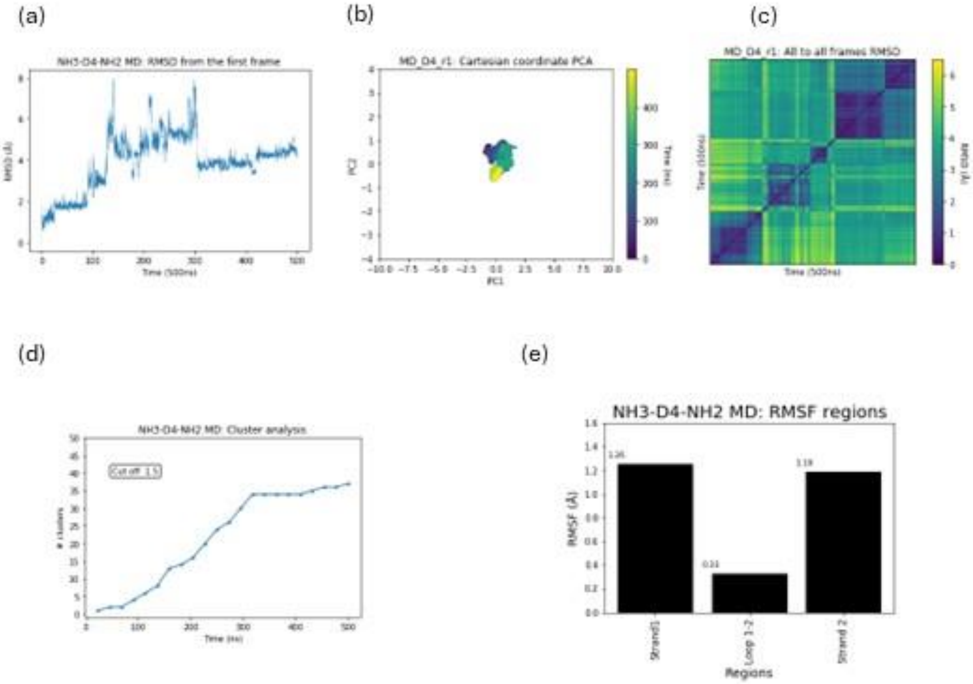


Figure A 14. Convergence analysis of NH3-D4-NH2 MD, 2nd replicate.

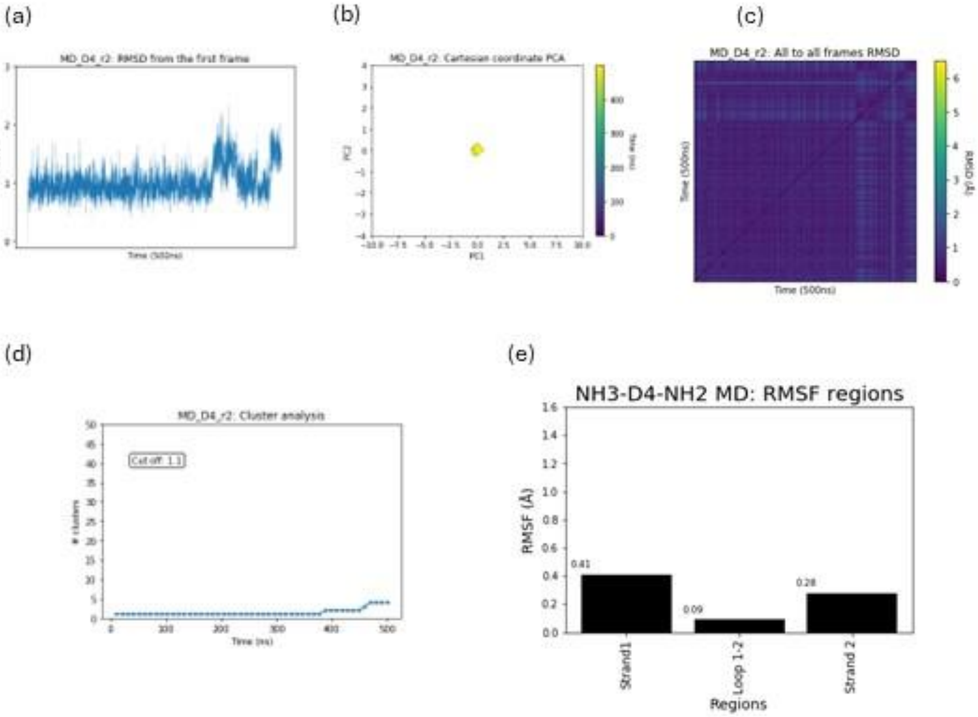


Figure A 15. Convergence analysis of NH3-D4-NH2 MD, 3rd replicate.

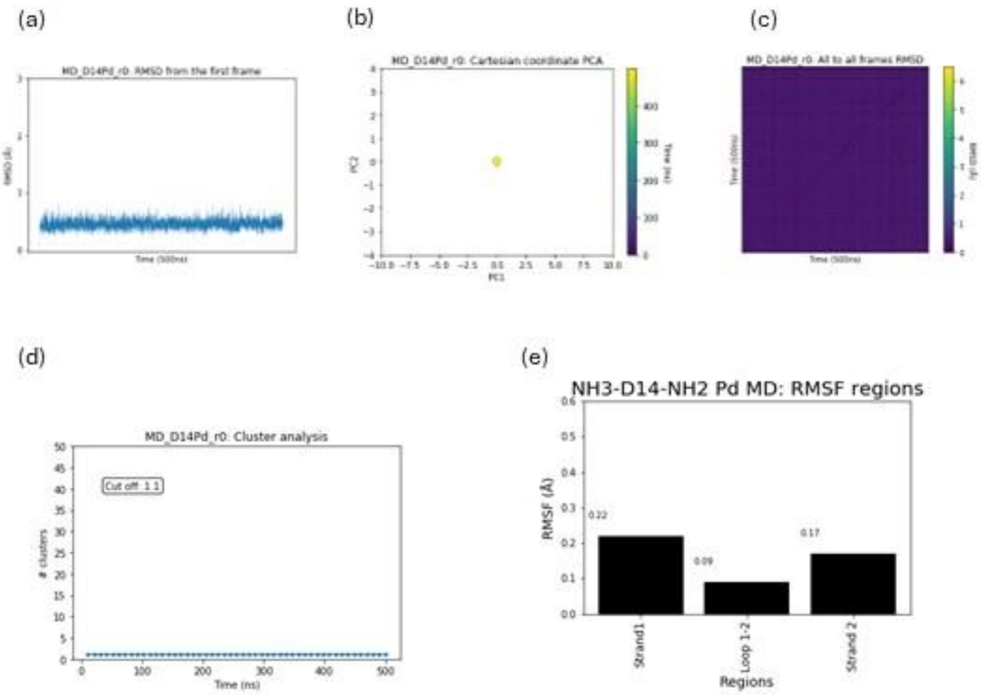


Figure A 16. Convergence analysis of [Pd(II)]NH3-D14-NH2 MD, 1st replicate.

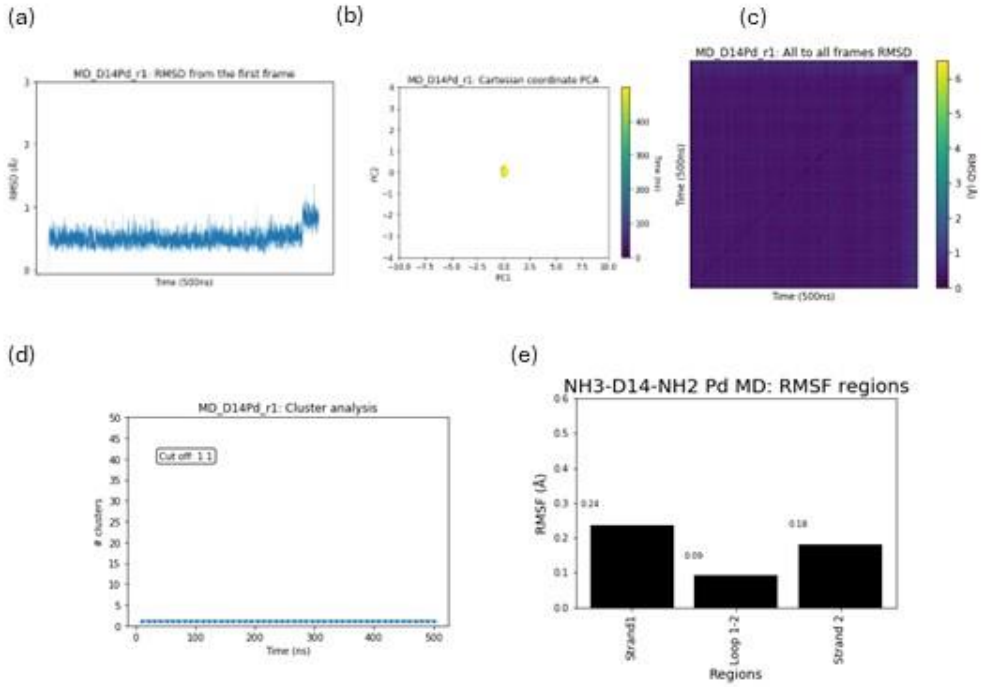


Figure A 17. Convergence analysis of [Pd(II)]NH3-D14-NH2 MD, 2nd replicate.

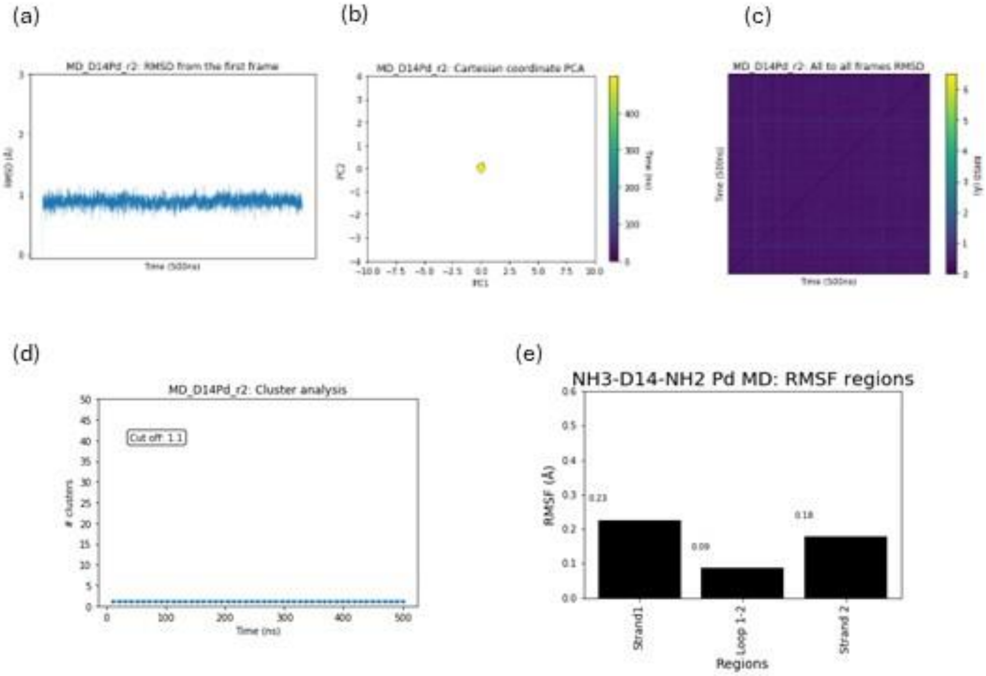


Figure A 18. Convergence analysis of [Pd(II)]NH3-D14-NH2 MD, 3rd replicate.

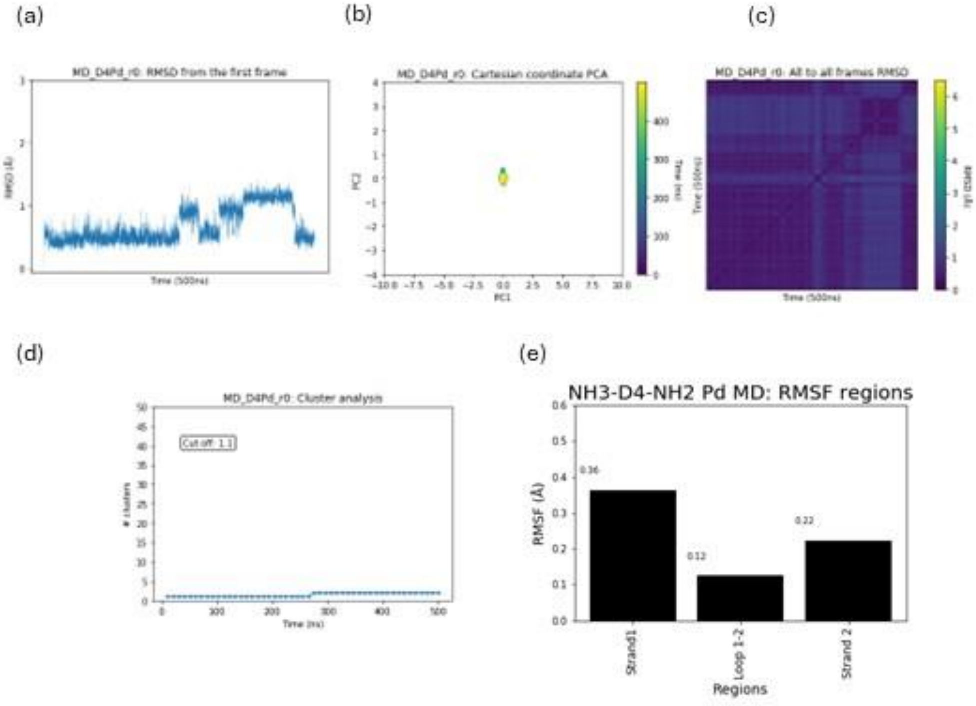


Figure A 19. Convergence analysis of [Pd(II)]NH3-D4-NH2 MD, 1st replicate.

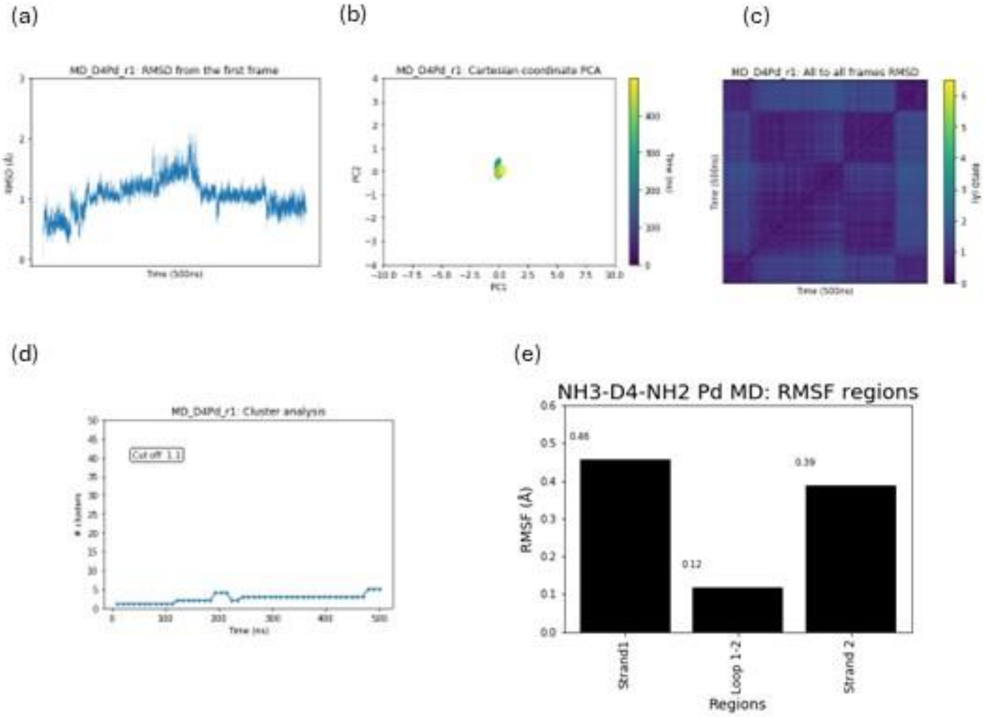


Figure A 20. Convergence analysis of [Pd(II)]NH3-D4-NH2 MD, 2nd replicate.

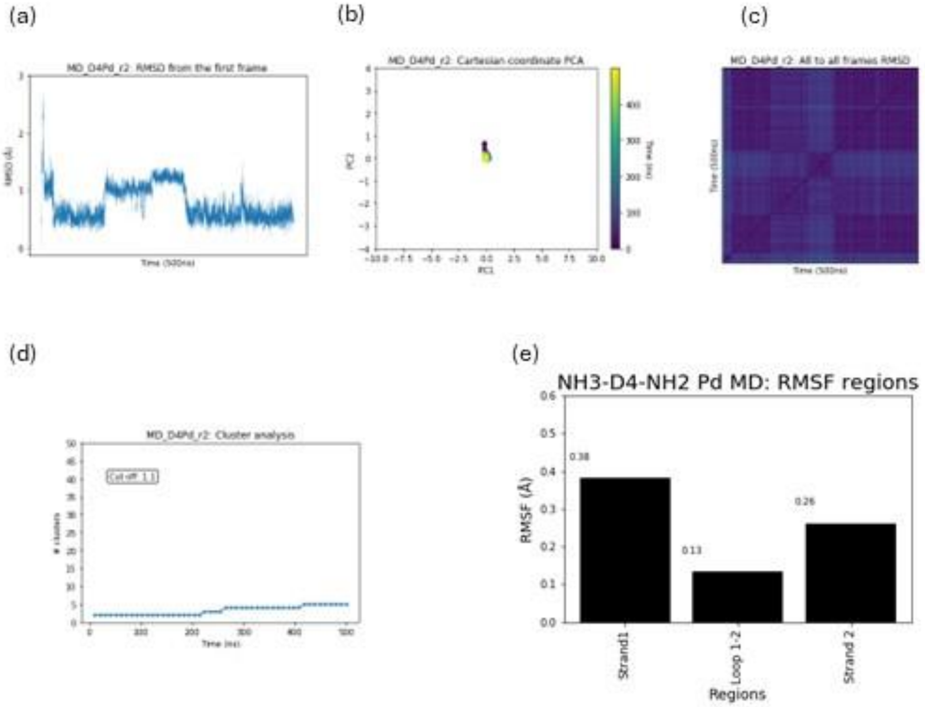


Figure A 21. Convergence analysis of [Pd(II)]NH3-D4-NH2 MD, 3rd replicate.

APPENDIX B

Chapter 5: Supplementary information

B1. Convergence of 4-methoxybenzoic acid-bound CYP199A4 simulations

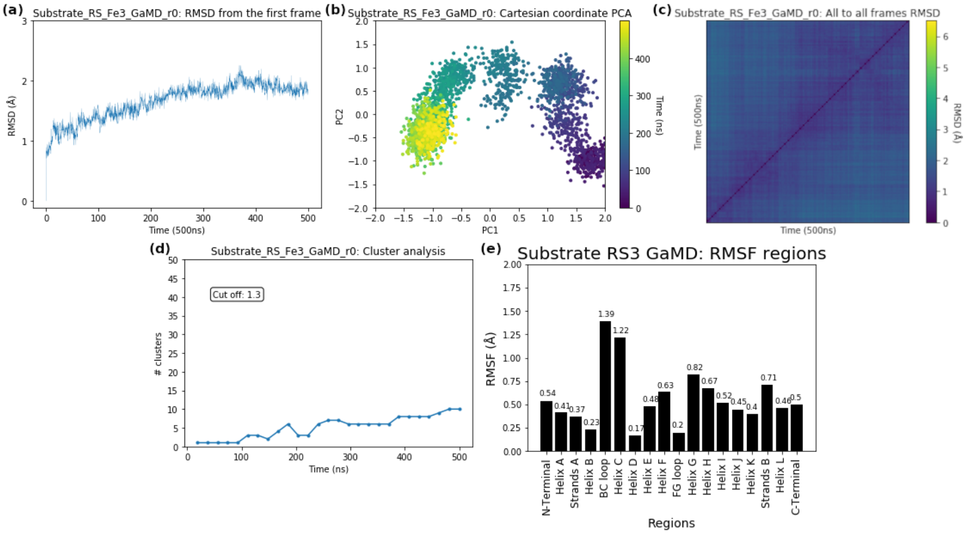


Figure A 22. Convergence analysis of GaMD simulation of CYP199A4 bound to substrate 4-methoxybenzoic acid in RS, 1st replica.

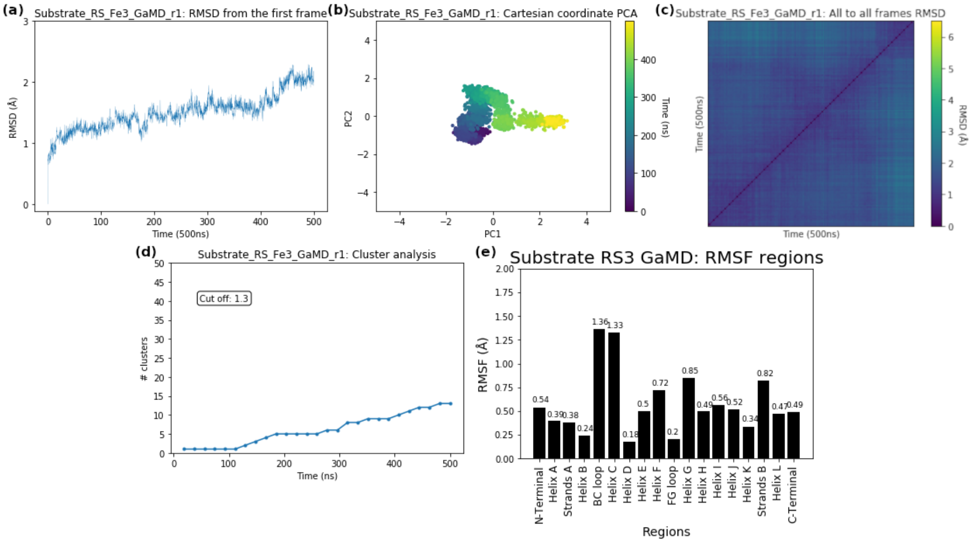


Figure A 23. Convergence analysis of GaMD simulation of CYP199A4 bound to substrate 4-methoxybenzoic acid in RS, 2nd replica.

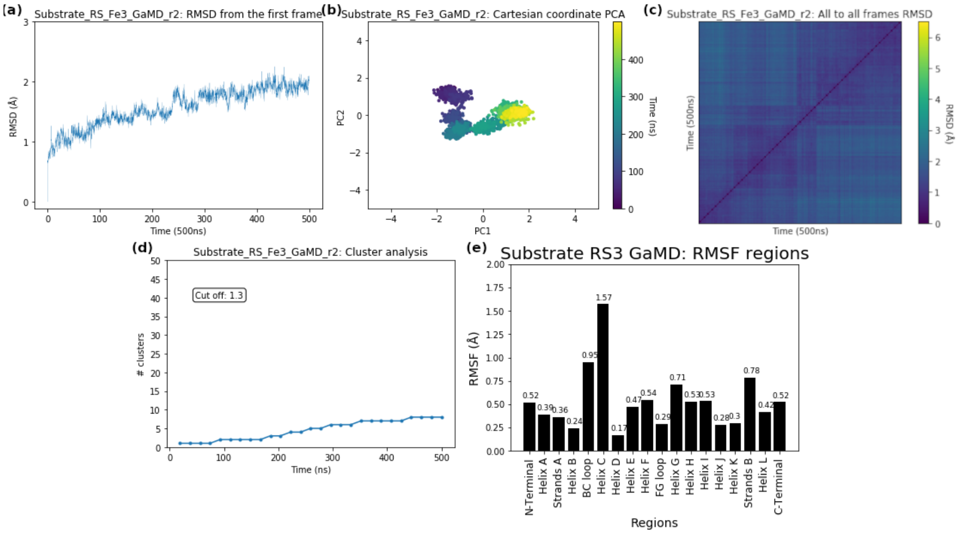


Figure A 24. Convergence analysis of GaMD simulation of CYP199A4 bound to substrate 4-methoxybenzoic acid in RS, 3rd replica.

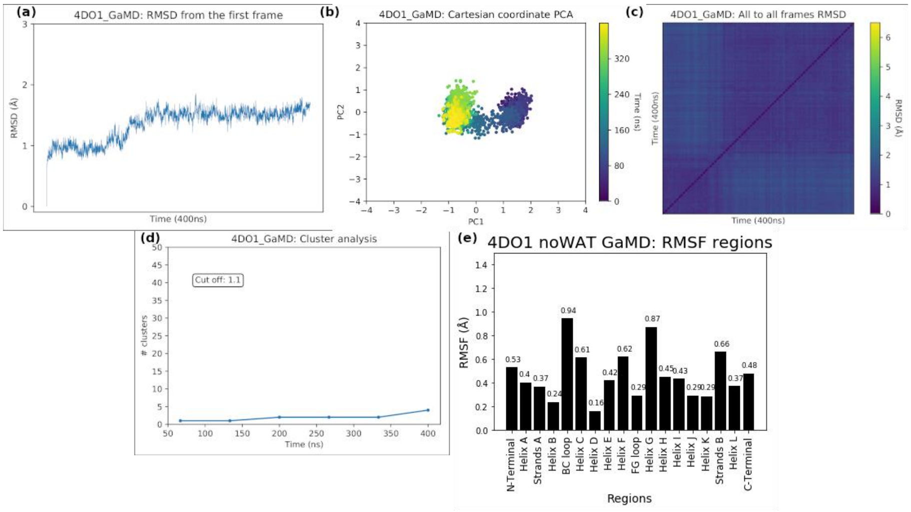


Figure A 25. Convergence analysis of GaMD simulation of CYP199A4 bound to substrate 4-methoxybenzoic acid in S1, 1st replica.

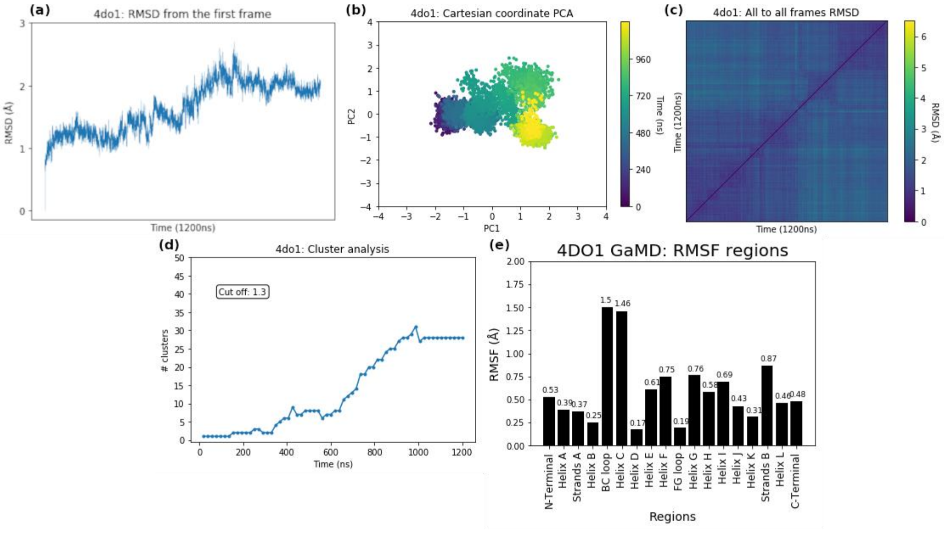


Figure A 26. Convergence analysis of GaMD simulation of CYP199A4 bound to substrate 4-methoxybenzoic acid in S1, 2nd replica.

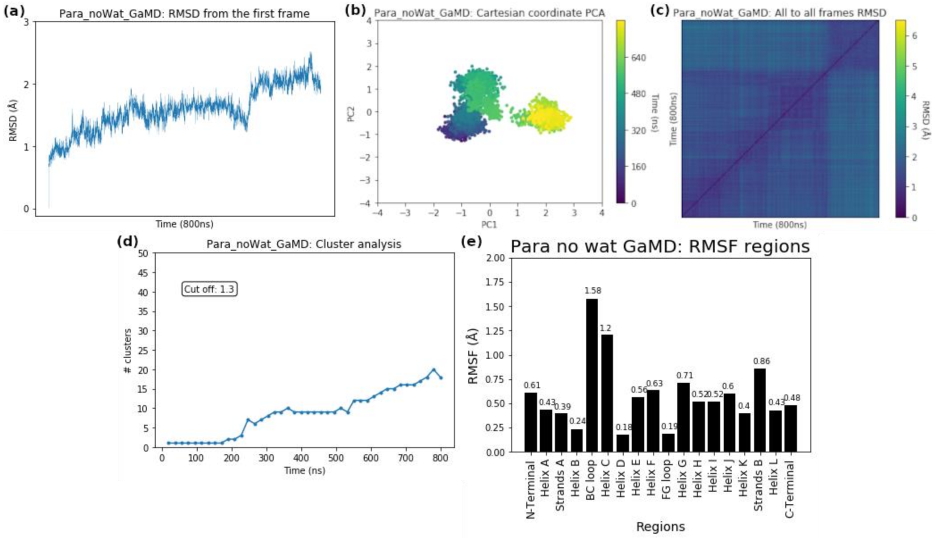


Figure A 27. Convergence analysis of GaMD simulation of CYP199A4 bound to substrate 4-methoxybenzoic acid in S1, 3rd replica.

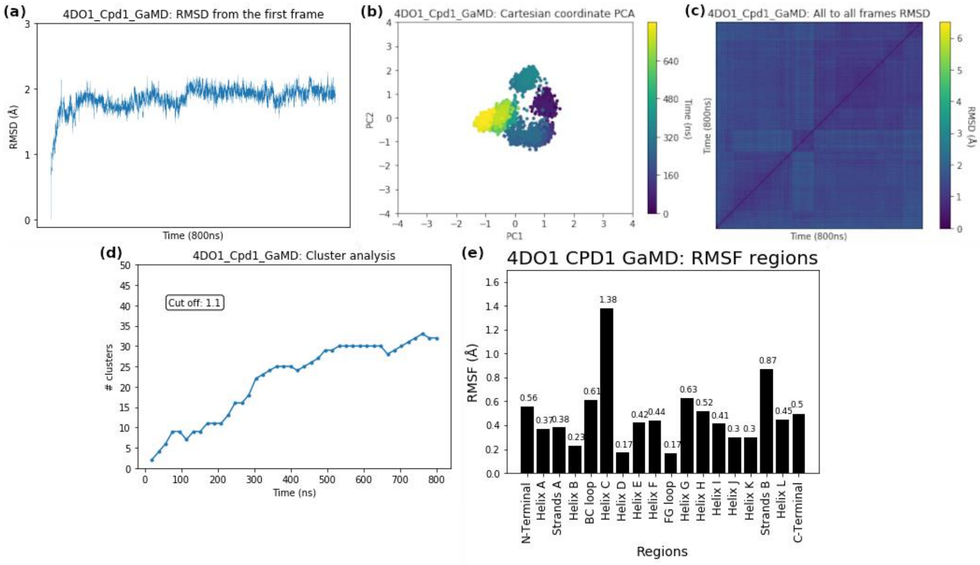


Figure A 28. Convergence analysis of GaMD simulation of CYP199A4 bound to substrate 4-methoxybenzoic acid in CPD1, 1st replica.

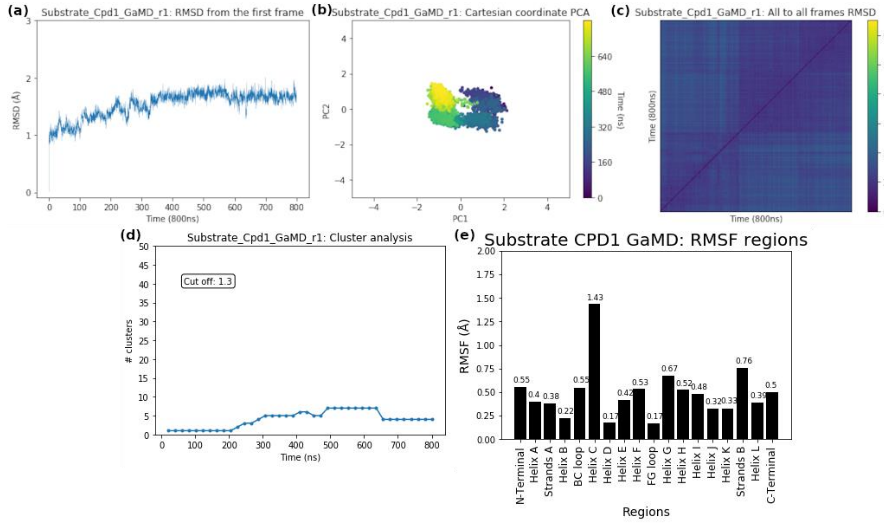


Figure A 29. Convergence analysis of GaMD simulation of CYP199A4 bound to substrate 4-methoxybenzoic acid in CPD1, 2nd replica.

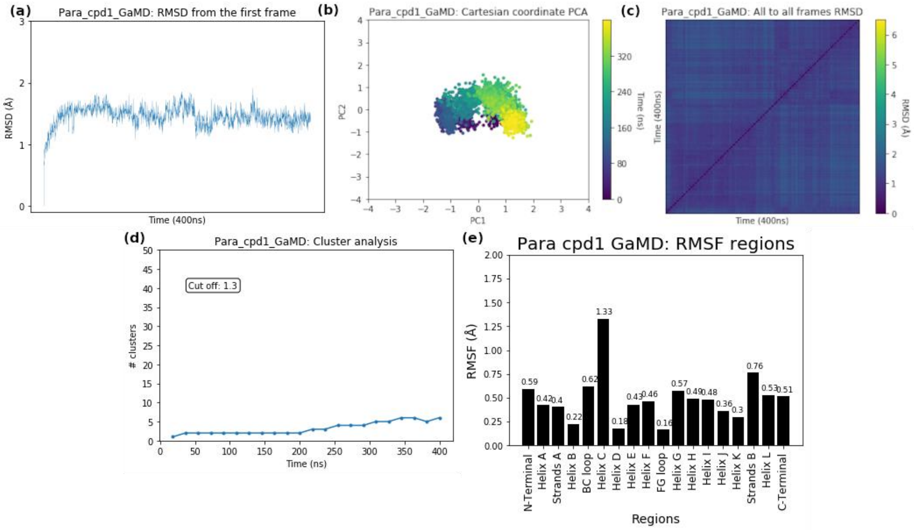


Figure A 30. Convergence analysis of GaMD simulation of CYP199A4 bound to substrate 4-methoxybenzoic acid in CPD1, 3rd replica.

B2. Convergence of 3-methoxybenzoic acid-bound CYP199A4 simulations

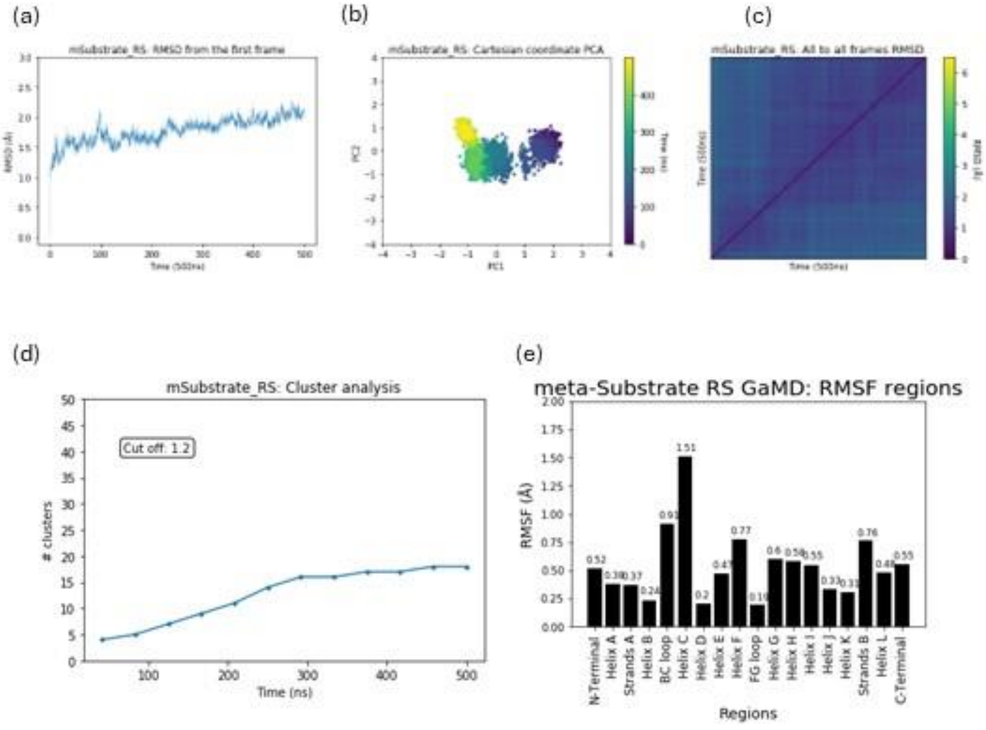


Figure A 31. Convergence analysis of GaMD simulation of CYP199A4 bound to substrate 3-methoxybenzoic acid in RS, 1st replica.

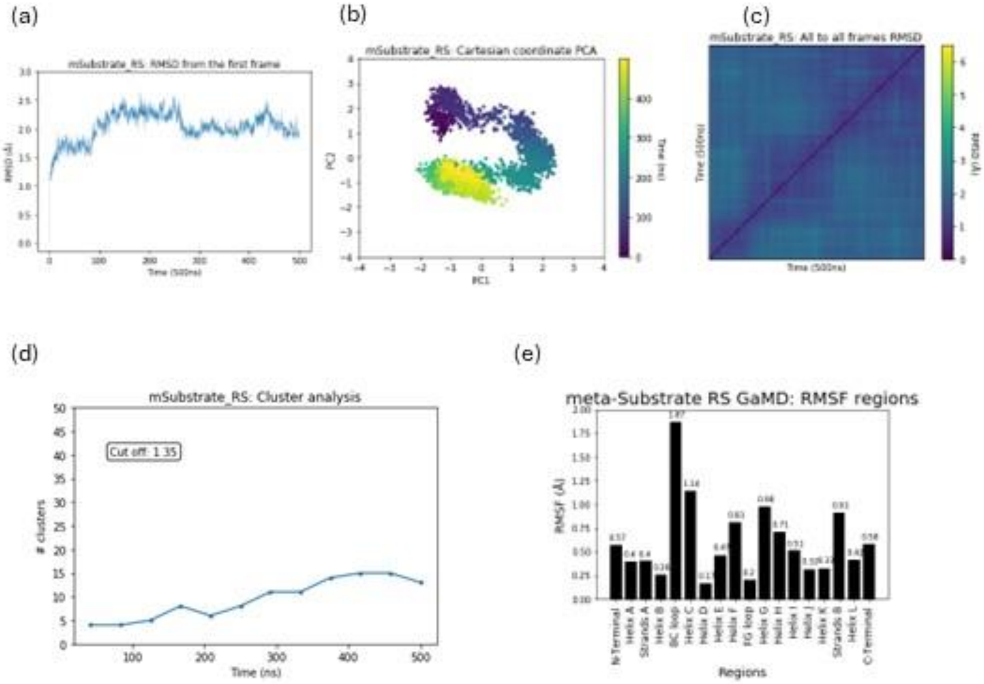


Figure A 32. Convergence analysis of GaMD simulation of CYP199A4 bound to substrate 3-methoxybenzoic acid in RS, 2nd replica.

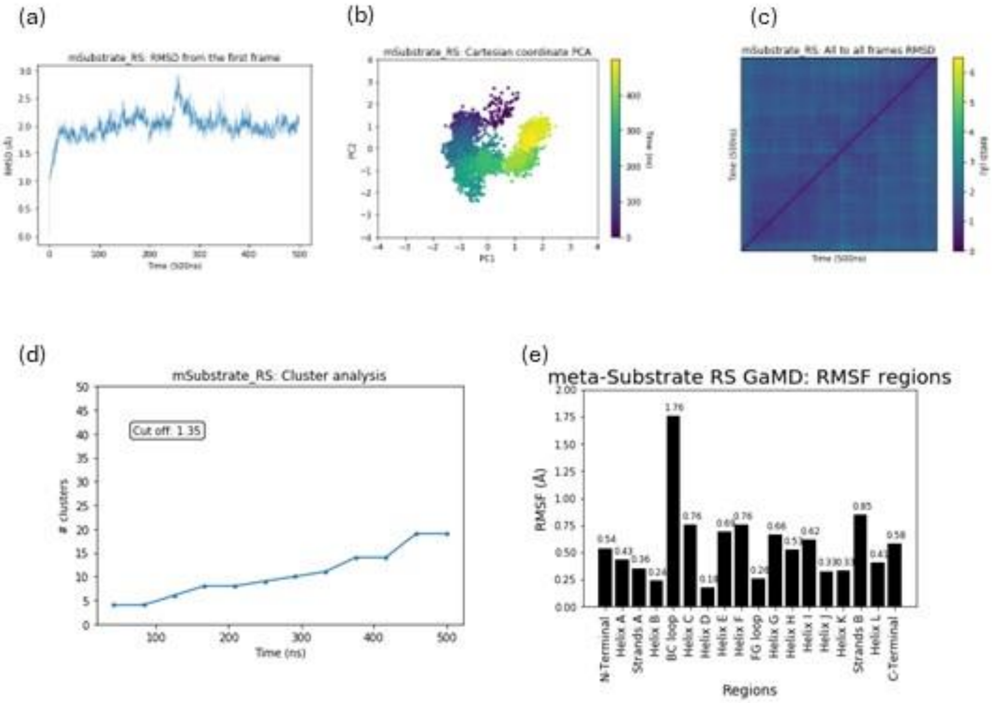


Figure A 33. Convergence analysis of GaMD simulation of CYP199A4 bound to substrate 3-methoxybenzoic acid in RS, 3rd replica.

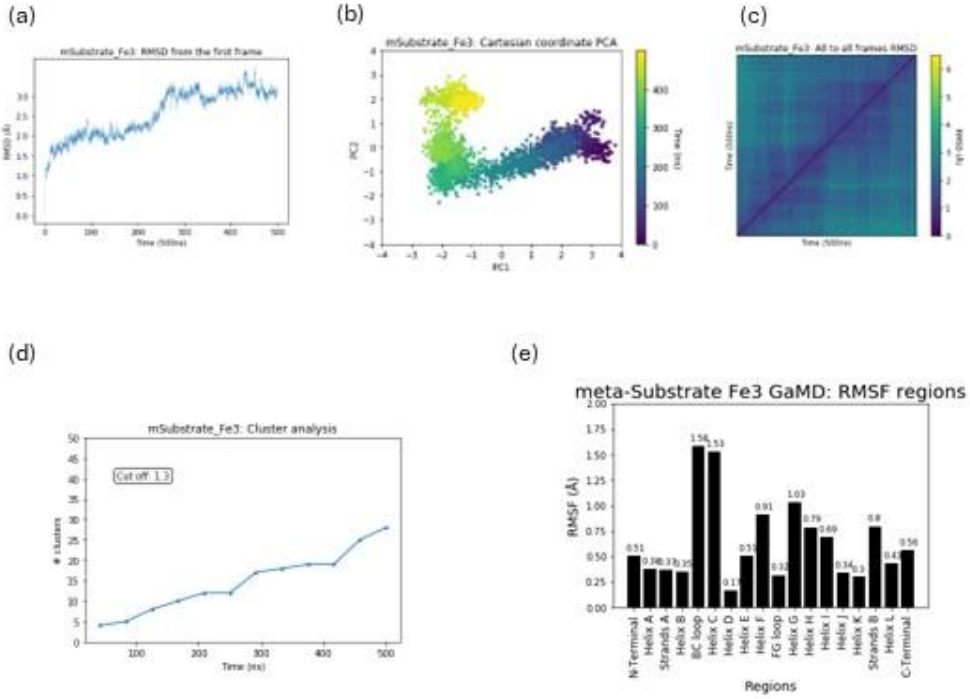


Figure A 34. Convergence analysis of GaMD simulation of CYP199A4 bound to substrate 3-methoxybenzoic acid in S1, 1st replica.

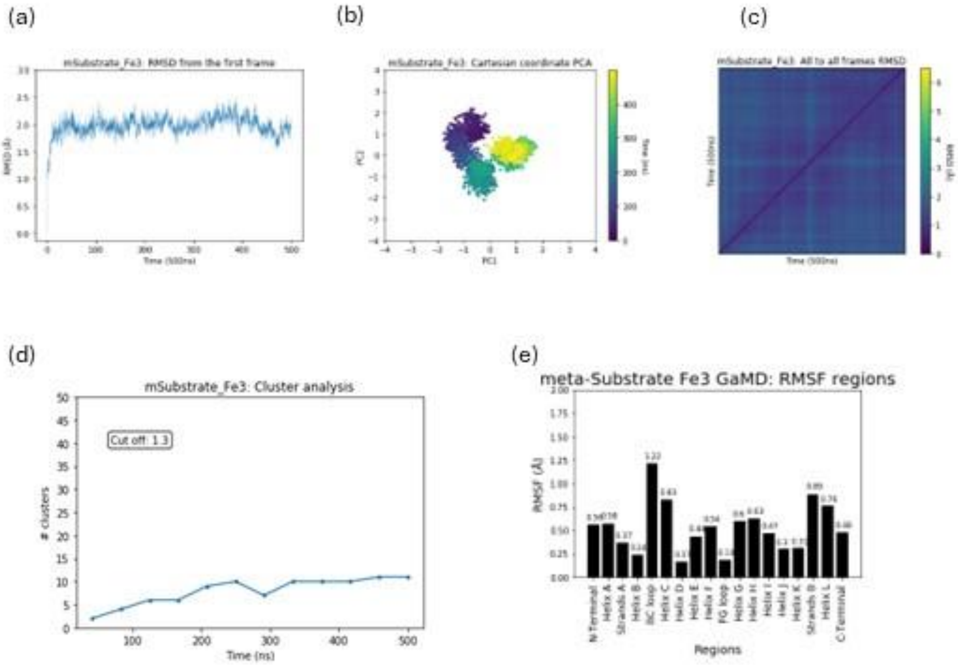


Figure A 35. Convergence analysis of GaMD simulation of CYP199A4 bound to substrate 3-methoxybenzoic acid in S2, 2nd replica.

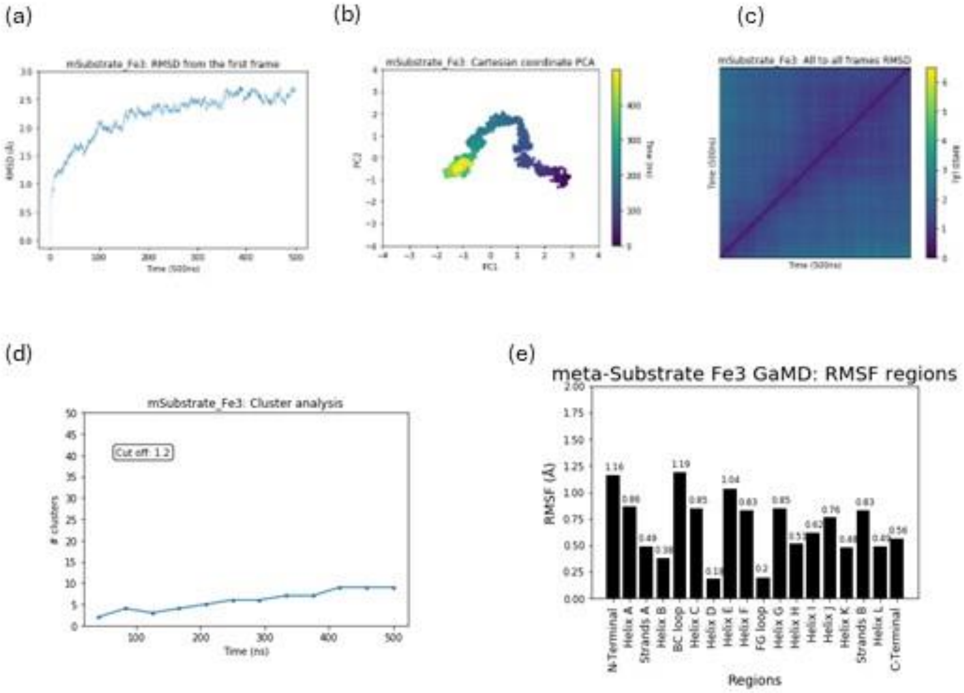


Figure A 36. Convergence analysis of GaMD simulation of CYP199A4 bound to substrate 3-methoxybenzoic acid in S1, 3rd replica.

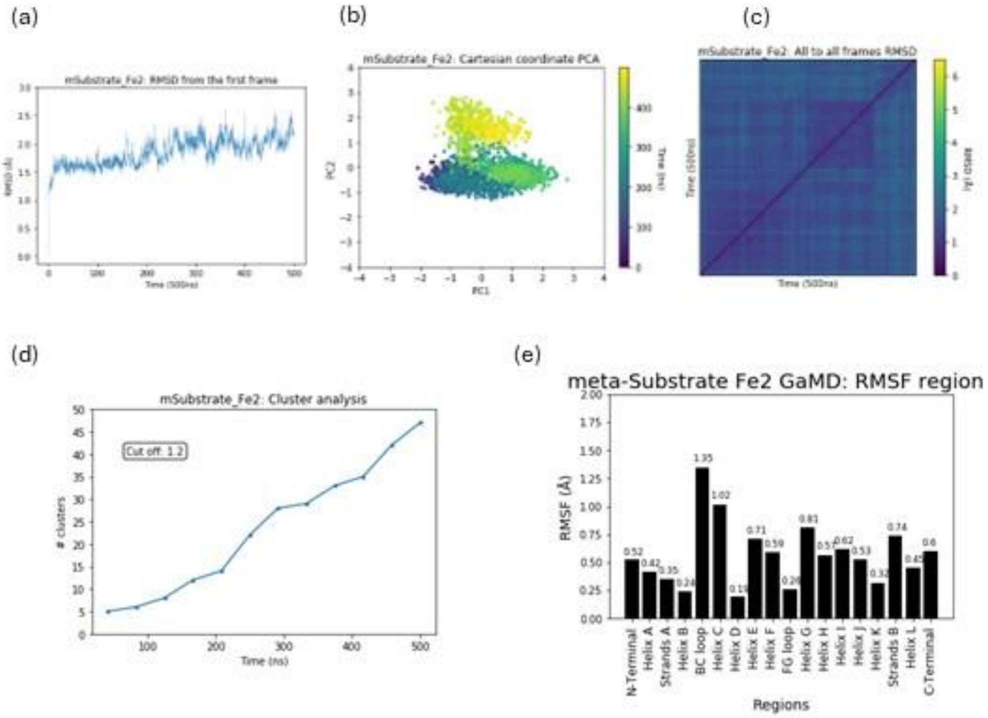


Figure A 37. Convergence analysis of GaMD simulation of CYP199A4 bound to substrate 3-methoxybenzoic acid in S2, 1st replica.

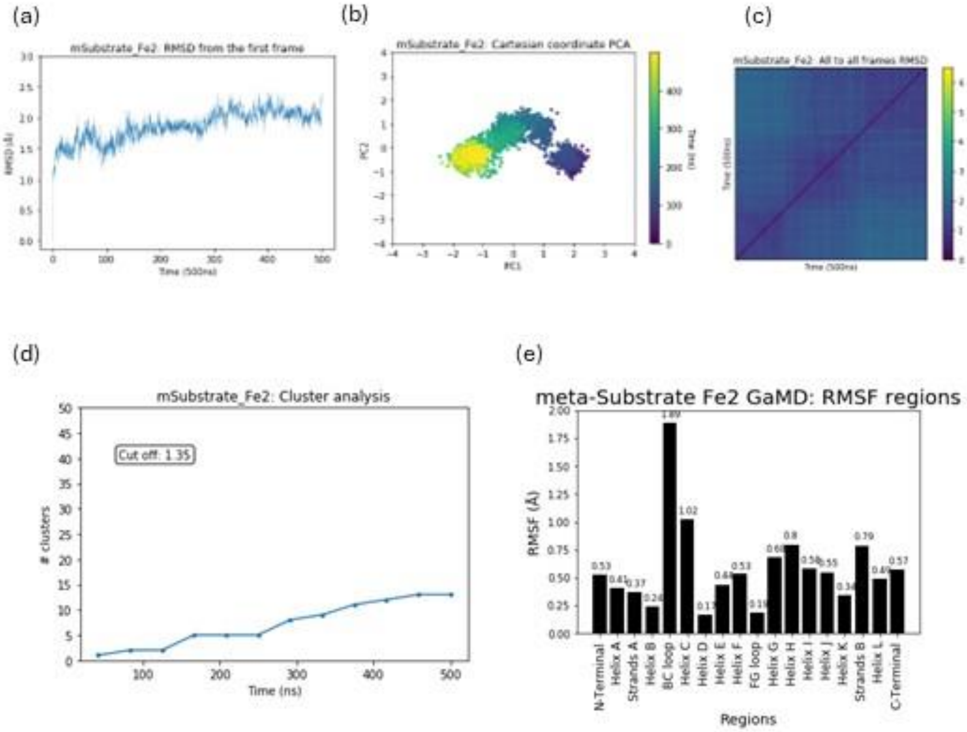


Figure A 38. Convergence analysis of GaMD simulation of CYP199A4 bound to substrate 3-methoxybenzoic acid in S2, 2nd replica.

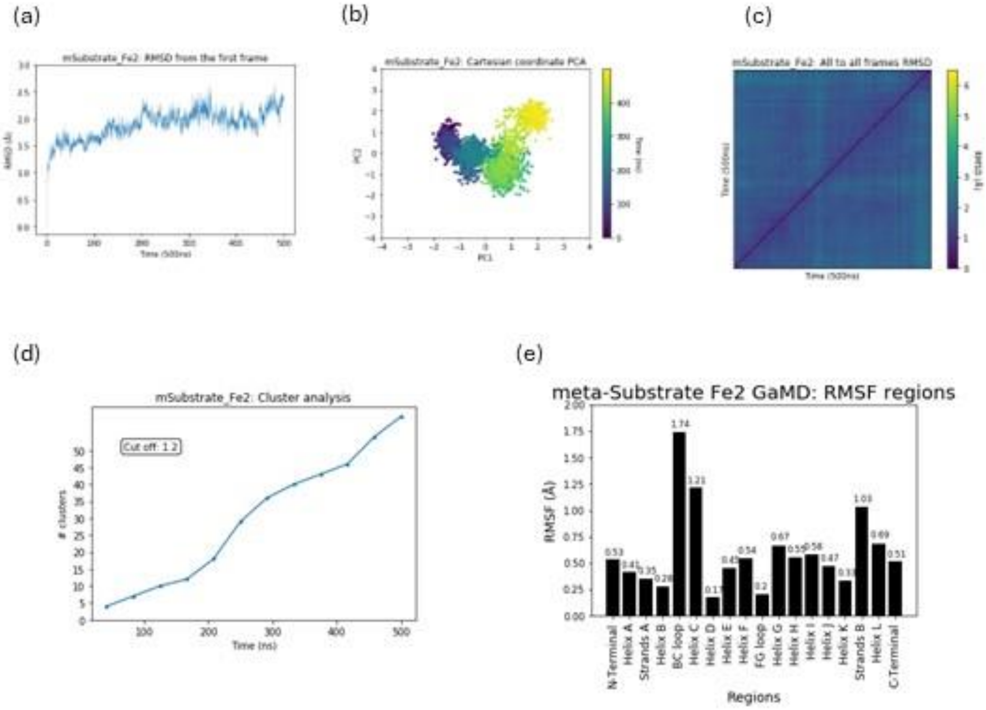


Figure A 39. Convergence analysis of GaMD simulation of CYP199A4 bound to substrate 3-methoxybenzoic acid in S2, 3rd replica.

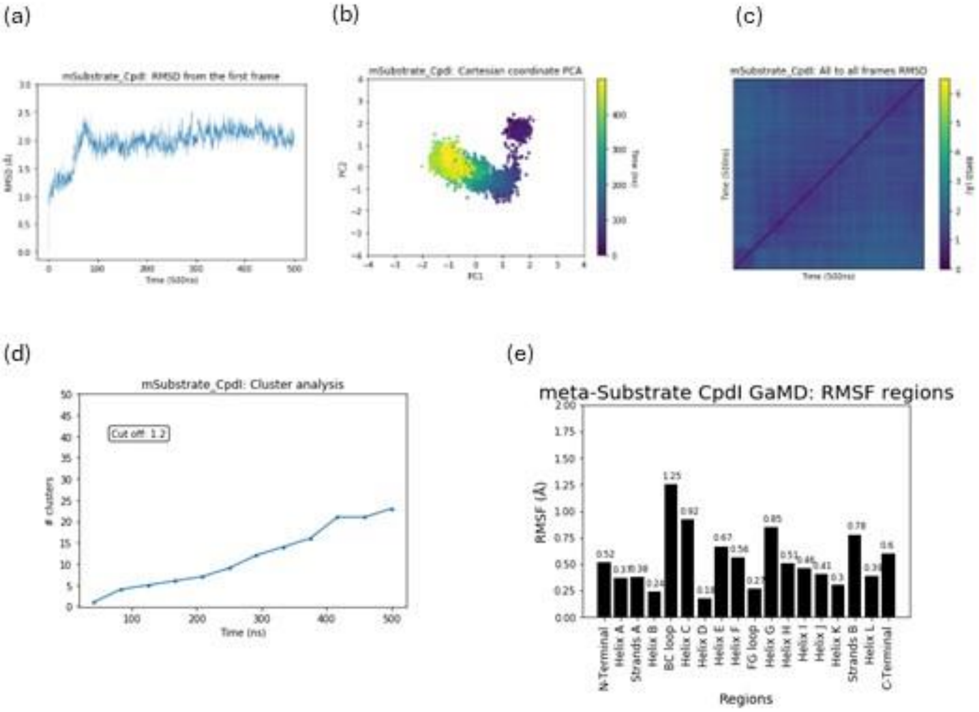


Figure A 40. Convergence analysis of GaMD simulation of CYP199A4 bound to substrate 3-methoxybenzoic acid in CPD1, 1st replica.

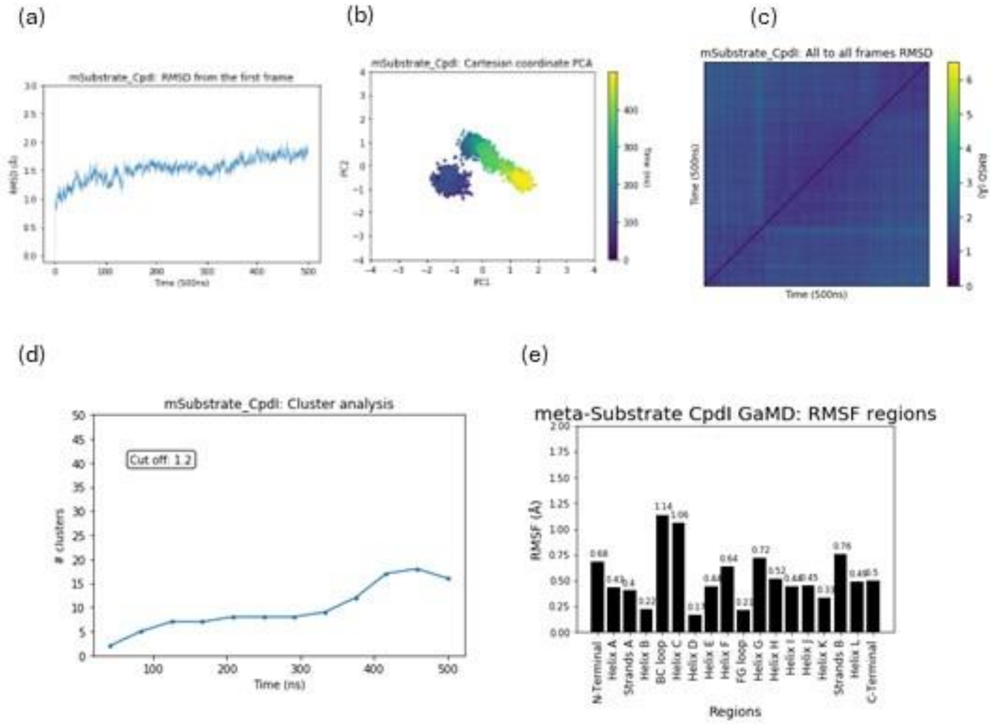


Figure A 41. Convergence analysis of GaMD simulation of CYP199A4 bound to substrate 3-methoxybenzoic acid in CPD1, 2nd replica.

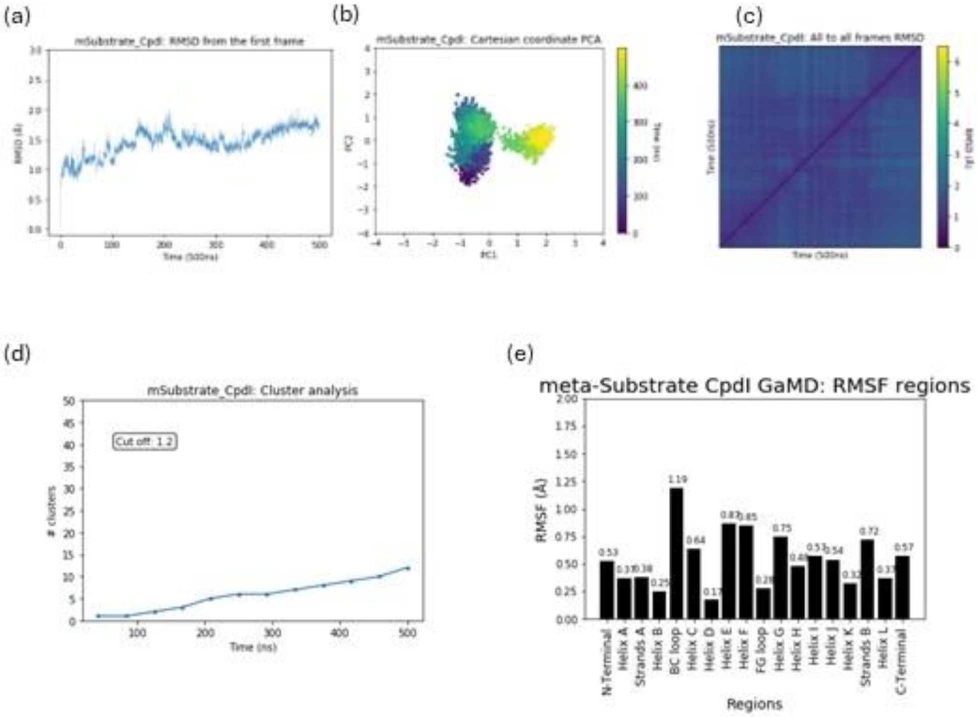


Figure A 42. Convergence analysis of GaMD simulation of CYP199A4 bound to substrate 3-methoxybenzoic acid in CPD1, 3rd replica.

B3. Convergence of 4-hydroxybenzoic acid-bound CYP199A4 simulations

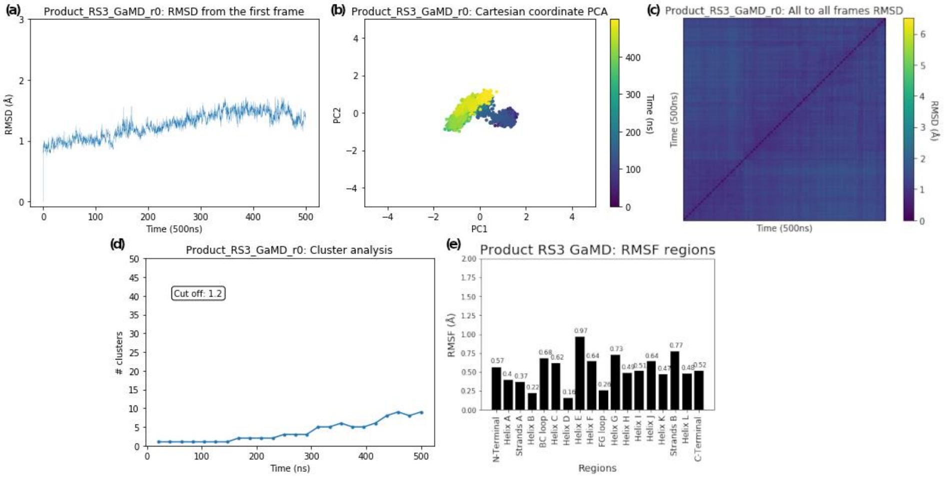


Figure A 43. Convergence analysis of GaMD simulation of CYP199A4 bound to product 4-hydroxybenzoic acid in RS, 1st replica.

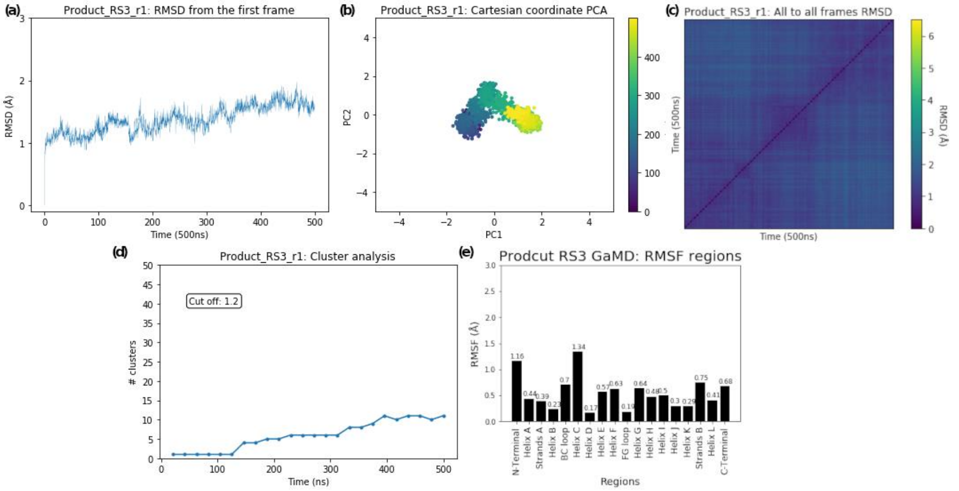


Figure A 44. Convergence analysis of GaMD simulation of CYP199A4 bound to product 4-hydroxybenzoic acid in RS, 2nd replica.

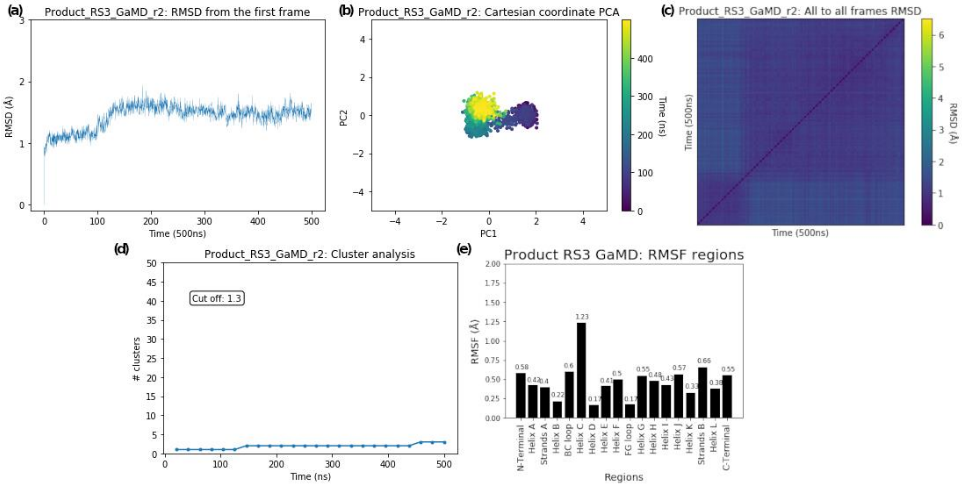


Figure A 45. Convergence analysis of GaMD simulation of CYP199A4 bound to product 4-hydroxybenzoic acid in RS, 3rd replica.

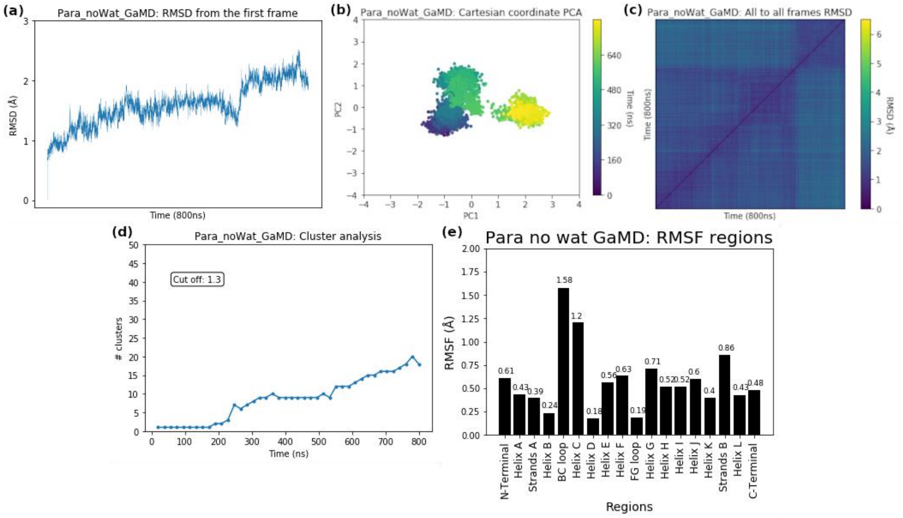


Figure A 46. Convergence analysis of GaMD simulation of CYP199A4 bound to product 4-hydroxybenzoic acid in S1, 1st replica.

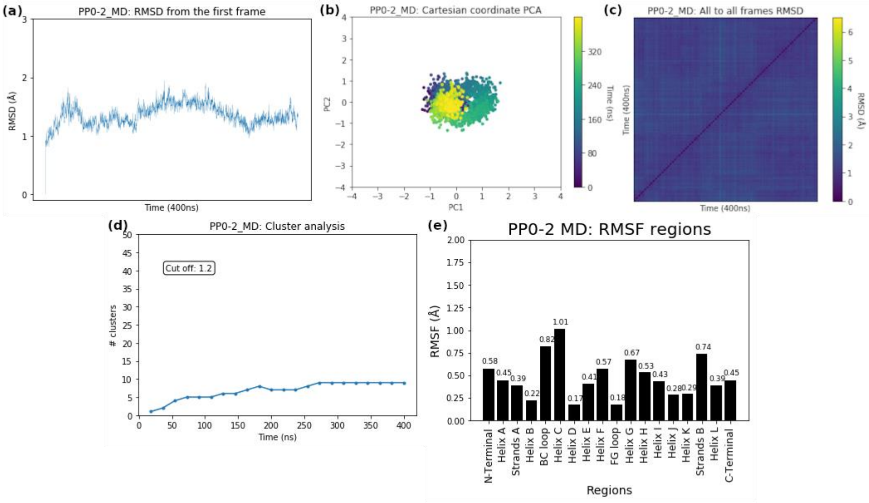


Figure A 47. Convergence analysis of GaMD simulation of CYP199A4 bound to product 4-hydroxybenzoic acid in S1, 2nd replica.

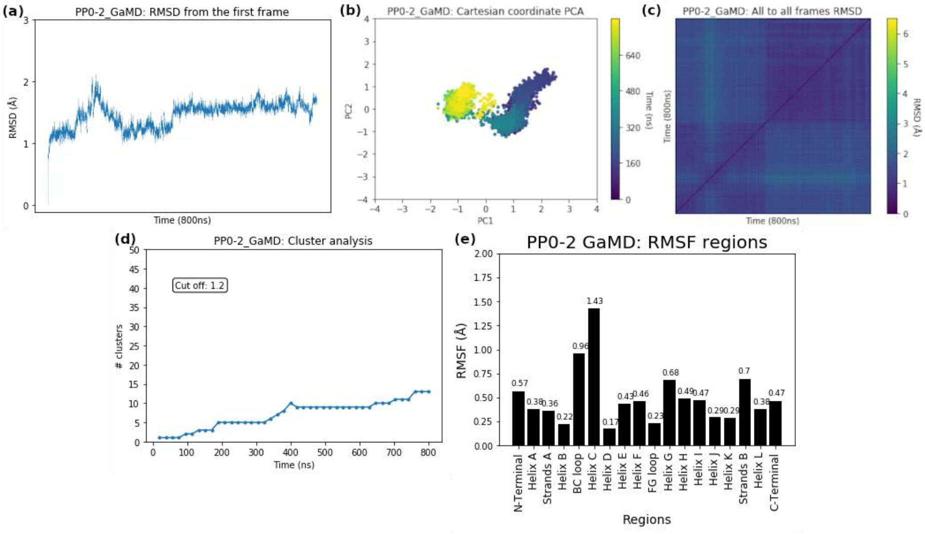


Figure A 48. Convergence analysis of GaMD simulation of CYP199A4 bound to product 4-hydroxybenzoic acid in S1, 3rd replica.

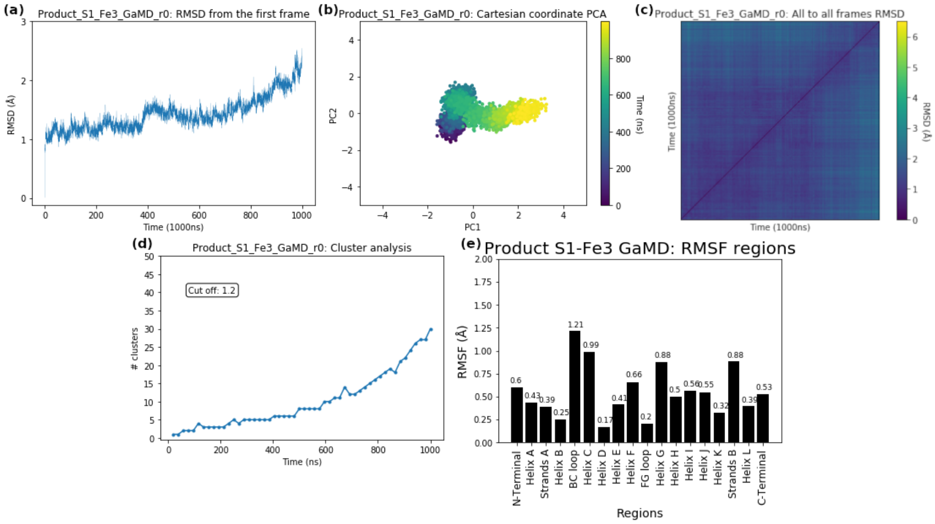


Figure A 49. Convergence analysis of GaMD simulation of CYP199A4 bound to product 4-hydroxybenzoic acid in S1 Fe(III), 1st replica.

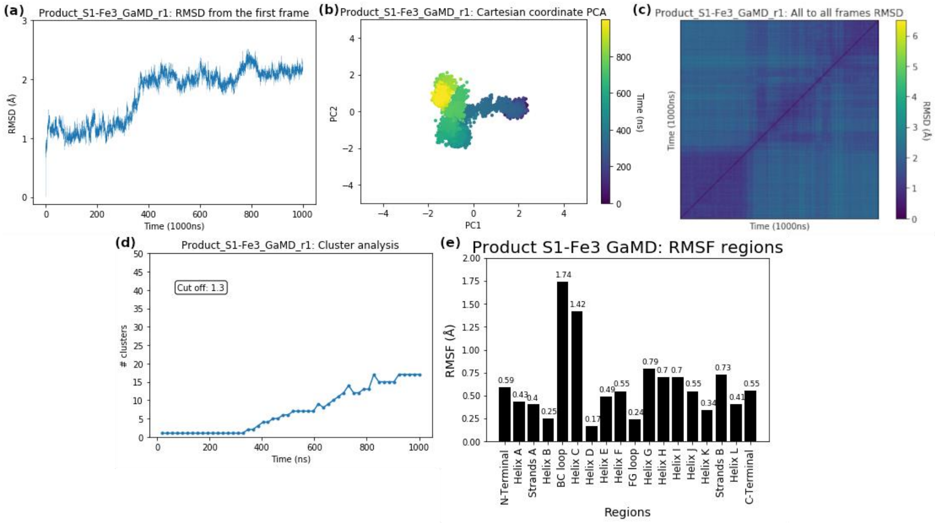


Figure A 50. Convergence analysis of GaMD simulation of CYP199A4 bound to product 4-hydroxybenzoic acid in S1 Fe(III), 2nd replica.

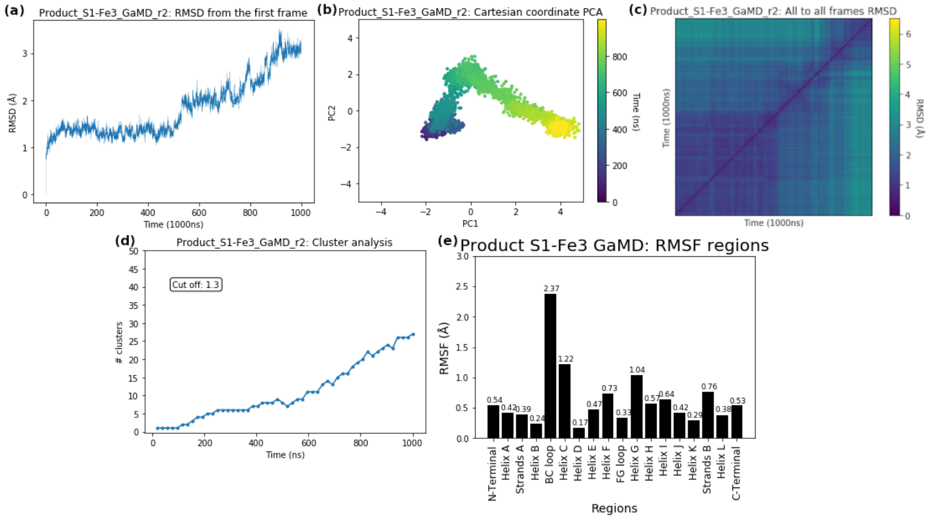


Figure A 51. Convergence analysis of GaMD simulation of CYP199A4 bound to product 4-hydroxybenzoic acid in S1 Fe(III), 3rd replica.

List of publications

1. Learte-Aymamí, S., Martínez-Castro, L., González-González, C., Condeminas, M., Martin-Malpartida, P., Tomás-Gamasa, M., Baúlde Sandra and Couceiro, J. R., Maréchal, J.-D., Macias, M. J., Mascareñas, J. L., & Vázquez, M. E. (2024). De novo engineering of Pd-metalloproteins and their use as intracellular catalysts. *JACS Au*, *4*(7), 2630–2639. /https://doi.org/10.1021/jacsau.4c00379

2. González-González, C., Martínez-Castro, L., Learte-Aymamí, S., Pose-Insua, C., Couceiro José R and Martin-Malpartida, P., Macias, M. J., Maréchal, J.-D., Mascareñas, J. L., & Vázquez, M. E. (2025). Streamlined identification of metallopeptides for intracellular catalysis using positionally addressable combinatorial libraries. *ACS Catal.*, *15*(10), 8624–8632. https://doi.org/10.1021/acscatal.5c00525

3. Podgorski, M. N., Martínez-Castro, L., Bruning, J. B., Campbell, E. C., Maréchal, J.-D., & Bell, S. G. (2025). Investigating the correlation between product release and solvation in cytochrome P450 enzymes. *ACS Catal.*, *15*(4), 2867–2884. https://doi.org/10.1021/acscatal.4c05873

4. Gómez-González, J., Martínez-Castro, L., Tolosa-Barrilero, J., Alcalde-Ordóñez, A., Learte-Aymamí, S., Mascareñas, J. L., García-Martínez, J. C., Martínez-Costas, J., Maréchal, J.-D., Vázquez López, M., & Vázquez, M. E. (2022). Selective recognition of A/T-rich DNA 3-way junctions with a three-fold symmetric

tripeptide. *Chem. Commun.*, *58*(56), 7769–7772. https://doi.org/10.1039/D2CC02874C

5. Illa, O., Olivares, J.-A., Gaztelumendi, N., Martínez-Castro, L., Ospina, J., Abengozar, M.-Á., Sciortino, G., Maréchal Jean-Didier and Nogués, C., Royo, M., Rivas, L., & Ortuño, R. M. (2020). Chiral Cyclobutane-Containing Cell-Penetrating Peptides as Selective Vectors for Anti-Leishmania Drug Delivery Systems. *Int. J. Mol. Sci.*, *21*(20). https://doi.org/10.3390/ijms21207502

Bibliography

1. Bleackley, M. R. & Macgillivray, R. T. A. Transition metal homeostasis: from yeast to human disease. *Biometals* **24**, 785–809 (2011).

2. Bhandary, S. *et al.* Manipulation of spin state of iron porphyrin by chemisorption on magnetic substrates. *Phys Rev B Condens Matter Mater Phys* **88**, (2013).

3. Pardo, A., De Lacey, A. L., Fernández Víctor M and Fan, H.-J., Fan, Y. & Hall, M. B. Density functional study of the catalytic cycle of nickel-iron [NiFe] hydrogenases and the involvement of high-spin nickel(II). *J. Biol. Inorg. Chem.* **11**, 286–306 (2006).

4. Isaac B. Bersuker. *Electronic Structure and Properties of Transition Metal Compounds: Introduction to the Theory. Jou* (John Wiley & Sons, Inc., 2010). doi:10.1002/9780470573051.

5. Pearson, R. G. Hard and soft acids and bases. *JACS* **85**, (1963).

6. Zhu, F. *et al.* Iron catalyzed organic reactions in water: A "nature-like" synthesis. *Curr. Opin. Green Sustain. Chem.* **40**, 100754 (2023).

7. Crichton, R. *Inorganic Biochemistry of Iron Metabolism: From Molecular to Clinical Consequences.* (John Wiley & Sons, Chichester, England, 2003).

8. Seoane, A. & Mascareñas, J. L. Exporting Homogeneous Transition Metal Catalysts to Biological Habitats. *European J. Org. Chem.* **2022**, e202200118 (2022).

9. Silva, R. T. C. *et al.* New Palladium(II) Complexes Containing Methyl Gallate and Octyl Gallate: Effect against Mycobacterium tuberculosis and Campylobacter jejuni. *Molecules* **28**, 3887 (2023).

10. Pereira, T. H. R. *et al.* Palladium (II) compounds containing oximes as promising antitumor agents for the treatment of osteosarcoma: An in vitro and in vivo comparative study with cisplatin. *Eur. J. Med. Chem.* **264**, 116034 (2024).

11. Miller, M. A. *et al.* Nano-palladium is a cellular catalyst for in vivo chemistry. *Nat. Commun.* **8**, 15906 (2017).

12. Martínez-Calvo, M. *et al.* Intracellular Deprotection Reactions Mediated by Palladium Complexes Equipped with Designed Phosphine Ligands. *ACS Catal.* **8**, 6055–6061 (2018).

13. Learte-Aymamí, S., Vidal, C., Gutiérrez-González, A. & Mascareñas, J. L. Intracellular Reactions Promoted by Bis(histidine) Miniproteins Stapled Using Palladium(II) Complexes. *Angew. Chem. Int. Ed Engl.* **59**, 9149–9154 (2020).

14. Filice, M. *et al.* Preparation of an immobilized lipase-palladium artificial metalloenzyme as catalyst in the heck reaction: Role of the solid phase. *Adv. Synth. Catal.* **357**, 2687–2696 (2015).

15. Kobayashi, Y., Murata, K., Harada, A. & Yamaguchi, H. A palladium-catalyst stabilized in the chiral environment of a monoclonal antibody in water. *Chem. Commun.* **56**, 1605–1607 (2020).

16. Bell, E. L. *et al.* Strategies for designing biocatalysts with new functions. *Chem. Soc. Rev.* **53**, 2851–2862 (2024).

17. Quin, M. B. & Schmidt-Dannert, C. Engineering of biocatalysts - from evolution to creation. *ACS Catal.* **1**, 1017–1021 (2011).

18. Vaissier Welborn, V. & Head-Gordon, T. Computational design of synthetic enzymes. *Chem. Rev.* **119**, 6613–6630 (2019).

19. Bertozzi, C. R. A decade of bioorthogonal chemistry. *Acc. Chem. Res.* **44**, 651–653 (2011).

20. Devaraj, N. K. The future of bioorthogonal chemistry. *ACS Cent. Sci.* **4**, 952–959 (2018).

21. Boldon, L., Laliberte, F. & Liu, L. Review of the fundamental theories behind small angle X-ray scattering, molecular dynamics simulations, and relevant integrated application. *Nano Rev* **6**, 25661 (2015).

22. Johnson, W. C. *Protein Secondary Structure and Circular Dichroism: A Practical Guide.*

23. Kay, L. E. NMR studies of protein structure and dynamics. *J. Magn. Reson.* **213**, 477–491 (2011).

24. Ubbink, M., Perrakis, A. & Jeschke, G. The contribution of modern EPR to structural biology. *Emerg Top Life Sci* **2**, 9–18 (2018).

25. Mazal, H. & Haran, G. Single-molecule FRET methods to study the dynamics of proteins at work. *Curr. Opin. Biomed. Eng.* **12**, 8–17 (2019).

26. Konermann, L., Pan, J. & Liu, Y.-H. Hydrogen exchange mass spectrometry for studying protein structure and dynamics. *Chem. Soc. Rev.* **40**, 1224–1234 (2011).

27. Klepeis, J. L., Lindorff-Larsen, K., Dror, R. O. & Shaw, D. E. Long-timescale molecular dynamics simulations of protein structure and function. *Curr. Opin. Struct. Biol.* **19**, 120–127 (2009).

28. O'Rourke, K. F., D'Amico, R. N. & Sahu Debashish and Boehr, D. D. Distinct conformational dynamics and allosteric networks in alpha tryptophan synthase during active catalysis. *Protein Sci.* **30**, 543–557 (2021).

29. Castelli, M. *et al.* Decrypting allostery in membrane-bound K-Ras4B using complementary in silico approaches based on unbiased molecular dynamics simulations. *J. Am. Chem. Soc.* **146**, 901–919 (2024).

30. Alberts, B. *et al.* Protein Function. in *Molecular Biology of the Cell. 4th edition* (Garland Science, 2002).

31. Mairbäurl, H. & Weber, R. E. Oxygen transport by hemoglobin. *Compr. Physiol.* **2**, 1463–1489 (2012).

32. Holding, D. R. Recent advances in the study of prolamin storage protein organization and function. *Front. Plant Sci.* **5**, 276 (2014).

33. Bruice, T. C. & Benkovic, S. J. Chemical basis for enzyme catalysis. *Biochemistry* **39**, 6267–6274 (2000).

34. Jorgensen, W. L. Rusting of the lock and key model for protein-ligand binding. *Science (1979)* **254**, 954–955 (1991).

35. Csermely, P., Palotai, R. & Nussinov, R. Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. *Trends Biochem. Sci.* **35**, 539–546 (2010).

36. Andreini, C., Bertini, I. & Rosato, A. Metalloproteomes: a bioinformatic approach. *Acc. Chem. Res.* **42**, 1471–1479 (2009).

37. Sigel, R. K. O. & Pyle, A. M. Alternative roles for metal ions in enzyme catalysis and the implications for ribozyme chemistry. *Chem. Rev.* **107**, 97–113 (2007).

38. Andreini, C., Putignano, V., Rosato, A. & Banci, L. The human iron-proteome. *Metallomics* **10**, 1223–1231 (2018).

39. Denisov, I. G., Makris, T. M., Sligar, S. G. & Schlichting, I. Structure and chemistry of cytochrome P450. *Chemical Reviews* vol. 105 2253–2277 (2005).

40. Persans, M. W., Wang, J. & Schuler, M. A. Characterization of maize cytochrome P450 monooxygenases induced in response to safeners and bacterial pathogens. *Plant Physiol.* **125**, 1126–1138 (2001).

41. Lu, K., Song, Y. & Zeng, R. The role of cytochrome P450-mediated detoxification in insect adaptation to xenobiotics. *Curr. Opin. Insect Sci.* **43**, 103–107 (2021).

42. Poulos, T. L. & Johnson, E. F. Structures of cytochrome P450 enzymes. in *Cytochrome P450: Structure, Mechanism, and Biochemistry, Third Edition* (ed. Ortiz de Montellano, P. R.) 3–32 (Plenum Publisher, New York, 2015). doi:10.1007/978-3-319-12108-6_1.

43. Girvan, H. M. & Munro, A. W. Applications of microbial cytochrome P450 enzymes in biotechnology and synthetic biology. *Current Opinion in Chemical Biology* vol. 31 136–145 (2016).

44. Cojocaru, V., Winn, P. J. & Wade, R. C. The ins and outs of cytochrome P450s. *Biochimica et Biophysica Acta - General Subjects* vol. 1770 390–401 (2007).

45. Coelho, P. S., Brustad, E. M., Kannan, A. & Arnold, F. H. Olefin cyclopropanation via carbene transfer catalyzed by engineered cytochrome P450 enzymes. *Science (1979)* **339**, 307–310 (2013).

46. McIntosh, J. A., Farwell, C. C. & Arnold, F. H. Expanding P450 catalytic reaction space through evolution and engineering. *Curr. Opin. Chem. Biol.* **19**, 126–134 (2014).

47. Zhang, J., Huang, X., Zhang, R. K. & Arnold, F. H. Enantiodivergent α-amino C-H fluoroalkylation catalyzed by engineered cytochrome P450s. *J. Am. Chem. Soc.* **141**, 9798–9802 (2019).

48. Liu, C. & Chen, X. Recent advances in the engineering of cytochrome P450 enzymes. *Catalysts* **15**, 374 (2025).

49. Li, H. *et al.* Regio- and stereo-selective 1β-hydroxylation of lithocholic acid by cytochrome P450 BM3 mutants. *Biotechnol. Bioeng.* **120**, 2230–2241 (2023).

50. Dubey, K. D. & Shaik, S. Cytochrome P450-The Wonderful Nanomachine Revealed through Dynamic Simulations of the Catalytic Cycle. *Acc. Chem. Res.* **52**, 389–399 (2019).

51. Mokkawes, T., Lim, Z. Q. & de Visser, S. P. Mechanism of melatonin metabolism by CYP1A1: What determines the bifurcation pathways of hydroxylation versus deformylation? *J. Phys. Chem. B* **126**, 9591–9606 (2022).

52. Schröder, G. C., Smit, M. S. & Opperman, D. J. Harnessing heme chemistry: Recent advances in the biocatalytic applications of cytochrome P450 monooxgenases. *Curr. Opin. Green Sustain. Chem.* **39**, 100734 (2023).

53. Hahn, K. W., Klis, W. A. & Stewart, J. M. Design and synthesis of a peptide having chymotrypsin-like esterase activity. *Science (1979)* **248**, 1544–1547 (1990).

54. Fukushima, H., Ohashi, S. & Inoue, S. Asymmetric synthesis catalyzed by poly(5-benzyl L-glutamate). *Makromol. Chem.* **176**, 2751–2753 (1975).

55. Zozulia, O., Marshall, L. R., Kim, I., Kohn, E. M. & Korendovych, I. V. Self-Assembling Catalytic Peptide Nanomaterials Capable of Highly Efficient Peroxidase Activity. *Chemistry (Easton)* **27**, 5388–5392 (2021).

56. Leone, L. *et al.* Peptides and metal ions: A successful marriage for developing artificial metalloproteins. *J. Pept. Sci.* **30**, e3606 (2024).

57. Koebke, K. J., Pinter, T. B. J., Pitts, W. C. & Pecoraro, V. L. Catalysis and electron transfer in DE Novo designed metalloproteins. *Chem. Rev.* **122**, 12046–12109 (2022).

58. Nicolis, S., Casella, L., Roncone, R., Dallacosta, C. & Monzani, E. Heme-peptide complexes as peroxidase models. *C. R. Chim.* **10**, 380–391 (2006).

59. Constable, E. C. Homoleptic Complexes of 2,2'-Bipyridine. in *Advances in Inorganic Chemistry* vol. 34 1–63 (Elsevier, 1989).

60. Coquière, D., Bos, J., Beld, J. & Roelfes, G. Enantioselective artificial metalloenzymes based on a bovine pancreatic polypeptide scaffold. *Angew. Chem. Int. Ed Engl.* **48**, 5159–5162 (2009).

61. Eom, H., Cao, Y., Kim, H., de Visser, S. P. & Song, W. J. Underlying role of hydrophobic environments in tuning metal elements for efficient enzyme catalysis. *J. Am. Chem. Soc.* **145**, 5880–5887 (2023).

62. Van Stappen, C. *et al.* Designing artificial metalloenzymes by tuning of the environment beyond the primary coordination sphere. *Chem. Rev.* **122**, 11974–12045 (2022).

63. Hansen, W. A. & Khare, S. D. Benchmarking a computational design method for the incorporation of metal ion-binding sites at symmetric protein interfaces. *Protein Sci.* **26**, 1584–1594 (2017).

64. Chalkley, M. J., Mann, S. I. & DeGrado, W. F. De novo metalloprotein design. *Nat Rev Chem* **6**, 31–50 (2021).

65. Nastri, F. *et al.* Engineering metalloprotein functions in designed and native scaffolds. *Trends Biochem. Sci.* **44**, 1022–1040 (2019).

66. Zastrow, M. L. & Pecoraro, V. L. Designing functional metalloproteins: from structural to catalytic metal sites. *Coord. Chem. Rev.* **257**, 2565–2588 (2013).

67. Smith, S. J. *et al.* Tunable helicity, stability and DNA-binding properties of short peptides with hybrid metal coordination motifs. *Chem. Sci.* **7**, 5453–5461 (2016).

68. Stellato, F. *et al.* Metal binding in amyloid beta-peptides shows intra- and inter-peptide coordination modes. *Eur. Biophys. J.* **35**, 340–351 (2006).

69. Portelinha, J. *et al.* Antimicrobial peptides and copper(II) ions: Novel therapeutic opportunities. *Chem. Rev.* **121**, 2648–2712 (2021).

70. Harford, C. & Sarkar, B. Amino terminal cu(II)- and Ni(II)-binding (ATCUN) motif of proteins and peptides: Metal binding, DNA cleavage, and other properties. *Acc. Chem. Res.* **30**, 123–130 (1997).

71. Michael Kormaník, J. *et al.* Design of Zn-binding peptide(s) from protein fragments. *Chembiochem* **26**, e202401014 (2025).

72. Miao, Y., Feher, V. A. & McCammon, J. A. Gaussian Accelerated Molecular Dynamics: Unconstrained Enhanced Sampling and Free Energy Calculation. *J. Chem. Theory Comput.* **11**, 3584–3595 (2015).

73. Qi, R., Wei, G., Ma, B. & Nussinov, R. Replica exchange molecular dynamics: A practical application protocol with solutions to common

problems and a peptide aggregation and self-assembly example. *Methods Mol. Biol.* **1777**, 101–119 (2018).

74. York, D. M. Modern alchemical free energy methods for drug discovery explained. *ACS Phys. Chem. Au* **3**, 478–491 (2023).

75. Aponte, E. A. *et al.* An introduction to thermodynamic integration and application to dynamic causal models. *Cogn. Neurodyn.* **16**, 1–15 (2022).

76. Bussi, G. & Laio, A. Using metadynamics to explore complex free-energy landscapes. *Nat. Rev. Phys.* **2**, 200–212 (2020).

77. You, W., Tang, Z. & Chang, C.-E. A. Potential mean force from umbrella sampling simulations: What can we learn and what is missed? *J. Chem. Theory Comput.* **15**, 2433–2443 (2019).

78. Alam, N. *et al.* High-resolution global peptide-protein docking using fragments-based PIPER-FlexPepDock. *PLoS Comput. Biol.* **13**, e1005905 (2017).

79. van Zundert, G. C. P. *et al.* The HADDOCK2.2 Web server: User-friendly integrative modeling of biomolecular complexes. *J. Mol. Biol.* **428**, 720–725 (2016).

80. Rodríguez-Guerra Pedregal, J., Sciortino, G., Guasp, J., Municoy, M. & Maréchal, J.-D. GaudiMM: A modular multi-objective platform for molecular modeling. *J Comput Chem* **38**, 2118–2126 (2017).

81. Leaver-Fay, A. *et al.* ROSETTA3: An Object-Oriented Software Suite for the Simulation and Design of Macromolecules. doi:10.1016/S0076-6879(11)87019-9.

82. Jones, G., Willett, P., Glen, R. C., Leach, A. R. & Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **267**, 727–748 (1997).

83. Richter, F., Leaver-Fay, A., Khare, S. D., Bjelic, S. & Baker, D. De novo enzyme design using Rosetta3. *PLoS One* **6**, e19230 (2011).

84. Sciortino, G., Rodríguez-Guerra Pedregal, J., Lledós, A., Garribba, E. & Maréchal, J.-D. Prediction of the interaction of metallic moieties with proteins: An update for protein-ligand docking techniques. *J. Comput. Chem.* **39**, 42–51 (2018).

85. Sumner, S., Söderhjelm, P. & Ryde, U. Effect of geometry optimizations on QM-cluster and QM/MM studies of reaction energies in proteins. *J. Chem. Theory Comput.* **9**, 4205–4214 (2013).

86. Chung, L. W. *et al.* The ONIOM method and its applications. *Chem. Rev.* **115**, 5678–5796 (2015).

87. Christoffel, F. *et al.* Design and evolution of chimeric streptavidin for protein-enabled dual gold catalysis. *Nat Catal* **4**, 643–653 (2021).

88. Dhakal, A., McKay, C., Tanner, J. J. & Cheng, J. Artificial intelligence in the prediction of protein-ligand interactions: recent advances and future directions. *Brief. Bioinform.* **23**, (2022).

89. Shah, M., Patel, M., Shah, M. & Patel Monali and Prajapati, M. Computational transformation in drug discovery: A comprehensive study on molecular docking and quantitative structure activity relationship (QSAR). *Intelligent Pharmacy* **2**, 589–595 (2024).

90. Gangwal, A. *et al.* Generative artificial intelligence in drug discovery: basic framework, recent advances, challenges, and opportunities. *Front. Pharmacol.* **15**, 1331062 (2024).

91. Bihani, V. *et al.* EGraFFBench: Evaluation of equivariant graph neural network force fields for atomistic simulations. *Digit. Discov.* **3**, 759–768 (2024).

92. Xie, F., Lu, T., Meng, S. & Liu, M. GPTFF: A high-accuracy out-of-the-box universal AI force field for arbitrary inorganic materials. *Sci. Bull. (Beijing)* **69**, 3525–3532 (2024).

93. Levine, I. N. *Physical Chemistry*. (Thomas Timp, 2009).

94. Peter Atkins, J. de P. Structure. in *Physical Chemistry* 241–620 (Oxford University Press, 2006).

95. Jensen, F. *Introduction to Computational Chemistry*. (John Wiley & Sons, Nashville, TN, 2017).

96. Sholl, D. & Steckel, J. A. *Density Functional Theory: A Practical Introduction*. (Wiley-Blackwell, Hoboken, NJ, 2009).

97. Kohn, W. & Sham, L. J. Self-consistent equations including exchange and correlation effects. *Phys. Rev.* **140**, A1133–A1138 (1965).

98. Perdew, J. P. Jacob's ladder of density functional approximations for the exchange-correlation energy. in *AIP Conference Proceedings* vol. 577 1–20 (AIP, 2001).

99. Cramer, C. J. *Essentials of Computational Chemistry: Theories and Models*. (John Wiley & Sons, 2013).

100. Muegge, I. & Martin, Y. C. A general and fast scoring function for protein-ligand interactions: a simplified potential approach. *J. Med. Chem.* **42**, 791–804 (1999).

101. Yang, C., Chen, E. A. & Zhang, Y. Protein-ligand docking in the machine-learning era. *Molecules* **27**, 4568 (2022).

102. Passerini, A., Lippi, M. & Frasconi, P. MetalDetector v2.0: predicting the geometry of metal binding sites from protein sequence. *Nucleic Acids Res.* **39**, W288–92 (2011).

103. Lin, X. *et al.* SuperMetal: A generative AI framework for rapid and precise metal ion location prediction in proteins. *bioRxivorg* 2025.03.21.644685 (2025).

104. Sánchez-Aparicio, J.-E. *et al.* BioMetAll: Identifying Metal-Binding Sites in Proteins from Backbone Preorganization. *J. Chem. Inf. Model.* **61**, 311–323 (2021).

105. Andreini, C., Cavallaro, G. & Lorenzini Serena and Rosato, A. MetalPDB: a database of metal sites in biological macromolecular structures. *Nucleic Acids Res.* **41**, D312–9 (2013).

106. Coin, I., Beyermann, M. & Bienert, M. Solid-phase peptide synthesis: from standard procedures to the synthesis of difficult sequences. *Nat. Protoc.* **2**, 3247–3256 (2007).

107. Vanier, G. S. Microwave-assisted solid-phase peptide synthesis based on the Fmoc protecting group strategy (CEM). *Methods Mol. Biol.* **1047**, 235–249 (2013).

108. Wang, X. *et al.* Copper-triggered bioorthogonal cleavage reactions for reversible protein and cell surface modifications. *J. Am. Chem. Soc.* **141**, 17133–17141 (2019).

109. Weiss, J. T., Carragher, N. O. & Unciti-Broceta, A. Palladium-mediated dealkylation of N-propargyl-floxuridine as a bioorthogonal oxygen-independent prodrug strategy. *Sci. Rep.* **5**, 9329 (2015).

110. Li, J. *et al.* Palladium-triggered deprotection chemistry for protein activation in living cells. *Nat. Chem.* **6**, 352–361 (2014).

111. Koepf, E. K., Petrassi, H. M., Sudol, M. & Kelly, J. W. WW: An isolated three-stranded antiparallel beta-sheet domain that unfolds and refolds reversibly; evidence for a structured hydrophobic cluster in urea and GdnHCl and a disordered thermal unfolded state. *Protein Sci.* **8**, 841–853 (1999).

112. Macias, M. J. *et al.* Structure of the WW domain of a kinase-associated protein complexed with a proline-rich peptide. *Nature* **382**, 646–649 (1996).

113. Sciortino, G., Garribba, E. & Maréchal, J.-D. Validation and applications of protein-ligand docking approaches improved for metalloligands with multiple vacant sites. *Inorg. Chem.* **58**, 294–306 (2019).

114. McGibbon, R. T. *et al.* MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.* **109**, 1528–1532 (2015).

115. Cirilio, A. D. RamachanDraw. *PyPI* Preprint at (2020).

116. Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585**, 357–362 (2020).

117. Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).

118. Subirós-Funosas, R., El-Faham, A. & Albericio, F. Aspartimide formation in peptide chemistry: occurrence, prevention strategies and the role of N-hydroxylamines. *Tetrahedron* **67**, 8595–8606 (2011).

119. Jäger, M., Dendle, M. & Kelly, J. W. Sequence determinants of thermodynamic stability in a WW domain–an all-beta-sheet protein. *Protein Sci.* **18**, 1806–1813 (2009).

120. Lättig-Tünnemann, G. *et al.* Backbone rigidity and static presentation of guanidinium groups increases cellular uptake of arginine-rich cell-penetrating peptides. *Nat Commun* **2**, (2011).

121. Li, S. *et al.* Hydrocarbon staple constructing highly efficient α-helix cell-penetrating peptides for intracellular cargo delivery. *Chem. Commun. (Camb.)* **56**, 15655–15658 (2020).

122. Frank, R. Spot-synthesis: an easy technique for the positionally addressable, parallel chemical synthesis on a membrane support. *Tetrahedron* **48**, 9217–9232 (1992).

123. Thiele, A., Stangl, G. I. & Schutkowski, M. Deciphering enzyme function using peptide arrays. *Mol. Biotechnol.* **49**, 283–305 (2011).

124. Biswas, R., Maillard, N., Kofoed, J. & Reymond, J.-L. Comparing dendritic with linear esterase peptides by screening SPOT arrays for catalysis. *Chem. Commun. (Camb.)* **46**, 8746–8748 (2010).

125. Minor Jr, D. L. & Kim, P. S. Measurement of the beta-sheet-forming propensities of amino acids. *Nature* **367**, 660–663 (1994).

126. Gnuplot 4.4: An Interactive Plotting Program. *Scribd.*

127. Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graph.* **14**, 27–28, 33–38 (1996).

128. Andreasen, M. *et al.* The importance of being capped: Terminal capping of an amyloidogenic peptide affects fibrillation propensity and fibril morphology. *Biochemistry* **53**, 6968–6980 (2014).

129. Hannemann, F., Bichet, A., Ewen, K. M. & Bernhardt, R. Cytochrome P450 systems–biological variations of electron transport chains. *Biochim. Biophys. Acta* **1770**, 330–344 (2007).

130. Harvey, J. N., Bathelt, C. M. & Mulholland, A. J. QM/MM modeling of compound I active species in cytochrome P450, cytochrome C peroxidase, and ascorbate peroxidase. *J. Comput. Chem.* **27**, 1352–1362 (2006).

131. Schyman, P., Usharani, D., Wang, Y. & Shaik, S. Brain chemistry: how does P450 catalyze the O-demethylation reaction of 5-methoxytryptamine to yield serotonin? *J. Phys. Chem. B* **114**, 7078–7089 (2010).

132. Shaik, S. *et al.* P450 enzymes: their structure, reactivity, and selectivity-modeled by QM/MM calculations. *Chem. Rev.* **110**, 949–1017 (2010).

133. Guengerich, F. P. & Munro, A. W. Unusual Cytochrome P450 Enzymes and Reactions. *Journal of Biological Chemistry* **288**, 17065–17073 (2013).

134. Larimer, F. W. *et al.* Complete genome sequence of the metabolically versatile photosynthetic bacterium Rhodopseudomonas palustris. *Nat Biotechnol* **22**, 55–61 (2004).

135. Wei, Y., Ang, E. L. & Zhao, H. Recent developments in the application of P450 based biocatalysts. *Curr. Opin. Chem. Biol.* **43**, 1–7 (2018).

136. Julsing, M. K., Cornelissen, S., Bühler, B. & Schmid, A. Heme-iron oxygenases: powerful industrial biocatalysts? *Current Opinion in Chemical Biology* vol. 12 177–186 Preprint at https://doi.org/10.1016/j.cbpa.2008.01.029 (2008).

137. Bhattacharya, S. S. & Yadav, J. S. Microbial P450 enzymes in bioremediation and drug discovery: Emerging potentials and challenges. *Curr. Protein Pept. Sci.* **19**, 75–86 (2018).

138. Behrendorff, J. B. Y. H. Reductive cytochrome P450 reactions and their potential role in bioremediation. *Front. Microbiol.* **12**, 649273 (2021).

139. Coleman, T., Chao, R. R., De Voss, J. J. & Bell, S. G. The importance of the benzoic acid carboxylate moiety for substrate recognition by CYP199A4 from Rhodopseudomonas palustris HaA2. *Biochim. Biophys. Acta* **1864**, 667–675 (2016).

140. Chao, R. R. *et al.* The Stereoselective Oxidation of para-Substituted Benzenes by a Cytochrome P450 Biocatalyst. *Chemistry (Easton)* **27**, 14765–14777 (2021).

141. Coleman, T. *et al.* Cytochrome P450 CYP199A4 from Rhodopseudomonas palustris Catalyzes Heteroatom Dealkylations, Sulfoxidation, and Amide and Cyclic Hemiacetal Formation. *ACS Catal.* **8**, 5915–5927 (2018).

142. Bell, S. G., Tan, A. B. H., Johnson, E. O. D. & Wong, L.-L. Selective oxidative demethylation of veratric acid to vanillic acid by CYP199A4 from Rhodopseudomonas palustris HaA2. *Mol. Biosyst.* **6**, 206–214 (2010).

143. Podgorski, M. N. *et al.* Investigation of the requirements for efficient and selective cytochrome P450 monooxygenase catalysis across different reactions. *J Inorg Biochem* **203**, (2020).

144. Podgorski, M. N. *et al.* Biophysical techniques to distinguish ligand binding modes in cytochrome P450 monooxygenases. *Biochemistry* **59**, 1038–1050 (2020).

145. Bell, S. G. *et al.* The crystal structures of 4-methoxybenzoate bound CYP199A2 and CYP199A4: Structural changes on substrate binding and the identification of an anion binding site. *Dalton Transactions* **41**, 8703–8714 (2012).

146. Li, H., Narasimhulu, S., Havran, L. M., Winkler, J. D. & Poulos, T. L. Crystal structure of cytochrome P450cam complexed with its catalytic product, 5-exo-hydroxycamphor. *J. Am. Chem. Soc.* **117**, 6297–6299 (1995).

147. Nagano, S. *et al.* Crystal Structures of Epothilone D-bound, Epothilone B-bound, and Substrate-free Forms of Cytochrome P450epoK. *Journal of Biological Chemistry* **278**, 44886–44893 (2003).

148. Li, S. *et al.* Substrate recognition by the multifunctional cytochrome P450 MycG in mycinamicin hydroxylation and epoxidation reactions. *J. Biol. Chem.* **287**, 37880–37890 (2012).

149. Murarka, V. C., Batabyal, D., Amaya, J. A., Sevrioukova, I. F. & Poulos, T. L. Unexpected differences between two closely related bacterial P450 camphor monooxygenases. *Biochemistry* **59**, 2743–2750 (2020).

150. Mast, N. *et al.* Structural basis for three-step sequential catalysis by the cholesterol side chain cleavage enzyme CYP11A1. *J. Biol. Chem.* **286**, 5607–5613 (2011).

151. DeVore, N. M. & Scott, E. E. Structures of cytochrome P450 17A1 with prostate cancer drugs abiraterone and TOK-001. *Nature* **482**, 116–119 (2012).

152. Shahrokh, K., Orendt, A., Yost, G. S. & Cheatham 3rd, T. E. Quantum mechanically derived AMBER-compatible heme parameters for various states of the cytochrome P450 catalytic cycle. *J. Comput. Chem.* **33**, 119–133 (2012).

153. Costa, G. J., Egbemhenghe, A. & Liang, R. Computational characterization of the reactivity of compound I in unspecific peroxygenases. *J. Phys. Chem. B* **127**, 10987–10999 (2023).

154. Duckett, D. R. & Lilley, D. M. The three-way DNA junction is a Y-shaped molecule in which there is no helix-helix stacking. *EMBO J.* **9**, 1659–1664 (1990).

155. Zell, J., Rota Sperti, F. & Britton Sébastien and Monchaud, D. DNA folds threaten genetic stability and can be leveraged for chemotherapy. *RSC Chem. Biol.* **2**, 47–76 (2021).

156. Tateishi-Karimata, H. & Sugimoto, N. Roles of non-canonical structures of nucleic acids in cancer and neurodegenerative diseases. *Nucleic Acids Res.* **49**, 7839–7855 (2021).

157. Frank, M., Johnstone, M. D. & Clever, G. H. Interpenetrated cage structures. *Chemistry (Easton)* **22**, 14104–14125 (2016).

158. Fujita, M., Fujita, N., Ogura, K. & Yamaguchi, K. Spontaneous assembly of ten components into two interlocked, identical coordination cages. *Nature* **400**, 52–55 (1999).

159. Dalton, D. M. *et al.* Supramolecular Ga4L6(12-) cage photosensitizes 1,3-rearrangement of encapsulated guest via photoinduced electron transfer. *J. Am. Chem. Soc.* **137**, 10128–10131 (2015).

160. Gouridis, G. *et al.* Structural dynamics in the evolution of a bilobed protein scaffold. *Proc. Natl. Acad. Sci. U. S. A.* **118**, e2026165118 (2021).